iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IV **Month of publication:** April 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

A Relative Evaluation of Clustering Algorithms

Sunil Bhutada[1], G Yamini[2], M Selvarasan[3], D Rohan Goud[4], D Srikanth[5]
[1]*Professor, Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad*
[2]*Assistant Professor, Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad*
[3, 4, 5]*B. Tech IV year, Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad*

Abstract: ***Clustering is a machine learning technique where the data is grouped into clusters. With the help of clustering algorithms the given data can be clustered into specific groups. This comes under unsupervised learning as the labels of the groups are not known. There are many clustering algorithms like KMeans, DBSCAN, Agglomerative clustering, Mean-shift etc. In this project, three algorithms are considered which are Kmeans, DBSCAN, Agglomerative clustering. These are some of the widely used algorithms for clustering. The main objective of this project is to relate these and find the algorithm which outperforms the rest in the given situation. We will analyze the algorithms based on the time of execution of the algorithm, the accuracy and well definedness of the groups or clusters formed after the execution of each algorithm. In this project, we will be using Jupyter Notebook, Python 3 and some libraries like Pandas, Numpy, Scikit-learn and Matplotlib. Four different datasets will be used with varying size and dimensionality. Each dataset will be processed and runned on the three respective algorithms and the validation factors will be noted and based on that the conclusion will be obtained.***
Keywords: KMeans, DBSCAN, Agglomerative clustering, Pandas, Python 3, Numpy, Scikit-learn and Matplotlib.

I. INTRODUCTION

In this project, the main objective is to evaluate and comapare the clustering algorithms with few datasets of different sizes and dimensionality. All the datasets will be later clustered into groups using the clustering algorithms individually. The performance and the accuracy of the algorithms will be evaluated relatively with the four different type of datasets based on the size and dimensionality.

Coming to the factors on which the evaluation will be considered are time taken for execution, accuracy and well definedness of the groups or the clusters that will be obtained after performing the operations on each individual algorithm and each dataset.

II. LITERATURE SURVEY

The entire project will be done on Jupyter Notebook as it works on IPython which gives better graphics and contains cells where one can execute a part of the code which will give us an upper hand in the relative evaluation. The code then can be converted to a python file or a pdf file depending on ones need. The entire execution will be saved properly hence by giving us a better way to ensure no missing or shuffling of data.

As mentioned above, the programming language used will be python version 3 as it has a large number of libraries for data analysis and machine learning which our project is primarily based on. The libraries which will be used for the relative evaluation of the algorithms will be Numpy and Pandas for creation and cleaning.

Matplotlib is another such library which will used in this project and will be used for plotting the results we get after the evalution. Another most important library that will be used in this project is Scikit-learn. This library contains all that we need for the machine learning related tasks in our project.

It contains all the clustering algorithms which is need for the project.

Time will be one factor and the time taken will be calculated using the module called time which is available readily in the python libraries. There are two other factors that will be considered in this project which are required to find out how well defined the clusters or the groups are after the operations.

The factors required are Silhouette coefficient and Davies-Bouldin Index, both will tell us how well the clusters or the groups are formed after performing the operations. These are used when the labels are not known that is basically for unsupervised learning such as clustering.

Coming to the datasets that are used in this project, we have used four datasets which are different from each other in some aspects like size and dimensionality. The four datasets are name numbered as dataset with respective number.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

