RESEARCH ARTICLE OPEN ACCESS

Prediction of Soil and Crop Yield by Big Data Analysis

Venkata Chennareddy, Ramanayagam S.
(Master of Computer Applications, Hindustan Institute of Technology and Science, Chennai.
Email: mvcreddy.898@gmail.com)
(Master of Computer Applications, Hindustan Institute of Technology and Science, Chennai.
Email: sramanayagam@hindustanuniv.ac.in)

--************************----------------------------------

Abstract:

The agriculture sector in India is facing a tough problem to increase crop productivity. More than 60 per cent of the crop still depends on rainfall. Crop yield depends on many factors including soil, climate, rainfall, fertilizers and pesticides. Many factors have different effects on agriculture. Recent developments in information technology for agriculture have become an interesting research area for assessing crop yields. The yield estimation problem is a major problem that needs to be addressed based on available data. it is possible for farmers and government agencies to gain information or knowledge that will help them make better decisions and make decisions that lead to production. The proposed approach is to give farmers instructions on other crops that are suitable for their land conditions in the area. Various data mining techniques can be used and evaluated in agriculture to predict future crop production. Various classification algorithms are applied to assess soil fertility. This paper focuses on the classification of soil fertility rate using K-Means, Random Tree and Apriori .

Keywords —Crop prediction, Agriculture, Yield prediction, Datamining,

--************************----------------------------------

1. INTRODUCTION

Agriculture is the backbone of the Indian economy. In India, for many reasons most farmers are not getting the expected crop yields. Agricultural yield depends mainly on climatic conditions. Rainfall conditions also affect rice cultivation. In this case, farmers need timely advice to assess future crop productivity and an analysis should be done to help farmers increase crop production in their crops.

Yield estimation is an important agricultural issue. Every farmer is interested to know how much yields he expects. In the past, yields were estimated by considering the farmer's previous experience on a particular crop. There is a lot of data in Indian agriculture. Data that is turned into information can serve many purposes well.

Data mining is widely applied to agricultural issues. Data mining is used to analyse large data sets and to establish useful classifications and patterns in data sets. The overall goal of the data mining process is to extract information from a data set and make it an understandable structure for further use. Data mining in marketing Understandable and unknown information for large data repositories and their use for key business decisions to support they are organized, strategic marketing strategies and calculating their success. Data mining facilitates discovery Correlation in data. It covers the fields of database and expertise Systems, Information Theories, Statistics, Mathematics, Logics and an array of connected fields. Prediction is a process of forecasting about the event whose original result has not been observed. Common place instance may be the estimation of some of the variables of interest at some specific future date. Use of commodities can be different between areas of applications.

The main objective of this paper is to create a user-friendly interface for farmers, which provides an analysis of rice production based on available data. Various data mining techniques have been used to estimate crop yields to increase crop productivity.

2. RELATED WORK

From the research article , the researcher expresses the large amount of data collected and stored for analysis. Proper use of this data often leads to significant gains in efficiency and hence economic benefits.

There are many applications of data mining techniques in the agricultural sector. Researchers have implemented the K - means algorithm to estimate pollution in the atmosphere, K neighbours are applied to simulate daily precipitation and other weather variables and analyse various changes of climate conditions using vector machines.

The researcher has proposed soil profile descriptions for classifying soils in conjunction with GPS-based technologies . They used the K-Means approach for soil classification. In a similar vein, crop classifications were made using hyper spectral data by adopting one of the data mining approaches , namely support vector machines. One of the researchers used extreme fuzzy cluster analysis to classify plants, soil, and residues

In agronomy, clustering techniques are found in grading apples before marketing. Weeds have been found on precision farming. Researchershave worked on the analysis of rainfall variability and its impact on crop productivity. The influence of seasonal climatic conditions such as rainfall and temperature variation on crop yield estimation has been considered by empirical crop modelling Furthermore, there are two Methods for investigating the impact of climate change on crop production, including crop adaptation and production function .

Researchers have found that reducing the growth stages of plants reduces the yield of winter wheat when temperatures rise and concludes that the complexity of a model depends on the level of detailed analysis. Or it is less detailed with estimates of humidity .

3. METHODOLOGY

In this paper the statistical method K-means clustering algorithm technique and data mining method namely Random Tree and Apriori algorithm were taken up for the estimation of crop yield analysis.

a) K-means Clustering Algorithm

The K-means algorithm is an iterative algorithm that attempts to divide the dataset into K-pre-defined different non-overlapping subgroups (clusters), where each data point belongs to the same group.
The work flow of K-means algorithm: -

- Specify the number of K groups.
- Start the centroids by changing the dataset first and then start by randomly selecting the K data points for the centroids.
- Repeat until there is no change in centroids. This means that the allocation of data points to the clusters does not change
- Calculate the sum of the squared distance between the data points and all the centroids.
- Assign each data point to a nearby cluster (centroid).
- Calculate the centroid for the clusters by taking the average of all data points for each cluster..

