International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 6, Issue 11
November 2017

Crop Yield Prediction Using Big Data Analytics

Ms. Ruchita Thombare
Student, Department of Information Technology, Pimpri Chinchwad College of Engineering,Pune, India
Ms. Shreya Bhosale
Student, Department of Information Technology, Pimpri Chinchwad College of Engineering,Pune, India
Mr. Prasanna Dhemey,
Student, Department of Information Technology, Pimpri Chinchwad College of Engineering,Pune, India
Ms.Anagha Chaudhari,
Assistant Professor, Department of Information Technology, Pimpri Chinchwad College of Engineering, Pune, India

ABSTRACT
With the digital advancements in the field of agriculture, a large amount of data is being produced constantly as a result; agricultural data has entered the world of big data. This big data or massive volumes of data have a wide variety that can be captured, analyzed and used for decision-making. This aims to gain insight into the Crop Yield Prediction Using Big Data Analysis and identify the related socio-economic challenges. In this project the analysis of huge data would be carried out using K-means clustering methods to analyze the best suited way of agriculture methods in that specific region and prediction of yield would be found by Apriori algorithms and this useful data would be again given to farmers for the better results of crop yields and green agriculture.

KEYWORDS
Agriculture, big data, crop yield prediction, K-means clustering, Apriori Algorithm.

INTRODUCTION
India is an indomitable country with more than billion plus people, and also one of the world's rapidly flourishing economies. Out of the huge population, 58.4% are innocent agricultural assemblage. India's recent accomplishments in crop yields while being inspiring, are still just 30% to 60% of the best crop yields reachable in the farms of developed as well as other developing countries [1]. India ranks second worldwide in farm productivity. Sectors like forestry and fisheries count for 13.7% of the GDP (gross domestic product) in 2013, about 50% of the workforce [1]. The economic contribution of agriculture to India's GDP is steadily declining with the country's broad- based economic growth[1]. India is the world's largest manufacturer of many fresh fruits and vegetables, milk, major spices, select fibrous crops such as jute, staples such as millets and castor oil seed. India is the second largest producer of wheat and rice, the world's major food staples[3].

LITERATURE SURVEY
It is analyzed from the literature review that with the digital advancements in the field of agriculture, a large amount of data is being produced constantly as a result agriculture data has entered the world of big data [2]. An initiative in Columbia has found that data-driven climate adaption could revive rice yields in Columbia [3]. From the research article, the researcher expresses that large amount of data which is collected and stored for analysis. Making appropriate use of these data often leads to considerable gains in efficiency and therefore economic advantages [1].The researchers implemented. K-Means algorithm to forecast the pollution in the atmosphere, the Nearest Neighbor is applied for simulating daily precipitations and other weather variables and different possible changes of the weather scenarios are analyzed using Support Vector Machines [2]. Soil profile descriptions were proposed by the researcher for classifying soils in combination with GPS based

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 6, Issue 11
November 2017

technologies [1]. They were applied K-Means approach for the soil classification. One of the researcher used an intensified K-Means cluster analysis for classifying plants, soil and residue regions of interest from GPS based color images[5]. Weeds were detected on precision agriculture. The researchers worked on rainfall variability analysis and its impact on crop productivity [4]. The effect of observed seasonal climatic conditions such as rainfall and temperature variability on crop yield prediction was covered through an empirical crop model [11]. Furthermore, there are approaches to investigate the impact of climate change on crop production which include the crop suitability approach and the production function approach.