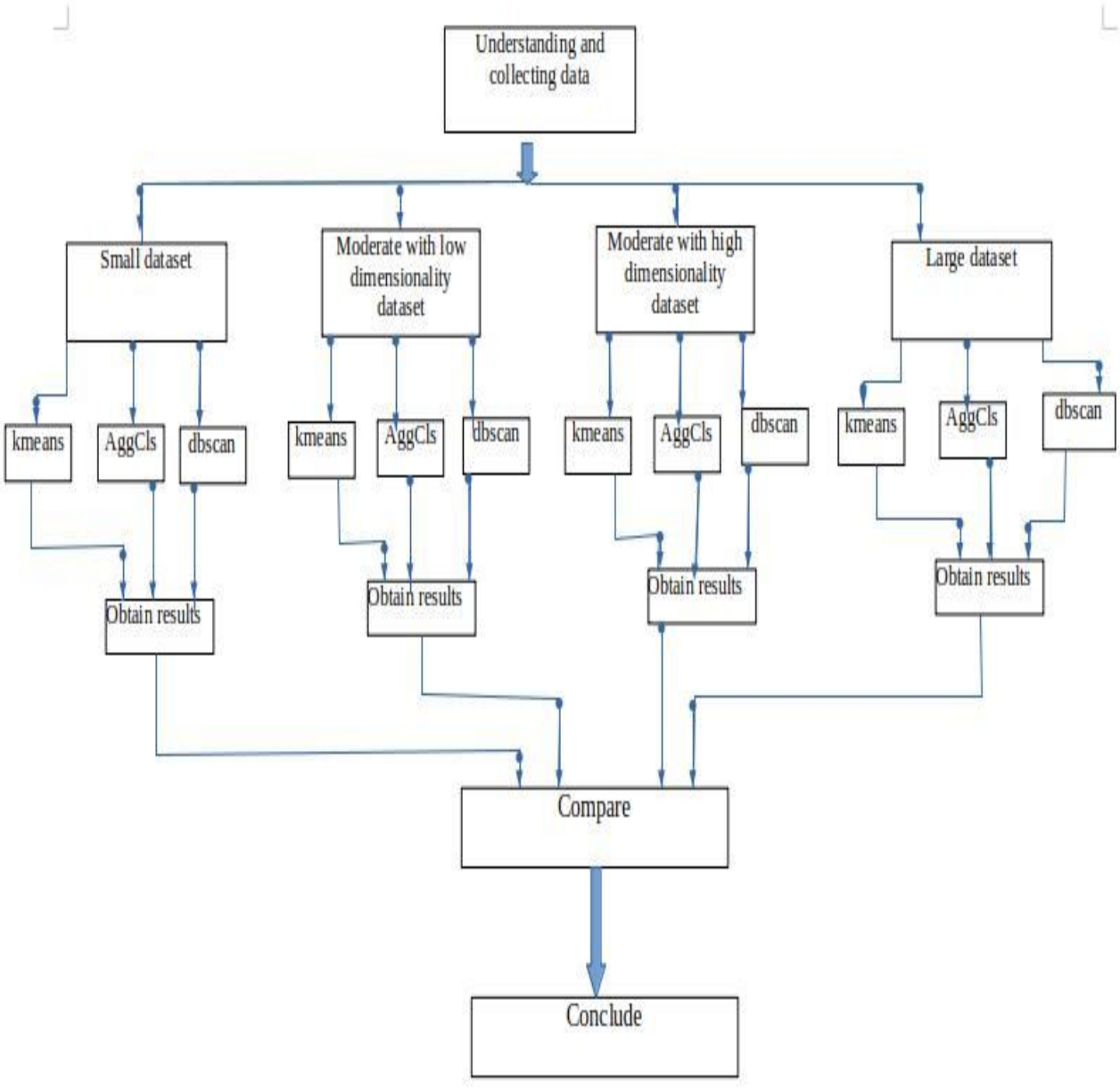
III. PROBLEM STATEMENT

Based on all these three factors the evaluation will be carried out and the a study will be carried out. This study will conclude in understanding which algorithm is suitable for which situation. As a conclusion to this evaluation the outcomes and the findings from the data which we will get based on those three factors the conclusion will be derived.

The first dataset is a small dataset with less sample data and this data is entirely created using random fuction and make blob function available in python libraries. The second data set is a moderately larger dataset but with lesser dimension the third dataset. The third dataset is a dataset which is similar to second dataset but with a little higher dimensions. The fourth dataset is also randomly created but is the largest with higher number of samples.

So after performing the evaluation using these different types of datasets, we will understand how the algorithm functions in different environment and get to know the algorithm which outperforms the rest in that respective situation or environment and derive a conclusion out of the data derived through the validation factors.

IV. PROCESS FLOW



V. METHODOLOGY

A. *Collecting And Understanding Data*

In this step, the datasets are collected according to the needs. As in our case we require datasets of different sizes and dimensions so that we can understand how these algorithms behave in different situations.

B. *Data Cleaning*

In this step the datasets are cleaned and processed. This step will mainly deal with organising the dataset. This step includes performing of operations like replacement of the missing values with appropriate values, deleting the repetitive rows and columns.

C. *Modeling*

In this step, the algorithms are applied to the datasets individually. There are four set of datasets and three algorithms. So each dataset is made to run on all the three algorithms and after modeling, the validation factors are obtained. Here by, this method is applied on all the remaining datasets and a set of 12 results are obtained.

D. *Data Visualization*

The following figure contains data of the large data set when plotted.

