Crop Yield Prediction Using Epsilon Density Based Prediction

D. Esther Rani
Assistant Professor of CSE, NBKRIST,
Vidya Nagar, Nellore Dist., Andhra Pradesh, India.
esther.divine@gmail.com
Dr.N.Sathyanarayana
Professor of IT, TKR College of Engineering & Technology,
Meerpet, Hyderabad Dist., Telangana, India.
nsn1208@gmail.com
Dr. B. Vishnu Vardhan
Professor of CSE and Vice-Principal., JNTUH College of Engineering,
Manthani, Karimnagar Dist., Telangana, India.
mailvishnu@yahoo.com

Abstract: Machine learning algorithms play a significant role in data analysis in many disciplines like Agriculture, Food, Medicine, and Twitter Data. Yield prediction is a significant agricultural problem that remains to be solved based on the available data. Earlier yield prediction is an exciting challenge, and this prediction is performed by considering farmers' knowledge of a specific field and crop. Machine learning techniques are used to increase the crop yield, where data is collected from different agricultural sectors. In machine learning, clustering plays a vital role. In this paper, various clustering techniques such as k-Means, Expectation-Maximization, Hierarchical Micro Clustering, Density-Based Clustering, Weight-based clustering are briefed, and a new clustering approach, Epsilon Density-Based Prediction(EDBP), is proposed for obtaining the best crop yield prediction.

Keywords: Clustering, Epsilon Density Based Prediction(EDBP),Expectation-Maximization, K Means, Machine Learning, Prediction.

1. Introduction

In today's world, agriculture plays a significant role in every country, and in India, its impact on the Indian economy is too high as it contributes about 17% to the total GDP (gross domestic production) and gives employment to more than60% of the population. India is continuously facing the problems of increasing crop production due to various climatic variations. Crop Yield Prediction is a significant problem, where the farmers need some prior information about the crop. To maximize crop production, provide preliminary information about the crop for the farmers. It can be done through machine learning techniques by analyzing the historical crop data. Machine learning serves to be a better choice in agriculture to predict future yield. To proceed with machine learning techniques, the dataset is to be collected from various agricultural sectors, and the collected data is used in training the model, which is used to learn how to classify to predict the crop yield.
Rice is an important food crop and most cultivated crop in India as well as in Asian countries. Rice is a primary diet in major parts of India. India ranks second after China in rice production. Clustering [8] plays a significant role in agricultural mining, where a large amount of the data is classified or grouped into a set of similar clusters.
In this present work, Epsilon Density Based Prediction (EDBP) clustering technique is implemented for crop yield prediction and to maximize the production by considering different parameters of rice.

2. Literature Review

[14] In this paper, they have discussed about predicting yield production. Yield prediction is very essential in agriculture areas for the farmers, as they couldn't predict the future crop yield because of natural calamities like drought, rainfall, temperature, weeds, insects, and so forth.

Data mining techniques like K-Means and Multiple Regression techniques are used to provide the solution for predicting yield production. This aims the data models to achieve high accuracy and a high generality in terms of yield prediction. [8] This paper summarizes the data mining strategies such as k-means, bi clustering, k nearest neighbour, Neural Networks, Support Vector Machine, and Naïve Bayes Classifier in the field of agriculture sector. Different mining strategies are discussed to deal with agricultural problems.

[15]In this paper, the authors discussed various soil types and predictions using the Multiple Linear Regression (MLR) approach for the specific region. This paper aims to find various models that obtain high accuracy and high generality in terms of yield prediction.

[16] In this paper, a survey of various mining techniques such as K-Means, K-Nearest neighbor (KNN) are used for the crop yield estimation.[20] In this paper, the survey is to provide a complete review of various forms of data mining techniques used in Precision Agriculture and discussed about Yield Prediction and soil classification strategies.

Clustering is the process of categorizing the data into classes or clusters so that objects within a cluster have high to one another but are very dissimilar to objects in other clusters. Various clustering methods are implemented and assessed by the researchers throughout the global agricultural fields for agricultural data analysis.

