Original Research Article

APPLICATION OF K-MEANS CLUSTERING ALGORITHM IN RICE PRODUCTION OF TAMIL NADU

ABSTRACT

Aim: (i) To Cluster the Rice data using K-Means clustering algorithm. (ii) To helps the study of crop yield prediction.
Study design: K-Means clustering technique is one of the most common exploratory data analysis used to get an intuition about the structure of the data.
Place and Duration of Study: Time Series crop data were collected from the season and crop report, Directorate of Economics and Statistics, Chennai for the period of 2015-2020.
Methodology: The machine learning algorithm of big data analytics method such as K-means clustering algorithm helps to predict the paddy yield accurately in Tamil Nadu. The performance of the technique is examined through the determinable value of k by Elbow method and Silhouette method which helps in the crop yield prediction.
Results: The observed results show that there is a positive relationship between area, production, area under irrigation, minimum temperature, and relative humidity and a close negative relationship with moisture and wind speed. Additionally, two clusters were identified with cluster 2 having the highest mean value, followed by 1. The identification of the highest mean clusters will guide farmers on where best to concentrate on when planting their crops in ordered to improve productivity and crop yield.
Conclusion: This study reveals a scalable, simple, and reduced method for correctly assessing rice production over a large area using publicly released multi-source data, which may have been used to calculate crop production in areas with rarely observed data and all around the world.

Keywords*: crop yield prediction, K-Means clustering, machine learning algorithm, rice, Tamil Nadu.*

1. INTRODUCTION

Although high input costs and water constrains may prevent world rice production from rising, the 2022 harvest is predicted to remain abundant amid public efforts to help the sector cope with profitability challenges. Thus, world rice utilization and trade may continue rising, while reserves remain ample (FAO, 2022).

Total rice utilization in 2022-23 is pegged at 522.0 million tones, only slightly above the 2021-22 high, as another sturdy expansion in food intake is forecast to be mostly outweighed by declines in non-food uses. To meet this forecast volume of use, global rice inventories would need to be drawn down, albeit by a small volume of 0.8 million tones. This would place world rice stocks at 191.6 million tones, their second highest level on record, largely due to accumulations in China and India.

Tamil Nadu, given its history as a pioneer in irrigation, is regarded as a major rice-producing state. But it does not figure as a key provider of rice at the all-India level if one is too furnished by the Food Corporation of India (FCI).

In the last five years, the contribution of Tamil Nadu to the Central pool hovered between 2.6% and 5%. The current Kharif Marketing season 2021-22 (October-September) and the year before have been particularly good. Yet, this does not take the state beyond the 5%mark. Considering Tamil Nadu's annual requirement of about 38 lakh tons of rice for the public distribution system (PDS), the state has to depend on supplies from the Central authorities or other states, as what gets procured by the authorities locally is barely sufficient for the PDS.

The K-Means algorithm is a method proposed by Macqueen in the year 1967. Under unsupervised learning, this is one of the oldest and most common clustering algorithms. It divides the dataset into partitions based on the dataset's mean value and processes iteratively until no more partitions are available. This survey studies the problems of and solutions to partition-based clustering, and more specifically the widely used k-means algorithm, which has been listed among the top 10 clustering algorithms for data analysis (Wu et.al.,2008). K-means is the most popular clustering formulation in which the goal is to maximize the expected similarity between data items and their associated cluster centroids. Although the k-means clustering algorithm itself performs well with compact and hyper-spherical clusters, and also interested in highlighting its limitations and suggesting solutions.

Machine learning is often used for predictive analytics, and it is a field that has grown more popular as the amount of data being generated increases. Python is a popular programming language that is becoming more and more prevalent in scientific computing and machine learning (Nabeel et.al. 2022).