SYSTEM ARCHITECTURE

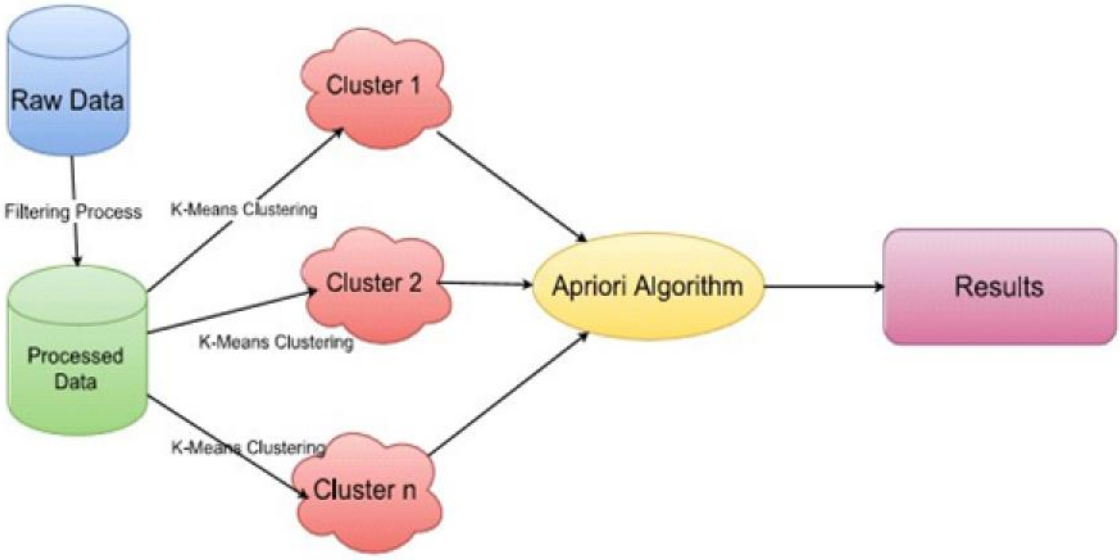

Fig. 1: System Architecture

- Module 1 : Data pre-processing
- Module 2 : K-means clustering
- Module 3 : Apriori
- Module 4 : Experimental Results

- **Module 1 : Data pre-processing**

The given data contains all the information regarding particular region's past 10 year's attributes such as weather, rainfall, soil type, soil fertility, average crop yield, pesticides used, crop type, water availability, electricity availability. This data is pre-processed accordingly which is further send to K-means clustering model for further operations on data.

1. Raw Data:

Agriculture production of different food grains from year 2003 to 2014 at all India level dataset provides us with State name, Season, Crop, Area, Production required for us to analyse the crop production in that area[9].

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 6, Issue 11
November 2017

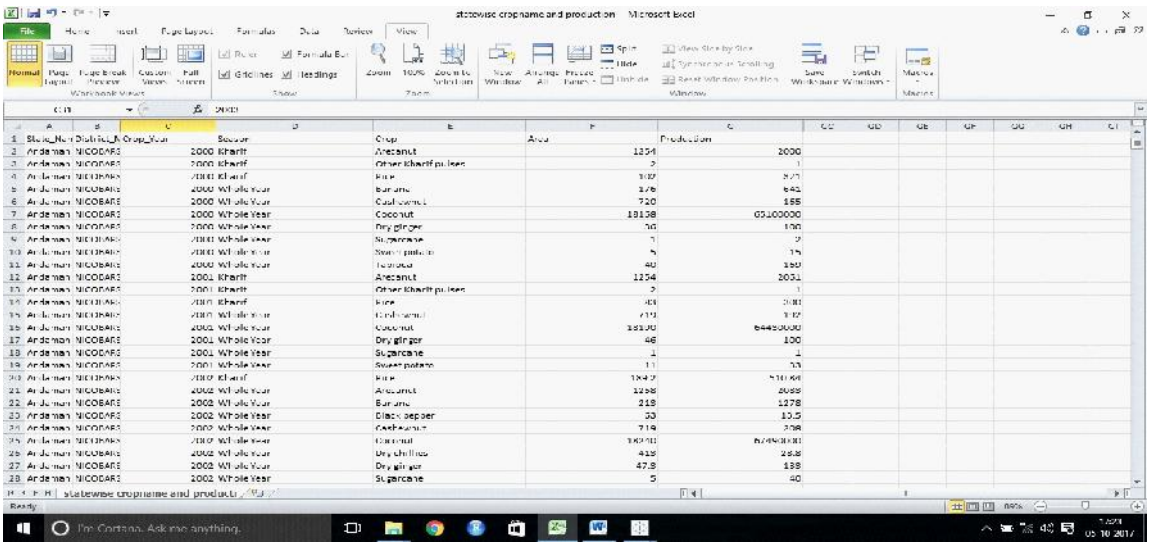

Fig. 2: Data Set No.1

2. Rainfall Dataset

Average annual rainfall is 300–650 millimeters (11.8–25.6 in), but is very unreliable; as in much of the rest of India, the southwest monsoon accounts for most precipitation.

In this rainfall dataset, rainfall in India's state and district is elaborated systematically. The join of this and previous dataset is taken to form effective dataset where district name is taken as common element in both the datasets[10].

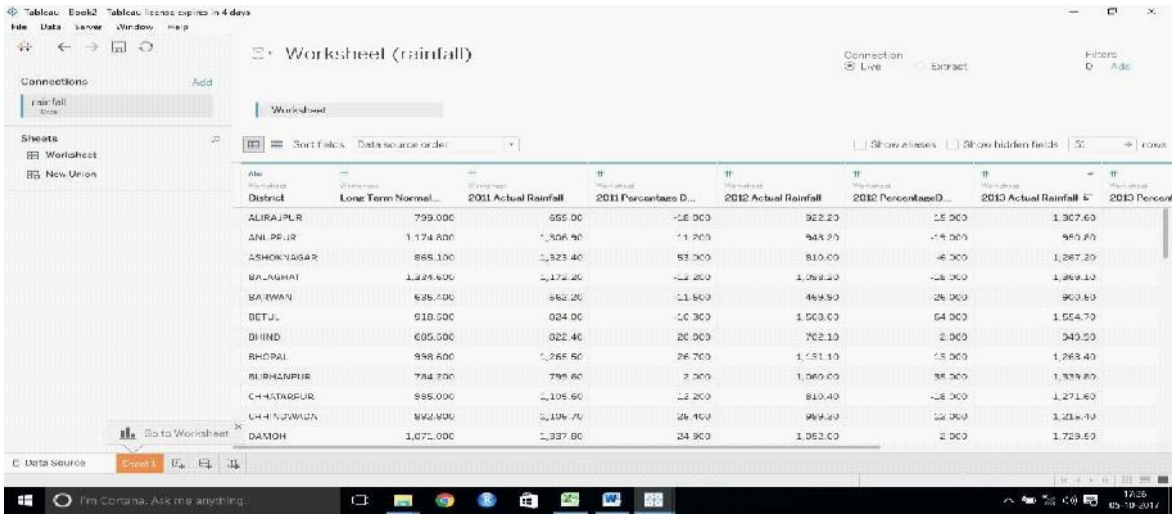

Fig. 3: Data set No.2

3. Tableau Tool

In 2020 the world will generate 50 times the amount of data as in 2011. And 75 times the number of information sources (IDC, 2011). Within these data are huge, unparalleled Opportunities for human advancement. But to turn opportunities into reality, people need the power of data at their fingertips. Tableau is building software to deliver exactly that. Tableau Software is an American computer software company headquartered in Seattle, WA,USA. It produces a family of interactive data visualization products focused on business intelligence.

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 6, Issue 11
November 2017

4. Tableau public:

Is for anyone who wants to tell stories with interactive data on the web. It is delivered as a service that allows you to be up and running overnight. With Tableau Public you can create amazing interactive visuals and publish them quickly, without the help of programmers or IT.

5. Data Join:

Data joining is a very common requirement in any data analysis. You may need to join data from multiple sources or join data from different tables in a single source. Tableau provides the feature to join the table by using the data pane available under Edit Data Source in the Data menu.

In Worksheet Fig.4 join of both datasets are taken using join function and data is displayed in the form of joins where rainfall and crop production is joined.

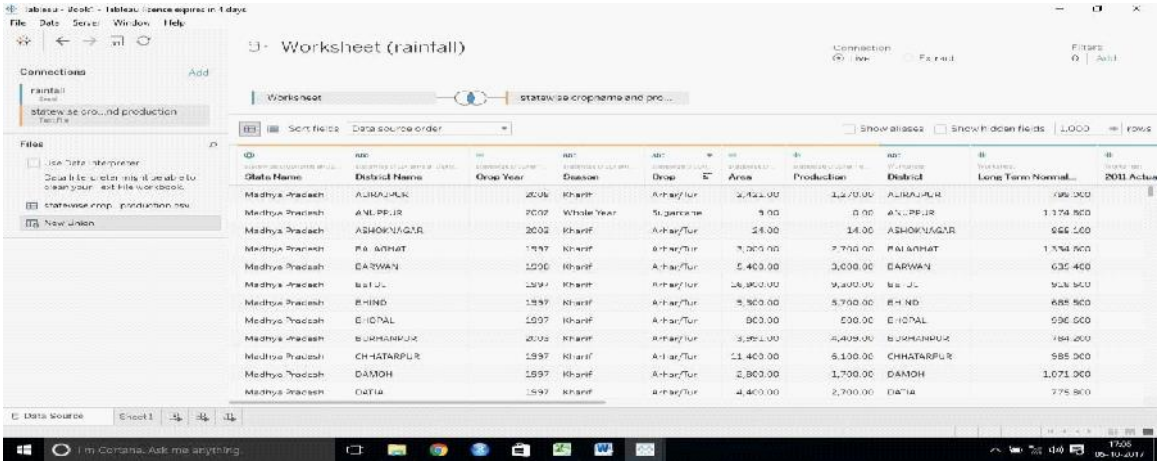

Fig. 4: Tableau Join

6. Tableau Analysis for Data Pre-processing:

As Tableau helps in analyzing lots of data over diverse time periods, dimensions, and measures, it needs a very meticulous planning to create a good dashboard or story.

Tableau is an analysis tool which helps us to analyses the given dataset in graphical manner . Here in Fig.6.2.2 the X-axis shows actual rainfall and production for a specific crop species and for specific district[9].

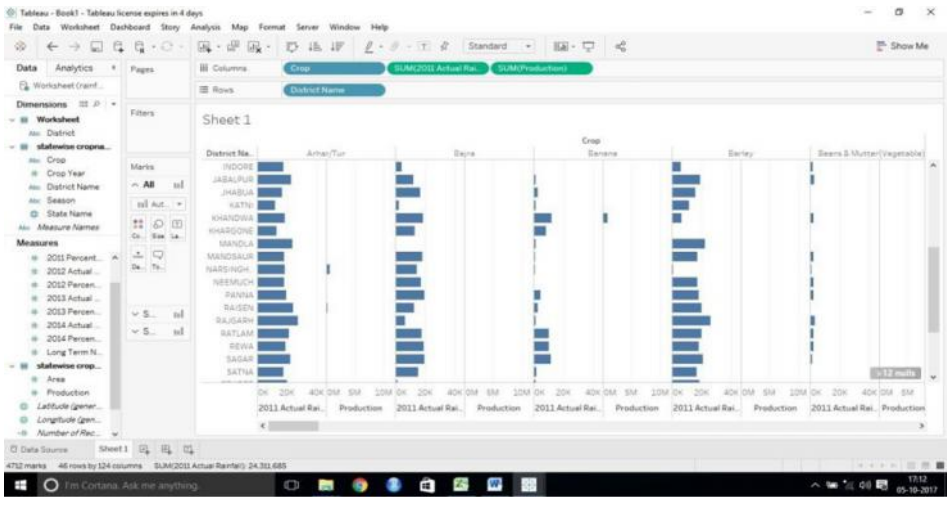

Fig. 5: Tableau Analysis

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 6, Issue 11
November 2017

From this rainfall and production sum of a district for specific area is obtained which are the essential results for analysis.

7. Processed Data:

Following is the input given to K-Means clustering algorithm. It is pre-processed according to the State name, district name, crop and the production in that area. Using the above input, k clusters will be formed. These clusters will formed by calculating the distance between the input dataset, number of clusters and the centroid. The distance is calculated using Euclidean

Distance Formula. The centroid can be any value from the dataset or random value around which the datasets have to be clustered

A	B	C	D
State_Name	District_Name	Crop	Production
Andaman and Nicobar Islands	Nicobar	Rice	321
Andaman and Nicobar Islands	Nicobar	Rice	300
Andaman and Nicobar Islands	Nicobar	Rice	511
Andaman and Nicobar Islands	Nicobar	Rice	90
Andaman and Nicobar Islands	Nicobar	Rice	73
Andaman and Nicobar Islands	Nicobar	Rice	12
Andaman and Nicobar Islands	Nicobar	Rice	10
Andaman and Nicobar Islands	Nicobar	Rice	318

Fig. 6: Processed Data Set

) Module 2 : K-Means clustering

K-means clustering is a partition-based cluster analysis method. According to this algorithm we firstly select k data value as initial cluster centers, then calculate the distance between each data value and each cluster center and assign it to the closest cluster, update the averages of all clusters, repeat this process until the criterion is not match. K-means clustering aims to partition data into k clusters in which each data value belongs to the cluster with the nearest mean [5]. According to the basic K-means clustering algorithm, clusters are fully dependent on the selection of the initial clusters centroids. K data elements are selected as initial centers; then distances of all data elements are calculated by Euclidean distance formula. Data elements having less distance to centroids are moved to the appropriate cluster. The process is continued until no more changes occur in clusters [k-1]. This partitioning clustering is most popular and fundamental technique. It is vastly used clustering technique which requires user specified parameters like number of clusters k, cluster initialization and cluster metric. First it needs to define initial clusters which makes subsets (or groups) of nearest points (from centroid) inside the data set and these subsets (or groups) called clusters. Secondly, it finds means value for each cluster and define new centroid to allocate data points to this new centroid and this iterative process will goes on until centroid does not changes [8].

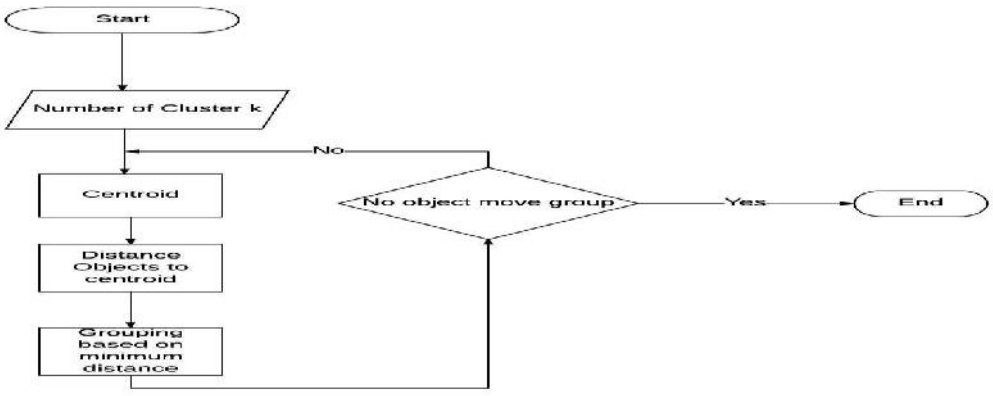

Fig. 7: K-means Clustering Flowchart.

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 6, Issue 11
November 2017

The processed data is given as input to the K-means clustering algorithms which forms clusters of data according to the centroid and sends the data to next module based on the attributes of agriculture [5].

IMPORTANT EQUATION:

Distance measure will determine how the similarity of two elements is calculated and it will influence the shape of the clusters [7].

1. The Euclidean distance is given by:

$$\text{Distance} = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

where, n is the number of dimensions (attributes) and $\mathbf{p_k}$ and $\mathbf{q_k}$ are, respectively, the $\mathbf{k^{th}}$ attributes (components) or data objects **p** and **q**.

K-MEANS CLUSTERING ALGORITHM

8. If k is given, the K-means algorithm can be executed in the following steps:

9. Partition of objects into k non-empty subsets

10. Identifying the cluster centroids (mean point) of the current partition.

11. Assigning each point to a specific cluster

12. Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum[7].