Clustering algorithms are used extensively in distinct areas of research, along with pattern recognition [24] [31], data mining [25] [31], classification [26] [31], data analysis and modeling [27] [31]. Clustering algorithms arise because of the need to find records that share similar characteristics in a given dataset, presently there are numerous fuzzy clustering algorithms consisting of Fuzzy C-Means (FCM) [11] [18], Possibilistic C-Means (PCM) [28] [31], Fuzzy Possibilistic C-Means (FPCM) [29] [18], and Possibilistic Fuzzy C- Means (PFCM) [30] [31]. Each data from data sets must be in reality determined to one cluster by means of the conventional clustering set of rules. Actually, the maximum matter is fuzzy in attribution, that is, the matter has no boundary absolutely. So, the theory of fuzzy clustering is more suitable to the essence of the matter. Now, the Fuzzy C-Means clustering method has become the most important clustering algorithm.

Hierarchical micro clustering algorithm, SWK k-Means algorithm, constrained k-Means algorithm, expectation-maximization algorithm, Bee Hive Algorithm, Density-Based, Weight Based are clustering techniques[4],[5],[6],[1],[9],[10]. Clustering techniques play a significant role in Information Retrieval and Text Mining [11,12]. A. Mucherino [13] applied Artificial Neural Networks, K Means, K Nearest Neighbours techniques on agricultural data sets to find important information.

The hierarchical micro-clustering algorithm introduces the clustering feature and the clustering feature tree. The clustering feature is a triplet that summarizes the information about the cluster of objects, such as the number of data points, the linear sum of data points, and the square sum of data points. The Clustering Feature Tree is a height-balanced tree that stores the clustering features for hierarchical clustering. The hierarchical micro-clustering algorithm gives a good quality cluster for a huge dataset, but it is expensive to update and store the cluster [4].

Kiri Wagstaff et.al developed HARVIST(Heterogeneous Agricultural Research Via Interactive, Scalable Technology) graphical interface that allows users to interactively run automatic classification and clustering algorithm in 2005. They have used a constrained k-means clustering algorithm for pixel clustering, which merges the concept of constraint-based and partitioning methods. It provides good quality of clusters for large datasets [5].

A Majid Awan et.al has developed an information system in 2007 for predicting Oil-Palm Yield from climate and plantation data. A weighted kernel k-means clustering algorithm is used, which

incorporated spatial constraints to spare spatial neighbourhood information. It shows the good quality of clusters for huge datasets [6].

Researchers in [10] used density-based clustering to estimate the crop yield prediction in three specific areas under some specific parameters and obtained good accuracy compared with other clustering techniques.

EM algorithm is an iterative refinement algorithm used to estimate the parameters of the probability distribution. It starts with initial estimates. Then, it iteratively refines the parameters based on the Expectation and Maximization step. It is best in handling real-world dataset but becomes sensitive to noise [7].

Weight-based clustering is adopted to estimate the crop yield prediction over a specific area and has provided the relationship between rainfall and production variables [9].

M. Gunasundari et.al suggested a crop yield prediction model, BeeHive algorithm, which predicts crop yield from patterns across multiple data sets. The outcome helps in identifying the areas of unusually high or low yield. [1].

3. Overview of Data

In this present paper, the proposed model aims to find out which districts of Telangana and Andhra Pradesh will give the best yield by computing the value of R^2.To observe the results, various agricultural information is required. The data for this model have been collected from the Govt. Organizations like the Indian Meteorological Department (IMD) for the variables such as Rainfall, Temperature, Pressure, Cloud cover. The data for the attributes like Yield, Area of Sowing, Fertilizers(Nitrogen, Phosphorous, Potassium) have been collected from the Directorate of Economics and Statistics Department. The attribute 'Year' indicates the year in which the data is available in Hectares. 'Rainfall' attribute indicates the average rainfall in the specified year in Millimetres. ’Pressure’ attribute specifies in the specified year in millimetres. ‘Cloud cover’ attribute specifies in the specified year in percentages. 'Area of Sowing' attribute specifies the total area sowed in the specified year for that region in Hectares. 'Yield' specifies in Kilogram per hectare. 'Production' attribute specifies the production of the crop in the specified year in Metric Tons. 'Fertilizers' specify in Tons in the specified year. The model is implemented for 35 years of data.