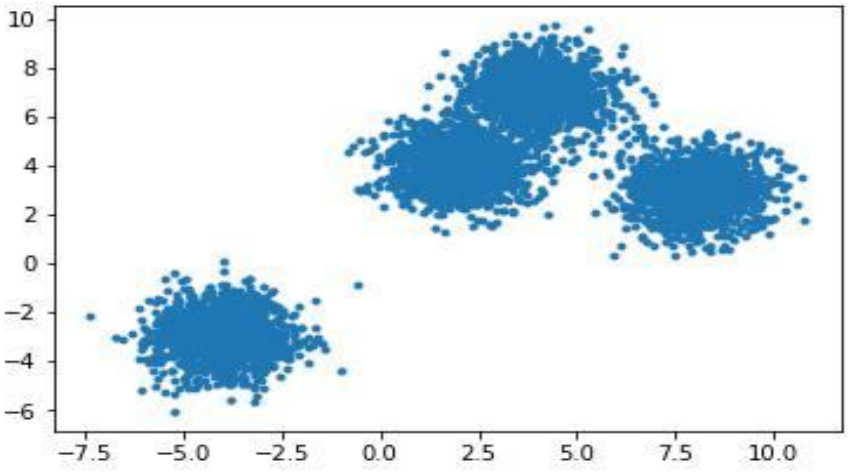


Figure: lrg_ds

The following figure contains the clusters obtained by performing Kmeans to the large dataset. The different colours show the different groups of clusters formed by the large dataset when kmeans is applied.

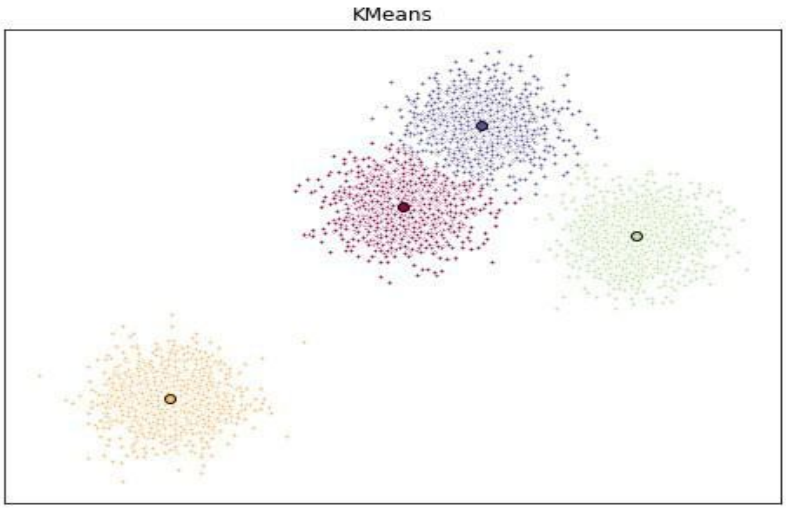


Figure:kmeans_large1

The following figure contains the clusters obtained by performing DBSCAN on a large dataset.

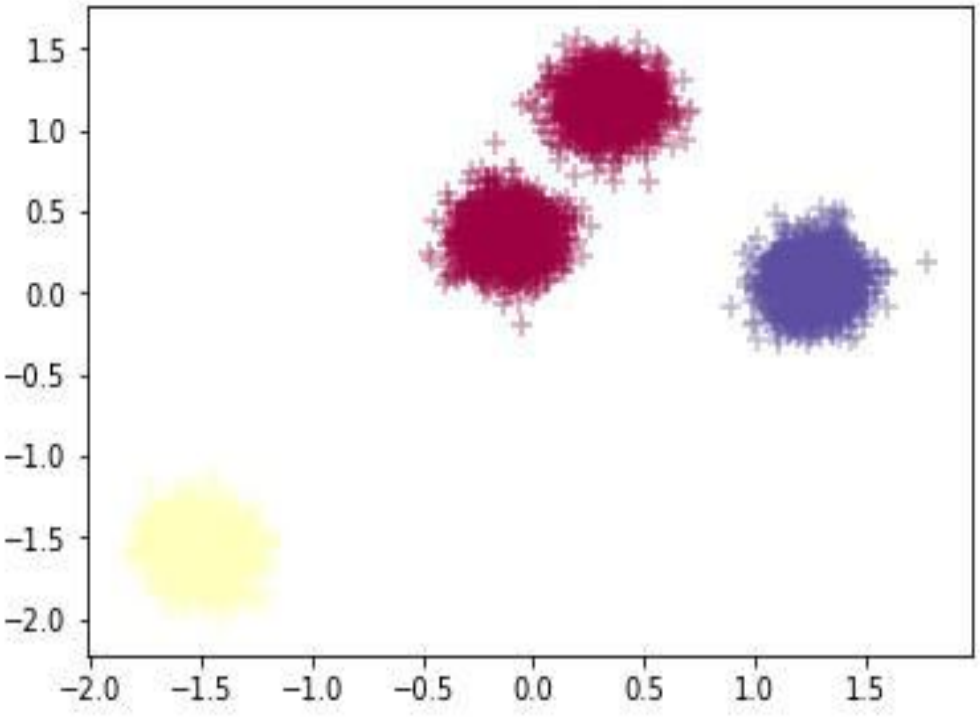


Figure:db_large1

The following figure contains the clusters obtained by performing Agglomerative clustering on a large dataset.