After re-allotting the points, find the centroid of the new cluster formed.

```
K-MEANS({x_1, ..., x_N}, K)
 1  (s_1, s_2, ..., s_K) ← SELECTRANDOMSEEDS({x_1, ..., x_N}, K)
 2  for k ← 1 to K
 3  do μ_k ← s_k
 4  while stopping criterion has not been met
 5  do for k ← 1 to K
 6       do ω_k ← {}
 7       for n ← 1 to N
 8       do j ← arg min_j' |μ_j' − x_n|
 9            ω_j ← ω_j ∪ {x_n}  (reassignment of vectors)
10       for k ← 1 to K
11       do μ_k ← (1/|ω_k|) Σ_{x∈ω_k} x  (recomputation of centroids)
12  return {μ_1, ..., μ_K}
```

Fig. 8: K-means Clustering Algorithm

Module 3: Apriori

The Apriori algorithm is the original algorithm of Boolean association rules of mining frequent item sets, raised by R. Agrawal and R. Srikan in 1994. Subsets of frequent itemsets are frequent item sets and the supersets of infrequent itemsets are infrequent item sets[6]. Association rules the data mining

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 6, Issue 11
November 2017

classification.

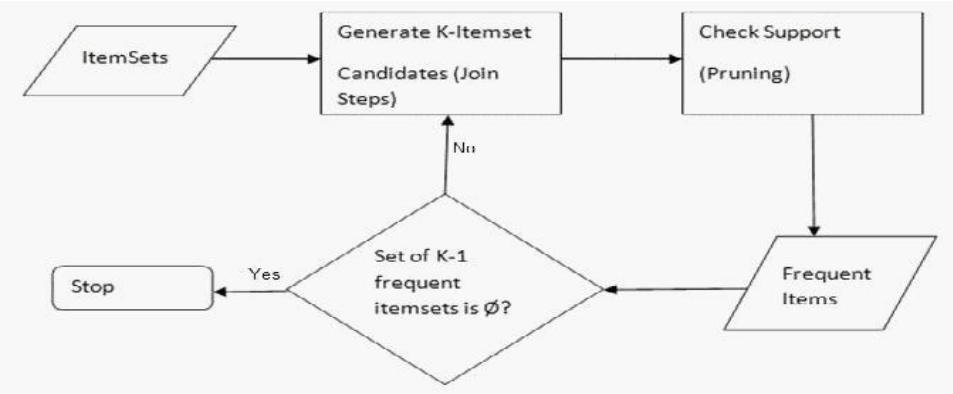

Fig. 9: Apriori working

The algorithm is used to find out all the frequent item sets. In the first iteration, item set A directly constitutes the first candidate itemset C1[6]. Assume that A={a1,a2,…,am}, then C1={{a1},{a2},…,{am}}. In the Kth iteration, firstly, the candidate item set Ck of this iteration emerges according to the frequent item set Lk-1 found in the last iteration. (The candidate item set's potential frequent item set and is the superset of the K-1th frequent item set. Item set with k candidate item sets is expressed as Ck, which was consisted by k frequent item sets Lk.) [6]Then distribute a counter which has a initial value equals to zero to ever item set and scan affairs in database D in proper order[6]. Make sure every affairs belongs to each item sets and the counter of these item sets will increase. When all the affairs have been scan, the support level can be gotten according to the actual value of |D| and the minimum support level of the certain Ck of the frequent item set[6]. Repeat the process until no new item occurs.

J Module 4: Experimental Results

Here, c1, c2, c3 are the clusters which are formed by giving input dataset, number of clusters and centroid. Clusters c1, c2, c3 show minimum, average and maximum crop yield prediction, of rice in Nicobar district of Andaman and Nicobar islands respectively.

