Crop Yield Prediction Using Data Analytics and Hybrid Approach

Ms. Shreya V. Bhosale
Student, Department of Information Technology,
Pimpri Chinchwad College of Engineering,
Pune, India
shreyab6541@gmail.com

Ms. Ruchita A. Thombare
Student, Department of Information Technology,
Pimpri Chinchwad College of Engineering,
Pune, India
ruchitathombare@gmail.com

Mr. Prasanna G. Dhemey
Student, Department of Information Technology,
Pimpri Chinchwad College of Engineering,
Pune, India
prasannadhemey@gmail.com

Ms. Anagha N. Chaudhari
Asst. Professor, Department of Information Technology,
Pimpri Chinchwad College of Engineering,
Pune, India
anagha.chaudhari@gmail.com

***Abstract*—Agricultural data is being produced constantly and enourmosly. As a result, agricultural data has come in the era of big data. Smart technologies contribute in data collection using electronic devices. In our project we are going to analyse and mine this agricultural data to get useful results using technologies like data analytics and machine learning and this result will be given to farmers for better crop yield in terms of efficiency and productivity.**

Keywords—Big data, K-Means clustering, Apriori, Naive Bayes (key words), Agriculture

I. INTRODUCTION

Indian Economy has Agriculture as its backbone. In India, agricultural yield is primarily depends on weather conditions. Rice cultivation is majorly depends on rainfall. In this context, timely advice to predict the future crop productivity and an analysis is to be made in order to help the farmers to maximize the crop production of crops [3]. Yield prediction is an important agricultural problem. Earlier Farmers used to predict their yield from past yield experiences. Thus, for such kind of data analytics in crop prediction, there are different techniques or algorithms, and with the help of those algorithms we can predict crop yield [3].

1. K-means Clustering
2. Apriori Algorithm
3. Naïve Bayes algorithm

Using all these algorithms and with the help of inter-relation between them, there are growing range of applications and the role of Big data analytics techniques in agriculture.

II. SYSTEM ARCHITECTURE

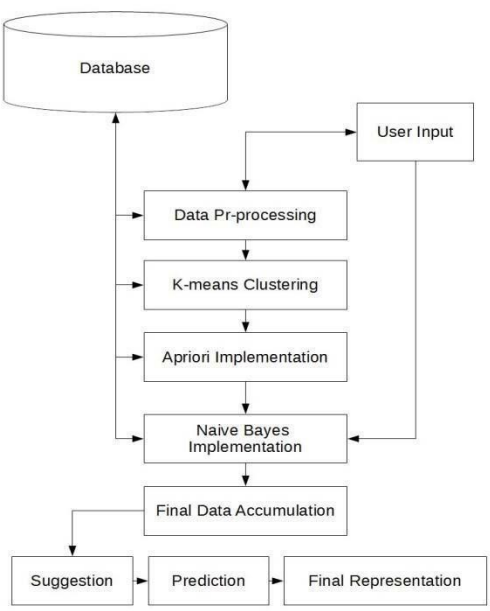

Fig 1. System Architecture

Module 1: Database

Database module contains attributes like district, land area, Soil type, Year, Season, Crop Name, Production and Rainfall. This database is used for analysis and further sent to Data Pre- processing.

A. Module 2: Data Pre-processing

In Data Pre-processing module data is cleaned and only necessary attributes are taken for further analysis.

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

B. Module 3:Techniques

Technique 1: K-Means Clustering

K-mean is an unsupervised, non-deterministic, numerical, iterative method of clustering. In k-mean each cluster is represented by the mean value of objects in the cluster [5]. Data objects are similar to one another within the same cluster and dissimilar to the objects in other clusters Cluster analysis is grouping a set of data objects into clusters [5].

K-means clustering is a partition-based cluster analysis method. In this algorithm we firstly select k data value as initial cluster centres, then calculate the distance between each data value and each cluster centre and assign it to the nearest cluster, update the averages of all clusters, repeat this process until the criterion is not match. K-means clustering partitions data into k clusters in which each data value belongs to the cluster with the nearest mean [11].

Our goal is to predict k centroids and a label c (i) c (i) for each data point. The k-means clustering algorithm is as follows:

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.

