International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-03, Issue-09, Dec 2017

Applying Data Mining Technique to Predict Annual Yield of Major Crops of Different Districts in Maharashtra

[1]Prof. B K Patil, [2]Praful S. Tayde

[1]Asst Professor, [2]P.G. Student, Department of Computer Science & Engineering, Everest College of Engineering & Technology, Aurangabad, Maharashtra, India

***Abstract*: - Agriculture is one of the major revenue producing sectors of India. The Agricultural yield is primarily depends on weather conditions, pests and planning of harvest operation. Climate and other environmental changes impact on agricultural economy of any country. Production of crop normally depends on factors like biology, climate, economy and geography, this factors lead to impacts on agriculture. Accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management. By applying different methodologies and techniques on yields of crops, it is possible to obtain information about which help to government organization for making good decisions and applying different policies, also it's helpful for farmers to make better plan to increase the production.**

Keywords: -* Agriculture, Data mining, crop productivity, *crop analysis, yield prediction, K-means, K-Nearest Neighbor (KNN);

I. INTRODUCTION

Agriculture is broadest economic sector of India. It plays as significant role in the overall socio-economic fabric of India As with increasing economic sector; the economic Contribution of farm is getting decreased. As we know agriculture is main economic sector of India which plays an important role for economic growth. India is the world's largest producer of many crops like fresh fruits and vegetables, milk, fabric crops, several other crops such as castor oil seeds. India is at second position in production of wheat and rice.

Maharashtra has typical monsoon climate, with hot, rainy and cold weather seasons. Tropical conditions prevail all over the state.

Summer: March, April and May are the hottest months.

Rainy Season: Rainfall starts normally in the first week of June. July is the wettest month in Maharashtra, while August too gets substantial rain. Monsoon starts its retreat with the coming of September from the state.

Winter: Winter is Cool dry spell and pleasant weather prevails from November to February.

Eastern part of Maharashtra sometimes receives some rainfall. Temperature varies between 12°C-34°C during this season.

Maharashtra's different district has varying climates and so it is very important to consider environmental factors of these separate areas. This will help to choose the best districts for cultivation of different type of crops.

Rainfall also varies from district to district and this has a huge impact on farming because while too little or too much rain can kill crops, the proper amount of rain leads to an ideal crop yield. With rainfall comes humidity and since rainfall varies from district to district so does humidity.

Humidity causes changes in the level of water that can be absorbed by atmosphere which can cause crops to remain too wet or too dry and so to get proper yield, a district with an ideal average annual rainfall and humidity is required. Pesticide is most important part in farming. Without pesticides crops would die significantly more due to insects and other pests leading to a sudden drop in yield. Too much pesticides may affect the crop on its own while too little may not get rid of pests. So, the amount of pesticides required by crops is a very important parameter.

Common specific problem that occurs is yield prediction. As early into the growing season as possible, a farmer is interested in knowing how much yield he is about to expect. [6]

In our project research, we have considered various environmental environmental (weather), biotic (pH, soil salinity) and the areas of production are the factors for crop in different districts of Maharashtra. Taking this into consideration we developed a database for various districts, for this we applies clustering techniques to partition regions, and then we apply suitable predictive algorithm to obtain crop yield predictions.

II. RELATED WORK

In [1] Raorane A.A. and Kulkarni .V. Research paper they focus on the data mining as the tool for yield estimation. They said actually accurate information about the nature of historical yield of crop is important modeling input, which are helpful to farmers & Government organization for decision making process in establishing proper policies.

Ramesh and Vardhan [4] deal with the challenge of predicting the yield of various crops. One approach to this problem is to employ data mining techniques. In this paper, different types of data mining methods were applied and then evaluated on the datasets we prepared.