4. Methodology

In this paper, the Epsilon Density Based Prediction (EDBP) clustering technique is used for estimating the crop yield analysis. The implementation procedure is as follows:

Let X = (x1,x2,x3,....., xn) be the data set with n attributes such as, Area of Sowing, Cloud Cover, Pressure, etc.

Step 1: Divide the data set into 70% training set and 30% test set using leave one out method without selecting the y dependent variable.

Step 2: Now, take each record of test set x_i to find the distance from each record of training using Euclidean distance.

Step 3: Fix a value of epsilon = 2,3,4,5, ..., n where n >0 and n <no of records in the training set.

Step 4: From step 2, We get a matrix for all records of the test called 'D.'

Each record of set 'D' is sorted in ascending order making the least distance to maximum distance.

Step 5: From these arranged records of 'D,' select the Epsilon Density, €=2,3,4.The dependent variable values of these Epsilon Density, € are selected as a set 'M' for each record in the test set. The mean value of these sets 'M' will be the predicted values.

Step 6: For all epsilon values, the R^2 is measured.

Step 7: The model is best for prediction for the epsilon values in which $R^2 > 0.6$.

5. Results and Conclusions

Kharif Season

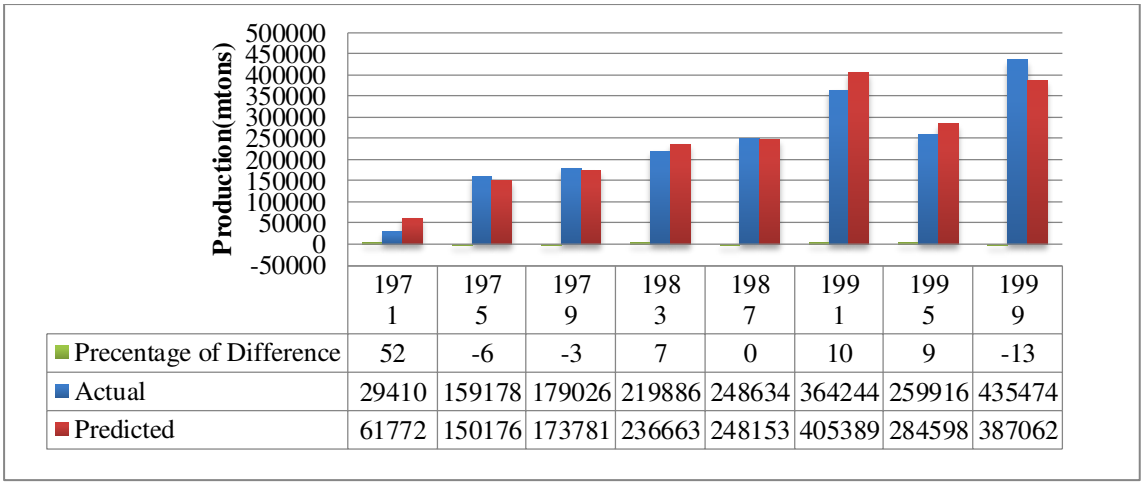
The model has obtained R^2in Kharif Season for the following € values in the below table.

