engineering proceedings

MDPI

Proceeding Paper

A Heuristic Evaluation of Partitioning Techniques Considering Early Type Galaxy Databases [†]

Prithwish Ghosh [1,*,‡] and Shinjon Chakraborty [2,‡]

[1] Department of Statistics, Visva Bharati, Santiniketan 731235, India
[2] Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata 700019, India; shinjonchakraborty07@gmail.com
[*] Correspondence: ghosh.prithwish1999@gmail.com

[†] Presented at the 4th International Electronic Conference on Applied Sciences, 27 October–10 November 2023; Available online: https://asec2023.sciforum.net/.

[‡] These authors contributed equally to this work.

Abstract: Galaxies are one of the most interesting and complex astronomical objects statistically due to their continuous diversification caused mainly due to incidents such as accretion, action, or mergers. Multivariate studies are one of the most useful tools to analyze this type of data and to understand various components of it. We study a sample of the local universe of Orlando 509 galaxies, imputed with Predictive Mean Matching(PMM) multiple imputation algorithm, with the aim of classifying the galaxies into distinct clusters through k-medoids and k-mean algorithms and in turn performing a heuristic evaluation of the two partitioning algorithm through the percentage of misclassification observed. From the clustering algorithms, it was observed that there were four distinct clusters of the galaxies with misclassification of about 1.96%. Also comparing the percentage of misclassification heuristically k-means is a superior algorithm to k-medoids under fixed optimal sizes when the said category of galaxy datasets are concerned. By considering that galaxies are continuously evolving complex objects and using appropriate statistical tools, we are able to derive an explanatory classification of galaxies, based on the physical diverse properties of galaxies, and also establish a better method of partitioning when working on the galaxies.

Keywords: galaxy; classification; clustering; machine learning

check for updates

Citation: Ghosh, P.; Chakraborty, S. A Heuristic Evaluation of Partitioning Techniques Considering Early Type Galaxy Databases. *Eng. Proc.* **2023**, *52*, 0. https://doi.org/

Academic Editor: Firstname Lastname

Published:

1. Introduction

A galaxy represents a vast and intricate system composed of stars and interstellar matter within the expanse of our universe. To effectively engage with these complex and dynamic databases. The repository of galaxy data encompasses an extensive array of information encompassing diverse aspects of galaxies, encompassing their morphological characteristics, photometric properties, spectral attributes, and more. While substantial research has been conducted in these specific domains, the comprehensive exploration of their "physical properties" remains a relatively uncharted territory.

Esteemed statisticians and physicists concur that multivariate techniques represent the most suitable approach for deriving meaningful insights from these astronomical databases. Among the array of partitioning techniques widely embraced in multivariate statistics, the *K-Means* and *K-medoid* methods emerge as notable contenders. As we navigate through our analysis, it becomes increasingly apparent that a heuristic comparison between these robust partitioning techniques can illuminate their relative strengths, particularly concerning the percentage of misclassification, all within the context of an assumed optimal number of clusters tailored to this specific category of astronomical data. This dataset was meticulously assembled by Ogando et al. in 2008 [1,2] and comprises a set of parameters that hold paramount significance for our study. Furthermore, we have enriched our dataset by incorporating supplementary parameters sourced from the Hyperleda database, enhancing the depth and breadth of our analytical endeavors.

2. Materials and Methods

2.1. Missing Value Imputations

To address the absence of data in the Galaxy dataset, we have employed the multiple imputation technique known as Predictive Mean Matching (PMM). In essence, PMM computes the anticipated value of the target variable Y based on the specified imputation model. Predictive mean matching is used in statistics and data analysis to impute missing values by matching them with the predicted means of similar observations, preserving the original data distribution and relationships.

2.2. Choice of Optimal Clusters

2.2.1. Elbow Plot

To ascertain the ideal number of partitions into which the data can be divided, the Distortion Plot Method stands as a widely embraced technique for determining this optimal value, often denoted as 'k'. This method computes the average sum of squared distances from the partition centers within the generated partitions. Essentially, the optimal number of clusters becomes evident when examining the graph for a distinct 'elbow-like' point [3].

2.2.2. Dunn Index

Ref. [4] The Dunn Index is a metric used to evaluate the quality of clustering results in unsupervised machine learning. It helps assess the separation between clusters and the compactness of data points within each cluster.

3. Formula

The Dunn Index is calculated using the following formula:

$$\text{Dunn Index} = \frac{\min(\text{Inter-cluster distances})}{\max(\text{Intra-cluster distances})}$$

where:

- Inter-cluster distances refer to the distances between different clusters.
- Intra-cluster distances refer to the distances within each cluster.

