iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VII **Month of publication:** July 2023

DOI: https://doi.org/10.22214/ijraset.2023.55065

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

Using Data Mining Techniques for Field Data

Punya H N[1], Dr. Bharathi[2]
[1]*Assistant Professor, Department Of Computer Science, Maharani's Science College for Women, Bangalore*
[2]*Assistant Professor, Department Of Electronics, Maharani's Science College for Women, Bangalore*

Abstract: The population of India is continuously increasing and to meet the food necessities of this growing population, agricultural yield should be boosted. Knowledge discovered from raw data is useful for many purposes. This paper aims to analyse the field data using data mining algorithms and to find useful information from the results of these techniques which would help to improve the agricultural yield. Various mining algorithms applied on agricultural data were studied. Data mining techniques applied in this paper include clustering algorithms- K- means, DBSCAN, EM, the results of these algorithms are analysed.
Keywords: Data Mining, DBSCAN, K-means, EM, WEKA

I. INTRODUCTION

The crop cultivation primarily depends on environmental factors such as rainfall, temperature and geographical topology of the particular region. Knowledge acquired from data is highly useful for many purposes. Data mining is a field in Information Technology that deals with finding unknown hidden patterns from the available data. Applying data mining algorithms helps to predict useful crop productivity related information .This paper aims to analyze such agricultural data using data mining techniques and consolidate the knowledge acquired from the result of data mining techniques. The comparison of results from different data mining algorithms will be made which will help in finding the most suitable algorithm for crop cultivation

II. BACKGROUND

Data mining in the field of crop cultivation is a recent research topic. Recent technologies are nowadays able to find abundant information on crop cultivation related activities, which can then be analyzed in order to find important information. India is agriculture based country. Crop yield depends on multiple different factors such as climate changes, soil type etc. Farmers are interested in knowing the crop yield beforehand. Traditionally, this process was dependent on experiences of farmers and it used to be limited only for a particular region.Data mining Algorithms can be helpful in predicting crop yield. Data mining Algorithms such as data classification and data clustering can be used for data analysis..Multiple data mining algorithms have been used to analyze agricultural data. Various algorithms including K- Means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are applicable to agricultural data.Suitable data models can be found out that can achieve a high accuracy in terms of yield prediction . The researchers' implemented K-Means algorithm to forecast the pollution in the atmosphere, the K Nearest Neighbor is applied for simulating daily rains and other weather variables and various changes of the weather conditions are analyzed using Support Vector Machines. Artificial Neural Networks can be used to analyze the patterns in soil data set .

Frequent pattern mining is also a data mining technique. A frequent pattern is a pattern that occurs frequently in a dataset and provides crucial information that was unknown before . Support vector machine is a binary classifier. It is able to disjoint classes. The basic idea behind it is to classify the sample data into linearly separable classes. It is a set of allied supervised learning methods used for classification and regression. It is used to access spatiotemporal characteristics of the soil moisture product .Decision tree is one of the popular classification algorithm that is currently used in data mining and machine learning. Decision tree involves algorithmic gaining of structured knowledge in the forms such as- concepts, decision trees and discrimination nets or production rules.

A Naïve Bayes classifier is a simple probabilistic classifier established on applying Bayes theorem with strong independence assumptions. Depending on the precise kind of probability model, Naïve Bayes classifier can be trained very proficiently in a supervised learningsettings. J48 is an open source java

implementation of the C4.5 algorithm in the weka data mining tool. C4.5 is a program that makes a decision tree based on the set of labelled input data. This decision tree can be tested against unseen labelled test data to tell how well it generalizes .

Partitioning algorithms specifies initial number of groups and iteratively altering objects among groups to conjunction. In contrast hierarchical algorithms combine and divide existing groups creating hierarchical structure that returns the order in which groups are combined or divided . Data clustering is an efficient unsupervised learning technique that deals with grouping unlabeled data into clusters. Clustering algorithms such as k-Means Clustering, Hierarchical Clustering, DBSCAN (Density Based Spatial Clustering of Applications with Noise) clustering, OPTICS (Ordering Points to Identify the Clustering Structure), STING (Statistical Information Grid). The WEKA (Waikato Environment for Knowledge Analysis) system provides a broad suite of facilities for applying data mining techniques to large data . Overview of the data used for analysisis given in the next section.

