A Study on Various Data Mining Techniques for Agriculture Crop Yield Prediction

[1]Mohanadevi M, [2]Dr.Vinodhini V
[1]MPhil Scholar, [2]Associate Professor
[1]Department of Computer Science, [2]Department of Information Technology
[1,2]Dr.N.G.P Arts and Science College, Coimbatore, Tamilnadu, India

Abstract: This Data mining a techniques of examining large pre-existing databases in order to generate new information .It also pays a vital role in the field of Agriculture crop yield analysis. Data mining proves to be fertile ground for future innovations in agricultural statistics. Use of data mining techniques has been increased in agriculture field due to enhancement in the technology. Agriculture yield analysis is a very complex and vast research area as it deals with large data sets including different factors viz. yields of various crops, meteorological parameters affecting crop yields, diseases, pests [1]. So farmers are always curious about yield prediction. Crop yield depends on various factors like soil, weather, rain, fertilizers and pesticides. Several factors have different impacts on agriculture, which can be quantified using appropriate statistical methodologies. Applying such methodologies and techniques on historical yield of crops, it is possible to obtain information or knowledge which can be helpful to farmers and government organizations for making better decision and policies which lead to increased production [2]. In this paper discussed about applying various data mining techniques for agriculture growing crop yield prediction.

Index Terms **– Data Mining Techniques, Agriculture, Crop Yield Analysis.**

I. INTRODUCTION

Agriculture is the main occupation and the crop production is a difficult phenomenon that is recommended by Agriculture input parameters changes according to the farmers and the fields. An interdisciplinary subfield of computer science was evolved in 1960's and initially statisticians used the term data fishing and later in 1990's the term Data Mining was used to describe it [3]. The overall goal of data mining is to extract information from the data sets and to transform it into understandable form involving intersection of artificial Intelligence, Machine Learning, statistics and database systems. In other word we can say data mining to be the analysis step of knowledge discovery in database.

Globally, day to day the requirement of food is escalating; hence the agricultural scientists, farmers, government, and researchers are tiresome to put extra attempt and use numerous techniques in agriculture for improvement in production. As an effect, the data generated in the field of agricultural data enhanced day by day. Even at present, a very only some farmers are really using the new methods, tools and techniques in agriculture for better production. Data mining can be used for forecasting the future trends of agricultural processes [4].Data mining techniques mainly divided into two types of tasks namely predictive and descriptive.

Predictive Tasks

Predictive data mining techniques are supervised learning techniques. These techniques are used to generate models from class labeled data. Such produced models can be used for classification or prediction. It includes data mining techniques like classification, regression, time series analysis for data analysis. Classification is a process of organizing data into predefined categories with class labels whereas regression tries to map a data item to a real valued prediction variable. Time series analysis technique includes prediction of future values based on the discovery of similar patterns over a particular time period.

Descriptive Tasks

To derive patterns that summarizes the underlying relationship between data. It is used to find human-interpretable patterns describing the data. These techniques can be used to generate useful patterns from unlabeled data. Such techniques include clustering, association rules, sequential pattern discovery for data analysis. Clustering is a technique of dividing data into different meaningful subsets called clusters. In clustering method, there are no such predefined classes occurred like in classification. Association rule mining method is one of the useful techniques of data mining to discover interesting and meaningful patterns that frequently occurs together in data. Sequential pattern discovery method includes the extraction of frequently occurring patterns in the data. It compares the different sequences and recovers the missing sequence numbers.

1.1 Data Mining Techniques in Agriculture

There are different data mining techniques and algorithms are available for crop yield prediction and estimation. An effective methodology for crop yield prediction can be built up by using following data mining Techniques.

1.1.1 Regression

Regression allows us to model the relationship between two or more variables using simple mathematical techniques. Regression method works on two types of variables viz. independent variables and dependent variables. In practical life, regression analysis are applied to predict profit, sales, credit rates, house values, crop yield, temperature, distance between two or more points etc. A regression model that predicts the crop yield could be developed based on observed data for many yields of that particular crop over a period of time. In addition to the value, the data might track the sowing area, temperature, rainfall, humidity, fertilizers used, number of pesticides and spray applied and so on [5].

1.1.2 Association Rule Mining

Association rule mining method is one of the most important and useful method of data mining to discover interesting and meaningful patterns among large amount of data. Association rules are in the form of IF – THEN statements which help to find the desired relationships occurred between the various instances of data stored in data warehouses and other information repositories. These rules are applied in various areas viz. medical diagnosis, logistics, marketing and agriculture. In agriculture research domain, association rule mining helps to discover useful information and generates important rules about crop yields based on the relationships between different crops and soil parameters. For this purpose, the various association rule mining algorithms like Apriori, Predictive Apriori and FP Growth algorithms are applied for agriculture yield data analysis [6].