A higher Dunn Index indicates better clustering, as it signifies greater inter-cluster separation and smaller intra-cluster distances.

- When the Dunn Index is high, it suggests that the clusters are well-separated and compact, indicating a good clustering solution.
- Conversely, a low Dunn Index implies that clusters are either too close to each other (poor separation) or data points within clusters are too spread out (low compactness).

3.1. Clustering(Partitioning) Algorithms and Discriminant Analysis

Clustering is a method that involves categorizing individuals with diverse characteristics based on their similarities or dissimilarities. In this study, several renowned algorithms have been employed, including the following:

3.1.1. K-Means

K-means clustering is a popular unsupervised machine learning technique used for data clustering and segmentation. It is a simple yet effective algorithm for partitioning a dataset into K distinct, non-overlapping clusters. The goal is to group similar data points together based on their feature similarity.

3.1.2. Algorithm

The K-means algorithm works as follows:

Algorithm 1 K-means Clustering

```
1. Initialize K cluster centroids randomly.
2. Assign each data point to the nearest centroid.
3. Recalculate the centroids as the mean of the data points in each cluster.
4. Repeat steps 2 and 3 until convergence (centroids no longer change significantly).
```

K-means clustering is a versatile and straightforward technique for clustering data. It is easy to implement and can be applied to various domains like here we used in the classification and clustering of galaxy diversification discovering hidden patterns and grouping similar data points together.

3.1.3. K-Medoids

We use this as a second algorithm to compare between them. The method is given below.

1. Initialize K medoids randomly.
2. Assign each data point to the nearest medoid.
3. For each cluster, select the data point that minimizes the total distance to other points in the same cluster as the new medoid.
4. Repeat steps 2 and 3 until convergence.

K-medoid clustering is a valuable technique for partitioning data into meaningful clusters. It's particularly useful when dealing with noisy or non-linear data.

3.1.4. The Linear Discriminant Analysis (LDA)

The primary objective of LDA is to find a linear combination of features that best separates two or more classes in a dataset. It aims to maximize the between-class variance while minimizing the within-class variance [5]. In LDA, key concepts include:

- Scatter matrices: Within-class and between-class scatter matrices.
- Eigenvectors and eigenvalues: Used to find the optimal linear transformation.
- Decision boundaries: Separating classes based on discriminant functions.

4. Results

Astronomy generates complex datasets, especially for galaxies. K-means and K-medoids are vital for:

- **Classification:** Grouping galaxies by attributes.
- **Structure Detection:** Identifying cosmic structures.
- **Outlier Detection:** Finding rare celestial objects.
- **Dimensionality Reduction:** Simplifying data for analysis.

These techniques help astronomers unveil patterns, understand celestial structures, and explore the universe's mysteries.

From the techniques used to find the optimal number of clusters i.e., *Elbow plot* and *Dunn Index* is 4 and 3 for *K-Means* and *K-Medoids* respectively. The Elbow plots and Value of the Dunn Index are given in Table 1, Figures 1 and 2.

Table 1. Dunn Index.

K	K Means	K Medoid
3	1.10833750	0.81437369
4	1.1747325	0.47461267 [1]

