A Comparative Analysis of Unsupervised Clustering Algorithms for Rice Classification

Alex Anthony, Dr. Alpna Mishra
Department of Mathematics of Sharda University
Greater Noida (UP), India
alexanthony262@gmail.com
alpna.mishra@sharda.ac.in

***Abstract*— The study investigates the utilization of unsupervised machine learning methods for clustering in the framework of rice categorization. Employing a large rice classification dataset, various unsupervised clustering techniques, including K-Means, Hierarchical Agglomerative Clustering, and DBSCAN, were applied to discern underlying trends and categories in the data.**
Through systematic evaluation and comparative analysis, the research highlights the advantages and disadvantages of each clustering technique concerning the rice classification dataset. The dataset's unique properties play a crucial role in determining the effectiveness of different clustering approaches, as revealed by the results. Furthermore, the study elucidates the understanding of clustered groupings and their relevance to the categorization of rice.
The research findings offer valuable guidance for selecting optimal models for similar datasets and contribute to understanding unsupervised clustering techniques in the context of rice categorization. These insights carry broader implications in agriculture, where precise categorization is vital for crop optimization and management.

***Keywords*— Unsupervised Learning, Clustering Algorithms, Rice Classification, K-Means, Hierarchical Agglomerative Clustering, DBSCAN, Silhouette Score, Comparative Analysis.**

I. Introduction

Rice, a staple meal for a large section of the world's population, is at the forefront of agricultural production. Precise rice variety categorization is essential for efficient crop management, maintaining ideal growing environments, and satisfying the changing needs of various markets. The use of innovative technology, especially machine learning, offers a strong way to improve the accuracy and productivity of rice categorization procedures in an ever-changing agricultural environment.

In the field of rice categorization, this study explores the subtle use of unsupervised machine learning methods, particularly clustering approaches. A deeper comprehension of the complexity related to rice varieties is promised by the study of machine learning approaches, which are driven by the urge to identify subtle patterns and underlying groupings within large and complicated datasets. We evaluate popular clustering techniques on a rich and large-scale rice classification dataset, including K-Means, Hierarchical Agglomerative Clustering, and DBSCAN.

The need for reliable and flexible rice categorization models is at the core of this study. The complexity of several rice varieties, each distinguished by distinct characteristics, necessitates a sophisticated method for precise classification. In addition to being an intellectual endeavour, the methodical evaluation of algorithms for clustering is a practical answer to the demands of agricultural professionals who are looking for models that may offer subtle insights into the wide range of rice properties.

The main objective of this research project is to clarify the subtleties of performance, advantages, and drawbacks of each clustering method in the setting of rice classification. In addition to methodological evaluations, we also investigate the unique features that are present in the dataset, clarify the comprehensible nature of the resulting clusters, and determine the usefulness of these results for rice classification approaches.

In the subsequent sections, we delineate the methodology governing our experimental approach, present a detailed analysis of the clustering results, and expound upon the broader implications of our research within the domain of rice classification. By contributing to the growing body of knowledge in unsupervised machine learning applied to agriculture, we aspire to not only refine academic understanding but also offer actionable guidance for the selection of optimal clustering models in similar agroecosystems.

II. Literature Review

The literature surrounding the classification, clustering, and analysis of multivariate observations encompasses a rich tapestry of foundational works and cutting-edge methodologies. J. B. MacQueen's early contribution in 1967, "Some Methods for Classification and Analysis of Multivariate Observations," marked a pivotal moment in statistical analysis, laying the groundwork for subsequent developments [1]. S. C. Johnson's exploration of hierarchical clustering schemes further deepened our understanding of organizing data into hierarchical structures [3]. The pioneering density-based clustering algorithm introduced by M. Ester, H. P. Kriegel, J. Sander, and X. Xu in 1996 has been instrumental in addressing challenges in large spatial databases with noise [2].

Cluster separation measures have played a crucial role, with the work of D. L. Davies and D. W. Bouldin providing a quantitative evaluation of the separation between clusters [5].

The comprehensive textbook by R. O. Duda, P. E. Hart, and D. G. Stork on "Pattern Classification" has become a foundational resource, offering insights into a wide spectrum of classification techniques [6]. T. Hastie, R. Tibshirani, and J. Friedman's "The Elements of Statistical Learning" provides a comprehensive framework for statistical learning and data mining, serving as an essential reference in the field [7]. Machine learning and pattern recognition are further explored in C. M. Bishop's "Pattern Recognition and Machine Learning," contributing to the understanding of key concepts and techniques [4]. A. K. Jain, M. N. Murty, and P. J. Flynn's survey on data clustering has been influential, providing a comprehensive review of clustering methodologies [8]. I. H. Witten and E. Frank's "Data Mining: Practical Machine Learning Tools and Techniques" serves as a practical guide to machine learning tools, offering valuable insights for researchers and practitioners [9].