III. BACKGROUND

The data used in this paper contains information about plantation, fruits and vegetables of 35 states of India including- Andhra Pradesh, Andaman Nicobar, Arunachal Pradesh, Assam, Bihar, Chandigarh, Chhattisgarh, Dadra and Nagar Haveli, Daman and Diu, Delhi, Goa, Gujarat, Haryana, Himachal Pradesh, Jammu and Kashmir, Jharkhand, The dataset contains total 4180 instances having eight attributes. They are Year, State, Crop type, and Crop name, Area, Production, Rainfall and Temperature. Following figure shows database schema. The data has been preserved from records in Agriculture Department,kunigal ,The collected data is been analyzed using WEKA

No.	1: YEAR Numeric	2: STATE Nominal	3: CROPTYPE Nominal	4: CROPNAME Nominal	5: AREA Numeric	6: PRODUCTION Numeric	7: RAINFALL Numeric	8: **TEMPERATURE** Numeric
1	2005.0	ANDAM...	PLANTATION	CASHEWNUT	0.0	0.0	2967.0	26.0
2	2005.0	ANDHR...	PLANTATION	CASHEWNUT	170.0	92.0	912.0	26.5
3	2005.0	ARUNA...	PLANTATION	CASHEWNUT	0.0	0.0	2782.0	20.0
4	2005.0	ASSAM	PLANTATION	CASHEWNUT	14.0	10.0	2818.0	23.0
5	2005.0	BIHAR	PLANTATION	CASHEWNUT	0.0	0.0	1256.0	25.5
6	2005.0	CHANDI...	PLANTATION	CASHEWNUT	0.0	0.0	617.0	23.5
7	2005.0	CHHATT...	PLANTATION	CASHEWNUT	0.0	0.0	1511.0	26.0
8	2005.0	D & N H...	PLANTATION	CASHEWNUT	0.0	0.0	2169.0	25.0
9	2005.0	DAMAN ...	PLANTATION	CASHEWNUT	0.0	0.0	911.0	25.0
10	2005.0	DELHI	PLANTATION	CASHEWNUT	0.0	0.0	617.0	25.0

Fig. 3.1 Dataset Overview

IV. RESULT ANALYSIS

A. K-means

In K-means algorithm clusters are formed based on centroids. On applying this algorithm, two clusters of data were formed. Clusters and their centroids w.r.t attributes are given below-

Table 5.1 Result of K-means algorithm

Attribute	Cluster 0	Cluster 1
Production	75.1519	308.1148
Rainfall	1803.281	1505.904

Means and standard deviations of clusters formed by DBSCAN and EM algorithmsare given below:

B. DBSCAN

Table 5.2 Result of DBSCAN algorithm

Attribute		Cluster 0	Cluster 1
Production	Mean	75.1519	308.1148
	Std. Dev	271.7347	838.2223
Rainfall	Mean	1803.281	1505.904
	Std. Dev	767.6352	723.2672

C. EM

Table 5.3 Result of EM algorithm

Attribute		Cluster0	Cluster1	Cluster2	Cluster3	Cluster 4	Cluster5
Production	Mean	97.7995	63.3811	44.6236	0	1404.2668	32.2018
	Std.Dev	114.4027	84.3918	56.0153	714.7723	1483.1565	38.5763
Rainfall	Mean	1164.493	3002.691	2786.752	1650.424	1476.171 6	1358.296
	Std.Dev	407.5313	38.5244	38.2653	873.6832	676.3877	453.2867