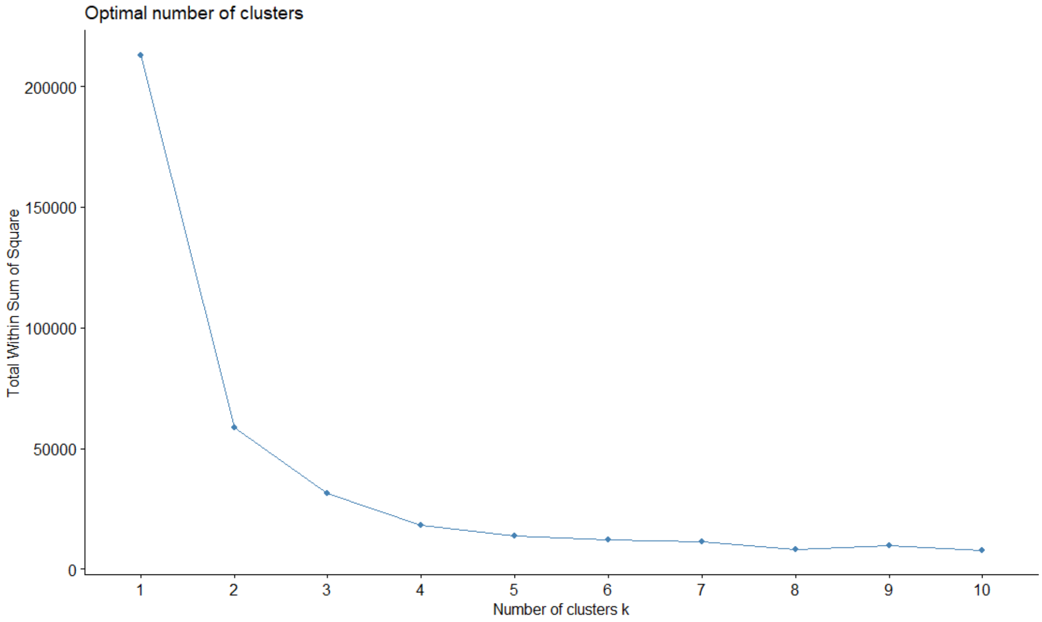

Figure 1. Elbow Plot for K means.

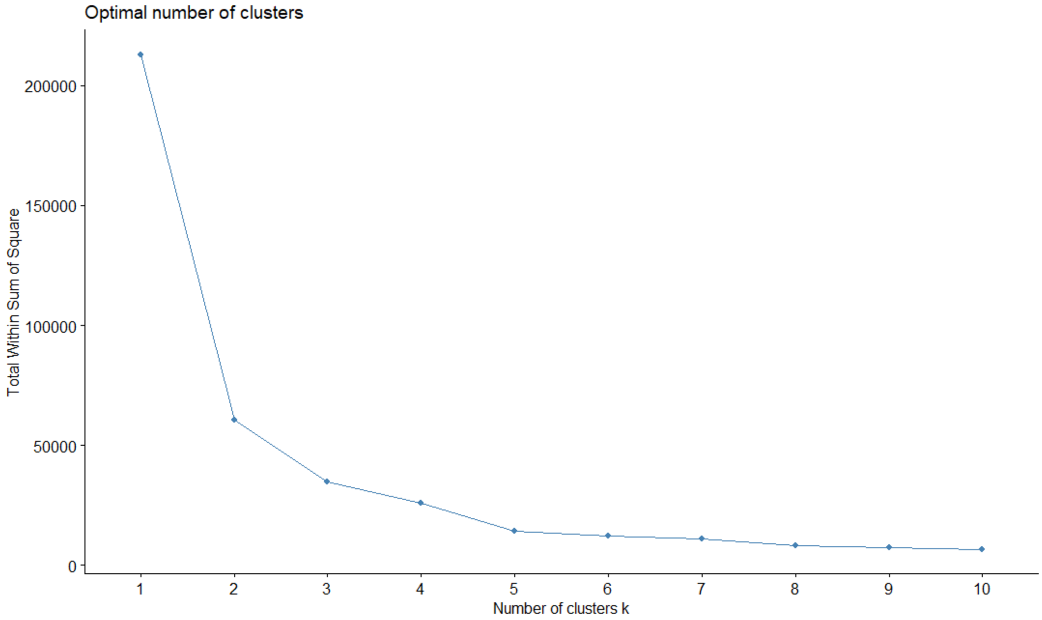

Figure 2. Elbow Plot for K Medoid.

Note: *From the shown Figure 3 it is quite evident that there are quite a few outliers present in the dataset and any procedure involving mean(K-Means) is not robust when handling outliers.*

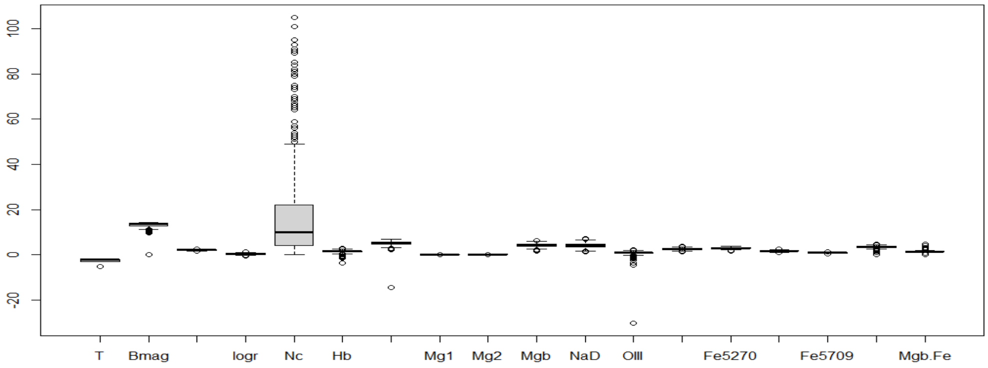

Figure 3. Box Plot.

