Prediction of Soil and Crop Yield using Big Data Analysis

E. Brumancia[1] Yogesh Dadhich[2]
[1,2]UG Student
[1,2]Department of Computer Science and Engineering
[1,2]Sathyabama Institute of Science and Technology, Chennai, India

Abstract— India is an essentially rural nation. It is the primary wellspring of salary for farmers, so farmers are constantly inquisitive about yield expectation. Harvest yield relies upon different components like soil, climate, downpour, composts and pesticides. A few elements impact sly affect agribusiness, which can be measured utilizing fitting factual techniques. Applying such approaches and methods on recorded yield of harvests, it is conceivable to get data or information which can be useful to ranchers and government associations for settling on better choice and arrangements which lead to expanded creation. The goal of the work is to look at different information mining strategies which gives the most extreme exactness. Information mining is just the way that helps to change over immense information into innovations and make them accessible to the ranchers. The immense measure of information can be used to mine chunk of information that can be valuable for ranchers and leaders to take compelling and brief choice. Right now, discussed some significant apparatuses and system handle and study enormous information.

Keywords: Agriculture, K-means, Random Forest, Prediction, etc.

I. INTRODUCTION

India is oldest country which still depends on agriculture. India is largest country with 70 percent of farmers still primarily depend on agriculture and 82 percent farmers small and marginal. Total production is 275 mt with 15.4 percent contribution in GDP of India. Agriculture is the backup of Indian economy. For the growth of nation farmers can be help in main stream if we find out the different methods in which we pay less and gain more production. For more production of crop the precision agriculture is modern and new technology which uses the old datasets of soil ,soil types, crop types ,land types from these large number of datasets we can predict the production which soil is suitable for which crop .It increases the production of crops and it makes more reliable and independent to farmers as well as they can contribute more to the economy . once we get to know which crop is perfect for our soil it helps to reduce the wastage of seeds and it reduces the uses of chemicals. which evolved to the healthy life it makes stronger to our health. Due to globalization we are not sure about weather conditions suddenly climate can change if we use the old datasets from satellite and analyzing the data in proper, we it help to save us from extreme loss of crop and pre protection of agriculture help us to achieve our ambitious goal of doubling the farmers income by 2022.

Climate change and agriculture are interlinked to each other. Global warming affects in different ways such as irregular rain and climate extremes (like heat waves, cold waves) changes in pests and rising of sea level. Increase of co2 content in environment is increasing and, in the research, it is found that the extreme level of weather and temperature will affect negatively.

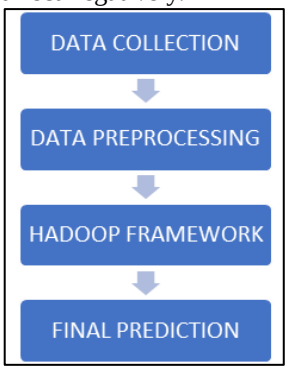

Fig. 1:

II. RELATED WORK

[1] Data mining is the investigation of extricating valuable data from the informational collections. At that point present examination centers around the use of Data mining procedures to foresee future creation of yields, for example, Rice, Wheat and Maize as for different parameters saw during the period (1950-2013). The parameters considered for the examination were precipitation, mean temperature, territory under water system, region, creation and yield. The relapse calculations utilized in the examination were Various Straight Relapse, Arbitrary Timberland Relapse and Multivariate Versatile Relapse Splines (Earth).

[2] This paper examines explore advancements led inside the most recent 15 years on AI based systems for exact harvest yield forecast and nitrogen status estimation. The paper reasons that the fast advances in detecting innovations and ML procedures will give practical and complete answers for better harvest and condition state estimation and dynamic. More focused on use of the sensor stages and ML methods, the combination of various sensor modalities and master information, and the improvement of crossover frameworks joining distinctive ML and sign handling systems are for the most part liable to be a piece of accuracy horticulture (Dad) sooner rather than later.