2. Repeat until convergence: {

For every i, set

$$c^{(i)} := \arg\min_j \|x^{(i)} - \mu_j\|^2.$$

For each j, set

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

}

Fig 2. K-Means Formulae

Technique 2: Apriori Implementation

The Apriori Algorithm is an algorithm used for mining frequent item sets for Boolean association rules [12]. Apriori Algorithm works as following:

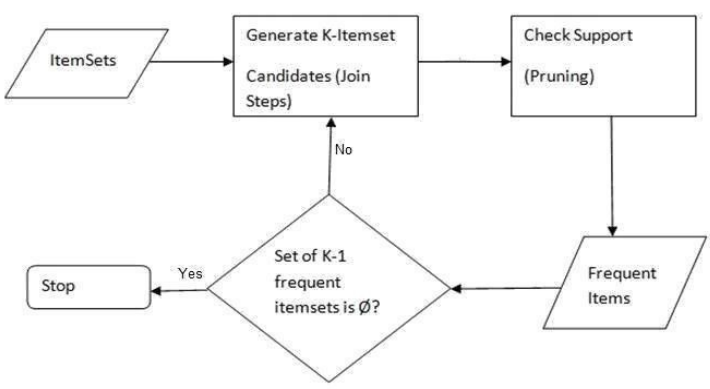

Fig 3. Apriori Flowchart

1: $k = 1$.
2: $F_k = \{ i \mid i \in I \wedge \sigma(\{i\}) \geq N \times minsup \}$. {Find all frequent 1-itemsets}
3: **repeat**
4: $k = k + 1$.
5: $C_k = \text{apriori-gen}(F_{k-1})$. {Generate candidate itemsets}
6: **for** each transaction $t \in T$ **do**
7: $C_t = \text{subset}(C_k, t)$. {Identify all candidates that belong to t}
8: **for** each candidate itemset $c \in C_t$ **do**
9: $\sigma(c) = \sigma(c) + 1$. {Increment support count}
10: **end for**
11: **end for**
12: $F_k = \{ c \mid c \in C_k \wedge \sigma(c) \geq N \times minsup \}$. {Extract the frequent k-itemsets}
13: **until** $F_k = \emptyset$
14: Result $= \bigcup F_k$.

Frequent Item sets: All the sets which contain the item with the minimum support (denoted by Li for ith item set). In our module prediction of crops is obtained as output [15].

Technique 3: Naive Bayes Implementation

Understanding and evaluating many learning algorithms Bayesian Classification is a useful technique. Practical learning algorithms and prior knowledge can be delivered with Bayes classification. Here observed data can be combined. It calculates explicit probabilities and it is vigorous to noise in input data [13]. The probability of crops that will be grown in that acre of land is calculated.

Mathematical Model for Naive Bayes Classification [17].
P (h/D) = (P (D/h) P (h))/ P (D) P (h):
Prior probability of hypothesis h
P (D): Prior probability of training data D
P (h/D): Probability of h given D
P (D/h): Probability of D given h

Fig 4. Naive Bayes Phases

C. Module 3: Final Accumulation

In Final Accumulation model all the results from K-Means clustering, Apriori and Naive Bayes are collected together and sent to next model (GUI).

D. Module 4: Graphical User Interface

1. Suggestion:

Suggestions are the crop names suggested to the farmer for better crop yield.

2. Prediction:

Prediction is result of Apriori and Naïve Bayes which predicts the crop yield in quintals.

3. Final Representation:

Final representation represents the graphical result of K-Means and Naive Bayes which is helpful for Analysis of crops in specified rainfall.

