**CLUSTERING ANALYSIS REPORT: HIERARCHICAL AND DBSCAN CLUSTERING ON THE WINE DATASET**

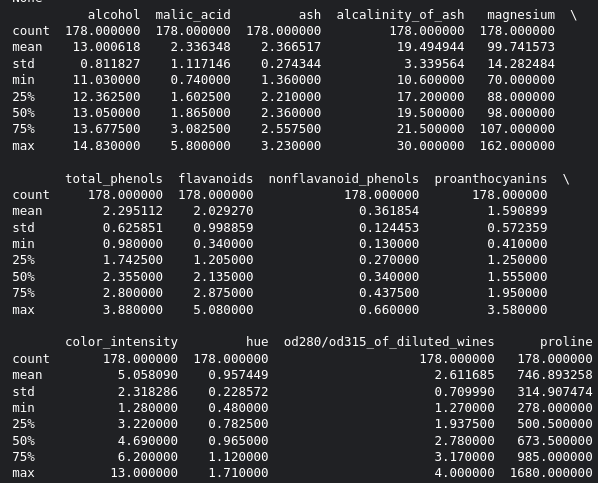
1. Introduction

The aim of this assignment is to compare two unsupervised learning algorithms, Hierarchical Clustering and DBSCAN, and consider them when analyzing the Wine dataset provided by Scikit-learn. The major research objective is to examine how such algorithms infer patterns and cluster similar data points when they do not have labels. Through the analysis of the clustering outcomes, the performance assessment based on a range of measures, and graphical depiction and evaluation of the clusters, we are going to draw a more detailed picture of the advantages and limitations of each of the methods with its application to the real-life dataset covering the chemical properties of wine samples (Putra, I. P. S. E., 2024).

2. Data Preparation and Exploration

The Wine dataset has 178 records that form the different drink samples but the features are 13 in number and represent the different chemical properties of the wine including alcohol content, malic acid concentration and color intensity. A basic analysis of the dataset after loading it by executing methods such as .head(), .info() and .describe() affirmed that there were no missing values and that all columns belonged to the variable type of float64 making all columns numeric. Because the clustering algorithms based on distance metrics are sensitive to the scale of features, the data is standardized with the use of StandardScaler of Scikit-learn. This made all the features comparable in determining the distances and this is imperative in the determination of reliable clusterings.

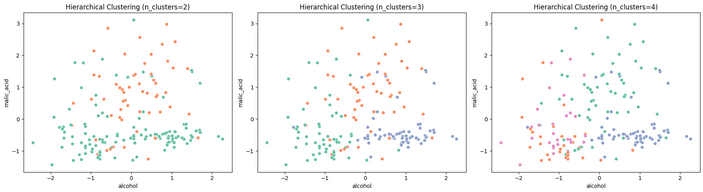
The data summary is shown below.

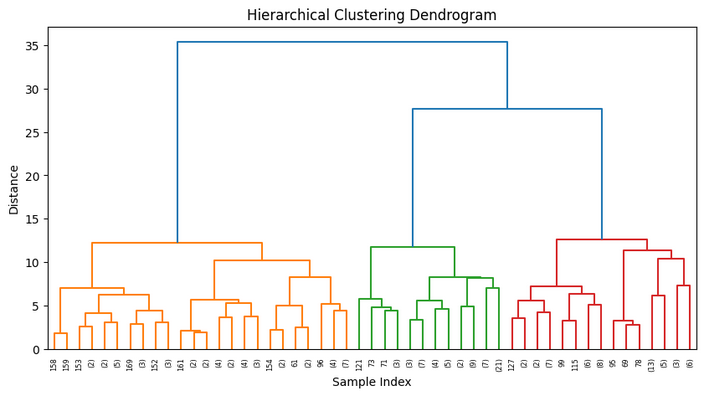


3. Hierarchical Clustering

The agglomerative hierarchical clustering was applied to the standardized data to study the same data in an agglomerative way. The cluster configurations tested included the use of two, three and four clusters. In each of the configurations, the scatter plot of first two standardized features, namely, alcohol and malic acid were used as x and y axes, respectively. Using two clusters the algorithm managed to separate the data into two different groups albeit with some variance of labels. Increasing the number of clusters to three gave a more transparent clustering and was more representative of what is sociologically expected in regard to the wine types. Using a fourth cluster resulted in further fragmentation, with some clusters taking on the aspect of being arbitrarily partitioned, explaining why there is a chance of overfitting the model to the noise in the dataset.

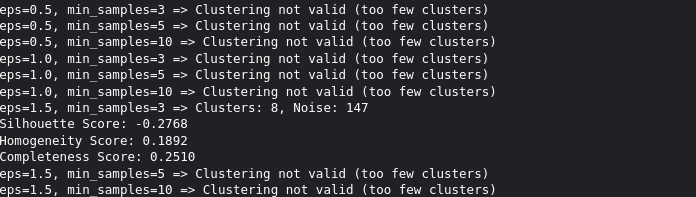
Along with the visualization of the clusters, a dendrogram was created on the basis of the linkage method provided by the SciPy hierarchy section. A dendrogram gave a complete depiction of the way individual data points are sequentially combined into huge clusters. By looking into the vertical gaps between successive links, it was more accessible to determine an appropriate number of clusters depending on the largest linkage jumps. This graphical representation helped buttress the preceding cluster choices and was an easy way to learn the hierarchical nature of the given dataset.