[3] Within this paper an attempt has been made to focus on the use of knowledge mining techniques in the field of agriculture. Calculation technique J48from C4.5 is used. Currently, information mining strategies are being developed which utilize past data such as a specific soil type, soil pH, ESP, EC.

[4] locale to give better harvest yield estimation for that district. This model can be utilized to choose the most magnificent harvests for the locale and furthermore its yield there by improving the qualities and increase of cultivating too. This guides ranchers to settle on the yield they might want to plant for the inevitable year. Forecast will help the related enterprises for arranging the coordination of their business.

[5] This paper gives a reason about how to discover encounters from exactness horticulture data through huge information approach. Right now, the important information in a successful way drives a structure turning to big computing difficulties in crop examination where data is remotely accumulated. For the capacity motivation behind colossal information accessibility in agribusiness, we are planning Hadoop structure in store for our jobs a gigantic volume of yield information. That job gives you a superior forecast for the ranchers to plant which sort of harvests to their homestead field dependent on their dirt substance to improve the efficiency. The arbitrary woods calculation is incorporated with the MapReduce programming model in Hadoop structure.

Information are assuming a significant job making great arranging and approaches for agrarian development and advancement. Populace development and environmental change are overall patterns that are expanding the significance of utilizing huge information science to improve farming. Add to that land debasement expanding peripheral land and loss of biodiversity are better arrangements with investigation of huge information science. Harvest information can be separate into bits and bytes it will give better investigation about the yield improvement by utilizing advance information examination instruments for improvement of farming. Here, talk about some significant apparatuses and strategies to deal with and study the enormous information.

III. METHODOLOGY

A. Overview

In the proposed system we are predicting the crop production using the big data analysis. To predict the present scenario, we collect the historical data by the help of these data's we predict that which soil is good for which crop and output will be in the format that it is accessible for the farmers.

There are two ways first analyzing the data later on classifying the data. First the raw data is filtered by filtering process and converted into processed data by the help of k means clustering we categorize into the clusters later we classify these clusters through classification algorithm random tree and apriori algorithm.

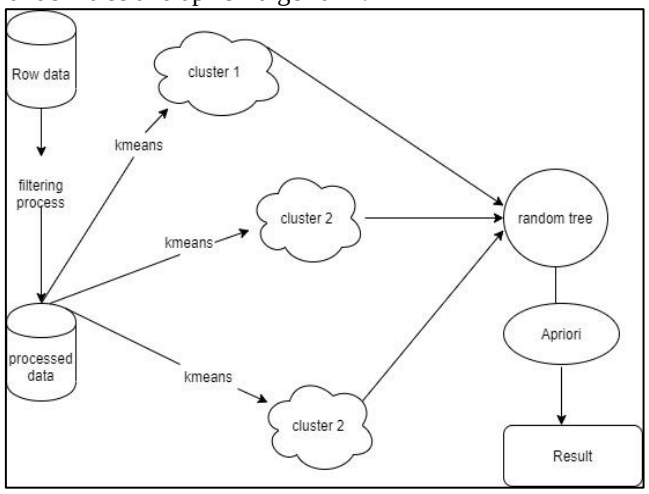

Fig. 2: Architecture

B. Data Collection

The data is collected from my.gov and Kaggle websites. These data gathered from online sources such as weather data, soil data, land data, temperature data and different types of crop data. These data are stored as a format of .csv file in excel sheet. The datasets may be containing more data's it can be in structured or unstructured format.