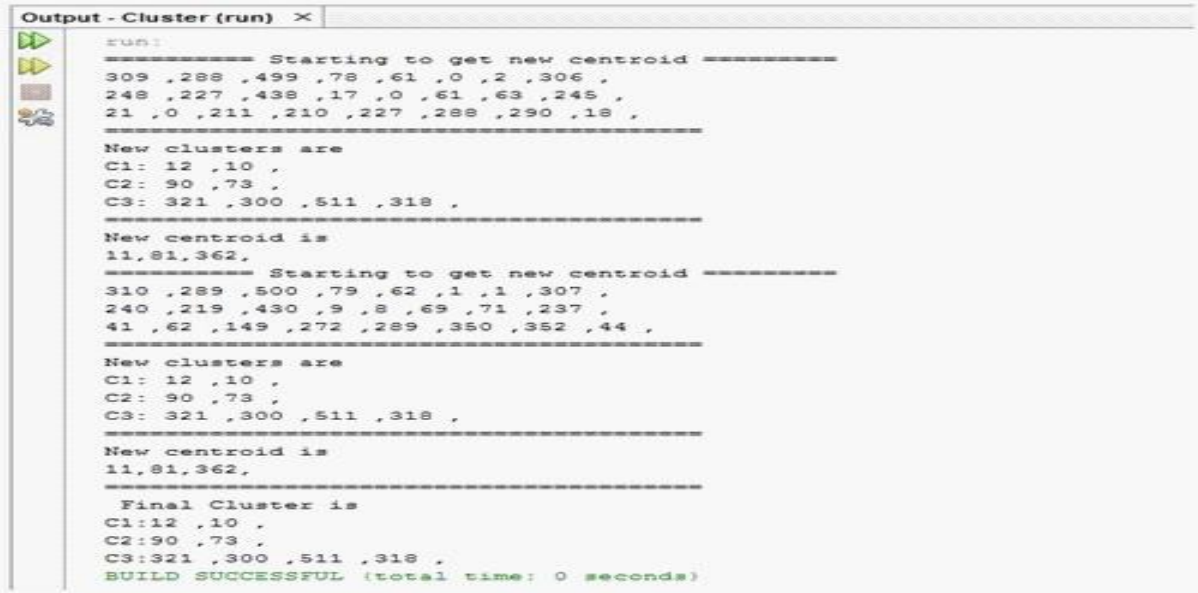

Fig. 10: Output clusters

International Journal of Computer & Mathematical Sciences
IJCMS
ISSN 2347 – 8527
Volume 6, Issue 11
November 2017