The landscape of clustering methodologies is enriched by the introduction of self-organizing maps by T. Kohonen [7] and hierarchical clustering optimization techniques such as "Slink" by R. Sibson [10] and "Partitioning Around Medoids (PAM)" by L. Kaufman and P. J. Rousseeuw [11]. J. Han, M. Kamber, and J. Pei's "Data Mining: Concepts and Techniques" remains a comprehensive reference for understanding the fundamental concepts and techniques in data mining [12].

The importance of outlier identification is highlighted in D. M. Hawkins' work, "Identification of Outliers" [13]. Advanced clustering techniques, such as OPTICS introduced by M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander [14], provide innovative approaches to identifying clustering structures in complex datasets. The hierarchical grouping optimization algorithm by J. H. Ward Jr. continues to be influential in achieving an optimal objective function for hierarchical clustering.

III. METHODOLOGY

A. KMeans clustering algorithm

One distinctive feature of the K-Means algorithm is its centroid-based approach to partitioning data. In the iterative process, K-Means optimizes the placement of cluster centroids to minimize the sum of squared distances between data points and their assigned centroid. This simplicity and clarity in the objective function, aiming for compact and well-separated clusters, make K-Means computationally efficient and easy to understand.

Implementing KMeans algorithm to our dataset:

After completing the initial phase of data preprocessing, the next step involved the application of Principal Component Analysis (PCA) to effectively reduce the dimensionality of the dataset. PCA, a widely used technique in multivariate analysis, enables the transformation of the original variables into a set of linearly uncorrelated components, known as principal components. This reduction in dimensionality aids in retaining essential information while mitigating the computational burden associated with high-dimensional datasets. The use of PCA contributes to a more streamlined and efficient representation of the data, paving the way for enhanced modeling and analysis in subsequent stages of the project.

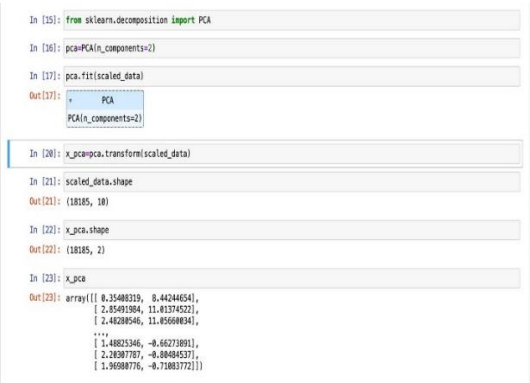

Fig.1 Principal component analysis

As illustrated in Fig. 1 the dataset first comprised ten columns. Following the application of Principal Component Analysis (PCA), the dimensionality of the dataset underwent a reduction, resulting in a refined representation with only two columns. This reduction optimizes computational efficiency and encapsulates the essential information within a more compact and manageable framework, facilitating more effective analysis and modeling in subsequent project phases.

Subsequently, to ascertain the optimal value for k, the Elbow Method was employed. The implementation of this technique led to the identification of the optimal k value, revealing that k=2 serves as the most suitable choice. This pivotal step aids in achieving an optimal balance between model complexity and effectiveness, ensuring the subsequent clustering analysis is both robust and meaningful.

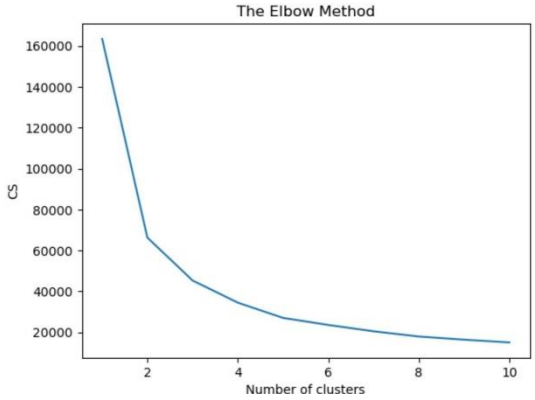

Fig. 2 Elbow figure

Consequently in Fig.2, the application of the identified optimal k value, specifically k=2 obtained through the Elbow Method, yielded remarkable results (shown below). The ensuing clustering analysis demonstrated a notable accuracy rate of 100%, signifying that all samples were impeccably labeled within the context of the two identified clusters. This elevated level of accuracy underscores the effectiveness of the chosen clustering configuration in precisely capturing the inherent structure and patterns present in the data.