III. ANALYSIS ANDRESULTS

A. Tableau Analysis:

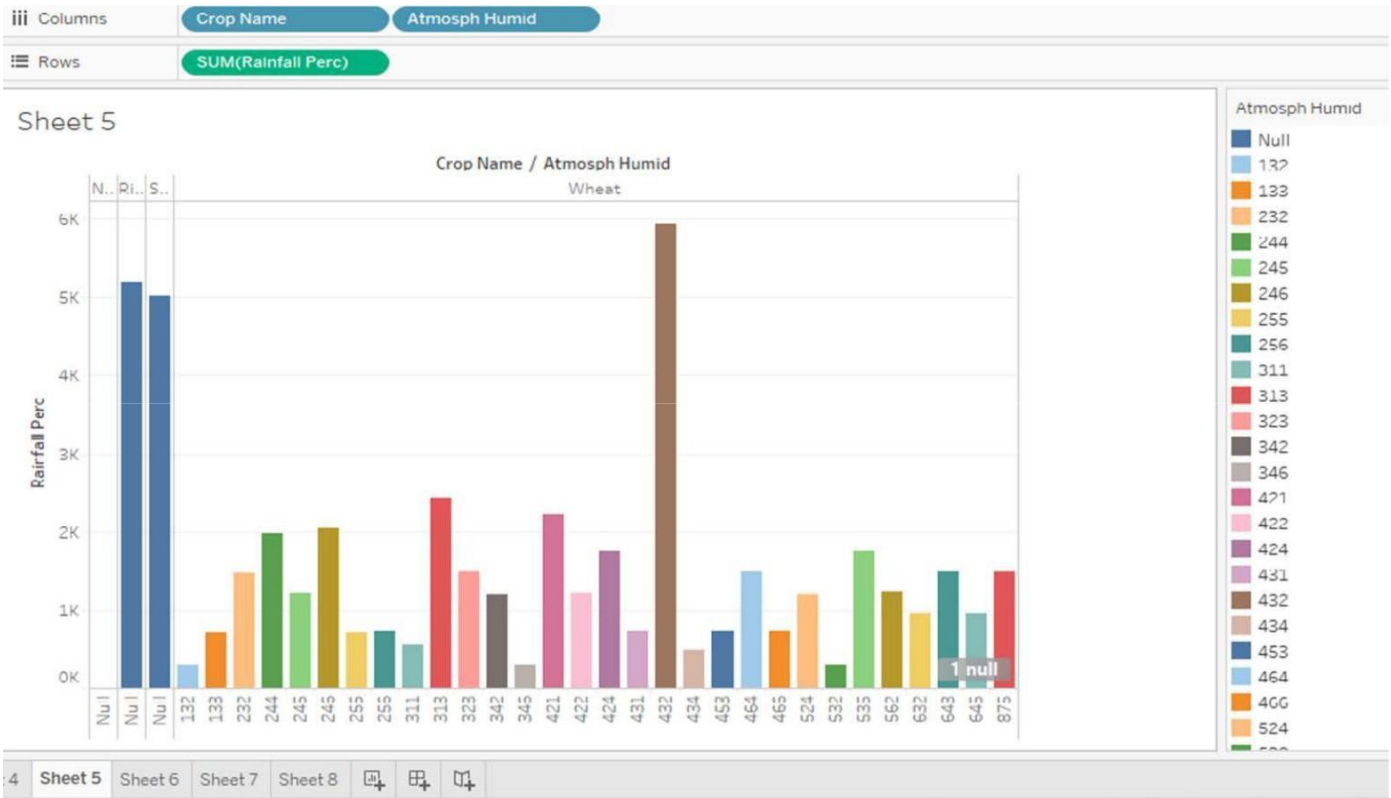

Fig 5. Tableau Analysis

For Analysis tableau tool is used in our project. Tableau tool defines the dataset used for analysis by using graphs and pie – charts which further can be classified using K-means clustering. The clustered data is used as input to Apriori and naive Bayes. Following are the results shown.

B. Input:

The input to our system consists of following attributes: Place (Area), Yield (in kilograms), Area of farm (in Acres), Soil Type, Rainfall level (in mm).All these parameters are used for segregation as well as analysis in K-Means, Apriori and Naive Bayes.

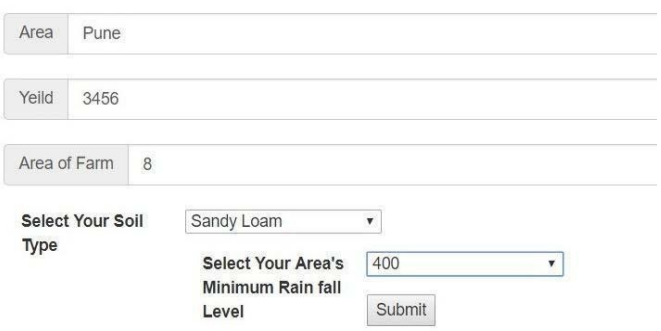

Fig 6. Input of System

C. K-Means Clustering Results

K-Means Clustering is more elaborated in graphs, Hence the following graph shows the K-Means result of our system. Here the purple dot is one centroid and second centroid is denoted by red dot. The X-axis represents Crops while the Y-axis represents Yield per Hectare.