1.1.3 K Nearest Neighbors (KNN)

KNN is a classification and regression method mostly used for pattern recognition and statistical estimates. KNN classifies the objects based on the distance functions. In case of continues variables Euclidean, Manhattan and Minkowski distance functions are used for calculation whereas in case of categorical variables or binary data, hamming distance is used for statistical calculation. During classification, KNN algorithm provides a class having highest frequency count among K most similar instances as an output. But in case of regression, KNN provides the output based on the mean or median of K most similar instances. KNN applied for resampling of weather variables in order to design a K Nearest Neighbors simulator for daily precipitation and other climate parameters [7].

1.1.4 K Means Clustering

K Means Clustering is an unsupervised learning algorithm of clustering technique. It partitions given instances of data into K clusters by using mean of cluster as key parameter. Each instance present in cluster is nearest to the mean of that cluster. Major application areas of K Means Clustering include image processing, market research, pattern recognition, medical data analysis, meteorological data analysis and agriculture. In agriculture research field, K Means Clustering method is capable to partition the samples of crop yields and weather parameters into different clusters which are helpful for agriculture yield analysis [8].

1.1.5 Decision Tree Induction

Decision tree method includes the learning of decision trees from class-labeled training data sets. A decision tree is a flowchart-like tree structure, where each internal node represents a test on an attribute, each branch denotes an outcome of the test, and each leaf node represents a class label. The uppermost node of the tree represents the root node. The attribute values of a given data sample are tested against decision tree for classifying that unknown data sample. Decision tree induction method has been used in the field of biomedical engineering, financial analysis, manufacturing and production. This technique can be applied on agricultural data set to predict the impact of climate parameters on crop productivity based on the relationship between crop and weather parameters [9].

1.1.6 Support Vector Machine (SVM)

SVM is a supervised learning method for classification of both linear and nonlinear data. It uses a nonlinear mapping to transform original training data into a higher dimension. SVM classifies the data by finding the hyper plane that maximizes margin width between any two classes. SVM technique has been used in many fields includes bioinformatics, multimedia, artificial intelligence, pattern recognition, agriculture and so on. A SVM based downscaling model applied to obtain the future projections of precipitations for meteorological sub divisions in India [10].

II. LITERATURE REVIEW

In [11] N. Gandhi et al (2016), presented the overview on utilization of machine learning system for Indian rice editing ranges. Machine learning systems can be used to enhance forecast of harvest yield under various climatic situations. This paper examines at the exploratory outcomes acquired by applying SMO classifier utilizing the WEKA apparatus on the dataset of 27 areas of Maharashtra state, India. Those dataset acknowledged to the rice trim yield forecast might have been sourced from

openly available Indian organization records. The parameters recognized for those review were precipitation, base temperature, Normal temperature, most extraordinary temperature Furthermore reference trim evapotranspiration.

In [12] Vikas Chawla et al (2016), propose a data-driven approach that is „gray box" i.e. that seamlessly utilizes expert knowledge in constructing a statistical network model for corn yield forecasting. Multivariate gray box model is developed on Bayesian network analysis to build a Directed Acyclic Graph (DAG) between predictors and yield. Starting from a complete graph connecting various carefully chosen variables and yield, expert knowledge is used to prune or strengthen edges connecting variables. Subsequently the structure (connectivity and edge weights) of the DAG that maximizes the likelihood of observing the training data is identified via optimization. The focus of this work is to construct a corn yield predictor at the county scale. Corn yield (forecasting) depends on a complex, interconnected set of variables that include economic, agricultural, management and meteorological factors. Conventional forecasting is either knowledge-based computer programs (that simulate plant-weather-soil-management interactions) coupled with targeted surveys or statistical model based. The former is limited by the need for painstaking calibration, while the latter is limited to unvariate analysis or similar simplifying assumptions that fail to capture the complex interdependencies affecting yield.

In [13] The researcher presents various data mining techniques and their application in the agriculture and allied disciplines, such as the Support Vector Machines, Artificial Neural Net-works, the K-Nearest Neighbour technique, the K-means, Iterative Dichotomiser 3 algorithms and Association Rule Mining technique. The researcher denotes that in foresting prediction of agricultural crops and animal production is relative a new area of interest, but however, there is no one best technique, and the appropriateness of the method chose depends on the data set being mined, the expected results and the researchers knowledge about the technique.