The objective function is:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^{i} - \mu_k\|^2 \qquad (1)$$

where wik=1 for data point xi if it belongs to cluster k; otherwise, wik=0. Also, μk is the centroid of xi’s cluster.

b) Random Tree

The algorithm can solve classification and regression problems. A random tree is a collection of tree predictors, also known as more forest in this category (the term was also introduced by L.Breiman). The classification works as follows: The classification of random trees takes the input feature vector, Classifies with each tree in the forest and provides a class label with a majority of "votes". Soil presented at different locations in the district and these areas are classified under different soil grades based on the physical and chemical factors of the soil using the Random Tree Classification
It shows the different types of Agri block and these blocks are splited according to the soil factors. The soil samples collected from various places of district are analysed and the factors of the soils are similar which were found in same series.

c) Apriori Algorithm

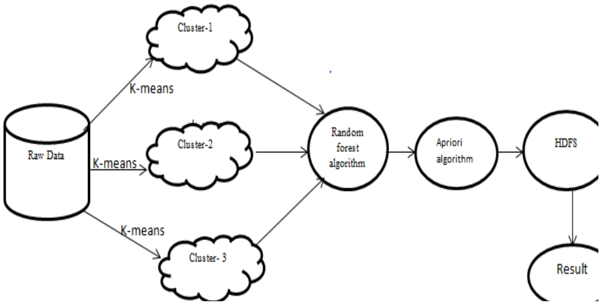
The Apriori algorithm is used for mining frequent itemset and devising association rules from a transactional database. The parameters “support” and “confidence” are used. Support refers to items’ frequency of occurrence; confidence is a conditional probability.
Items in a transaction form an item set. The algorithm begins by identifying frequent, individual items (items with a frequency greater than or equal

to the given support) in the database and continues to extend them to larger, frequent itemset
The apriori algorithm uses the *downward closure property* ,i.e., all the subsets of a frequent itemset are frequent, but the converse may not be true

The following are the main steps of the algorithm:

- Calculate the support of item sets (of size k = 1) in the transactional database (note that support is the frequency of occurrence of an itemset). This is called generating the candidate set.
- Prune the candidate set by eliminating items with a support less than the given threshold.
- Join the frequent itemset to form sets of size k + 1, and repeat the above sets until no more itemset can be formed. This will happen when the set(s) formed have a support less than the given support.

4. DESIGN AND DISCUSSION



We collect all the required dataset. Regarding the datasets the initial description and the point to be remembered is about the attributes. The dataset regarding the attributes which suits the project must be analyzed such that the entire results depend on the dataset collected and attributes containing in it. The data set collected are crops vs. seasons, crops vs. price(various years),seasons vs. temperature(various years).In datasets collected may contain various null values, inconsistent values and datasets may also be in different formats. In this process of the pre-processing, we will remove null values and inconsistent values. The null values reduce the accuracy of prediction of the data. And inconsistent data reduces and confuses the algorithm. In order to pre-process the data, we will be using the Hadoop Tool. The pre-processed data helps in more accurate prediction of the data. Map reduction technique in the Hadoop tool is used to get the data that's only required. Map reduction technique reduces the amount of data needed to be processed by the classification algorithm is reduced. Data classification and clustering: Data classification is the process (or) procedure of sorting and categorizing data into various types, forms which will help us more when we are analysing the data. Data classification helps us a lot to do the separation of data and classification of data according to data set requirements for various business motives or personal objectives. It is mainly a data management process. Data classification helps in the prediction of the data.

5. CONCLUSION:

It is used to assess the suitability of the crop for a particular soil type and to enhance the overall quality of the agricultural product. This helps the farmers to select a particular crop for the seeds depending on the weather conditionsAnd provides the information needed to choose the best climate for doing quality farming. The paper uses big data using the Hadoop platform to help deal with large datasets in the agrarian domain
. In this system I have proposed, we have different soils and different crops. Here I have predicted the soil fertility rate and which crop is suitable for the particular soil and which crop will give the more crop yield. This whole system id carried out by using the big data techniques. The future enhancement is we can predict the how much crop will give the how much yield will produce.

REFERENCES

[1]Jambekar, S., Nema, S. and Saquib, Z., 2018, August. Prediction of Crop Production in India Using Data Mining Techniques. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE.

[2] Chlingaryan, A., Sukkarieh, S. and Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and electronics in agriculture, 151, pp.61-69.

[3] Kumar, P., Kumar, A., Panwar, S., Dash, S., Sinha, K. and Ray, M., 2018. Role of big data in agriculture-A statistical prospective.

[4] Sahu, S., Chawla, M. and Khare, N., 2017, May. An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach. In 2017 International Conference on Computing Communication and Automation (ICCCA) (pp. 53-57). IEEE.

[5] Ramya, M., Balaji, C. and Girish, L., 2015. Environment change prediction to adapt climate-smart agriculture using big data analytics. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 4(5).

[6] Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh," Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique",Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, T.N., India. 6 - 8 May 2015. pp.138-145

[7] Veenadhari, S., Bharat Misra, D Singh, "Data mining Techniques for Predicting Crop Productivity – A review article", IJCST, International Journal of Computer Science and technology march 2011.

[8] M.S.PrasadBabu, N.V.Ramana Murty, S.V.N.L.Narayana, "A Web Based Tomato Crop Expert Information System Based on Artificial Intelligence and Machine Learning Algorithms", IJCSIT, Vol. 1 (1), 2010, 6-15.

[9] Leemans V, M F Destain, "A Real Time Grading Method of Apples Based on Features Extracted from Defects", J. Jood Eng., 2004, pages: 83-89.

[10] R J Brooks, "Simplifying Sirius: Sensitivity Analysis and Development of a Meta-Model for Wheat Yield Prediction", European Journal of Agronomy, vol. 14, 2001