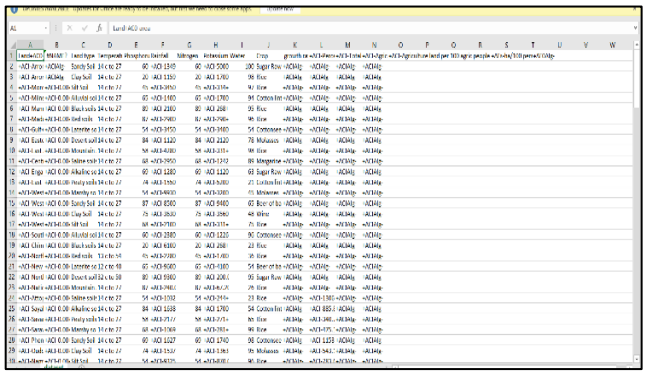

Fig. 3:

C. Data Preprocessing

Prediction of soil and crop yield data is preprocessed after the collection of various records. The data sets contain more than 150 number of datasets in which many of them are missing records. In the processing of data, it removes all the unnecessary data. The data consists of various columns as crop yield, area, temperature, irrigation area and year.

D. K-means Clustering

k- mean clustering algorithm is used to cluster the various data of nearest mean. It helps us to reduce the dimensions and to improve the feature learning.

Input: D= {t1, t2, t3……Tn} //Set of elements
k //Number of desired clusters
Output: K //set of clusters.
Algorithm: assign initial values for means
m1, m2, …… mk
Repeat assign each item ti to the cluster which has closest mean;
calculate new mean for each cluster;
until convergence criteria is met.

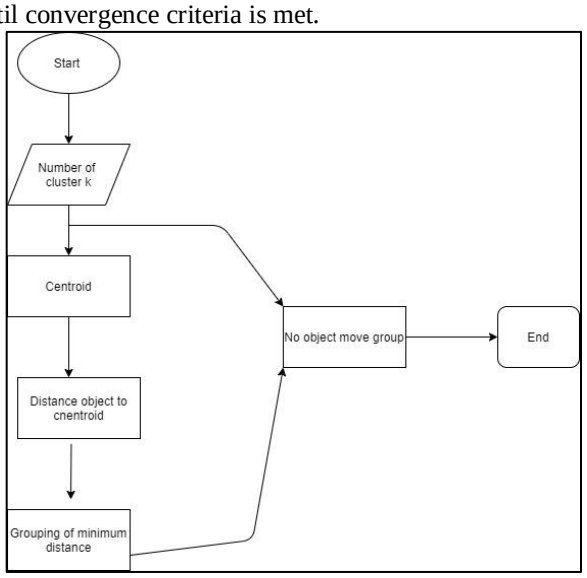

Fig. 4: K-means clustering

E. Random tree:

The random tree helps in to make our prediction more accurate because it classifies and completes the regression task. It provides more accuracy to our project. The data we clustered using k- means clustering if any set of data is missing it will handle those values. We have large amount of data the random tree creates more decisions and merge them to obtain a more accurate prediction.

The soil presented in various places of district and these places are classified under different soil series based on the soil physical and chemical factors of soil using Random Tree Classifier. It shows the different types of Agri block and these blocks are splinted according to the soil factors. The soil samples collected from various places of district are analyzed and the factors of the soils are similar which were found in same series.

Require: Input: D dataset – features with a target class
for ∀features do
for Each sample do
Execute the Decision Tree algorithm
end for
Identify the feature space f1, f2, . . ., fx of dataset UCI.
end for
Obtain the total number of leaf nodes l1, l2, l3, . . ., ln with its constraints (10)
Split the dataset D into d1, d2, d3, . . ., dn based on the leaf nodes constraints. (11)
Output: Partition datasets d1, d2, d3, . . ., dn

F. Apriori Algorithm:

The apriori algorithm is used to make our project more reliable by this classification algorithm it is used for mining different frequent item sets. It associates the frequent item sets and consider the least minimum support item. It also helps to maintain more accuracy to the large number of datasets.