In [14] Yethiraj systematic review on data mining techniques in the agricultural domain established that there are numerous algorithms that can be employed to recognize patterns in the data and aid in future prediction in production .This is as well affirmed by Barghavi and Jyothi who claims that data mining techniques can be employed in a given agricultural dataset to gain information, though this highly depends on the amount of data used. This clearly implies that accuracy of the gained information can be enhanced by increasing the size of the data-set. Such an aspect may enhance the valid patterns verification as compared to the conventional statistical analysis.

In [15] Bhatia and Anu Gupta used an agricultural data warehouse to mine quantitative association rules ion the dataset. The researchers compared different association rule techniques including FP-tree growth algorithm, Dynamic Item Set, Pincer-Search Algorithm, Partition Algorithm and Apriori Algorithm. The authors found the FP-Tree growth algorithm to generate the best results, as the Katter bears the least time complexity, making it more optimal.

In [16] D. B. Lobell and C. B. Field discussed about the impacts of recent warming on the production of major crops in the world. Average global yields of six major crops for time period of 42 years were taken into consideration for results. Multiple linear regressions have been performed using global yields as response variables and climate parameters viz. minimum & maximum temperature and rainfall as predictor variables. Analyses suggest that increased atmospheric temperature had negative impact on global yields of several major crops. There was a clear decline in yields of wheat, maize and barley crops with respect to temperature rise in the past. As climate changes, farmers would accommodate such cropping systems in order to minimize the negative impacts of warming.

In [17] D. W. Parvin, S. W. Martin, F. Cooke, Jr., and B. B. Freeland, Jr. studied the effect of harvest season rainfall on cotton yield at 22 locations in the Delta area of Mississippi from year 1991 to 1993 and 2002. For this purpose Regression analysis were performed to estimate the relationship between yield as the dependent variable and time and rainfall as independent variable. After the analysis, results indicate that during the harvest season increase in rainfall decreases the cotton yield and rainfall during the late season results in greater yield reduction.

In [18] F. Khan and D. Singh presented the implementation of association rule mining methodology for analysis of agricultural data set in order to generate rules to discover the relationships between different crop yields. The data sets of five different crops from Bhopal district in Madhya Pradesh were collected for analysis work. The parameters like soil type, PH value of the soil and cropping season were into consideration for results generation. The results generated through Apriori algorithm are further compared with results obtained by FP Growth method.

In [19] S. Veenadhari, Dr. B. Mishra and Dr. C. D. Singh presented in their paper about soybean productivity modeling using Decision Tree Algorithms. Bayesian classification and rule accuracy together with decision trees applied to study the impact of climate variables on crop productivity. Data sets of meteorological data of Bhopal district for 20 years were collected for results generation. Results obtained after analysis indicate that the productivity of soybean crop was mostly influenced by Relative humidity followed by rainfall and temperature variables.

In [20] P. Gwimbi and T. Mundoga have been presented the impact of climate change on cotton production under rain fed conditions in Gokwe, Zimbabwe. The dataset was taken of 25 years and a survey of 50 farmers in Gokwe district for proposed work. Significant Climate pattern were generated using rainfall and temperature data were statistically correlated to cotton yield using Statistical Package for the Social Sciences (SPSS) software package. This correlation provided evidence of the relationship

between rainfall and temperature variability and cotton production over that time period. The results generated by SPSS tool indicate that rainfall has positive impact on cotton yield but increase in temperature follows the considerable decrease in the cotton yield.

In [21] D. Ramesh and B. Vardhan (2015), exhibited a short Investigation about crop yield prediction utilizing Density based clustering technique and Multiple Linear Regression (MLR) for the selected region. A recent development in Information Technology for horticulture field has turned into an intriguing exploration region to anticipate the harvest yield. The issue from claiming yield forecast may be a significant issue that remains will be tackled In view of accessible information. Diverse information mining strategies are utilized also assessed in agribusiness assessing what's to come year's harvest creation. At first the factual model Multiple Linear Regression strategy is connected on existing information. The impacts so gotten were checked also investigated utilizing the information mining framework to be particular Density-based grouping strategy. In this system the aftereffects of two strategies were looked at as expressed by particular locale.

In [22] S. Dahikar and S. Rode (2014), proposed simulated neural system approach for horticultural product yield expectation. By considering different circumstances of climatologically marvels influencing neighborhood climate conditions in different parts of the world. These climate conditions directly affect edit yield. Different examines have been done investigating the associations between extensive scale climatologically marvels and product yield. Shown to be intense devices for displaying and expectation, to expand their adequacy. Edit forecast procedure is utilized to foresee the appropriate harvest by detecting different parameter of soil and furthermore parameter identified with climate parameters. For that reason we are utilized Artificial Neural Network (ANN).Authors inferred that ANN is valuable apparatus for product expectation. In this paper incorporates the parameter of their provincial soil parameter. At that point it is examine by utilizing encourage forward back engendering ANN. Break down in tangle lab ANN way to deal with make it more effective.

