Original Research Article

A Comparative Study of Multivariate Clustering Algorithms for Climate Variables and Chickpea Production in Dharwad District of Karnataka, India

ABSTRACT

India's agriculture sector holds immense significance in its economy, and this study delves into how different clustering techniques can enhance our comprehension of the interplay between weather patterns and chickpea production in Dharwad district,specificallyDensity-Based Spatial Clustering of Applications with Noise(DBSCAN), K-Means, and Hierarchical clustering methods to discern underlying patterns within climate variables and chickpea yield data. Leveraging historical weather data spanning an extensive 42-year period and crop yield records were employed to train and validated. The findings illuminate that K-Means clustering consistently outperforms DBSCAN and Hierarchical clustering when evaluated through a variety of validity indices, including the Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index, and the determination of the optimal number of clusters. In essence, India's pivotal agricultural sector dynamics and the intricate relationship between climatic factors and chickpea production in Dharwad district are elucidated effectively by the superiority of K-Means clustering in this study.

Keywords:*Density-Based Spatial Clustering of Applications with Noise, K-MeansClustering, Hierarchical Clustering, Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index*

INTRODUCTION

India's economy relies heavily on agriculture, especially during covid-19, when it played a pivotal role in contributing to the country's GDP. However, changes in climate conditions can have detrimental effects on the growth of the agricultural sector (Mishra*et al.*, 2021). Climate change poses a danger to long-term growth. The agriculture sector is the most vulnerable to climate change compared to other industries. Consequently, agricultural vulnerability to climate change, many rural people live in extreme poverty and face food insecurity. Understanding the causes and impacts of climate change is vital for developing adaptation and mitigation programs to ensure sustainable agricultural output and eliminate hunger (Rasul and Sharma, 2016). Climate specifically temperature, precipitation and distributionis a significant element in agricultural production. The consequences of shifting climate on agricultural productivity cause widespread concern worldwide (Choudhury*et al*., 2017).

The primary origin of this concern is the potential harm and gains from climate change impacts on agriculture such as productivity and farmer return which will affect domestic and international policies, resource allocation and food security. In comparison to the advantages, there has been a greater emphasis on potential harms. Higher carbon levels in the atmosphere along with high temperatures altered precipitation patterns and potentially increased frequency of extreme events like droughts and floods would reduce yields and increase risks in agricultural output in numerous regions across the globe (Kumar *et al*., 2018).

Rainfall, temperature, humidity, sunlight, clouds, wind speed, evapotranspiration and other weather factors have a big impact on farming. They greatly affect how crops grow, develop and produce yields. They also influence the occurrence of pests and diseases, the amount of water needed and the necessary fertilizer (Sawan, 2018). This is due to variations in nutrient mobilization caused by water stress and the timing and efficacy of preventive measures, scientific practices and cultural activities associated with different crops. Weather anomalies can exacerbate crop damage and soil erosion (Anon, 2015). Furthermore, weather factors influence the quality of crop products as they progress from the field to storage and subsequently to the market resulting in economic fluctuations.

Clustering algorithms are a fundamental component of machine learning and data analysis, serving as powerful tools for grouping data points with shared characteristics or features. The primary objective of clustering is to unveil underlying patterns and structures within a dataset, effectively organizing data points into clusters where members exhibit greater similarity to one another than to those in different clusters(Sarker, 2021).This process enables the identification of natural groupings within data, facilitating insights and decision-making in various domains. By utilizing mathematical and statistical techniques, clustering algorithms aim to optimize the arrangement of data points based on predefined criteria, making them invaluable in tasks such as customer segmentation, image recognition, and anomaly detection. With diverse clustering methods available, each tailored to different data types and scenarios, clustering algorithms play a pivotal role in extracting meaningful information from complex datasets and contributing to data-driven solutions across a wide range of applications (Ahmed*et el.*, 2016)

MATERIALS AND METHODS

Nature of study area

Dharwad is situated in the North-Western part of Karnataka, with geographic coordinates at 15˚27’30” North latitude and 75˚00’30” East longitude. It resides at an elevation of 741 meters above sea level and experiences a tropical wet climate. The district receives an average annual rainfall of 864 mm, and the average temperature hovers around 24.1˚C. The climate is characterized by mild heat during the summer months of April and May, while the rest of the year tends to be pleasant. Interestingly, the hottest season begins in December, which is also the coldest month in Dharwad.The study area is shown in Fig. 1.

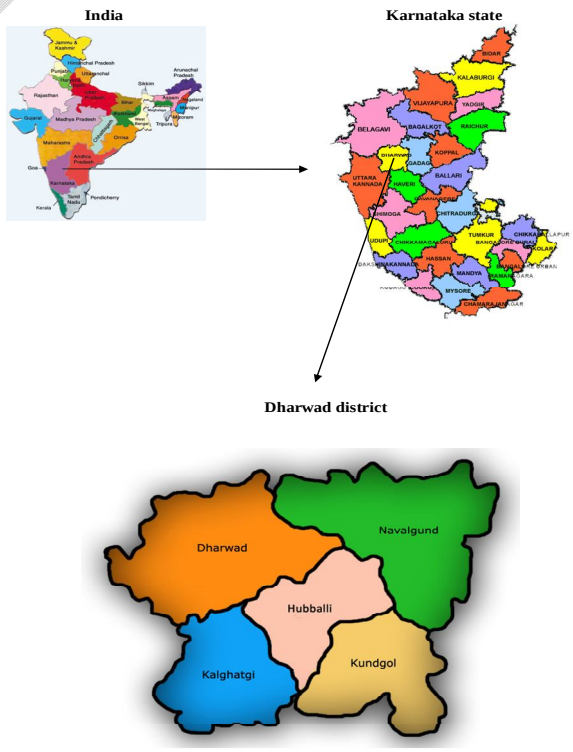

Fig. 1. Map showing study area

Nature and sources of data

The current study is based on a dataset comprising secondary data on various weather parameters, including Temperature (°C), Relative humidity (%), Rainfall (mm), Wind

speed (km/hr), and latent heat flux (W/m2). This dataset spans a 41-year period from 1980 to 2021 and was collected from multiple sources, including the Department of Agrometeorology at UAS, Dharwad, and CMIP-6 (Climate Model Intercomparison Project) with a resolution of 0.5 x 0.5 degree. The data was then analysed and sorted to the specific locations using the FERRET software in the Linux platform. Crop yield data were collected from the District Statistical Office, Dharwad and Directorate of Economics and Statistics, Bangalore

Clustering algorithms

Clustering algorithms are computational techniques in data analysis and machine learning to categorize similar data points into clusters based on their closeness. Clustering aims to recognize patterns within the data without prior knowledge of the groups. The training dataset comprises data on rainfall, relative humidity, wind speed, temperature, solar radiation, latent heat flux and production of major pulse crops in Dharwad district. This algorithm was written using Python language with pandas, matplotlib, and seaborn libraries.

DBSCAN Clustering algorithm

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm starts by inputting a dataset of spatial data points and defining key parameters, namely the radius (ε) and the minimum number of points (MinPts) within ε to constitute a core point. All data points are initialized as unvisited and marked as noise initially (Wibisono*et el.*, 2021). Core points are identified by assessing the number of points within ε, and clusters begin by selecting an unvisited core point and expanding to include all reachable points within ε. The process is repeated until all data points have been visited, and any remaining unvisited points are labeled as noise. DBSCAN is particularly useful as it doesn't require a predetermined number of clusters, making it adept at automatically determining cluster structures based on data density and parameter settings(Wang and Chen, 2020).

K-Means clustering algorithm

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning data into groups or clusters. It aims to divide a dataset into K distinct clusters, where each data point belongs to the cluster with the nearest mean (centroid). The algorithm works iteratively to find the optimal cluster centers by minimizing the within-cluster sum of squares.

K-Means clustering begins by selecting the desired number of clusters, K, to identify within the dataset. Cluster centroids are initialized by randomly choosing K data points from the dataset or using methods like K-means++. Data points are then assigned to the nearest cluster centroid based on Euclidean distance calculations. After assigning data points, the cluster centroids are updated by computing the mean of all points within each cluster. Steps 3 and 4 are iteratively repeated until a convergence condition is met, typically when cluster assignments and centroids remain stable or after a specified number of iterations (Morissette and Chartier, 2013). The final outcome includes K cluster centroids and the assignments of data points to their respective clusters, representing the K-Means clustering result.

Hierarchical clustering algorithm

Hierarchical clustering is a method of cluster analysis that builds a hierarchy of clusters by repeatedly merging or dividing them. It is used to group similar data points or objects into clusters based on their similarity, forming a tree-like structure called a dendrogram. Hierarchical clustering is particularly useful when the underlying data does not have a clear number of clusters, and you want to explore different levels of granularity in the grouping (Aarthi, 2019).

Hierarchical clustering initiates by treating each data point as an individual cluster. Pairwise distances are computed between clusters or data points using a chosen distance metric. The two closest clusters are then merged, reducing the total cluster count by one. After each merge, the distance matrix is updated, and distances between the newly merged cluster and remaining clusters are recalculated. This merging process is iteratively repeated until only one cluster remains, forming a dendrogram that visually represents the hierarchy of clustering. To determine a specific number of clusters, the dendrogram can be cut at a height corresponding to the desired cluster count or based on the problem's context(Mahmoud and Williams, 2016). Finally, cluster labels are assigned to data points according to the chosen cutting point, providing a hierarchical clustering result.

Clustering validation

Clustering validation is an important step in evaluating the quality and effectiveness of clustering algorithms. It helps assess the performance of the clustering algorithm and determine the optimal number of clusters or the best clustering solution (Majumdar*et el.*, 2017)

a) Silhouette Index

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The score quantifies how close a sample is to its own cluster compared to other clusters. The Silhouette Coefficient ranges from -1 to 1, where a value close to 1 indicates well-separated clusters, a value close to 0 indicates overlapping or ambiguous clusters, and a value close to -1 indicates incorrect or misclassified clusters.

$$S(i) = \frac{(b(i)-a(i))}{(\max\{(a(i),b(i))\}} \quad (1.1)$$

Where,

a(i) is the average dissimilarity of i^{th} object to all other objects in the same cluster

b(i) is the average dissimilarity of i^{th} object with all objects in the closest cluster.

b) Calinski-Harabasz Index:

The Calinski-Harabasz Index measures the ratio of between-cluster dispersion to within-cluster dispersion. It evaluates the compactness and separation of clusters. The index has no predefined range. A higher Calinski-Harabasz score indicates better-defined and more compact clusters.

c) Davies-Bouldin Index:

The Davies-Bouldin Index measures the average similarity between clusters, considering both cluster compactness and separation. Lower values indicate better-defined and more separated clusters. The index has no predefined range. A lower Davies-Bouldin index implies better clustering.

RESULTS AND DISCUSSION

The Silhouette Index for DBSCAN, K-Means and Hierarchical clustering was identified as 0.43, 0.53 and 0.12, respectively. The Calinski-Harabasz Index for DBSCAN, K-Means and Hierarchical clustering algorithms was identified as 15.25, 81.28 and 2.21,respectively. Whereas the Davies-Bouldin Index for DBSCAN, K-Means and Hierarchical clustering

algorithms was identified as 1.11, 0.83 and 0.88, respectively.

Table 1: Comparison of various clustering algorithms based on climatic variables and production of chickpea in Dharwad district

Validity index	DBSCAN Clustering	K -Means Clustering	Hierarchical Clustering
Silhouette Coefficient	0.43	**0.53**	0.12
Calinski-Harabasz Index	15.25	**81.28**	2.21
Davies-Bouldin Index	1.11	**0.83**	0.88
Number of clusters formed	1	3	3
Number of outliers	5	-	-

Table 2: Clustering algorithms based on climatic variables and production of chickpea in Dharwad district

Clustering	No. of cluster	Years	Average production
DBSCAN clustering	**I (High)**	1980, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019	23114.62
K-Means clustering	**I (High)**	1981, 1989, 1991, 1993, 1996, 2020, 2021	32270.43
	II (Medium)	1982, 1985, 1987, 1994, 1995, 2000, 2001, 2002, 2003, 2004, 2007, 2008, 2010, 2011, 2014, 2015, 2016, 2017	23970.83
	III (Low)	1980, 1983, 1984, 1986, 1988, 1990, 1992, 1997, 1998, 1999, 2005, 2006, 2009, 2012, 2013, 2018, 2019	22748.65
Hierarchical clustering	**I (High)**	1981, 1991, 1993, 2020, 2021	36228.00
	II (Medium)	1980, 1982, 1983, 1985, 1987, 1989, 1994, 1995, 1996, 2000, ,2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2010, 2011, 2014, 2015, 2016, 2017, 2019	24121.50
	III	1984, 1986. 1988, 1990, 1992, 1997, 1998, 2005,	21849.15