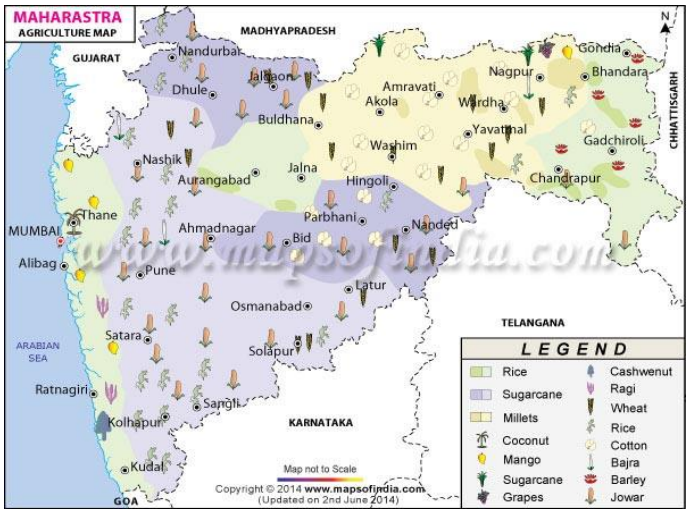
In [5] Murynin et al. study the dependency between the prediction and the accuracy of the forecast. The linear model is selected as a basic approach of yield prediction. Then, the model is extended with non-linear attributes in order to improve the accuracy of the prediction. The extensions take into consideration long-term technological advances in agricultural productivity as well as regional variations in yields. The accuracy of the model has been estimated based on the time period between the moment of the forecast formation and the time of harvest.

Jeysenthil.KMS, Manikandan.T, Murali.E implemented "Third Generation Agricultural Support System Development Using Data mining". They had implemented various classification techniques of data mining and apply them to a soil. In data mining conception, clustering and classification technique produced better solution to the farmers about their cultivation.[2]

III. MOTIVATION

Maharashtra is the third largest state in area and second largest state in population of India. It has an area of 307,713 sq. km. with 36 districts, 358 blocks and 43711 villages and a population of 112,372,972. The 45% population of the state is urban.

Agriculture is the mainstay of the state of Maharashtra. Maharashtra's economy is predominantly agrarian. It is the main occupation of the people. Both food crops and cash crops are grown in the state. Principal crops include rice, jowar, bajra, wheat, pulses, turmeric, onions, cotton, sugarcane and several oil seeds including groundnut, sunflower and soybean. The state has huge areas, under fruit cultivation of which mangoes, bananas, grapes, and oranges are the main ones. [6]



The total irrigated area which has been used for crop cultivation is 33, 500 square kilometers. The agriculture in state is predominantly rain-fed.

The state has 24 per cent of drought—prone area of the country. However state has potential for growth in agricultural sector in spite of challenges.

IV. DATA SET

The dataset used in this project has been collected from Department of agriculture of Maharashtra State.

From the dataset, we pre-processed and selected only the attributes which are important for our project rainfall, temperature humidity etc. And also cultivated area for every crop considered according to the districts. Also collected data from http://www.mahaagri.gov.in/, http://krishi.maharashtra.gov.in, http://www.maharain.gov.in

V. INPUT VARIABLES

.**The Climatic variables:**

I) Max Temperature: Temperature variation in year puts a great impact in that year's crop production. Hence we consider both the maximum as well as the minimum temperature.

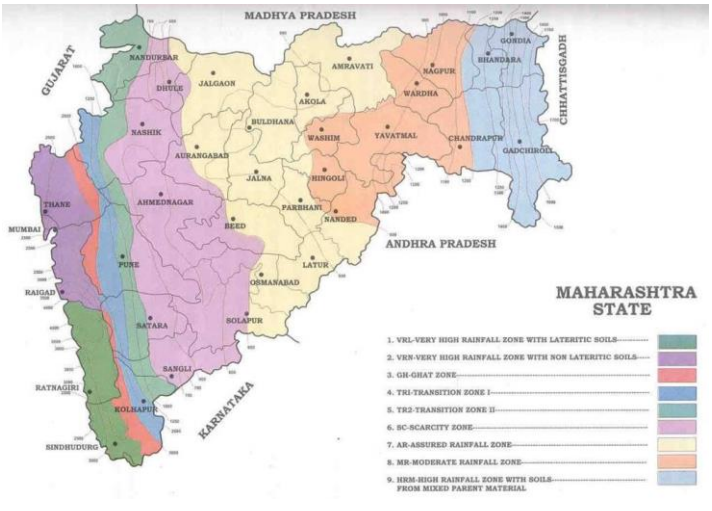
II) Min Temperature: The average yearly minimum Temperature considered in Celsius.