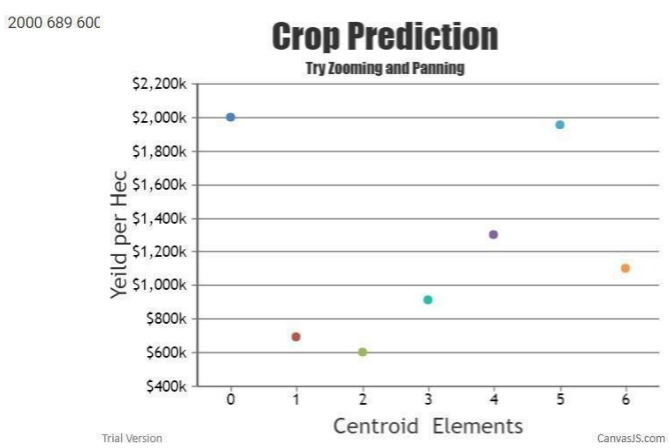

Fig 7. K-Means result graph

D. Apriori Results:

Apriori uses support to mine frequent item set [12]. Here in the following diagram Crops and their support are calculated as shown. And finally crops having maximum support are selected for further results.

```
1: [Black gram], support: 0.4
Added in Lis [Black gram]
2: [Sugarcane], support: 0.6
Added in Lis [Sugarcane]
3:  [Rice], support: 0.6
4: [Moong (green Gram)], support: 0.4
5: [Sugarcane, Black gram], support: 0.4
6: [Black gram, Rice], support: 0.4
7: [Moong (green Gram), Black gram], support: 0.3
8: [Sugarcane, Rice], support: 0.6
9: [Moong (green Gram), Sugarcane], support: 0.4
10: [Moong (green Gram), Rice], support: 0.4
11: [Sugarcane, Black gram, Rice], support: 0.4
12: [Moong (green Gram), Sugarcane, Black gram], support: 0.3
13: [Moong (green Gram), Sugarcane, Rice], support: 0.4
#################
Appriori List Crops [Black gram]
Appriori List Crops [Sugarcane]
```

Fig 8. Apriori Results

E. Naive Bayes Results

Naive Bayes uses probability of crop being grown in those circumstances. Hence in the following result probabilities are calculated and most probable crops are chosen for further final accumulation.

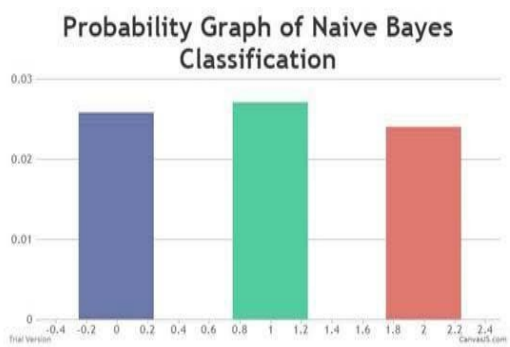

Fig 9. Naive Bayes Classification Graph

```
Final Crop after Navie Bayes Calculation|Jowar
Final Crop after Navie Bayes Calculation|linseed
Final Crop after Navie Bayes Calculation|banana
Final Crop after Navie Bayes Calculation|Jowar
Final Crop after Navie Bayes Calculation|linseed
Final Crop after Navie Bayes Calculation|banana
0.023148147389292717
0.0243055559694767
0.026145223528146744
From 1st ListArhar/Tur
From 1st ListSesamum
From 1st ListGram
From 1st ListUrad
Final Output after Complete Analysis|Black gram
Final Output after Complete Analysis|Sugarcane
Final Output after Complete Analysis|Jowar
Final Output after Complete Analysis|linseed
Final Output after Complete Analysis|banana
Final Output after Complete Analysis|Arhar/Tur
Final Output after Complete Analysis|Sesamum
Final Output after Complete Analysis|Gram
Final Output after Complete Analysis|Urad
List from DB|[Bean.SoilInfoBean@474a9779]
Rainfall not in Range forBlack gram
List from DB|[Bean.SoilInfoBean@b7dbc3d]
Rainfall not in Range forSugarcane
List from DB|[Bean.SoilInfoBean@16bf0f0f]
Yield for Jowar is6976
List from DB|[Bean.SoilInfoBean@608f6766]
Yield for linseed is8000
List from DB|[Bean.SoilInfoBean@d903cbe]
Rainfall not in Range for banana
List from DB|[Bean.SoilInfoBean@13e9f0d]
Rainfall not in Range for Arhar/Tur
List from DB|[Bean.SoilInfoBean@2a2cb1f]
Rainfall not in Range for Sesamum
List from DB|[Bean.SoilInfoBean@63428020]
```