The clusters thus formed by k-means and k-medoids considering the optimal number of clusters to be 3 and 4 are shown in the Figures 4–7.

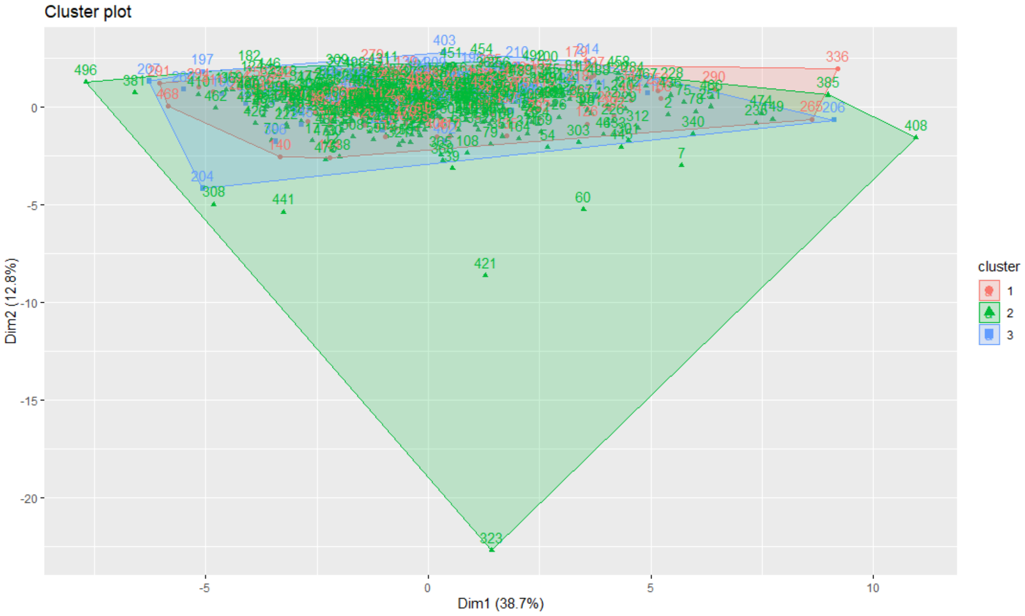

Figure 4. K means k = 3 cluster.

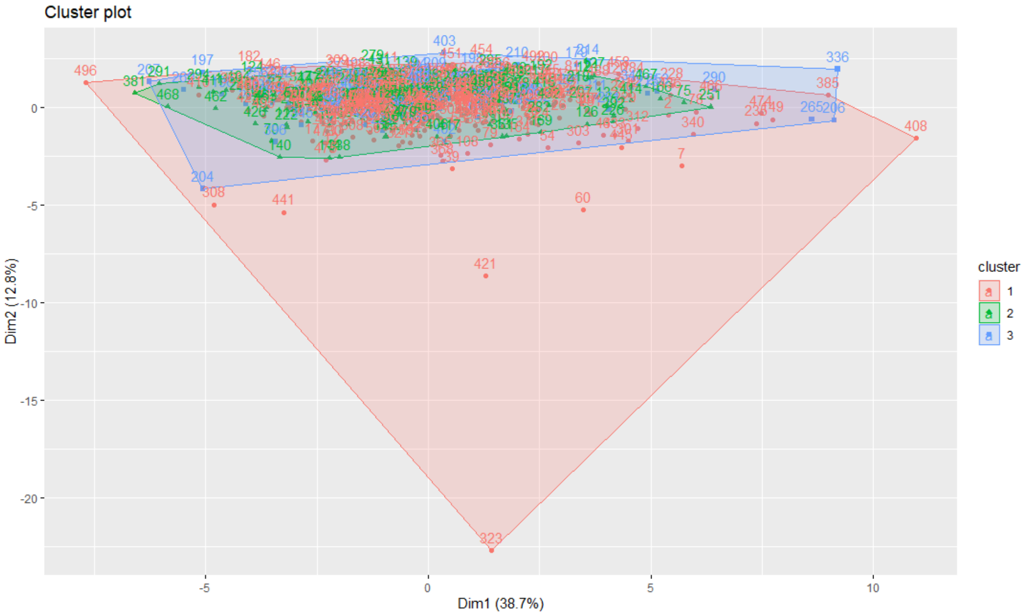

Figure 5. K Medoid k = 3 cluster.