The Tableau software is still quite unknown in the agricultural area and among agricultural specialists and farmers although it can provide them with a very good and useful Method for their work, knowledge and education (Smiciklas, 2012; Ganchev, 2017). Tableau Public is a free service that lets anyone publish interactive data visualizations to the web. The visualizations can be embedded into web pages and blogs, they can be shared via social media or email. The massive global demand for food will keep rising for at least another 40 years when population and consumption increases continue. Growing competition for land, water, and energy, as well as overpopulation, will have an impact on our ability to produce food, as will the essential need to mitigate the food system's environmental impact. Climate change's consequences are also a threat. However, the globe can produce more food while also ensuring that it is distributed more efficiently and reliably. To secure sustainable and equitable food security, a multidimensional and integrated global strategy is required.

2. MATERIALS AND METHODS

ML approaches can be used to estimate future crop yields, weather forecasts, pesticide and fertigation rates, and revenue generation, among other things. Techniques for machine learning (ML) can be supervised or unsupervised. It is known as supervised machine learning (ML) when the training data is acquired from a long time ago and utilized for training in order to learn how to identify future yield predictions.

Clustering is a common unsupervised yield prediction approach. A cluster is a subset of objects that are similar in nature. Things that are highly similar are placed in the same cluster, whereas objects that are distinct are placed in other clusters. This technique is used

to group data that has no previous knowledge of the data. The value of K indicates how many clusters the data is divided into. To define K centers, the K number is employed. These centers are spaced apart, and each data point is associated with the cluster with the closest centroid. Iteratively, the procedure of locating new K centers and allocating data samples to the clusters with the closest centroid is carried out until the data samples can no longer move clusters. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or the squared error function:

$$J = \sum_{k}^{j=1} \sum_{n}^{i=1} |x_{(i)} - c_j|^2$$

Where J is the objective function

k- Number of clusters

n- Number of cases

$|x_{(i)} - c_j|^2$ the distance function; x_i denotes case i & c_j denotes centroid for cluster j

The following stages of the K-Means clustering algorithm are

Step 1: First, to provide the number of clusters, K, that need to be generated by this algorithm.

Step 2: Next, choose K data points at random and assign each to a cluster. Briefly, categorize the data based on the number of data points.

Step 3: The cluster centroids will now be computed.

Step 4: Iterate the steps below until we find the ideal centroid, which is the assigning of data points to clusters that do not vary.

4.1 The sum of squared distances between data points and centroids would be calculated first.

4.2 At this point, we need to allocate each data point to the cluster that is closest to the others (centroid).

4.3 Finally, compute the centroids for the clusters by averaging all of the cluster's data points.

3. RESULTS AND DISCUSSION

The Big data analytic tool python was used for the analysis. Various methods were employed to find the number of k in k-means clustering algorithm. The findings of the study are discussed below.

Table.1- Basic statistics of the input data

	Mean	Minimum	Maximum	SD	CV	Skewness	Kurtosis
AREA	56237.5	235.4	186461.8	57542.56	102.32	0.96	2.57
AUI	52538.19	0.8	186461.8	55197.3	105.06	1.14	3.02
RAINFALL	891.31	517.9	1410.72	223.91	25.12	0.68	2.68
PRODUCTION	203217	839.4	633702.2	196871.2	96.88	1.02	2.69
MIN TEMP	16.41	12.03	23.72	2.92	17.78	0.82	3.26
MAX TEMP	40.56	33.42	43.43	2.42	5.97	-1.48	4.53
RH	69.98	64.5	79.71	3.66	5.23	0.72	3.15
MOISTURE	0.55	0.39	0.65	0.06	11.38	-0.52	3.26
WIND SPEED	5.15	3.82	6.35	0.58	11.24	-0.09	2.84

*AUI-Area Under Irrigation; MIN TEMP-Minimum Temperature; MAX TEMP-Maximum Temperature; RH-Relative Humidity