Fig 10. Naive Bayes Result

F. Final Result

Final result contains crop names which is suggested in that region for specified rainfall as well as land of farmer in acres. The predicted yield as crop count attribute is displayed in kg/acre format. The attribute yield describes the average production of that crop in 1 acre.

Crop	Crop Count According to land (kg/acr)	Rainfall Required	Yeild (quin/acr)
Jowar	6976	400	3
linseed	8000	200	4

Fig 11. Final Result

IV. CONCLUSION

We have studied and identified the problems faced by the farmers in India also collected and studied Agricultural datasets available online. Study of Data Visualization tool: Tableau, let us join two datasets on a year parameter. We studied K-means clustering which was used to create clusters. Data stored in clusters facilitated fast search in less time based on cluster hypothesis. The work will help farmers to increase the yield of their crops.Storage of big data in clusters by using K-means clustering algorithm, reduce it to appropriate/valid content using the algorithm. Apriori algorithm helped to count frequently occurring features which helped to predict crop yield for specific location. Also implemented Naive Bayes algorithm for finding out the exact crop. Thus, we implemented a system which will predict the crop name and approximate yield in a particular farm.

REFERENCES

[1] Athmaja S., Hanumanthappa M, “Applications of Mobile Cloud Computing and Big data Analytics in Agriculture Sector: A survey”, October 2016.

[2] P. Surya, Dr. I. Laurence and M. Ashok Kumar, “The role of big data analytics in agriculture sector: A survey”, March 2016

[3] D Ramesh, B Vishnu Vardhan, “Analysis of crop yield prediction using data mining techniques”, January 2015

[4] CCAFS, “Big Data for climate-smart agriculture”, 2015.

[5] Monika Sharma, Jyoti Yadav “A Review of K-mean Algorithm”, 2015

[6] Jorquera H, Perez R, Cipriano A, Acuna G, "Short Term Forecasting of Air Pollution Episodes", In: Zannetti P (eds) Environmental modeling , WIT Press, UK, 2001.

[7] 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.

[8] 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.

[9] Verheyen K, Adriaens D, Hermy M, Deckers S, "High-resolution continuous soil classification using morphological soil profile descriptions", Geoderma Vol.101, 2001, pages: 31-48.

[10] Ms.P. Kanjana Devi, “Enhanced Crop Yield Prediction and Soil Data Analysis Using Data Mining”

[11] Odilia Yim, Kylee T. Ramdeen, “Hierarchical Cluster “Analysis: Comparison of Three Linkage Measures and Application to Psychological Data”, October 2015

[12] Jiao Yabing,”Research of an Improved Apriori Algorithm in Data Mining Association Rules”, 2013

[13] Tellaeche A, X P Burgos Artizzu, G Pajares, A Ribeiro, "A Vision-Based Classifier for Weeds Detection in Precision Agriculture through the Bayesian and Fuzzy K-Means Paradigms", Adv.Soft. Comp., 2008, pages: 72-79.

[14] Mehta D R, Kalola A D, Saradava D A, Yusufzai A S, "Rainfall Variability Analysis and Its Impact on Crop Productivity - A Case Study", Indian Journal of Agricultural Research, Volume 36, Issue 1, 2002, pages : 29-33

[15] KavyashriD,,dwgeek.com/mining-frequent-itemsets-apriori-algorithm.html/, March 2018

[16] V.Vidhya Rani,Dr. I. Elizabeth Shanthi “A Study on Efficient Data Mining Approach on Compressed Transaction”