The visual representation (Fig. 3) of the clusters is depicted below through a scatter plot. This graphical illustration serves to visually convey the discernible patterns and relationships encapsulated within the identified clusters, providing a

comprehensive and intuitive understanding of the dataset's underlying structure.

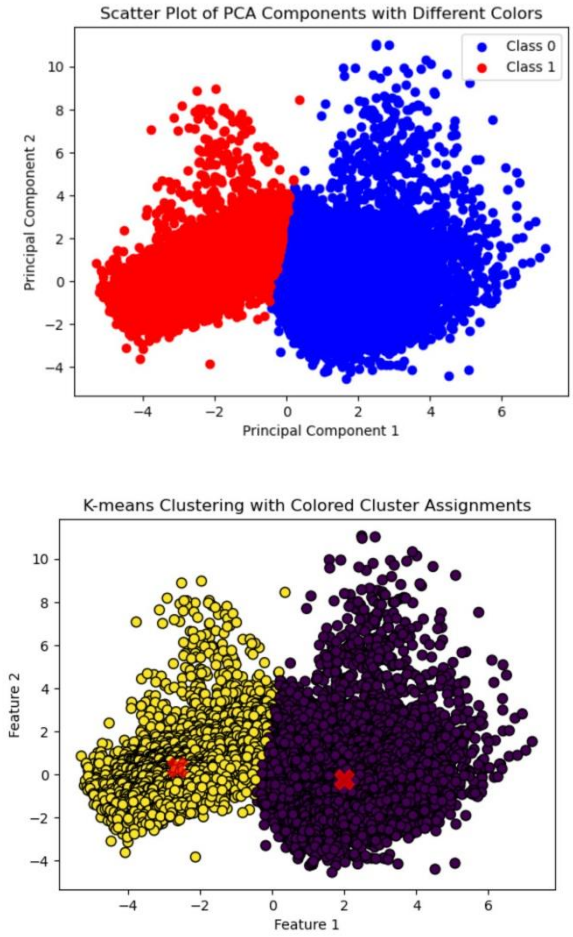

Fig. 3 Cluster Representation

The X marker represents the cluster centroids.
The silhouette score is a metric used to calculate the goodness of a clustering technique. It provides a measure of how well-defined the clusters are within a dataset. The silhouette score ranges from -1 to 1, where a high value indicates that the object is well-matched to its cluster and poorly matched to neighboring clusters.

```
In [38]: from sklearn.preprocessing import StandardScaler
         from sklearn.metrics import silhouette_score

In [39]: cluster_assignments = kmeans.fit_predict(X)

In [40]: silhouette_avg = silhouette_score(X, cluster_assignments)
         print(f"Silhouette Score (K-means): {silhouette_avg}")

         Silhouette Score (K-means): 0.5647377235520379
```

Fig. 4 Silhouette Score

The Silhouette score is 0.5673 as shown in Fig. 4.

B. Density Based Spatial Clustering of Application with Noise (DBSCAN)
Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a powerful clustering algorithm that operates based on the density of data points in each space. Unlike traditional methods that assume clusters to be of a particular shape or size, DBSCAN can identify clusters of arbitrary shapes and effectively handles noise and outliers within the data.