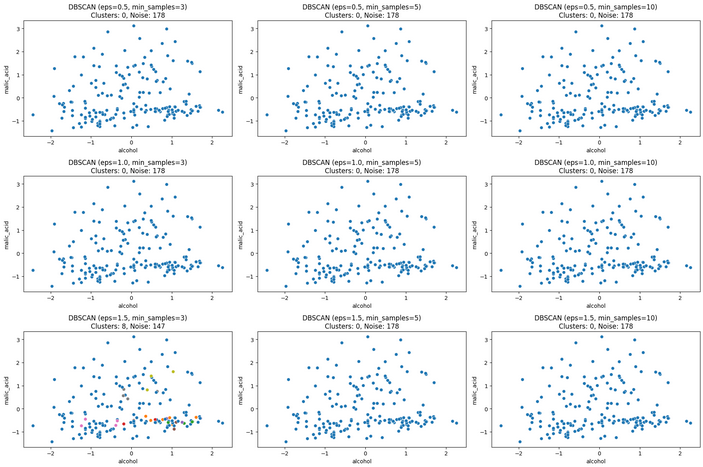
4. DBSCAN Clustering

In order to assess the density-based clustering, the same standardized data was under the combinatorial DBSCAN algorithm. The combination of the values of some parameters, namely eps (neighborhood radius) and min\_samples (the minimum points to be in one cluster forming a dense region) were experimented with. The majority of the combinations did not give any substantial clusters or too many points were regarded as noise or almost all the data were placed in the same cluster. Out of all possible settings, only a single combination of parameters, eps = 1.5 and min\_samples = 3, gave a legal clustering result. In this setting, there were eight clusters detected, which is an accurate result, but only 147 out of 178 data points were not considered noise, reflecting the parameter sensitivity of the DBSCAN algorithm type, and its inapplicability to datasets that may not have the clusters well-separated in terms of density.



Subsequent analysis of this DBSCAN output contained a negative value of the silhouette score, which implies that clustering was inadequately compact and discrete. The scores of homogeneity and completeness were also low since it indicated that the resulting clusters were not well structured in line with the internal structure of data. The same outcome indicates that DBSCAN, as accurate as it can be in some situations, is not necessarily an optimal method to use when working with datasets such as the Wine data, where clusters are not as well separated by dense areas, but are instead separated by slight deviations in numerous features.

The plot below shows the outcomes of DBSCAN clustering the results of a data set (probably the data set with the wine as a subject) with different eps (0.5, 1.0, 1.5) and the different min sample (3, 5, 10) parameters illustrated in the 3x3 grid. The scatter graphs in each subplot represent the plot of two features (e.g., alcohol and malic acid) with the points being of a different color, depending on the cluster they belong to, with the noise points identified as -1. The clusters and noise points quantities differ considerably: the smaller the eps value and the larger the min\_samples (e.g., eps=0.5, min\_samples=10), the fewer clusters and the more noise there is (e.g., 0 clusters, 178 noise points); with more restrictive clustering conditions, fewer clusters are detected. On the other hand, larger eps and smaller min\_samples (e.g., eps=1.5, min\_samples=3) generates numerous clusters (e.g., 14 clusters) and decreased noise, the result being a looser clustering parameter. It emphasizes the fact that DBSCAN is highly dependent on these parameters, and the ideal parameters in many cases may be a compromise between the number of identified clusters, on the one hand, and the amount of noise, on the other hand, depending on the intrinsic mechanistic structure of the dataset.



5. Analysis and Insight

The two clustering methods were compared and thus slightly differed in terms of performance and interpretability. Hierarchical clustering also gave meaningful groupings that always made sense, and it was clearly represented using dendrograms and scatter plots. Its scores remained more or less constant under varying n\_clusters and gave us knowledge on the dataset structure in suggesting nature separation and subgroups. Conversely, the DBSCAN was unable to create coherent clusters, owing to the characteristics of the dataset, which does not have well-separated density areas, which the algorithm can be applied to identify. The large amount of noise points and bad evaluation numbers even in the instances when clusters were generated indicated that DBSCAN did not reflect the actual structure of the data. Although the DBSCAN is a solid algorithm to detect arbitrarily shaped clusters and deal with noise when working with datasets of differing density, its usefulness quickly goes down in the face of more uniform or linearly separable data such as this Wine dataset.

6. Conclusion

Using this clustering analysis, one could see that the nature of the dataset has to dictate the choice of algorithm. Data on Wine dataset was more effectively and interpretable with hierarchical clustering which produced stable and informative clusterings, which were comparable with the domain expectations. Its dendrogram visualization helped to add value to it, as it helped to choose the best number of clusters. On the other hand, DBSCAN was unsuccessful in this experiment because the null hypothesis (of clear density-based separation) was not validated by the dataset. This led to an assignment of dense and noisy clusters. The task illustrates both the significance of preprocessing of data and tuning of parameters and visualization in clustering processes, and how cluster analysis calculations based on various unsupervised learning algorithms can arrive at wildly different results using the same data due to the nature of the data (Fuchs, M., 2022).

Reference

Fuchs, M., & Höpken, W. (2022). Clustering: hierarchical, k-means, DBSCAN. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications* (pp. 129-149). Cham: Springer International Publishing.

Putra, I. P. S. E., Wahyudi, I. P. A. T., Priadinata, I. P. B., Keniten, I. B. N. S. B., Sugiartawan, I. P., & Sudipa, I. G. I. (2024, December). WINE QUALITY CLUSTERING USING DBSCAN ON ALCOHOL AND MALIC ACID CHARACTERISTICS. In *Proceeding International Conference on Information Technology, Multimedia, Architecture, Design, and E-Business* (Vol. 3, pp. 238-244).