In [23] N.K. Newlands and L. Townley-Smith (2010), proposed Bayesian system approach for foreseeing vitality edit yield. Common asset issues regularly should be demonstrated utilizing information that is frequently fragmented at various spatial and fleeting scales with various levels of instability. Fluctuation because of atmosphere, soil, nuisances and administration choices add to facilitate auxiliary and useful many-sided quality of oversaw environments. Bayesian systems are perfect for such circumstances by empowering symptomatic thinking on contingent conditions to survey display auxiliary and also parameter vulnerability. They apply Bayesian systems to the issue of providing territorial with an ideal, hearty supply of biomass from vitality crops. Crops have diverse ideal atmosphere, water and supplement necessities, and affectability to outrageous climate, intrusive vermin and different effects. They test an improved model form in southern Manitoba, Western Canada. We analyze affectability of ideal respect planting/reap timing under verifiable climate, water, supplements and extraordinary occasion/bug misfortune inconstancy. We look at changed classifiers in getting a system arrangement and examine future work to apply the model at higher spatial and transient determination.

S.No	Author	Crop under Study	Parameter under study	Techniques applied
1	E.M.Adamgbe and F.Ujoh [7]	Maize	Rainfall	Correlation and Regression
2	S.Kaul [13]	Rice and Jower	Maximum and Minimum temperature ,rainfall, fertilizers usage and human labor	Regression
3	D.R.Mehta,A.D. Kalola,D.A Saradava and A.S.Yusufzai [5]	Groundnut,Pearl Millet,Sorghum and Cotton	Rainfall	Correlation and Regression
4	D.W.Parvin S.W.Martin,F.Cooke,Jr., and B.B. Freeland,Jr., [6]	Cotton	Rainfall	Regression
5	D.B.Lobell and C.B. Field [3]	Wheat,Rice,Maize,Soybean,Barley and Sorghum	Maximum and minimum termperature,rainfall	Multiple Linear Regression
6	P. Gwimbi and T.Mundoga [11]	Cotton	Rainfall and Temperature	Statistical Package for the Social Sciences

7	R.Dehgahi,A.Joniyas and M.D.Latip [12]	Wheat	Temperature and Rainfall	Variance Analysis
8	S.S.Hussain.M.Mudasser. M.M sheikh and N.Manzoor [14]	Wheat and Barley	Temperature and Rainfall	Regression
9	F.Khan and D. Singh [8]	Jower,Bajra,Rice,Soybean	Soil type,ph value of soil and Crop growing season	FP –Growth Algorithm and Apriori algorithm
10	S.Veenadhari,Dr.B.Mishra and Dr.C.D.Singh [17]	Soybean	Rainfall ,Temperature, evaporation and relative humidity	Decision tree and Bayesian Classification

Table 1: Various techniques and parameters used for agriculture yield prediction

III. DISCUSSION

An overview of data mining parameters on different crops. Different data mining techniques and algorithms which have been used for agriculture yield analysis are presented in this paper. In cited literature, data selection was carried out independently by researchers to generate results. We presented the qualitative overview of effect of various parameters on different crops along with the details of techniques applied in the form of table 1. In addition to this, we also tried to summarize the applications of different data mining techniques in weather forecasting.

IV. CONCLUSION

In the view of this, there are certain climate and Temperature parameters responsible for variable crop yields. There is a growing number of applications of data mining techniques in agriculture and it's a growing amount of data that are currently available from many resources. Various data mining techniques are available for analysis of different weather parameters with respect to different crop yields. By using these techniques one can build up a methodology for pre harvest crop forecasting. To find out the effect of meteorological parameters on a crop, a combination of two or more data mining methods can be applied to get better results.

REFERENCES

[1] Dr.Rupindes Singh and Gurpreet Sindh," Review: Role of Data Mining in Agriculture yield Analysis", International conference on soft computing Application in wireless communication (SLAWL 2017).

[2] P.Kanjana Devi,S,Sghenbagavadivu," Enhance crop yield prediction and soil data analysis using data mining," International Journal of Modern Computer Science (IJMCS) ISSN:2320-7868, Volume 4,Issue 6,Dec 2016.

[3] Dr.Devesh katiyar,Dr.Vinodani Katiyar,Dr.Shakuntala misra,"Data Mining techniques usedin Agriculture",International Journal jof Engineering Technology Science and Research (IJETSR) ISSN:2394-3386 Volume 4,,Issue 7 July-2017.