District	€=2	€=3	€=4	€=5
Adilabad	0.78	0.568	0.745	0.791
Chittoor	0.506	0.363	0.474	0.635
East Godavari	0.113	0.33	0.448	0.367
Karimnagar	0.835	0.844	0.832	0.828
Khammam	0.725	0.769	0.725	0.699
Kurnool	0.503	0.569	0.602	0.56
Nalgonda	0.547	0.476	0.618	0.586
Nellore	0.316	0.303	0.359	0.108
Nizamabad	0.002	0.001	0.028	0.09
West Godavari	0.008	0.289	0.019	0.015

Table1.R^2 values for different € in various districts during kharif

Year	Actual	Predicted	Percentage of difference
1971	29410	61772	52
1975	159178	150176	-6
1979	179026	173781	-3
1983	219886	236663	7
1987	248634	248153	0
1991	364244	405389	10
1995	259916	284598	9
1999	435474	387062	-13

Table 2. Actual production and Predicted values for the sample data of Karimnagar district during Kharif Season.



	1971	1975	1979	1983	1987	1991	1995	1999
Precentage of Difference	52	-6	-3	7	0	10	9	-13
Actual	29410	159178	179026	219886	248634	364244	259916	435474
Predicted	61772	150176	173781	236663	248153	405389	284598	387062

Figure1: Comparison between Actual production and Predicted values for the sample data of Karimnagar district during Kharif Season.

Rabi Season

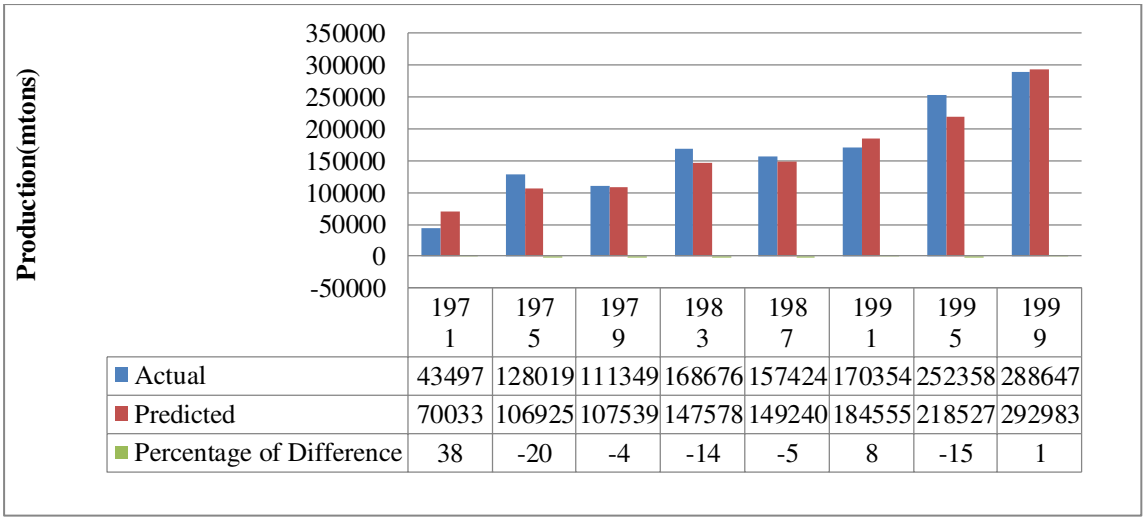
The model has obtained R^2in Rabi Season for the following € values in the below table.

District	€=2	€=3	€=4	€=5
Adilabad	0.13	0.095	0.059	0.04
Chittoor	0.933	0.915	0.866	0.861
East Godavari	0.872	0.843	0.872	0.923
Karimnagar	0.923	0.853	0.865	0.872
Khammam	0.627	0.695	0.709	0.745
Kurnool	0.705	0.74	0.76	0.749
Nalgonda	0	0.01	0.026	0.033
Nellore	0.821	0.809	0.854	0.869
Nizamabad	0.317	0.232	0.177	0.136
West Godavari	0.739	0.795	0.742	0.769

Table 3: R^2 values for different € in various districts during Rabi

Year	Actual	Predicted	Percentage of difference
1971	43497	70033	38
1975	128019	106925	-20
1979	111349	107539	-4
1983	168676	147578	-14
1987	157424	149240	-5
1991	170354	184555	8
1995	252358	218527	-15
1999	288647	292983	1

