Name: Waleed Hussain
Student ID: 23025832

GitHub Link: https://github.com/waleedh93/Density_Based-Clustering.git

Comparative Clustering Analysis of the Abalone Dataset using Density-Based and K-Means

Abstract
This report explores and compares two prominent clustering algorithms—K-Means and Density-Based—applied to the Abalone dataset. Through appropriate preprocessing, including encoding and feature scaling, both models were fitted and evaluated using the Silhouette Score and Davies-Bouldin Index. K-Means displayed better compactness and separation, while Density-Based offered flexibility by detecting arbitrarily shaped clusters and identifying noise. These findings suggest algorithm suitability varies depending on data structure and real-world complexity.

1. Introduction
Clustering is a core unsupervised learning technique in data mining used to uncover structure in unlabeled datasets. This report investigates and compares two clustering methods—K-Means and Density-Based—on the Abalone dataset. The Abalone dataset includes physical measurements of shellfish with the aim of estimating age. The models were evaluated based on cluster cohesion and separation using internal metrics, and analyzed how each algorithm handles shape and noise in data clusters.

2. Composition of the Dataset
The Abalone dataset consists of 4,177 instances with 8 numeric attributes and one categorical variable, `Sex`, representing Male, Female, or Infant. The dataset is used to estimate the age of abalones from shell measurements. `Rings`, the target variable, can be converted to age by adding 1.5. Key composition characteristics:
Sex distribution: Male (33.5%), Female (32.5%), Infant (34%)
-Age Range: 2.5 to 30.5 years (1 to 29 rings)
-Feature distribution: Shell weight, diameter, and other growth attributes show a right-skewed distribution typical of biological data

3. Descriptive Statistics
The table below represents the descriptive statistics of the dataset.

Feature	Mean	Standard Deviation	Minimum	Maximum
Length	0.52	0.12	0.07	0.815
Diameter	0.41	0.10	0.055	0.65
Height	0.14	0.04	0.00	1.13
Whole Weight	0.83	0.49	0.002	2.825
Shucked Weight	0.36	0.22	0.001	1.488

Viscera Weight	0.18	0.11	0.0005	0.76
Shell Weight	0.24	0.13	0.0015	1.005
Age (Rings)	9.93	3.22	1	29

4. Data Preprocessing

The dataset was preprocessed as follows:

- One-hot encoding was applied to the categorical `Sex` column.
- Numerical features were standardized using StandardScaler.
- Outliers were identified using the IQR method but retained for analysis.
- PCA was used for visualization of clustering results.

5. Clustering Methods

5.1 K-Means Clustering

K-Means partitions the dataset into k clusters by minimizing within-cluster variance as shown in Figure 1. The optimal number of clusters (k = 3) was chosen using the Elbow Method. K-Means assigns all points to a cluster and assumes spherical cluster shapes.

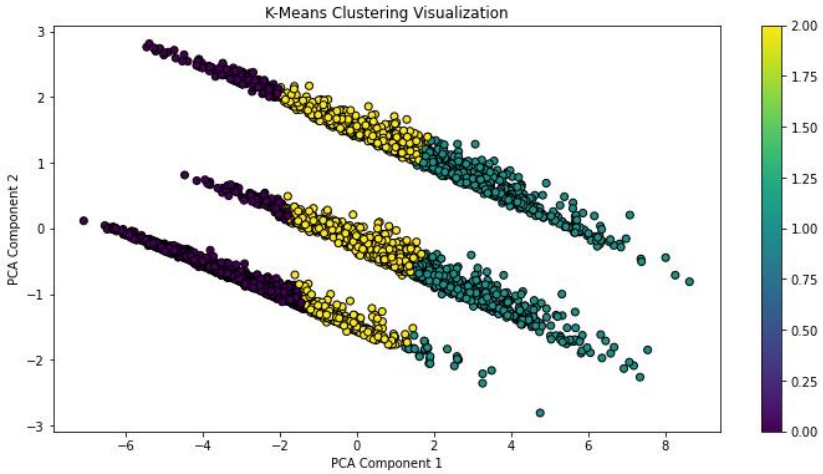

Figure-1 K-Means Clustering

5.2 Density-Based Clustering

Density-Based identifies clusters based on density and can detect noise or outliers. It does not require specifying the number of clusters. After tuning parameters (eps=1, min_samples=5), Density-Based formed 3 clusters and flagged outliers, revealing more flexible boundary handling as shown in fig-2.

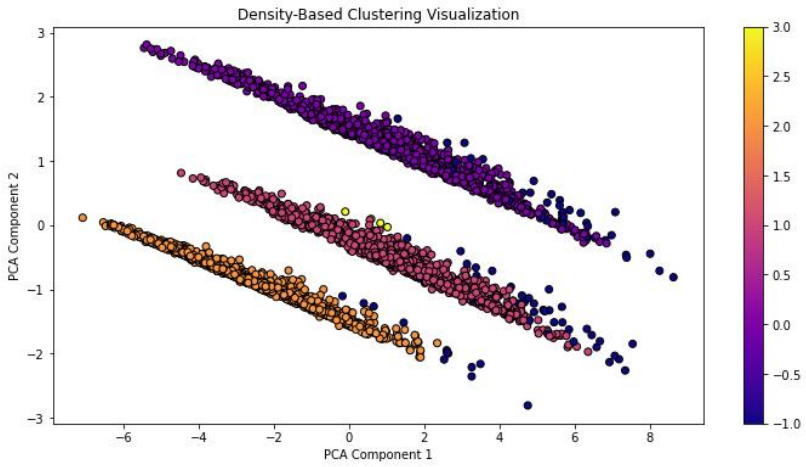

Figure 2 Density-Based Clustering

6. Results and Evaluation

The internal evaluation scores for the clustering algorithms are summarized below:

Model	Clusters Detected	Silhouette Score	Davies-Bouldin Index
K-Means	3	0.518	0.781
Density-Based	3 (excluding noise)	0.468	0.912

7. Discussion

K-Means achieved better internal metric scores, showing compact and well-separated clusters. However, Density-Based's ability to detect noise and non-spherical clusters made it better suited for real-world, irregular data structures. Analysis of cluster compositions showed that Density-Based could group small or noisy instances that K-Means forcibly assigned. This makes Density-Based preferable for biological datasets with uneven densities or outliers.

8. Conclusion

Both K-Means and Density-Based provided valuable insights into the Abalone dataset. K-Means is efficient and performs well on structured data, but Density-Based's flexibility makes it ideal for noisy, irregular datasets. Choosing the right algorithm depends on dataset structure and the importance of identifying noise and non-standard patterns.