III) Rainfall: The year that has the highest average rainfall leads for maximum crop yield production in that Year. The average yearly rainfall was considered by calculating average from the monthly rainfall (mm) of each district.

IV) Humidity: Average yearly humidity for each district is considered in percentage.

V) Average Sunshine: This attribute was considered in hours as a yearly average for each district.

The amount of sunshine received on areas each year greatly effects the production of green crops as it directly affects the photo-synthesis process in plants.



b) The biotic input attributes:

i) Max pH: pH value is important because it indicate how acidic soil is. It's scaled is defined by a value of 7, where soil pH above 7 meaning alkaline and below 7 meaning acidic. Crop production is highly affected by the variations of pH in soil.

ii) Min pH: Minimum pH of a district's soil.

iii) Soil Salinity: Soil salinity means the amount or content of salt in soil. It measure in MMHOS/cm, it has ranges (<2), (2-4), (4-8) and (8-15). Salinization process increase salt content. Too high soil salinity can cause a Detrimental effect towards crop production and Yield.

c) Central Area:

i) Irrigated Area:
Crop production is depend on the actual area of land that has been irrigated throughout the year. Hence irrigated area is considered for the selected districts in hectares.

ii) Cultivated Area:
The area that has been used to cultivate each crop also regulates the amount of Production of the crop. Areas were taken in Hectare unit.

VI. METHODOLOGY

The method of our project is initially divided into two major parts: (1) Clustering (2) Classification.

We have considered a total of 36 districts of Maharashtra. In order to group the districts into distinct clusters, the assumption that we had to use was that the districts containing the similar values of relevant attributes should belong to the same cluster. According to this assumption, we categorized our selected attributes for the consideration of clustering the districts as follows:

1) Cluster Type-1 is based on the following attributes: Rainfall, minimum temperature, maximum temperature, humidity and sunshine. These are the environmental or climatic attributes considered for our research. The degree of similarity of the collection of these attributes should indicated distinct clusters for the selected districts.

2) Cluster Type-2 is based on the following attributes: soil pH and soil salinity. As discussed earlier, these biotic factors contribute largely towards the prediction of the crops.

3) Cluster Type-3 is based on irrigated area. Clustering is based on the area attributes for each district was considered because we can obtain the separate clusters based on distinct ranges of areas that were irrigated for each district.
4) Cluster Type-4 is based on the individual crop yields of rice, potato and wheat. This type of clustering was considered in order to classify the districts into separate clusters with similar crop yields and after analysis of the results, to see whether they exhibit a pattern related to effects from the selected attributes.

K-Means Clustering:
The k-means algorithm captures the insight that each point in a cluster should be near to the center of that cluster. It works like this: first we choose k, the number of clusters we want to find in the data. Then, the centers of those k clusters, called centroids, are initialized in some fashion. We recalculate each centroid's location as the mean (center) of all the points assigned to its cluster. We then iterate these steps until the centroids stop moving or equivalently until the points stop switching clusters. Clustering was used upon the selected districts according to the categorized types mentioned previously.

Clustering results were separately written to Excel files for each cluster type (1 to 4) for the convenience of result showcasing and analysis.

B. Prediction of crop yields using classification techniques:

In our project, we determined prediction results for yields of selected crops for the selected districts in India. The predictions results were obtained according to the selected input attributes using appropriate classification and regression models.
The following classification/regression models were used to obtain the crop yield prediction results:
a) Linear Regression: It is a statistical measure that can be used to determine the strength of the relationship between one dependent variable and a series of other changing variables known as independent variables (regular attributes). If independent variable contains multiple input attributes like in our research (rainfall, sunshine hours, humidity, pH etc), then it is termed as multiple linear regressions. Linear regression provides a model for the relationship between a scalar variable and one or more explanatory variables. This is done by fitting a linear equation to the observed data [3].