Table.2-Correlation matrix explains the relationship of each attribute

	AREA	AUI	RAINFALL	PRODUCTION	MIN TEMP	MAX TEMP	RH	MOISTURE	WIND SPEED
AREA	1								
AUI	0.974344	1							
RAINFALL	0.150175	0.188639	1						
PRODUCTION	0.953986	0.983321	0.164263	1					
MIN TEMP	0.501719	0.452877	0.187824	0.424701	1				
MAX TEMP	-0.16763	-0.12179	-0.44873	-0.0975	-0.65113	1			
RH	0.266343	0.27625	0.589079	0.281655	0.599716	-0.79444	1		
MOISTURE	-0.16984	-0.04331	-0.20722	0.023081	-0.57021	0.481446	-0.32312	1	
WIND SPEED	0.49453	0.453936	0.011299	0.427629	0.682252	-0.34471	0.437283	-0.4466	1

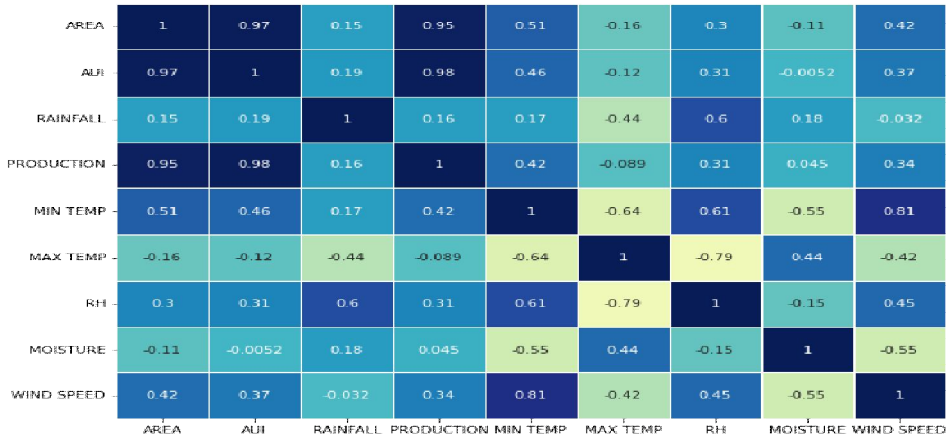

Figure.1 Correlation matrix

Figure 1 represents the correlation matrix uses correlation to show the relationship between each property. It is evident that there is a negative correlation between the maximum temperature and other variables, which is why the ideal temperature range for paddy cultivation is often between 35°C and 45°C. Others are almost all connected favorably with each of the characteristics meant by directly proportional.

Box plot

Box plot was the tool for the graphical representation of the numerical data by their quartiles. The standardized way of displaying the data with minimum, maximum, median, first quartile and third quartile. Box plot was also used to find the outliers present in the dataset.

Figure 2 Box plot

The Box plot of the given variables are figured. It finds the variables area and area under irrigation contains outliers shown in figure 2.

Determining number of k in k means

Using Elbow technique, Silhouette method, and Gap statistic method, which demonstrates that the plots display a bend in the graph is the appropriate number of clusters called k in k means algorithm.