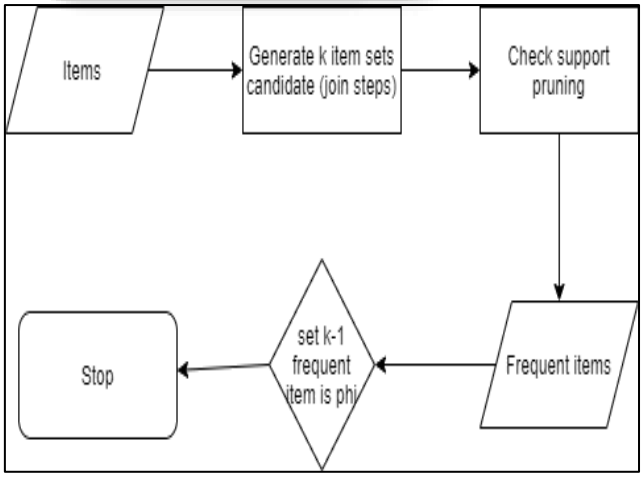

Fig. 5: Apriori classification

G. Hadoop Framework:

The data is loaded into the hdfs and processed the dataset by the map and reducing technique model. It handles all types of structured and unstructured data. The mapping part is done by the mapper class and the reducing part is completed by the reducer class. We are using large amount of different data it helps us to process this huge amount of data in our program and later we can implement on our driver. These data are generally sorted by the mapper class and used as input data for the reducer class. The duplicate data is removed by reducer from the key value pair.

The essential qualities of the MapReduce is to blend the information <key,value> pair with another rundown of <key, value> pair.

The information is handled by mapping capacity to create the halfway arrangement of <key,value> pair.

The reducer merger the yield made from mapper to outline a smaller game plan of characteristics.

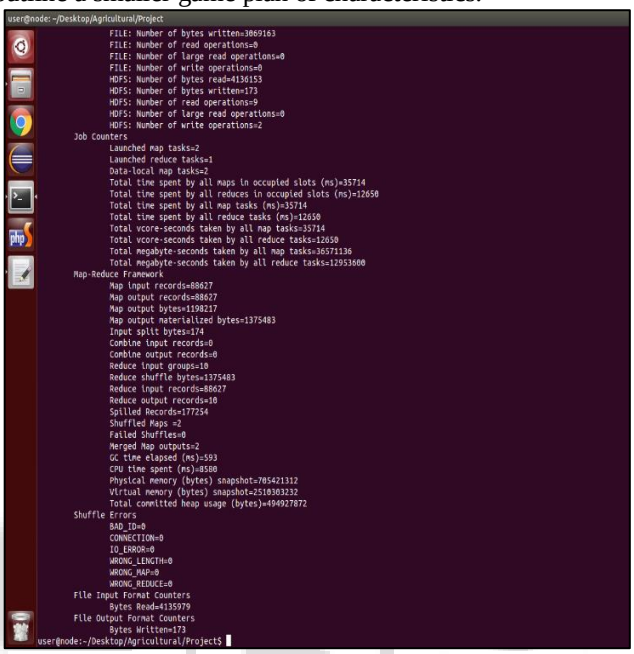

Fig. 6: Map reducing technique

The above figure mainly contains two tasks first is maping and second is reducing first it takes a set of data and it changes to another set of data and then this set of reduce in the key value pair. The file input counter reads the bytes is 4135979 after map and reducing file output format counters read the bytes is 173.

IV. Results and Discussion

This system how to make horticulture efficient by anticipating and consequently improve the harvest yields by utilizing soil data. The paper tells about how we can yield more production of crops based on the historical data of soil. This also helps to the farmers to match which crop is suitable according to the weather conditions and gives the data to cultivate best crop. The paper utilizes enormous information utilizing Hadoop stage which assists with managing the huge measure of datasets in rural space.