Table 4. Actual production and Predicted values for the sample data of Karimnagar district during Rabi Season



	1971	1975	1979	1983	1987	1991	1995	1999
Actual	43497	128019	111349	168676	157424	170354	252358	288647
Predicted	70033	106925	107539	147578	149240	184555	218527	292983
Percentage of Difference	38	-20	-4	-14	-5	8	-15	1

Figure 2: Comparison between Actual production and Predicted values for the sample data of Karimnagar district during Rabi Season.

In this paper, the prediction of the crop (rice) is carried out by using an Epsilon Density-based clustering method. From Table1 and Table 3, we can conclude that, the model is best for both the Karimnagar and the Khammam districts. It is also observed that model prediction is good in Kharif season for Adilabad, Karimnagar and Khammam districts. Similarly, during Rabi, the model suits best for Chittoor, East Godavari, Karimnagar, Khammam, Kurnool, Nellore, and West Godavari districts.

For different Epsilon values, analysis is carried out during Kharif and Rabi for all the districts of Telangana and Andhra Pradesh., and estimations were achieved an estimation process is evaluated only on a single variable called production.

6. Conflicts of Interest

The authors declare no conflict of interest.

7. Authors Contribution

D. Esther Rani, Dr.N.Sathyanarayana, and Dr. B. Vishnu Vardhan contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

8. Acknowledgment

The authors wish to thank the Jawaharlal Nehru Technological University, Hyderabad, India. The authors are highly motivated by the University authorities for providing the provision in order to publish the papers in International research domain.

9. References

[1] M. Ananthara, T. Arunkumar, and R. Hemavathy, "Cry: An improved crop yield prediction model using beehive clustering approach for agricultural data sets," *In: Proc. of IEEE International Conf. On Pattern Recognition, Informatics and Medical Engineering (PRIME),* pp.473-478,2013.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, Second Edition,Elsevier Science, 2006.

[3] R. A.A. and K. R.V, "Review - role of data mining in agriculture," *International Journal of Computer Science and Information Technologies(IJCSIT)*, vol. 4(2), no. 0975-9646, pp. 270-272,2013.

[4] J. Y. Hwanjo Yu and J. Han, in Classifying Large Data Sets Using SVMs with Hierarchical Clusters. *ACM New York*, USA 2003, pp.306-315.

[5] D. M. Kiri L. Wagstaff and S. R. Sain, " Harvist: A system for agricultural and weather studies using advanced statistical methods," *BEACON Library*,2005.

[6] A.Awan and M.Md.Sap," A framework for predicting oil-palm yield from climate data," *International Journal of Information and Mathematical Sciences,* 2012.

[7] S. Kim and Wilbur, " An EM clustering algorithm which produces a dual representation," *In: IEEE 10th International Conf. On,Machine Learning and Applications and Workshops (ICMLA),* 2011 vol. 2, pp.90-95

[8] E.Manjula*, S.Djodiltachoumy "Analysis of Data Mining Techniques for Agriculture Data" *International Journal of Computer Science and Engineering Communications* Vol.4, Issue.2, April2016

[9] D Ramesh, B Vishnu Vardhan "Crop Yield Prediction Using Weight Based Clustering Technique' *International Journal of Computer Engineering and Applications,* Volume IX, Issue IV, April 15.

[10] B. Vishnu Vardhan, D. Ramesh and O. Subhash Chander Goud, "Density Based Clustering Technique on Crop Yield Prediction" *International Journal of Electronics and Electrical Engineering*, Vol.2, No.1, March 2014.

[11] Steinbach M, G. Karypis, V. Kumar, "A Comparison of document clustering techniques",*In Proceedings of the 6th ACM SIGKDD Conf. On World Text Mining*, Boston, MA, 2000.