EXISTING CROP PRODUCTION SYSTEM

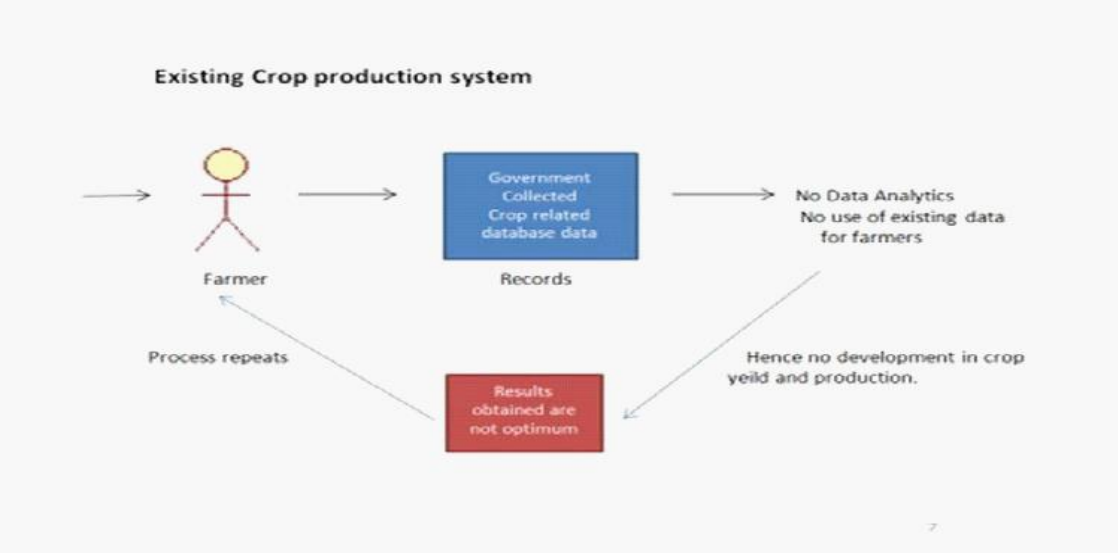

Fig. 11: Existing Crop Production System. PROPOSED CROP PRODUCTION SYSTEM

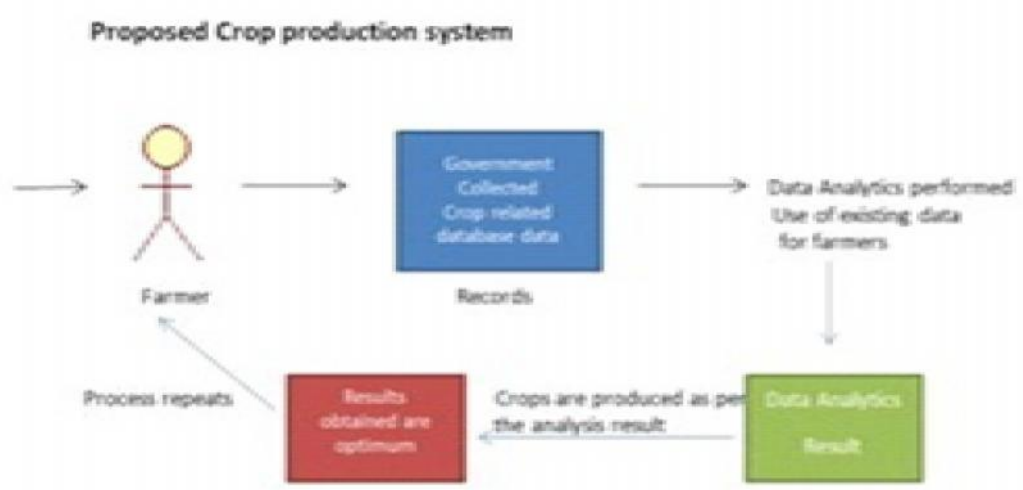

Fig. 12: Proposed Crop Production System. CONCLUSION

We have studied Dataset (Agriculture production of different food grains from year 2003 to 2014 at all India level and Rainfall of India) of agriculture and to this data we have applied K-Means Clustering algorithm to take reduced data as input and then store this data into clusters. Data stored in clusters will facilitate fast search in less time based on cluster hypothesis. The proposed work will help farmers to increase the yield of their crops. The future work involves the following:

1. Storage of big data in clusters by using various clustering algorithms, reduce it to appropriate/valid content using K-Means clustering algorithm
2. Apriori algorithm helps to count frequently occurring features which helps to predict crop yield for specific location.

REFERENCES

[1] Ruchita Thombare ,Shreya Bhosale,Prasanna Dhemey, Anagha Chaudhari,"*Big Data Aanalytics for climate smart Agriculture*",April 2017.
[2] Athmaja S., Hanumanthappa M, "*Applications of Mobile Cloud Computing and Big data Analytics in Agriculture Sector: A survey*", October 2016.
[3] P. Surya, Dr. I. Laurence and M. Ashok Kumar, "*The role of big data analytics in agriculture sector: A survey*", March 2016
[4] Tripathi S, Srinivas V V, Nanjundiah R S, "*Downscaling of Precipitation for Climate Change Scenarios: A Support Vector Machine Approach*", J Hydrol, 2006, pages : 621-640.
[5] Rajagopalan B, Lall U, *"A K-Nearest Neighbour Simulator for Daily Precipitation and Other Weather Variables*", Wat Res Res 35(10), 1999, pages : 3089-3101.
[6] Jiao Yabing,"*Research of an Improved Apriori Algorithm in Data Mining Association Rules"*,2013
[7] Jyoti Yadav, Monika Sharma, "*A Review of K-mean Algorithm*", International Journal of Engineering Trends and Technology, Volume 4, Issue 7, July 2013
[8] Ms.P. Kanjana 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] https://data.gov.in/catalog/agriculture-production-stock-yield
[10] https://data.gov.in/catalog/rainfall-india