b) k-NN: The k-nearest neighbour algorithm compares a given test example with training examples which are similar. Each example denotes a point in dimensional space. Thus, all of the training examples are saved in an n-dimensional pattern space. K is a positive integer, usually small. For our purpose, the basic k-NN algorithm was applied. It first finds the k examples from the training set that are closest to the unknown example. Then it takes the most common occurring classification for the k examples.

c) Neural Net: An artificial neural network (ANN) is mathematical model or computational model inspired by the structure and functional aspects of biological neural networks for instance in our brains. In most cases an ANN is an adaptive system that modifies its structure based on external or internal information that flows through the network during the learning

phase. The basic neural network model consists of three layers: the input layer, the hidden layer and an output layer [3].

VII. RESULTS

Clustering Results:

A. Test Case 1:

Module Name: Import Dataset		
Action	Input	Expected Output
Select dataset	Upload Action	Dataset Upload Successfully
Result : Success		

Table 1: Test case 1

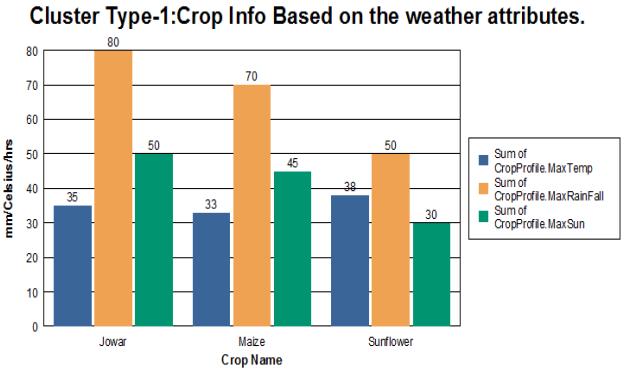
B. Test Case 2:

Module Name: Clustering		
Action	Input	Expected Output
Select pre-processing button	Cluster making	Dataset pre-processed Successfully
Result : Success		

Table 2: Test case 2

Average Result:

After applying K-means algorithm on the dataset then cluster will be generated and data is pre-processed. So these districts consider only at time of prediction. This graph shows the average of temperature, cloud, humidity, rainfall for pre-processed districts.



VIII. RECOMMENDED SYSTEM

After getting all the result graph charts and tables, then we have write a program which takes into account the necessary tables from our results to post process the data and give the best three possible crops in order of preference to choose from for farming across all the major agricultural districts. If there are not any feasible choices the program simply outputs „NONE". These recommendations are based on a combination of annual yield of that crop species per hectare area of a district.

IX. FUTURE WORKS

In our project we considered the 5 environmental variables, 2 biotic variables and also 2 area related variables to determine crop yield in the different districts. In the near future, geospatial analysis will be added to our data processing model to improve accuracy and also implement a better geographical data.

X. CONCLUSION

It is found that by applying different methodologies and techniques on yields of crops, it is possible to predict crop yield information, which helps to government organization for making better decisions and applying different policies, also it's helpful for farmers to make better plan and increase their production. It's lead to increasing the Countries' overall profit. In our project we found that the accurate prediction of different specified crop yields across different districts will help to farmer. From this farmers will plant different crops in different districts.

REFERENCES

[1] Raorane A.A., Kulkarni R.V."Data Mining: An effective tool for yield estimation in the agricultural sector"

[2]Jeysenthil.KMS, Manikandan.T, Murali.E "Third Generation Agricultural Support System Development Using Data Mining" International Journal of Innovative Research in Science, Engineering and Technology Vol. 3, Issue 3, March 2014

[3] Ye, Nong; Data Mining: Theories, Algorithms, and Examples, CRC Press, 2013.

[4] D Ramesh , B Vishnu Vardhan."Data Mining Techniques and Applications to Agricultural Yield Data". International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 9, September 2013,pp.3477-3480.

[5] Alexander Murynin, Konstantin Gorokhovskiy and Vladimir Ignatie "Efficiency of crop yield forecasting depending on the moment of prediction based on large remote sensing data set" retrieved from http://worldcompproceedings.com/proc/p2013/DMI8036.pdf

[6] http://agricoop.nic.in/sites/default/files/Maharashtra