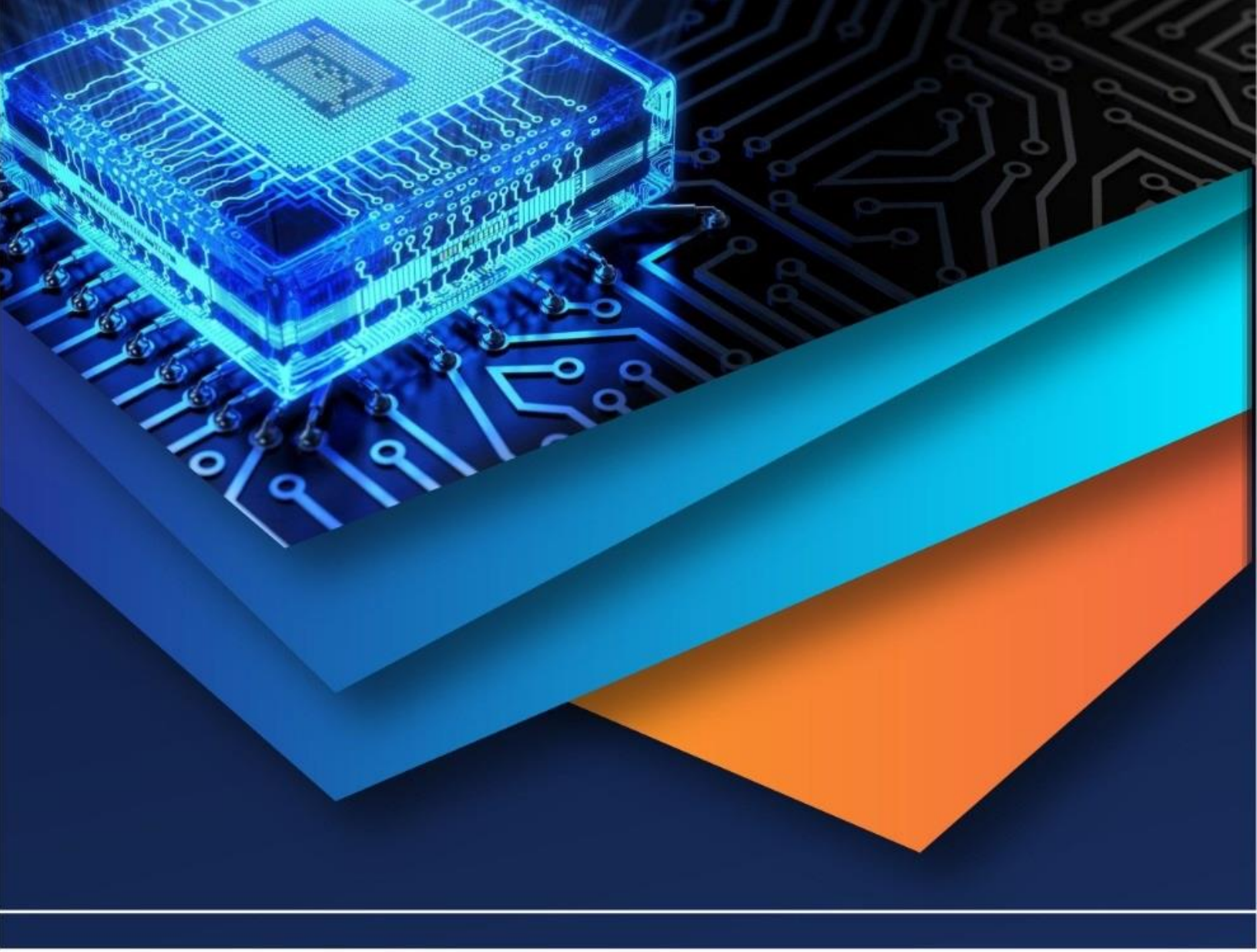
Result analysis shows that production tends to increase when rainfall ranges from 1405.904mm to 1562.3756mm.DBSCAN algorithm gives similar results as base algorithm K-means, whereas EM gives more specific production values on given rainfall and temperature range as compared to K-means and DBSCAN.

V. CONCLUSION

In this paper certain data mining algorithms were adopted to cluster the data that shows relevance with desired attributes. K-means clustering algorithm is adopted as base algorithm. DBSCAN and EM algorithms are also applied to data. DBSCAN showed similar behavior to K-means algorithm. Future work aimed at applying advanced mining techniques to larger dataset such as one of the big data techniques.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pie, “Data Mining Concepts and Techniques”,Morgan Kaufmann, ASIN B0058NBJ2M
[2] Dr. Jean-Claude Franchitti, “Data Mining Session 6 – Mining Frequent Patterns, Association, and Correlations” Adapted from course textbook resources Data Mining Concepts and Techniques (2nd Edition)
[3] Andrew Smith, Neil Alldrin, Doug Turnbull, “Clustering with EM and K-Means” International Journal of Advance Research in Computer and Communication Engineering
[4] “The Institute connecting the dots with Big Data” September 2014, www.theinstitute.ieee.org.in
[5] Mr. Osama Abu Abbas, “Comparison between Data Clustering Algorithms” The International Arab Journal of Information Technology Volume 5, No. 3.
[6] Aastha Joshi, Ranjeet Kaur, “A Review: Comparative Study of Various Clustering Techniques in Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, ISSN: 2277 128X, Issue 3.
[7] Sally Jo Cunningham and Geoffrey Holmes, “Developing Innovative Applications in Agriculture Using Data Mining”, Department of Computer Science, University of Waikato,Hamilton, New Zealand.
[8] Hongjun LU, Ling Feng and Jiawei Han, “Beyond Intratransaction Association Analysis: Mining Multidimensional Intertransaction Association Rules”, ACM Transactions on Information Systems, Vol. 18, October 2000.
[9] Vaishali, A., Harsh, K., Anil, K.A, 2016, Performance Analysis of the Competitive learning Algorithms on Gaussian Data in Automatic Cluster Selection”, 2016 Second International Conference on Computational Intelligence & Communication Technology.
[10] Eibe F, Mark A.H, IanH.W. 2016, “The WEKA Work bench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.
[11] Lichman, M.,2013, “UCI Machine Learning Repository” [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY