Implementation of K-Means Clustering Technique in Banana Production of Tamil Nadu, India

ABSTRACT

Aim: The main objectives of this study are to make use of the K-Means clustering approach to cluster the Banana data and to assist with crop yield prediction.
Study Design: One of the methods of Big Data Analytics K-Means clustering is usedto cluster the data set.
Place and Duration of study: So far, the period 2010-2020, time series data were collected from the season and crop report, Directorate of Economics and Statistics, Chennai.
Methodology: The horticulture industry has a significant impact on India's economic development. In the globe, after China, India ranks second in terms of fruit and vegetable production. Compare to the various fruits Mango and banana are one of the most abundant fruits in India. So, the Banana dataset were collected and dataset were clustered using the K-Means clustering technique and the optimum number of clusters were identify using the elbow approach.
Results: According to these results from this study, there is positive relationship between the Area, Soil moisture, Maximum Temperature, Relative Humidity and negative relationship between Rainfall, Wind Speed and Minimum temperature related Banana production. Using K-Means clustering it divides the given dataset into three clusters in which cluster 3 contains high Banana production afterwards two and one.
Conclusion: The selection of the most productive clusters is going to tell farmers on where to focus their efforts while planting crops in order to enhance productivity and crop production.

Keywords: Yield prediction; K-Means clustering algorithm; Banana; Tamil Nadu.

1.INTRODUCTION

Agriculture is the backbone of the Indian economy. Almost 69 percent of the country's population is still dependent on agriculture and horticulture. The contribution of GDP to the agricultural sector is increasing since however, a considerable shift in agricultural composition, demonstrating a transition from cropland to horticulture, livestock, and fisheries, plays a vital impact. The horticulture industry has a significant impact on India's economic development. In the globe, after China, India ranks second in terms of fruit and vegetable production. The largest yielding fruits in India are the Mango and Banana. Mango, banana, and jackfruit (named Mukkani in Tamil) were the most common fruits grown in Tamil Nadu [1, 2].

The process for finding previously undiscovered and possibly intriguing patterns in enormous databases is collectively referred to as data mining. Data mining is a method of finding and collecting relevant information and associated knowledge from a huge database using artificial intelligence, statistics, mathematics, and machine learning.

Data is typically pre-processed through data cleaning, data integration, data selection, and data transformation before being prepared for mining. Data mining may be done on a variety of databases and information repositories, but the sorts of patterns that can be discovered are determined by data mining functionalities such as class/concept description, association, correlation analysis, classification, prediction, cluster analysis, and so on [3].

Clustering is a data mining process that has evolved into a useful tool for tackling complex computer science and statistics issues. Clustering is the process of categorising data points into two or more groups so that data points in the same group are more similar to each other than data points in other groups, based only on the information associated with the data points [4, 5].

Macqueen proposed the K-Means algorithm in 1967. This is one of the oldest and typically used unsupervised learning methods for clustering. It splits the dataset into partitions based on the mean

value regarding the dataset and repeatedly processes until no more partitions are available[6]. The two basic types of metrics are distance and similarity measures. Distance measures are used to assess if two objects are similar or dissimilar. When dealing with high-dimensional and sparse information, it only organises the data as a crisp set and has limitations[7].

2. MATERIALS AND METHODS

2.1 The Data Source and Research Area

This study was conducted in Tamil Nadu and this study is based on secondary data. Dataset on Banana were collected from the season and crop report,Directorate of Economics and Statistics, Chennaifor the period of 2000-2020 for 28 districts of Tamil Nadu. The variables which are used for this study are area, production, productivity, rainfall, maximum temperature, minimum temperature, relative humidity, soil moisture and wind speed.

2.2 Statistical model

K-Means Clustering

Machine learning (ML) techniques have been utilised to forecast future agricultural yields, weather forecasts, pesticide and fertigation rates, and other things. They are two main techniques in machine learning such as supervised and unsupervised learning. Our study is mainly focus on the unsupervised machine learning. The uses of unsupervised machine learning are to identify the structure and hidden patterns in the dataset. Generally we use unlabelled dataset in the unsupervised machine learning [8].

K-means clustering technique is one of the methods under the unsupervised learning technique. The presence of the data objects in the dataset are clusters into a group by using the clustering technique. We also say clustering process as a learning by observation. After performing the clustering technique if observe the dataset. The objects within the clusters are highly similar and objects across the clusters are highly dissimilar. K-Means methods select k items at random to represent the k initial cluster centres. The next step is to take each point in a given data set and associate it with the nearest centre based on the object's proximity to the cluster centre using Euclidean distance [9, 10].

The methodology to carry out K-means clustering may be described as follows:

Step1: To select the number of clusters k that you want to create and then initialise K cluster centroids at random.
Step2: Based on a distance measure, often Euclidean distance, assign each data point to the nearest centroid. This phase creates K clusters.
Step3: Calculate the mean of all the data points allocated to each cluster to recalculate the centroids of the K clusters.
Step4: Steps 2 and 3 need to be repeated until convergence is obtained. Convergence occurs when the centroids no longer differ appreciably or when the number of iterations reaches a maximum.
Step5: The procedure ends when convergence has been attained and the final cluster centroids indicate the cluster centres. Each data point is part of the cluster indicated by the centroid nearest to it.

3.RESULTS AND DISCUSSION

RStudio, a big data analytics tool, was implemented for the analysis.Several approaches were used for identifying the number of clusters in the K-means clustering algorithm. But in our study, we use elbowapproach to determine the optimum number of clusters. The study's findings are reviewed in more detail below. The descriptive statistics are given in the Table1.

Table 1. Descriptive Statistics

	Area	Production	Productivity	Rainfall	Min temp	Max temp	RH	Moisture	Wind Speed
Mean	3470.7	42599.6	138393.1	996.2	16.2	40.2	71.1	0.6	5.3
Minimum	73.0	4818.0	2860.0	232.9	10.7	31.8	59.2	0.4	3.7
Maximum	70170.0	799968.0	1122220.0	2498.1	24.4	44.9	83.1	0.7	7.0
SD	4739.8	33183.1	146366.6	373.6	3.1	2.6	4.2	0.1	0.6
CV	136.6	77.9	105.8	37.5	18.9	6.4	5.9	10.7	11.6

Skewness	8.1	20.3	1.8	1.1	0.7	-1.1	0.4	-0.4	0.0
Kurtosis	102.5	464.1	4.4	1.5	-0.1	0.6	0.0	0.4	0.0

** RH-Relative Humidity, Min-temp- Minimum temperature, Max-temp-Maximum temperature*

Table 2 show the correlation matrix with their respective significant level. It shows that the area, production, relative humidity and soil moisture are positively influenced and significant at 1% level and rainfall, minimum temperature and wind speed are negatively influenced and significant at 1% level.