In below figure the prediction of yield the crop is shown using big data analysis. According to the given crop it predicts the soil and nutrition.


```
Prediction and enhancement of crop yield by big data analysis

    Crop          Land              Nitrogen          Phosphorus          Potassium

    Rice          Black-Soil                2.1   2.081967213114754   2.064516129032258
                  Red-Soil    2.032258064516129   2.1666666666666667   2.098360655737705
                  Sandy-Soil  2.0655737704918034   2.0483870967741935   2.1333333333333333

CottonSeed        Black-Soil             2.03125   2.0153846153846153                 2.0
                  Red-Soil    1.9696969696969697               2.046875   2.0307692307692307
                  Sandy-Soil   2.096774193548387   2.1475409836065573   1.8857142857142857

 Sunflower        Black-Soil  1.9411764705882353   1.8571428571428572   1.8985507246376812
                  Red-Soil    2.0307692307692307   2.1666666666666665   2.1129032258064515
                  Sandy-Soil                2.0               2.03125   2.1129032258064515

user@node:~/Desktop/Agricultural/Crop$
```

Fig. 7: Prediction of soil and crop yield

Fig. 8:

In above figure the x axis represents datasets and y axis represents overall purity. Where it tells about the crop, nutrition and temperature of the soil.

V. CONCLUSION

This framework the best approach to make farming efficient by anticipating and consequently improve the harvest yields by utilizing soil data. The paper presents a substitution Argo calculation which is utilized to foresee the appropriateness of a harvest for a particular soil type and upgrades the general nature of farming creation. This additionally helps the ranchers to pick a particular harvest to plant relying on the climatic condition and gives essential data to choose the least complex climate to attempt to quality cultivating. The paper utilizes huge information utilizing Hadoop stage which assists with influencing the monstrous measure of datasets in rural space.

ACKNOWLEDGMENT

Sathyabama institute of science and technology.

REFERENCES

[1] Suvidha Jambekar; Shikha Nema; Zia Saquib "Prediction of Crop Production in India Using Data Mining Techniques" 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)
[2] Anna Chlingaryana, Salah Sukkarieha, Brett Whelanb — Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, Computers and Electronics in Agriculture 151 (2018) 61–69, Elisver, 2018.
[3] Prakash Kumar, Anil Kumar, Sanjeev Panwar, Sukanta Dash, Kanchan Sinha, Vipin Kumar Chaudhary and Mrinmoy Ray "Role of big data in agriculture- A statistical prospective" Agric. Res. New Series Vol. 39 (2): 210-215 (2018).
[4] Pooja M C et al, —Implementation of Crop Yield Forecasting Using Data Mining, International Research Journal of Engineering and Technology (IRJET), 2018, p-ISSN: 2395-0072
[5] Shriya Sahu et al," An Efficient Analysis of Crop Yield Prediction Using Hadoop framework based on random Forest approach, International Conference on Computing, Communication and Automation, 2017.
[6] Bask J., Sudarshan, A., Trivedi D., M.S. Santhanam. 2004. "Weather Data Mining Using Independent Component Analysis". J. of Machine Learning Research 5: pp. 239- 253
[7] P. Surya, Dr. I. Laurence and M. Ashok Kumar, "The role of big data analytics in agriculture sector: A survey", March 2016
[8] Ms. Kanaan Devi, "Enhanced Crop Yield Prediction and Soil Data Analysis Using Data Mining", International Journal of Modern Computer Science, Volume 4, Issue 6, December 2016
[9] Anon (2010). Agricultural Statistics at a Glance 2010, Ministry of Agriculture Govt. of India
[10] Soil, Big Data and the Future of Agriculture Conference, Canberra, 25 June, 2015.
[11] Venkata eddy Konasani, Mukul Biswas and Praveen Krishnan Koleth, "Fraud Management using Big Data Analytics", A Whitepaper by Trend wise Analytics.
[12] Radha Krishna Murthy "Crop Growth Modelling and its Applications in Ag in Agricultural Methodology"
[13] J. O. Chan, "An Architecture for Big Data Analytics", Communications of the IIMA, vol. 13, no. 2, (2013), pp. 1-14
[14] J. M. Chambers, "Software for Data Analysis: Programming with R (Statistics and Computing)", Springer, (2012).
[15] AP AgTech Summit 2017: Progressive Farmer, Smart Farming, 15-17 Nov, 2017,Vizag, www.apagtechsummit2017.in/