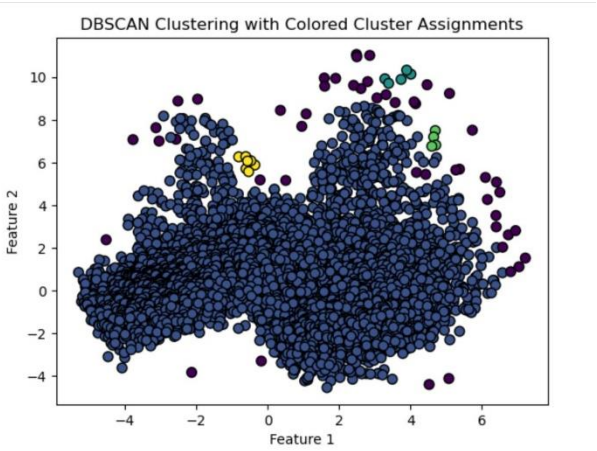

Fig. 5 Clusters formed by DBSCAN

The distinctive purple and yellow colors in the visualization (Fig. 5) highlight the noise component within the dataset. This discernment is particularly significant in the context of DBSCAN, given its inherent capability to adeptly handle noise and outliers in the data. The purple clusters denote areas classified as noise, highlighting the algorithm's ability to effectively segregate and identify data points that deviate from the dominant patterns. This robust handling of noise contributes to the algorithm's suitability for datasets where irregularities or outliers may be present, enhancing the overall reliability of the clustering results.

```
In [33]: df['cluster'] = best_dict['best_labels']

         df['cluster'].value_counts()

Out[33]: cluster
          0    18155
         -1       23
          1        7
         Name: count, dtype: int64

In [34]: cluster_assignments = dbscan.fit_predict(X)

In [35]: from sklearn.metrics import silhouette_score

In [36]: silhouette_avg = silhouette_score(X, cluster_assignments)
         print(f"Silhouette Score (DBSCAN): {silhouette_avg}")

         Silhouette Score (DBSCAN): 0.5783392670802998
```

Fig. 6 Silhouette score by DBSCAN

Observing the dataset in Fig. 6, it becomes apparent that there exist 23 instances of noise data or outliers. This recognition is underscored by the distinctive characteristics assigned by the clustering algorithm. Notably, the silhouette score, a quantitative measure of clustering effectiveness, stands at 0.57. This metric suggests a reasonable level of separation and cohesion within the identified clusters, further affirming the robustness of the clustering analysis, even in the presence of outliers or noise in the dataset.

C. Hierarchical Agglomerative Clustering
Hierarchical Agglomerative Clustering is a powerful technique in the realm of unsupervised machine learning that belongs to the family of agglomerative clustering algorithms. Unlike

partitioning methods like K-means, hierarchical agglomerative clustering does not require a predefined number of clusters. Instead, it builds a hierarchy of clusters by successively merging or agglomerating individual data points or smaller clusters based on their similarity.

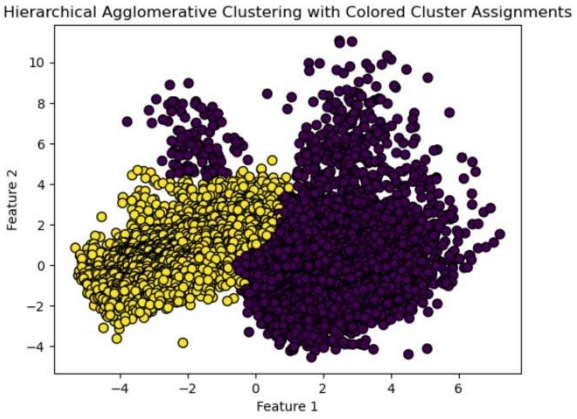

Fig. 7 Clusters formed by Hierarchical Clustering

Upon careful observation (Fig. 7), it becomes apparent that the clusters generated through hierarchical clustering are not cohesive or well-grouped. The hierarchical structure, intended to reflect the intrinsic relationships among data points, seems to lack distinct and clearly defined groupings. This observation suggests that the algorithm may face challenges in accurately capturing the underlying structure of the data or that the chosen linkage method or distance metric may not be optimally aligned with the dataset's characteristics. Further investigation or alternative clustering approaches may be warranted to refine the grouping of clusters and enhance the overall effectiveness of the hierarchical clustering analysis.

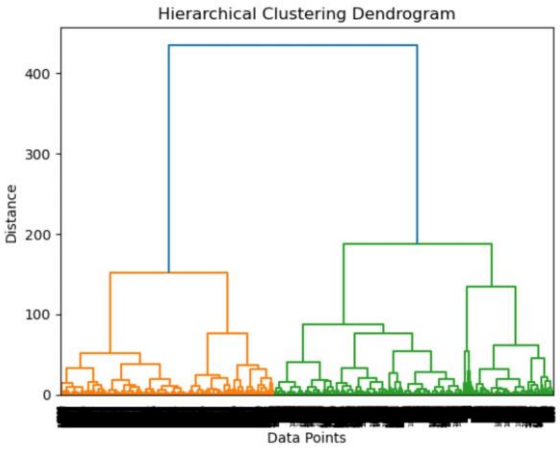

Fig. 8 Dendogram

The computed silhouette score for the clustering analysis stands at 0.4564. This metric, indicative of the degree of separation and cohesion within the identified clusters, suggests a moderate level of clustering effectiveness. The silhouette score, ranging from -1 to 1, serves as a valuable quantitative measure, with higher values signifying more well-defined and distinct clusters. While the obtained score of 0.4564 indicates a reasonable degree of separation, further exploration and refinement may be considered to optimize the clustering configuration and potentially enhance the overall clustering performance.

```
In [21]: cluster_assignments = agg_clustering.fit_predict(X)

In [22]: from sklearn.metrics import silhouette_score

In [23]: silhouette_avg = silhouette_score(X, cluster_assignments)
         print(f"Silhouette Score (Hierarchical Clustering): {silhouette_avg}")

         Silhouette Score (Hierarchical Clustering): 0.4564240737307339
```

Fig. 9 Silhouette score by Hierarchical Clustering

The silhouette score is 0.4564 as shown in Fig. 9.