Table 2.Correlation coefficient matrix with significant level

		Area	Production	Productivity	Rainfall	Min Temp	Max Temp	RH	Soil Moisture	Wind Speed
Area	Pearson Correlation	1	.052	.610**	-.157**	-.107**	-.075	.187**	.078	-.050
	Sig. (2-tailed)		.207	.000	.000	.009	.069	.000	.058	.223
Production	Pearson Correlation	.052	1	.234**	-.007	-.080	.053	.024	.135**	-.081*
	Sig. (2-tailed)	.207		.000	.868	.053	.195	.564	.001	.049
Productivity	Pearson Correlation	.610**	.234**	1	-.194**	-.158**	-.070	.135**	.098*	-.121**
	Sig. (2-tailed)	.000	.000		.000	.000	.088	.001	.017	.003
Rainfall	Pearson Correlation	-.157**	-.007	-.194**	1	.066	-.253**	.231**	.057	.046
	Sig. (2-tailed)	.000	.868	.000		.112	.000	.000	.168	.262
Min Temp	Pearson Correlation	-.107**	-.080	-.158**	.066	1	-.567**	.440**	-.474**	.798**
	Sig. (2-tailed)	.009	.053	.000	.112		.000	.000	.000	.000
Max Temp	Pearson Correlation	-.075	.053	-.070	-.253**	-.567**	1	-.690**	.409**	-.380**
	Sig. (2-tailed)	.069	.195	.088	.000	.000		.000	.000	.000
RH	Pearson Correlation	.187**	.024	.135**	.231**	.440**	-.690**	1	.103*	.343**
	Sig. (2-tailed)	.000	.564	.001	.000	.000	.000		.013	.000
Soil Moisture	Pearson Correlation	.078	.135**	.098*	.057	-.474**	.409**	.103*	1	-.463**
	Sig. (2-tailed)	.058	.001	.017	.168	.000	.000	.013		.000
Wind Speed	Pearson Correlation	-.050	-.081*	-.121**	.046	.798**	-.380**	.343**	-.463**	1
	Sig. (2-tailed)	.223	.049	.003	.262	.000	.000	.000	.000	

**. Correlation is significant at the 0.01 level (2-tailed), *. Correlation is significant at the 0.05 level (2-tailed).

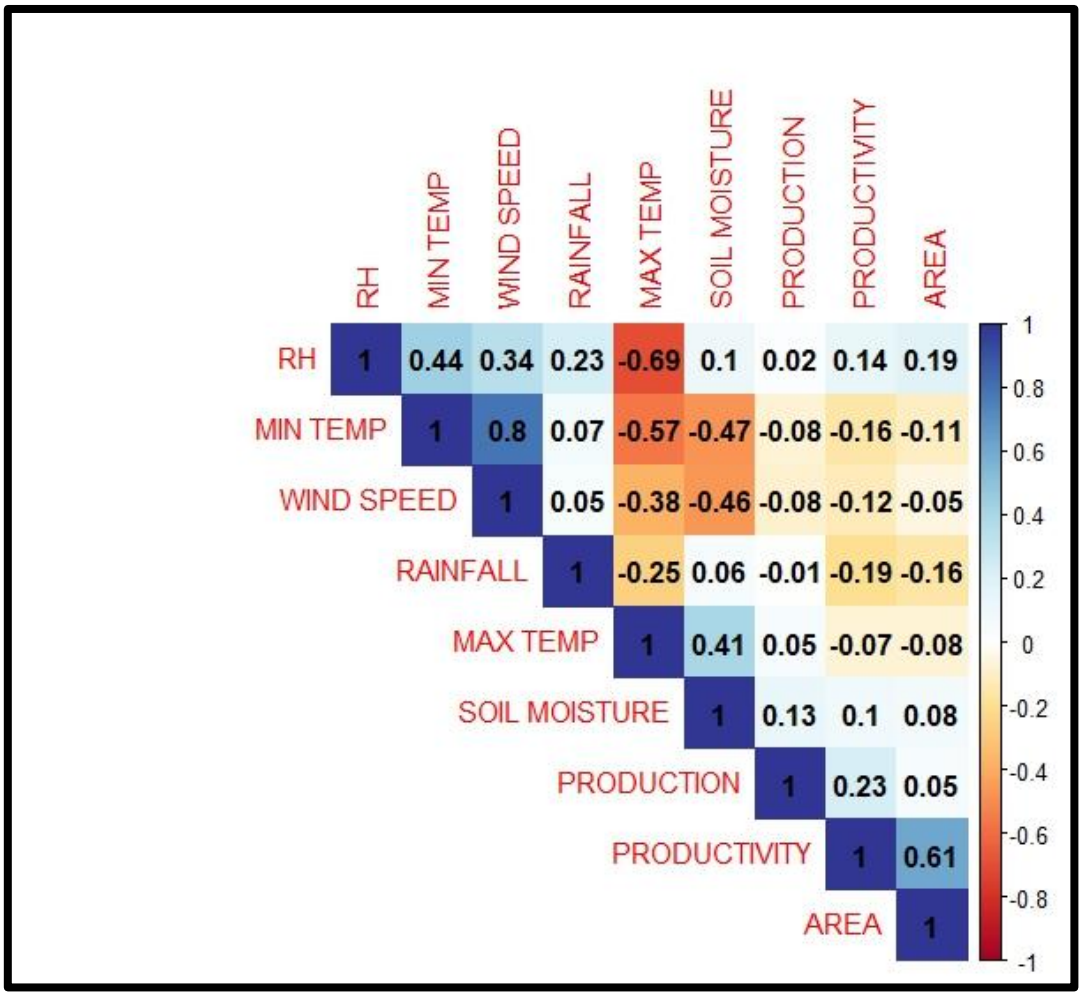

Fig.1. Correlation matrix

To obtain a deeper understanding into the data that is being utilised, we may display graphs like the correlation matrix which is one of the most crucial concepts which provide us a lot of information about how variables are associated to each other and the contribute to each of them have on the other. In the Fig.1 it shows that the relationship between each parameter. It indicates that there is a positive relationship between the Soil moisture, Production, and RH and negative relationship between Maximum temperature, Rainfall, Wind speed, Minimum temperature related to Banana Productivity. Also, Banana production only Relative Humidity, Maximum temperature, Productivity and Soil Moisture shows the positive relationship and others variables are showing the negative relationship.The positive correlation indicates that the given variables posses direct relation whereas the negative correlation indicated inverse relation among variables.

3.1 K-Mean Clustering Algorithm

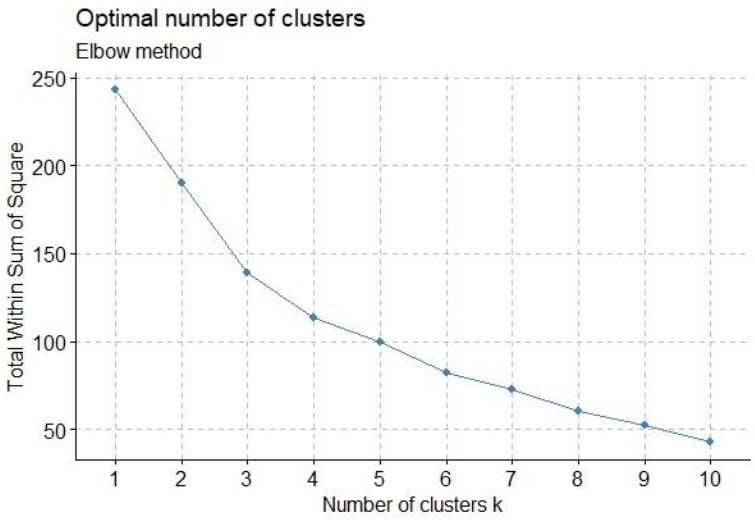

Fig. 2. Elbow curve

Table 3. Cluster Means

Variables	Cluster 1	Cluster 2	Cluster 3
Area	-0.482	-0.235	1.604
Production	0.051	-0.523	0.196
Productivity	-0.435	-0.406	1.575
Rainfall	0.061	0.629	-0.604
Min-temp	-0.248	1.875	-0.507
Max-temp	0.492	-1.963	-0.167
Rh	-0.427	1.307	0.410
Soil moisture	0.246	-1.641	0.355
Wind speed	-0.194	1.553	-0.453

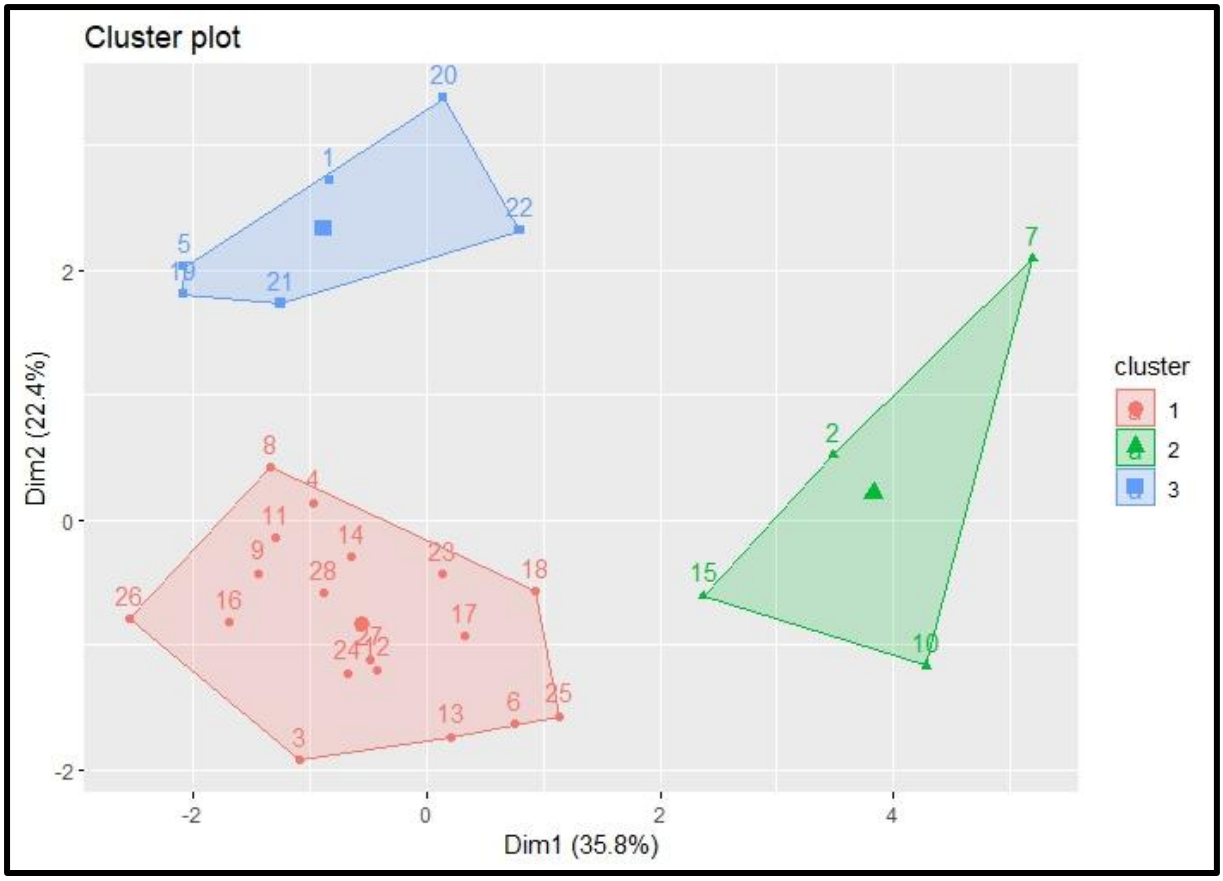

Fig.3. K-Means Cluster plot

By applying K-Means clustering algorithm for this given Banana data it divides the given data into three clusters. The elbow approach is used to figure out the optimal number of clusters. For this given data the elbow occurs at three so we choose the optimum number of clusters is three. Table 4 indicates that the 28 districts are divided into three cluster based on the elbow approach. In which 18 districts are come under cluster one, 4 districts come under cluster 2 and 6 districts come under cluster three. Table 3 shows the cluster means results in which cluster1 contains highest Banana production and cluster 3 contains highest Banana productivity.