[12] Dhillon I, J Fan, Y Guan, "Efficient Clustering of Very Large Document Collection", *In: Data Mining for Scientific and Engineering Applications, Kluwer Academic Publishers*, USA, 2001.

[13] Mucherino, P. Papajorgji, P M Pardalos, "A Survey of Data Mining Techniques Applied to Agriculture", *Operational Research: An International Journal* 9(2), 2009.

[14] S.Kavitha, D.Geetha, M.Gomathi, R.Suresh Kumar "Agricultural Analysis for Next Generation High Tech Farming In Data Mining" *International Journal of Scientific Development and Research* Vol 1,Issue 10, October2016.

[15] Perpetua Noronha1, Divya.J2, Shruthi.B.S3 "Comparative Study of Data Mining Techniques in Crop Yield Prediction". *International Journal of Advanced Research in Computer and Communication Engineering*, Vol 5, Sept issue 2, October2016.

[16] Ms.Kalpana.R, Dr.Shanthi.N ,Dr.Arumugam.S"A Survey on Data Mining Techniques in Agriculture" *International Journal of Advances in computer science and Technology,* Vol 3 issue.8 August2014.

[17] E.Manjula*, S.Djodiltachoumy "Analysis of Data Mining Techniques for Agriculture Data" *International Journal of Computer Science and Engineering Communications* Vol.4, Issue.2, April 2016.

[18] Mohammad Motiur Rahman, NaheenaHaq and Rashedur M Rahman "Comparative Study of Forecasting Models on Clustered Region of Bangladesh to Predict Rice Yield",In Proc. Of *IEEE Technology(ICCIT)*,Dhaka,2014.

[19] Tejas S. Mehta Dr. Dhaval R. Kathiriya "Survey of Data Mining Techniques in Precision Agriculture" *International Journal of Scientific Research*,Volume: 4, Issue: 7,July2015

[20] George KarypisEui-Hong (Sam) Han Vipin Kumar "*International Conf. On Computer and Information* CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling" University of Minnesota,2011.

[21] Leo Breiman Adele Cutler "Random Forest-Decision tree" University of California Berkeley January 2001.

[22] P. Berkhin, "A Survey of Clustering Data Mining Techniques" *Springer Press* (2006) 25-72

[23] J. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", *Springer*, 1981.

[24] K. Hirota, W. Pedrycz, "Fuzzy Computing for data mining," In: *Proceeding of the IEEE*, Vol 87(9), 1999, pp1575-1600.

[25] N. S. Iyer, A. Kendel, and M. Schneider, "Feature-based fuzzy classification for interpretation of mamograms," Fuzzy Sets andSystems, Vol.114,2000, pp. 271-280.

[26] X. Chang, W. Li, and J. Farrell, "A C-means clustering based fuzzy modeling method," In: *The Ninth IEEE International Conf.on,Fuzzy Systems*, vol.2, 2000, pp.937-940.

[27] R. Krishnapuram and J.M. Keller, "A possibilistic approach to clustering," In: Proceedings of *IEEE Transactions on Fuzzy Systems*, vol.1, no.2,pp. 98,110,May 1993.

[28] N.R.Pal,K.Pal,andJ.C.Bezdek,"Amixedc-meansclusteringmodel,"In: Proceedings of *sixth IEEE International Conf. On FuzzySystems,*vol.1, pp.11,21 vol.1, 1-5 Jul1997.

[29] N.R.Pal,K.Pal,J.M.Keller,andJ.C.Bezdek,"APossibilisticFuzzyc-MeansClusteringAlgorithm,"In: Proceedings of *IEEE Transactions on Fuzzy Systems*, vol.13, no.4, pp.517,530, Aug.2005.

[30] E. Rubio, O. Castillo and P. Melin, "A new Interval Type-2 Fuzzy Possibilistic C-Means.

[31] Clustering Algorithm", In: Proceedings of *IEEE Transactions on Fuzzy Systems*,2015.