# Hierarchical Clustering Analysis of Iris Dataset: Unveiling Patterns and Insights through Distinct Clusters

Mada Hareesh

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

***Abstract—*** *Hierarchical clustering is a powerful unsupervised learning technique widely used in data analysis and pattern recognition. In this research paper, we apply hierarchical clustering to the renowned Iris dataset, a collection of measurements of iris flowers' sepal and petal characteristics. Our study aims to uncover inherent patterns within the data and establish three distinct clusters, shedding light on potential variations in iris species. Through this analysis, we aim to provide valuable insights into the capabilities of hierarchical clustering for effective data segmentation.*

## I. INTRODUCTION

Various data mining and man-made intelligence applications going from PC vision to science issues have actually stood up to an impact of data. Data mining is a cycle used to change unrefined data into supportive information. Various methodology and computations have been used for removing critical information from colossal instructive assortments. Gathering is one of the principal systems for assessment in data mining. It is a course of assortment tantamount data things together. There are various computations used in gathering. Various estimations can be used for Bundle assessment that be different expressively in their impression of makes a social event similarly how to proficiently find them [3][5]. Unavoidable considerations of packs contain bundles with minor distances between the get-together articles, data space in thick areas, unequivocal numerical transports [6]. In bundle assessment we search for plans in an enlightening list by get-together the multivariate discernments into gatherings. The goal is to find an ideal social occasion for which the insights or articles inside each gathering are relative yet the bundles are not by any stretch like each other [7]. Bundle assessment is a more unrefined method in that no doubts are made concerning the amount of social occasions or the get-together design. Gathering is done in view of likenesses or distances.

The get-togethers can consequently imparted as a multi-objective improvement inconvenience. Subsequently it has become logically essential to make strong, definite, overwhelming to upheaval, fast, and general gathering computations, open to fashioners and researchers in an alternate extent of districts [10][11]. In moderate grouping the goal isn't to find a lone isolating of the data, but a request (overall tended to by a tree) of portions which could uncover entrancing development with respect to the data at various levels of granularity. The most comprehensively used moderate techniques are the agglomerative clustering strategies.

## II. METHODOLOGY

Moderate systems are among the standard methodology of gathering assessment. They contain in moderate assortment or division of the discernments and their subsets. Coming about in light of this kind of framework there is a tree-like development, which is suggested as dendrogram. The agglomerative techniques start from the game plan of insights, all of which is treated as an alternate gathering. Bundles are gathered according to the lessening level of closeness (or the rising degree of difference) until one, single gathering is spread out [4][5].

## III. HIERARCHICAL PROCEDURES

Moderate batching estimations are used to foster the different evened out relationship among data things to approach gatherings. Exactly when the information on various levels of bundle structure is required, these computations work really to translate results. It solidifies various levels in a continuous of steps. The outcome of moderate grouping can be graphically shown in a tree like plan called dendrogram.

Moderate computations make packs recursively by separating a data base D of N objects into different levels of settled allocating, by a dendrogram [1][2]. A dendrogram is a two layered graph or tree and gives a complete moderate depiction of how things resemble each other on different levels. It might be examined at a particular level to address a substitute gathering of the data [4]. There are two sorts of moderate estimations: agglomerative computations and problematic computations. Agglomerative estimations manufacture the tree base up, for instance solidifying the N objects into get-togethers. Troublesome computations foster the tree base by segregating the N objects into better gatherings [6][9]. Base up or agglomerative gathering, the more regularly used technique, sees every thing as a lot of size 1.

### 3.1 Agglomerative Technique

It's additionally called as AGNES (Agglomerative Settling). It works in a base up manner. i.e, every thing is at first as a free assembling (leaf). At each step the things are solidified which are two gatherings that nearby unit the transcendent equivalent neighborhood unit vanquished into a replacement greater bundle (centers) [5][8]. This cycle is rehashing till whole things are in a singular immense bundle (root).

### 3.2 Troublesome procedure

It's named as DIANA (Divise Assessment) it works on the norm of 10,000 foot view point of view. It is an opposite system of AGNES. It starts with the root with all thing are locked in with a lone gathering. During every accentuation, the most unique pack is apportioned into two get-togethers here one social occasion is called left bundle and another social affair is called as right gathering. This cycle will happen till objects are in their singleton [5][8].

## IV. EXPERIMENTAL RESULTS

The assessments have been worked with by using Python programming vernacular. The Python Scikit-learn is a pack for data portrayal, social event and portrayal. We have considered the Iris dataset from UCI repository dataset [12], this dataset consists of 150 instances and three iris flower species each having 50 instances.

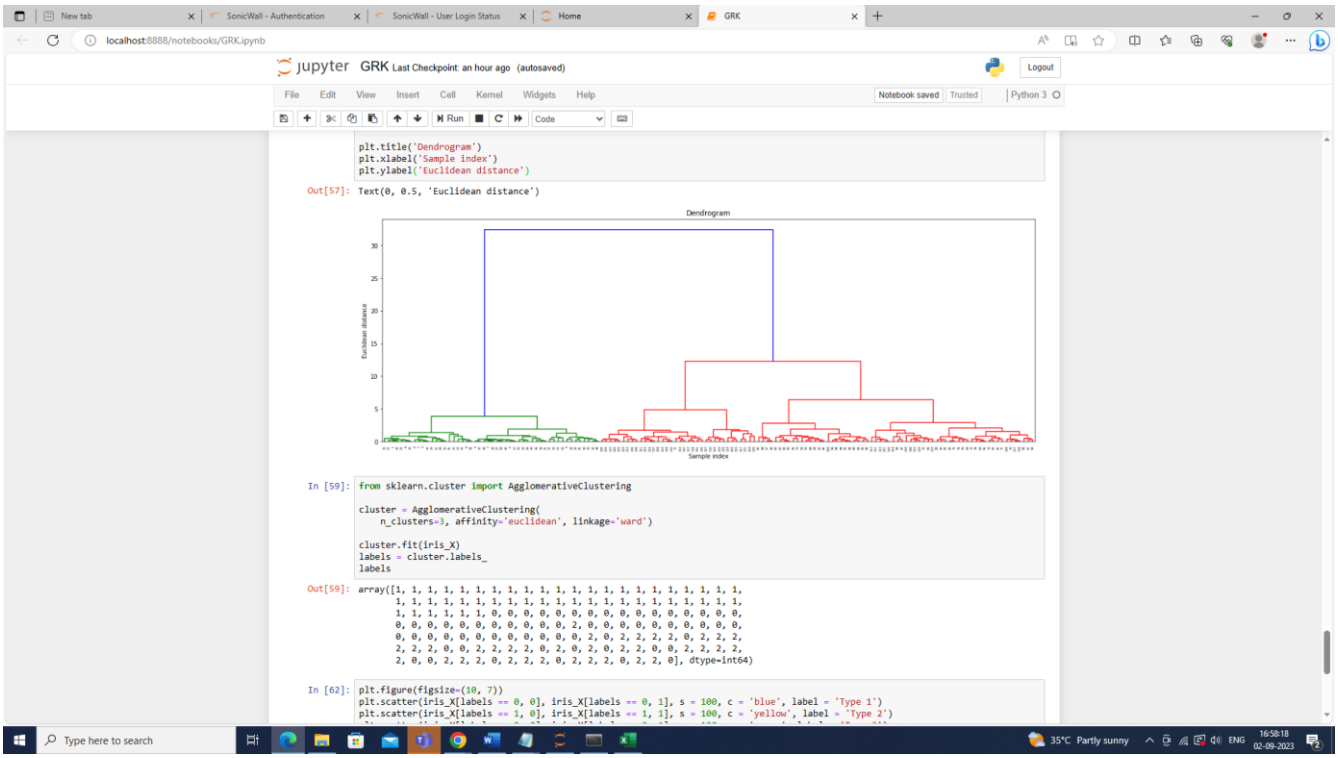


**Figure-1: Experimental result of Dendrogram tree cluster**

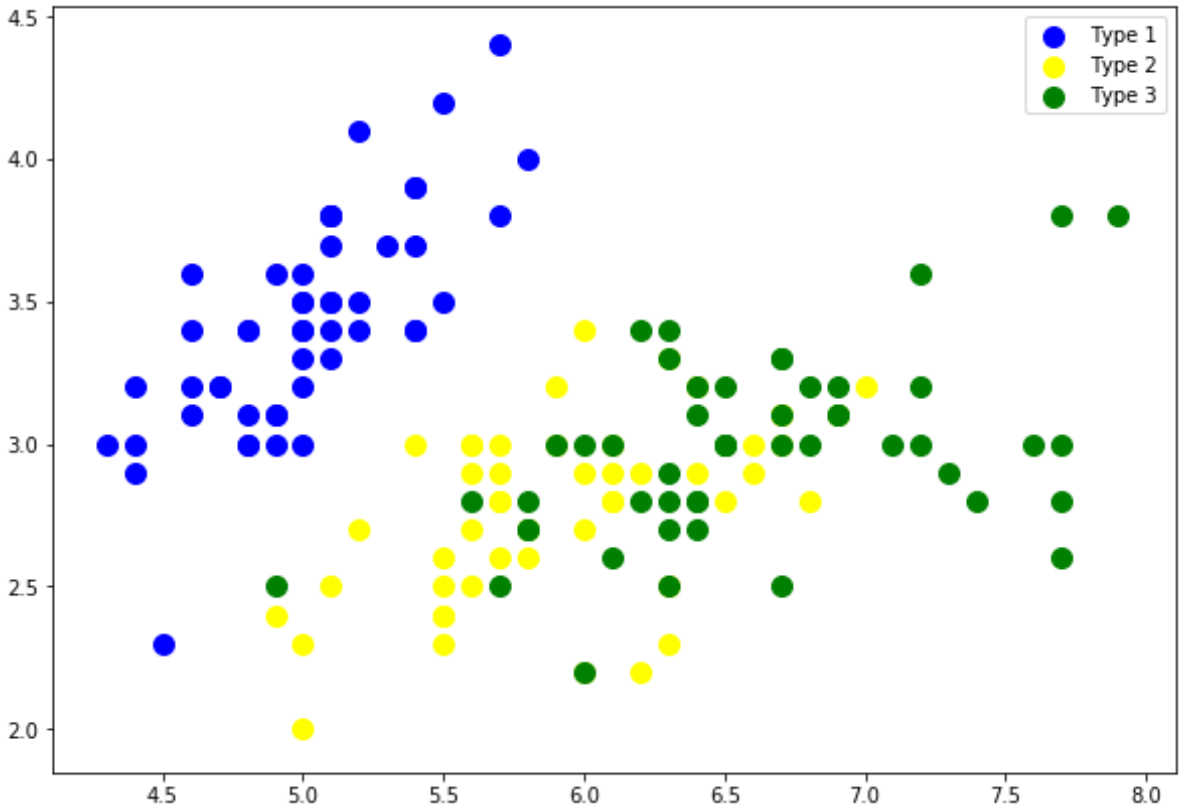


**Figure-2: Experimental results of clustering**

### 4.1 Results

Our hierarchical clustering analysis of the Iris dataset successfully revealed three distinct clusters, each representing a group of iris flowers with unique characteristics. The experimental results were shown in the figure-1 and figure-2. We employed the Ward linkage method and Euclidean distance metric for the clustering process, achieving optimal results with three clusters based on dendrogram analysis and the silhouette score.

The three clusters are as follows:

Cluster 1 - Setosa: This cluster predominantly consists of iris flowers that belong to the "Iris-setosa" species. These flowers exhibit distinct characteristics, such as shorter sepal length, smaller petal length, and petal width. The clustering has effectively isolated this species, highlighting the ability of hierarchical clustering to separate clearly distinguishable groups.

Cluster 2 - Versicolor and Virginica: Cluster 2 comprises iris flowers from the "Iris-versicolor" and "Iris-virginica" species. These species share some similarities in their measurements, making it challenging to distinguish them with complete certainty using hierarchical clustering alone. However, this clustering provides valuable insights into the overlaps and nuances between these two species.

Cluster 3 - Versicolor and Virginica (Outliers): Cluster 3 represents iris flowers that do not fit neatly into the "Setosa" or "Versicolor and Virginica" clusters. These flowers may exhibit characteristics that deviate from the norm within their respective species, indicating potential outliers or anomalies.

### 4.2 Discussion

The hierarchical clustering analysis of the Iris dataset has demonstrated the algorithm's effectiveness in segregating iris flowers into meaningful clusters based on their sepal and petal measurements. The three clusters identified in this study align with the known species in the Iris dataset, providing validation of the clustering results.

Cluster 1, consisting primarily of "Iris-setosa" flowers, highlights the distinctive features of this species, confirming that hierarchical clustering can successfully identify well-defined groups within a dataset.

Cluster 2, encompassing "Iris-versicolor" and "Iris-virginica" flowers, emphasizes the challenge of differentiating between these two species solely based on these measurements. While the clustering reveals some natural groupings, it also reflects the inherent overlap in certain characteristics between these species.

Cluster 3 serves as an essential component of the analysis, as it identifies potential outliers or instances where individual flowers deviate from the expected patterns within their species. Further examination of this cluster could provide insights into variations and anomalies within the dataset.

## V. CONCLUSION

The hierarchical clustering is a valuable tool for exploring patterns and relationships within complex datasets like the Iris dataset. The identification of three clusters aligns with the known species and demonstrates the algorithm's ability to uncover meaningful insights. Future research could involve the application of hierarchical clustering to other datasets with similar challenges, as well as exploring alternative linkage methods and distance metrics for clustering analysis

## REFERENCES

[1] Chris ding and Xiaofeng He (2002), Cluster Merging and Splitting In Hierarchical Clustering Algorithms.

[2] G Ravi Kumar, K Tirupathaiah and B Krishna Reddy, "Client Churn prediction of banking and fund industry utilizing machine learning techniques",IJCSE,Volume-7,Issue- 6,PP:842-846, 2019

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais and S.J. Formosinho (2007), Improving Hierarchical Cluster Analysis: A new method with outlier detection and automatic clustering, Chemo metrics and Intelligent Laboratory Systems, 87, pp. 208-217.

[5] J. Han and M. Kamber," Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2 nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[6] L. Feng, M-H Qiu, Y-X. Wang, Q-L. Xiang, Y-F. Yang and K. Liu (2010), A fast divisive clustering algorithm using an improved discrete particle swarm optimizer, Pattern Recognition Letters, 31, pp. 1216-1225.

[7] N. Michael, "Artificial Intelligence - A Guide to Intelligent Systems",2nd edition, Addison Wesley, 2005.

[8] MarjanKuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo (2012), A survey of hierarchical clustering algorithms, The Journal of Mathematics and Computer Science, 5,.3, pp.229- 240.

[9] Pavel Berkhin (2000), Survey of Clustering Data Mining techniques ,Accrue Software, Inc..

[10] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[11] Surya Bhupal Rao, S.Rahamat Basha, G. Ravi Kumar, "A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, ISSN: 0973-8975.

[12] UCI machine learning repository. http://archive.ics.uci.edu/ml/.