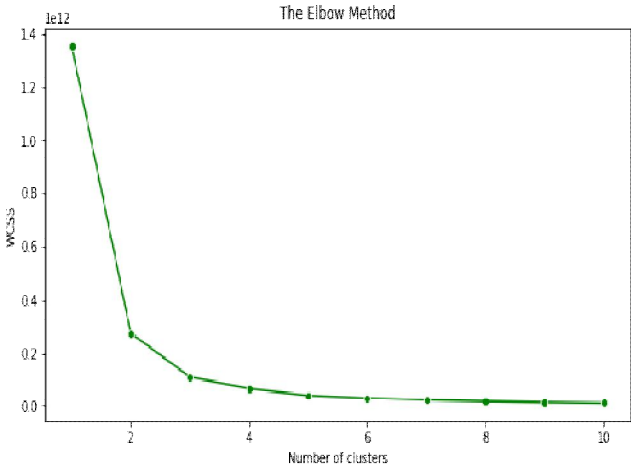

Figure 3 Within-Cluster-sum of squares

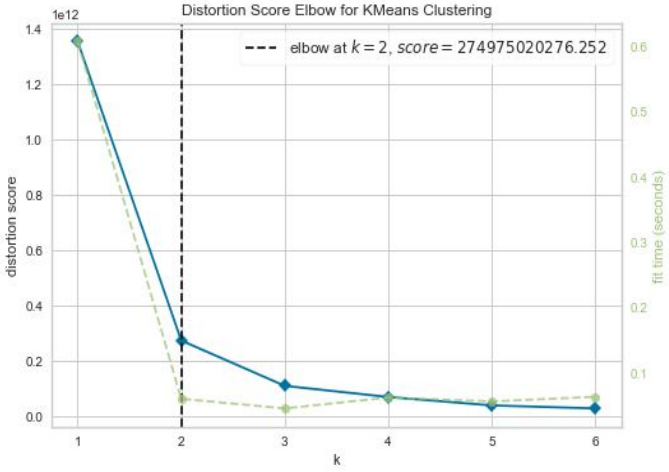

Figure 4 Elbow curve i.e., within sum of squares

Figure (3) & Figure (4) displaying Within-Cluster-sum of squares, Elbow curve i.e., within sum of squares, Average Silhouette and Gap statistic. The majority of these methods recommend employing 2 clusters for the final analysis and result extraction, which is the number of ideal clusters. The optimal number of k is 2, as shown by these charts of several approaches. Then employing the optimum number of clusters from the k-means method to divide the data. The final number of clusters 2, as estimated by k means clustering used to project future rice production in Tamil Nadu is shown in Fig. 6.

Fig

Figure 5.projection of future rice production

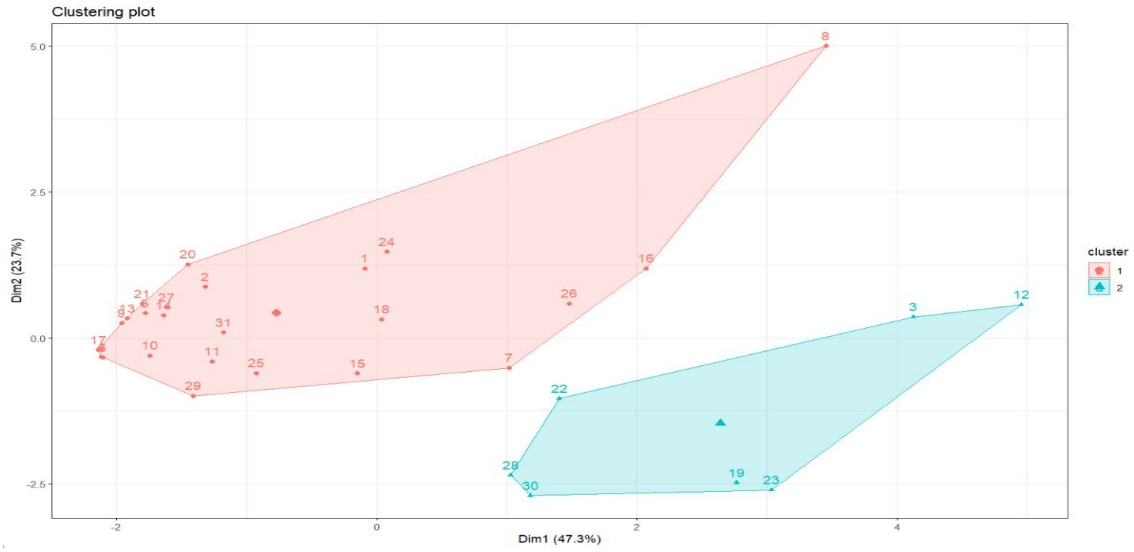

Figure 6.K-Means clustering used to project future rice production in Tamil Nadu

Table.3-Cluster means:

Variables	Cluster 1	Cluster 2
AREA	-0.583	1.06
AUI	-0.567	1.03
RAINFALL	-0.276	0.503
PRODUCTION	-0.589	1.072
MIN TEMP	-0.485	0.882
MAX TEMP	0.333	-0.606
RH	-0.462	0.84
MOISTURE	0.11	-0.201
WIND SPEED	-0.389	0.708

In these results, Table.3 clearly depicts that clusters data for 31 districts into 2 clusters based on the k value determined by the methods that was specified. Cluster 1 contains low mean values for all the variables except maximum temperature and moisture. Cluster 2 has the high mean value for most of the attribute containing districts, Cuddalore, Nagapattinam, Thanjavur, Tiruvallur, Thiruvarur, Tiruvannamalai and Villupuram. By the result, it was determined that the districts depicted in Figures (5) & (6) Cluster 2 aid in the correct

projection of future rice production as well as other crop production. The obtained results of cluster 2 shows that coastal region of Tamil Nadu had a high production of rice, where during the summer and monsoon seasons, both heavy rainfall and majorly irrigation sources provide ideal conditions for the cultivation of rice.

4. CONCLUSION

We also looked into the possibility of predicting rice yields in advance, and found that the best forecast could be made two/one month prior to single/double rice maturity. Our results reveal a scalable, simple, and reduced method for correctly assessing rice production over a large area using publicly released multi-source data, which may have been used to calculate crop production in areas with rarely observed data and all around the world. The government could improve their production techniques by acquiring additional farmers or creating new profitable agriculture plans based on crop yield forecast using this K-Means approach.

REFERENCES

Chougule, Archana, Vijay Kumar Jha, and Debajyoti Mukhopadhyay. 2019. "Crop suitability and fertilizers recommendation using data mining techniques." In *Progress in Advanced Computing and Intelligent Engineering*, 205-213. Springer.

Dash, B, Debahuti Mishra, Amiya Rath, and Milu Acharya. 2010. "A hybridized K-means clustering approach for high dimensional dataset." *International Journal of Engineering, Science and Technology* 2 (2):59-66.

Hayatu, Ibrahim Hassan, Abdullahi Mohammed, Barroon Ahmad Isma'eel, and Sahabi Yusuf Ali. 2020. "K-means clustering algorithm based classification of soil fertility in north west Nigeria." *FUDMA JOURNAL OF SCIENCES* 4 (2):780-787.

MacQueen, J. 1967. "Classification and analysis of multivariate observations." 5th Berkeley Symp. Math. Statist. Probability.

Manjula, E, and S Djodiltachoumy. 2017. "A model for prediction of crop yield." *International Journal of Computational Intelligence and Informatics* 6 (4):298-305.

Paidipati, Kiran Kumar, Christophe Chesneau, BM Nayana, Kolla Rohith Kumar, Kalpana Polisetty, and Chinnarao Kurangi. 2021. "Prediction of Rice Cultivation in India—Support Vector Regression Approach with Various Kernels for Non-Linear Patterns." *AgriEngineering* 3 (2):182-198.

Slonim, Noam, Ehud Aharoni, and Koby Crammer. 2013. "Hartigan's K-means vs. Lloyd's K means–is it time for a change?" Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI).

Suresh, A, P Ganesh Kumar, and M Ramalatha. 2018. "Prediction of major crop yields of Tamilnadu using K-means and Modified KNN." 2018 3rd International Conference on Communication and Electronics Systems (ICCES).

Wu, Xindong, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, and Philip S Yu. 2008. "Top 10 algorithms in data mining." *Knowledge and information systems* 14 (1):1-37.

Yadav, Jyoti, and Monika Sharma. 2013. "A Review of K-mean Algorithm." *Int. J. Eng. Trends Technol* 4 (7):2972-2976.

Zubair, Md, Asif Iqbal, Avijeet Shil, Enamul Haque, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2020. "An efficient K-means clustering algorithm for analysing COVID-19." International Conference on Hybrid Intelligent Systems.