Table 4.Districts convergence to respective clusters

Id	Districts	Clusters
1	COIMBATORE	3
2	CUDDALORE	2
3	DHARMAPURI	1
4	DINDUGAL	1
5	ERODE	3
6	KANCHPURAM	1
7	KANNIYAKUMARI	2
8	KARUR	1
9	MADURAI	1
10	NAGAPATTINAM	2
11	NAMAKKAL	1
12	NILGIRI	1
13	PERAMBALUR	1
14	PUDUKKOTAI	1
15	RAMANATHAPURAM	2
16	SALEM	1
17	SIVAGANGAI	1
18	THANJAVUR	1
19	THENI	3
20	THUTHUKUDI	3
21	THIRUCHIRAPALLI	3
22	TIRUNELVELI	3
23	THIRUVALLUR	1
24	THIRUVANNAMALAI	1
25	THIRUVARUR	1
26	VELLORE	1
27	VILLUPURAM	1
28	VIRUDHUNAGAR	1

4. CONCLUSION

Clustering is a data mining method which serves to obtain knowledge and retrieving information. It is one of the unique approaches used in the crop dataset and has a significant advantage in crop prediction. But the main disadvantage of this technique is random initialization of cluster centres. And this is carried out by the K-Means clustering technique which is used for the calculated value for initialising cluster centres and also for determining the number of clusters. Cluster analysis was conducted to figure out the various patterns related with the districts diminishing Banana productivity. Whereas the elbow approach was used for calculating the number of clusters, the K-Means algorithm has been shown to be useful for identifying similar groups. This study points out districts that have been properly grouped based on the specified variables.

ACKNOLEDGEMENT

I thank to Directorate Economics and Statistics for providing Dataset to me and also thank to Department of Physical Sciences and Information Technology TNAU, Coimbatore.

COMPETING INTEREST

Authors hereby declare that there is no competing interest.

REFERENCES

1. Sujatha, P., *Hybrid Statistical Models for Forecasting Yield of Mango and Banana in Tamil Nadu, India.* Asian Journal of Agricultural Extension, Economics & Sociology, 2021. **39**(11): p. 168-174.
2. Rathod, S. and G. Mishra, *Statistical models for forecasting mango and banana yield of Karnataka, India.* Journal of Agricultural Science and Technology, 2018. **20**(4): p. 803-816.
3. Dash, B., et al., *A hybridized K-means clustering approach for high dimensional dataset.* International Journal of Engineering, Science and Technology, 2010. **2**(2): p. 59-66.
4. Aggarwal, D. and D. Sharma, *Application of clustering for student result analysis.* International Journal of Recent Technology and Engineering, 2019. **7**(6): p. 50-53.
5. Groenendyk, D., et al., *A k-means clustering approach to assess wheat yield prediction uncertainty with a HYDRUS-1D coupled crop model.* 2014.
6. Abirami, B., et al., *Application of K-means Clustering Algorithm in Rice Production of Tamil Nadu, India.* International Journal of Environment and Climate Change, 2022. **12**(11): p. 1348-1355.
7. Khairani, N.A. and E. Sutoyo, *Application of k-means clustering algorithm for determination of fire-prone areas utilizing hotspots in West Kalimantan Province.* Int. J. Adv. Data Inf. Syst, 2020. **1**(1): p. 9-16.
8. Aldino, A., et al. *Implementation of K-means algorithm for clustering corn planting feasibility area in south lampung regency*. in *Journal of Physics: Conference Series*. 2021. IOP Publishing.
9. Santhanam, T. and M. Padmavathi, *Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis.* Procedia Computer Science, 2015. **47**: p. 76-83.
10. Kusak, L., et al., *Apriori association rule and K-means clustering algorithms for interpretation of pre-event landslide areas and landslide inventory mapping.* Open Geosciences, 2021. **13**(1): p. 1226-1244.