(Low)	2009, 2012, 2013, 2018

The number of clusters formed by DBSCAN, K-Means and Hierarchical clustering algorithms was identified as 1, 3 and 3, respectively.The DBSCAN clustering algorithm detected a total of five outliers, which were observed in the years 1981, 1991, 2007, 2020 and 2021. This information is presented in Table 1. For DBSCAN clustering, only one cluster (cluster-I) was formed, with an average production of 23114.62 Tons. In the case of K-Means clustering, most years were assigned to cluster-I, with an average production of 32270.83 Tons, followed by cluster-II and cluster-III with average productions of 23970.83 Tons and 22748.65 Tons, respectively. In the Hierarchical clustering, most years fell into cluster-II, with an average production of 24121.50 Tons, followed by cluster-III and cluster-I with average productions of 21849.15 Tons and 36228.00 Tons, respectively (Table 2).

K-Means clustering algorithm was identified as the most effective when compared with Silhouette Index, Calinski-Harabasz Index and Davies-Bouldin Index. There were an equal number of clusters were formed in the case of Hierarchical clustering and K-Means clustering, followed by DBSCAN clustering. It can be concluded that DBSCAN clustering can also be utilized for detecting five outliers when there is a deviation from the normal data, as shown in Table 1.

Similar results were obtained by Shobha N and Asha T (2017) while studying agricultural meteorological patterns, where K-Means and Hierarchical clustering techniques were employed to extract patterns of air temperature, relative humidity, rainfall, and pan evaporation. In their study, K-Means clustering was found to be the superior clustering algorithm based on criteria such as Connectivity, Silhouette width and Dunn index.The obtained results aligned with the findings of a study conducted by Shwetha (2009), which reported a cluster analysis incorporating groundwater and rainfall. The rainfall and groundwater level were classified based on intensity and the depth of water table in each variable was categorized into five clusters formed based on squared Euclidean distance measures,but the present study advances by incorporating more sophisticated algorithms, focusing on chickpea production, and introducing a more comprehensive evaluation and handling of outliers