[4] Hetal Patel Research Scholar Charusat, Dharmendra Patel Assistant Professor Charusat, Changa in (2014) A Brief survey of Data Mining Techniques Applied to Agricultural Data.

[5] S. Kaul. (2001). Bio-Economic Modelling of Climate Change on Crop Production in India. [Online]. Available: www.ecomod.org/files/papers/370.pdf.

[6] F. Khan and D. Singh (2014), "Knowledge Discovery on Agricultural Dataset Using Association Rule Mining," *International Journal of Emerging Technology and Advanced Engineering*, Vol. 4 Issue 5, pp. 925-930, May 2014.

[7] B. Rajagopalan and U. Lall, "A K–Nearest-Neighbor Simulator for Daily Precipitation and Other Weather Variables," *Water Resources Research*, Vol. 35, No. 10, pp. 3089–3101, October 1999.

[8] D. Ramesh and B. V. Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data," *International Journal of Advanced Research in Computer and Communication Engineering*, ISSN: 2319-5940, Vol. 2, Issue 9, pp. 3477-3480, 2013.

[9] S. Veenadhari, Dr. B. Mishra and Dr. C. D. Singh, "Soybean Productivity Modelling using Decision Tree Algorithms," *International Journal of Computer Applications*, Vol. 27, No. 7, pp. 11-15, August 2011.

[10] S. Tripathi et al., "Downscaling of Precipitation for Climate Change Scenarios: A Support Vector Machine Approach," *Journal of Hydrology*, 330, pp. 621-640, 2006.

[11] N.Gandhi, L.J. Armstrong, O. Petkar and A. Tripathy, "*Rice Crop Yield Prediction in India using Support Vector Machines*", IEEE The 13thInternational Joint Conference on Computer Science and Software Engineering (JCSSE), Thailand, 2016.

[12] Vikas Chawla, Hsiang Sing Naik, Adedotun Akintayo, "*A Bayesian Network approach to County-Level Corn Yield Prediction using historical data and expert knowledge*", Data Science for Food, Energy and Water, 2016.
[13] Yethiraj NG (2012) Applying Data Mining Techniques in the field of agriculture and allied sciences. Inter J Business Intelligents 1: 2.
[14] Barghavi P and Jyothi S (2009) Applying nave bayesdata mining technique for classification of agricultural land soils, Inter J Computer Sci Network Security 9: 117-122.
[15] Bhatia J, Gupta A (2014) Mining of Quantitative Association Rules in Agricultural Data Warehouse: A Road Map. Inter J Info Sci Intelligent System 3: 187-198.
[16] D. B. Lobell and C. B. Field, "Global scale climate–crop yield relationships and the impacts of recent warming," *Environmental Research Letters*, Vol. 2, pp. 1-7, 2007.
[17] D. W. Parvin, S. W. Martin, F. Cooke, Jr., and B. B. Freeland, Jr., "Effect of Harvest Season Rainfall on Cotton Yield," *Journal of Cotton Science*, Vol. 9, pp. 115–120, 2005.
[18] F. Khan and D. Singh (2014), "Knowledge Discovery on Agricultural Dataset Using Association Rule Mining," *International Journal of Emerging Technology and Advanced Engineering*, Vol. 4 Issue 5, pp. 925-930, May 2014.
[19] S. Veenadhari, Dr. B. Mishra and Dr. C. D. Singh, "Soybean Productivity Modelling using Decision Tree Algorithms," *International Journal of Computer Applications*, Vol. 27, No. 7, pp. 11-15, August 2011.
[20] P. Gwimbi and T. Mundoga, "Impact of Climate Change on Cotton Production under Rainfed Conditions: Case of Gokwe," *Journal of Sustainable Development in Africa*, ISSN: 1520-5509, vol. 12, No. 8, pp. 59-69, 2010.
[21] D. Ramesh and B. Vardhan, "*Analysis of crop yield prediction using data mining techniques*", International Journal of Research in Engineering and Technology, vol. 4, no.1, pp. 47-473, 2015.
[22] S. Dahikar and S. Rode, "*Agricultural crop yield prediction using artificial neural network approach*",International Journal of Innovative Research in Electrical, Electronic Instrumentation and Control Engineering, vol. 2, no. 1, pp. 683-686, 2014.
[23] N.K.Newlands, L. Townley-Smith, "*Predicting Energy Crop Yield Using Bayesian Networks*", In Proceedings of the Fifth IASTED International Conference, Vol. 711, pp. 014-106, 2010.