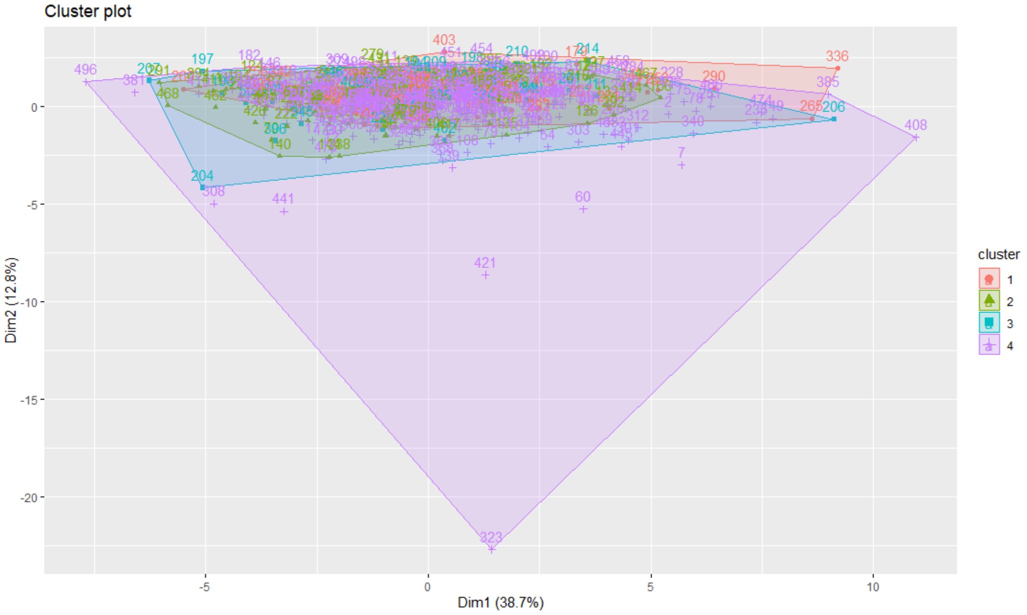

Figure 6. K means k = 4 cluster.

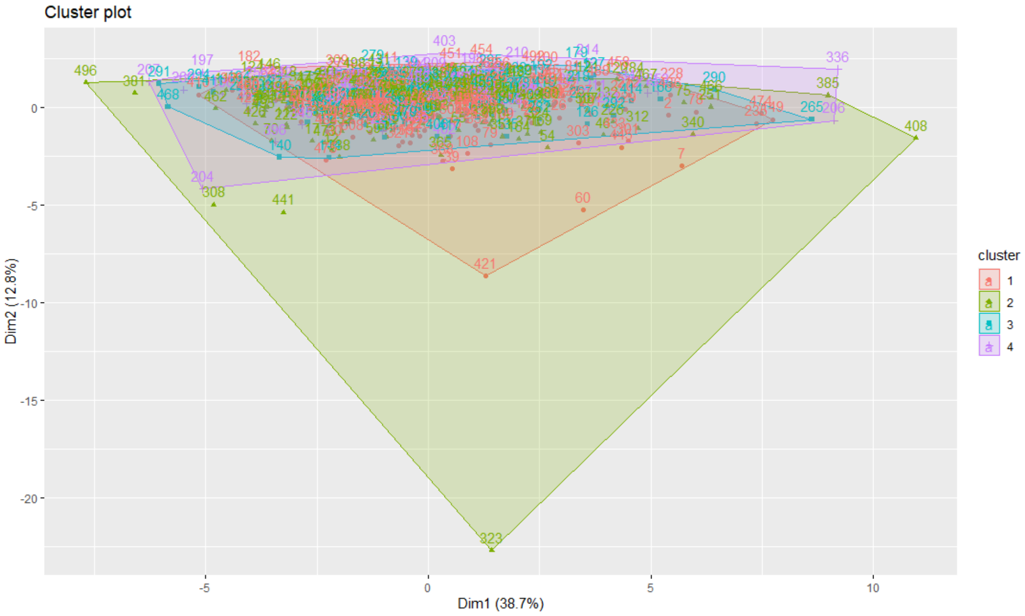

Figure 7. K Medoid k = 4 cluster.