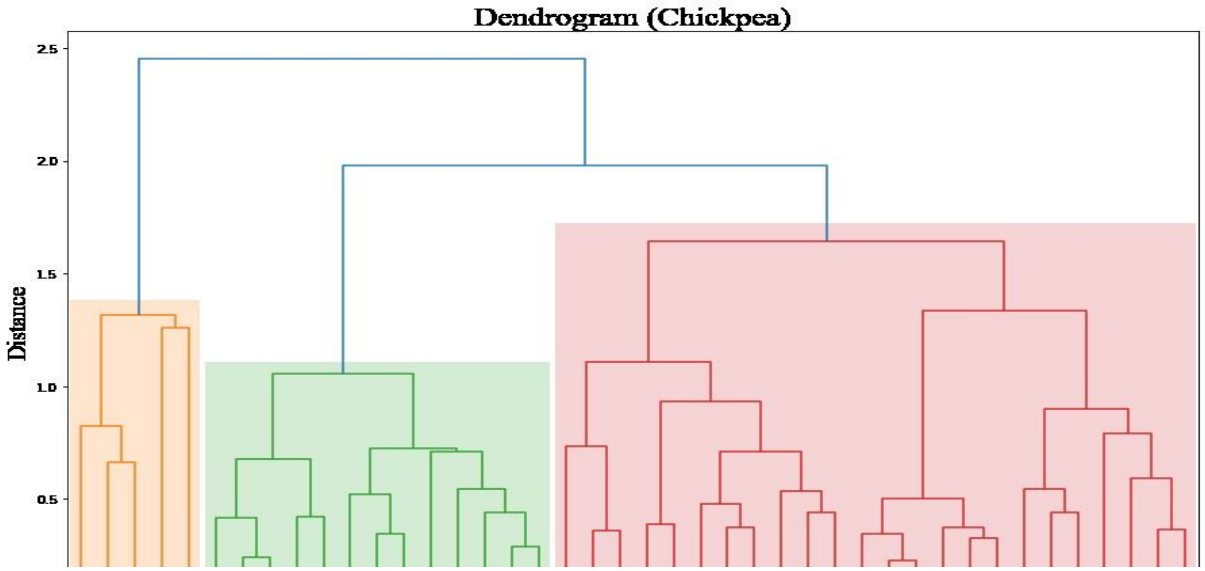

Fig. 2: Dendrogram

CONCLUSION

In this comprehensive study exploring the intricate relationship between climatic factors and chickpea production in Dharwad district, three prominent clustering techniquesDBSCAN, K-Means, and Hierarchical clusteringwere rigorously employed. The results unequivocally highlight the supremacy of the K-Means clustering method, which consistently outperformed its counterparts across various validity indices, including the Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index, while revealing the optimal formation of three distinct clusters. These findings shed valuable light on the dynamics of India's pivotal agricultural sector and the profound impact of climate variables on chickpea production in Dharwad. It is evident that K-Means clustering stands as a robust tool for uncovering hidden patterns within this intricate interplay of agricultural and meteorological data. However, it is crucial to acknowledge that the choice of clustering algorithm should remain adaptable to the specific dataset and research objectives, as demonstrated by the nuanced nature of this study's findings, and as echoed in similar research endeavours.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

Conference disclaimer:

Some part of this manuscript was previously presented and published in the conference: Cutting-Edge Solutions in Science-Agriculture, Technology,Engineering and Humanities (CSATEH-2024)- 6 International Conference dated from 24th -26th August 2024 in Uttarakhand, India, Web Link of the proceeding: https://agetds.com/wp-content/uploads/2024/06/Circular-CSATEH-2024-Nainital.pdf

REFERENCES

Aarthi R, 2019, Cluster analysis of extreme rainfall seasons in a particular region of Tamil Nadu districts. Journal of the Gujarat Research Society, 21(17): 94-114.

Ahmed M, Mahmood AN and Islam MR, 2016, A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278-88.

Anonymous, 2015, Climate change and food security: risks and responses. https://www.fao.org.

Chowdhury R B, Moore G A, Weatherley A J and Arora M, 2017, Key sustainability challenges for the global phosphorus resource, their implications for global food security and options for mitigation. Journal of Cleaner Production, 140(1): 945-963.

Kumar P, Tokas J, Kumar N, Lal M and

Singal H R, 2018, Climate change consequences and its impact on agriculture and food security. International Journal of chemical studies, 6(6): 124-133.

Mahmoud A and Williams G, 2016, Detecting, classifying, and tracing non-functional software requirements. *Requirements Engineering*,21:357-381.

Majumdar J, Naraseeyappa S and Ankalaki S, 2017, Analysis of agriculture data using data mining techniques: application of big data. Journal of Big Data, 4(1): 20.

Mishra A, Bruno E and Zilberman D, 2021 Compound natural and human disasters: Managing drought and COVID-19 to sustain global agriculture and food sectors. *Science of the Total Environment*, 7(5): 142-210.

Morissette L and Chartier S, 2013, The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, 9(1):15-24.

Rasul G and Sharma B, 2016, The nexus approach to water–energy–food security: an option for adaptation to climate change. *Climate policy*, 16(6):682-702.

Sarker IH, 2021, Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160.

Sawan ZM, 2018, Climatic variables: Evaporation, sunshine, relative humidity, soil and air temperature and its adverse effects on cotton production. *Information processing in agriculture*,5(1):134-148.

Shobha N and Asha T, 2017, Monitoring weather-based meteorological data: clustering approach for analysis. *International conference on innovative mechanisms for industry applications*, 75-81.

Shwetha K S, 2009, A statistical study on the impact of rainwater harvesting on farming economy. *M.Sc. (Agri.) Thesis*, University of Agricultural Sciences, Dharwad, Karnataka (India).

Wang S and Chen C, 2020, Short-term wind power prediction based on DBSCAN clustering and support vector machine regression. *5th International Conference on Computer and Communication Systems*, 941-945.

Wibisono S, Anwar M T, Supriyan A and Amin H A, 2021, Multivariate weather anomaly detection using DBSCAN clustering algorithm. *In Journal of Physics: Conference Series*, 18(1): 12-77.