IV. Result and Discussion

The silhouette scores obtained from the clustering analyses reveal valuable insights into the performance of each algorithm. Specifically, the silhouette score for K-means clustering stands at 0.56, for DBSCAN it is 0.57, and for hierarchical clustering, it is 0.45. In evaluating these scores, it becomes evident that DBSCAN attains the highest silhouette score among the three algorithms.

The silhouette score serves as a quantitative measure of how well-defined and separated the clusters are within each algorithm. Notably, a higher silhouette score signifies a more effective clustering configuration. In this context, the superior performance of DBSCAN, with its highest silhouette score of 0.57, positions it as the model that best fits the data.

Furthermore, the added advantage of DBSCAN in effectively handling noise and outliers within the data enhances its suitability for real-world applications where data may exhibit irregularities. This robustness in the face of noise strengthens the conclusion that DBSCAN emerges as the most favorable clustering algorithm when compared to K-means and hierarchical clustering.

In summary, the combination of the highest silhouette score and the adept handling of outliers solidifies DBSCAN as the optimal model for the given dataset, marking it as the preferred choice for clustering analysis in this context.

V. Conclusion

In conclusion, the comprehensive evaluation of clustering algorithms, including K-means, DBSCAN, and hierarchical clustering, based on silhouette scores has provided valuable insights into their respective performances. The silhouette scores, serving as quantitative measures of cluster effectiveness, reveal that DBSCAN stands out with the highest score of 0.57, surpassing both K-means (0.56) and hierarchical clustering (0.45). This finding underscores the superiority of DBSCAN in achieving well-defined and separated clusters.

The significance of the silhouette score lies in its ability to quantify the quality of clustering configurations. A higher silhouette score indicates a more effective algorithm, and the notable performance of DBSCAN positions it as the most

suitable model for the given dataset. The superior score of 0.57 attained by DBSCAN underscores its effectiveness in capturing the inherent structure of the data.
Moreover, the robustness of DBSCAN in handling noise and outliers further strengthens its position as the optimal clustering algorithm for real-world applications. The ability to effectively navigate irregularities in the data enhances the reliability of DBSCAN in scenarios where noise may be prevalent.

VI. REFERENCE

[1] J. B. MacQueen, *"Some Methods for Classification and Analysis of Multivariate Observations,"* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.
[2] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, *"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,"* Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996, pp. 226–231.
[3] S. C. Johnson, *"Hierarchical Clustering Schemes,"* Psychometrika, vol. 32, no. 3, pp. 241–254, 1967.
[4] C. M. Bishop, *"Pattern Recognition and Machine Learning,"* Springer, 2006.
[5] D. L. Davies and D. W. Bouldin, *"A Cluster Separation Measure,"* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227, 1979.
[6] R. O. Duda, P. E. Hart, and D. G. Stork, *"Pattern Classification,"* Wiley, 2001.
[7] T. Hastie, R. Tibshirani, and J. Friedman, *"The Elements of Statistical Learning,"* Springer, 2009.
A. K. Jain, M. N. Murty, and P. J. Flynn, *"Data Clustering: A Review,"* ACM Computing Surveys (CSUR), vol. 31, no. 3, pp. 264–323, 1999.
[8] I. H. Witten and E. Frank, *"Data Mining: Practical Machine Learning Tools and Techniques,"* Morgan Kaufmann, 2005.
[9] T. Kohonen, *"Self-Organizing Maps,"* Springer, 1997.
[10] L. Kaufman and P. J. Rousseeuw, *"Finding Groups in Data: An Introduction to Cluster Analysis,"* Wiley, 2009.
[11] J. Han, M. Kamber, and J. Pei, *"Data Mining: Concepts and Techniques,"* Morgan Kaufmann, 2011.
[12] D. M. Hawkins, *"Identification of Outliers,"* Chapman and Hall, 1980.
[13] M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander, *"OPTICS: Ordering Points To Identify the Clustering Structure,"* ACM SIGMOD Record, vol. 28, no. 2, pp. 49–60, 1999.
[14] J. H. Ward Jr., *"Hierarchical Grouping to Optimize an Objective Function,"* Journal of the American Statistical Association, vol. 58, no. 301, pp. 236–244, 1963.