Figures, Tables

To comprehensively compare both clustering algorithms, we formed clusters using both three and four optimal cluster numbers for each algorithm. We evaluated the comparison based on the percentage of misclassification within the procedure using the optimal cluster numbers. The results of the discriminant analysis are presented in a tabular format. Tables 2–5. K-means clustering exhibits superior performance compared to K-medoids when using three optimal clusters, with a misclassification rate of approximately 2.36% for K-means and 8.25% for K-medoids. The trend continues with four optimal clusters, where K-means maintains its advantage with a misclassification rate of around 1.96% versus 11.19% for K-medoids. In summary, K-means outperforms K-medoids overall, even in the presence of outliers, for the galaxy dataset.

Table 2. K means where K = 3.

	Predicted Cluster1	Predicted Cluster2	Predicted Cluster3
Actual Cluster 1	86	0	3
Actual Cluster 2	18	365	0
Actual Cluster 3	1	0	36

Table 3. K medoids where K = 3.

	Predicted Cluster1	Predicted Cluster2	Predicted Cluster3
Actual Cluster 1	303	36	0
Actual Cluster 2	0	101	6
Actual Cluster 3	0	0	63

Table 4. K means where K = 4.

	Pred Cluster1	Pred Cluster2	Pred Cluster3	Pred Cluster4
Actual Cluster 1	43	0	0	0
Actual Cluster 2	0	31	0	0
Actual Cluster 3	1	0	101	0
Actual Cluster 4	0	0	9	324

Table 5. K medoids where K = 4.

	Pred Cluster1	Pred Cluster2	Pred Cluster3	Pred Cluster4
Actual Cluster 1	187	30	0	0
Actual Cluster 2	9	139	14	0
Actual Cluster 3	0	0	76	4
Actual Cluster 4	0	0	0	50

5. Conclusions

From the results and findings of the work, we can observe there are four distinct clusters of galaxies in the local universe of Orlando (2008) based on their collective physical characteristics. The approximate mean values of the parameters in those robust clusters are also included in the study, which would give us a heuristic idea about the physical characteristics of a newly observed galaxy provided it falls into one of the three robust clusters. Additionally, there is about 1.96% misclassification in the data which indicates the high accuracy of the clustering. The misclassification that occurred while clustering for a given optimal number of clusters($k = 3$ and $k = 4$) can be unanimously inferred that k-means perform better than k-medoids under this category of galaxy database. Also, the misclassification with optimal no of clusters for k-means (k = 4) and k-medoids (k = 3) also serves as a reasonable indication of the superiority of the k-means algorithm over k-medoids considering galaxy data.

Author Contributions: Conceptualization, P.G. and S.C.; methodology, P.G. and S.C.; software, P.G.; formal analysis, S.C. and P.G.; investigation, P.G.; resources, S.C.; data curation, S.C. and P.G.; writing—original draft preparation, P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: We selected the sample of 509 early-type galaxies in the local Universe of Ogando et al. (2008). To describe the galaxies, we took from Ogando et al. (2008) [1,6].

Acknowledgments: We are thankful to the Professors for their constant support and motivation, resulting in the successful completion of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Davoust, D.; Fraix-Burnet, T.; Chattopadhyay, A.K.; Chattopadhyay, E.; Thuil- lard, M. A six-parameter space to describe galaxy diversification. *Astron-Omy Astrophys.* **2012**, *545*. https://doi.org/10.1051/0004-6361/201218769.10
2. Nigoche-Netro, A.; Aguerri, J.A.L.; Lagos, P.; Ruelas-Mayorga, A.; Sánchez, L.J.; Muñoz-Tuñón, C.;Machado, A. The intrinsic dispersion in the Faber-Jackson relation for early-type galaxies as function of the mass and redshift. *Astron. Astrophysics* **2011**, *534*, A61.
3. Ghosh, P.; Chakraborty, S. Classification and Distributional properties of Gamma Ray Bursts. In Proceedings of the 16th International Conference MSAST, 21–23 December 2022 ; Volume 11, p. 148.
4. Dunn†, J.C. Well-Separated Clusters and Optimal Fuzzy Partitions. *J. Cybern.* **1974**, *4.1*, 95–104. https://doi.org/10.1080/01969727408546059.

5. Ghosh, P.; Chakraborty, S. Spectral Classification of Quasar Subject to Redshift: A Statistical Study. *Comput. Sci. Math. Forum* **2023**, *7*, 43. https://doi.org/10.3390/IOCMA2023-14418
6. Guy, W.; Ottaviani, D.L. H_γ and H_δ absorption features in stars and stellar populations. *Astrophys. J. Suppl. Ser.* **1997**, *111.2*, 377.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.