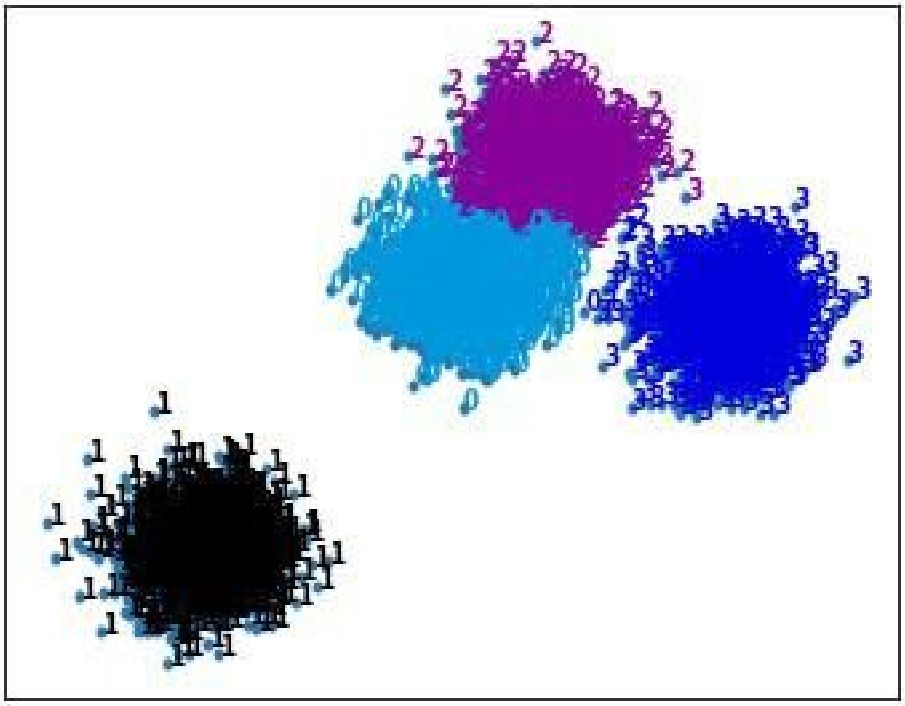


Figure:aggclus_lrg

VI. READINGS

As the three algorithm are applied on the four different datasets individually, the following table consists of the information gathered from the same.

Algorithms	Time-taken	Silhouette Coefficient	Davies Bouldin Score
KMeans Dataset : Small	0.05745339393615723 seconds	0.6087775972281565	0.5268672479471802
KMeans Dataset: Moderate with low dimentionality	0.045387983322143555 seconds	0.2715651796983224	1.1107019507889693
KMeans Dataset: Moderate with high dimentionality	0.07019853591918945 seconds	0.5622416593374175	0.6504296721369606
Kmeans Dataset: Large	0.11935281753540039 seconds	0.6755568831918008	0.459778248397348
Agglomerative Clustering Dataset : Small	0.0037217140197753906 seconds	0.5148419208775662	0.6712878080813125
Agglomerative Clustering Moderate with low dimentionality	0.002177715301513672 seconds	0.2600042305496735	1.1043927445833928
Agglomerative Clustering Dataset: Moderate with high dimentionality	0.038579463958740234 seconds	0.6554448410562985	0.5046129762201294
Agglomerative Clustering Dataset: Large	1.0374982357025146 seconds	0.6738473473908594	0.46309045330674675

DBSCAN Dataset : Small	0.0028192996978759766 seconds	0.6146751296815381	0.4271358834934774
DBSCAN Dataset: Moderate with low dimentionality	0.0032682418823242188 seconds	0.3020158181818765	2.8791561659173674
DBSCAN Dataset: Moderate with high dimentionality	0.017862558364868164 seconds	-0.02331933188904454	1.9033777795599338
DBSCAN Dataset: Large	0.16966462135314941 seconds	0.73411423147305	0.3947038286416833

VII. CONCLUSION

A. On large datasets agglomerative clustering is slower when compared to Kmeans and DBSCAN as the time taken for execution is a lot more than the rest.

B. Agglomerative Clustering is an efficient algorithm interms of accuracy and well definedness as both Silhouette Coefficient and Davies Bouldin Score prove it for every type of dataset.

C. On a general note Kmeans is a slower algorithm when compared Agglomerative Clustering and DBSCAN as time taken for Kmeans is more than the both most of the time except when the Agglomerative Clustering takes more time than for large datasets.

D. When time of implementation is the criteria then the best clustering algorithm is DBSCAN as it always outperforms the rest.

E. When it comes to the accuracy and well definedness of the clusters Agglomerative Clustering is the preferable one out of the three and then comes DBSCAN and later Kmeans according to the obtained Silhouette Coefficient and Davies Bouldin Score.

F. As per data analytics, time is a factor of importance but more than that it is the accuracy and well definedness of the clusters we derive.

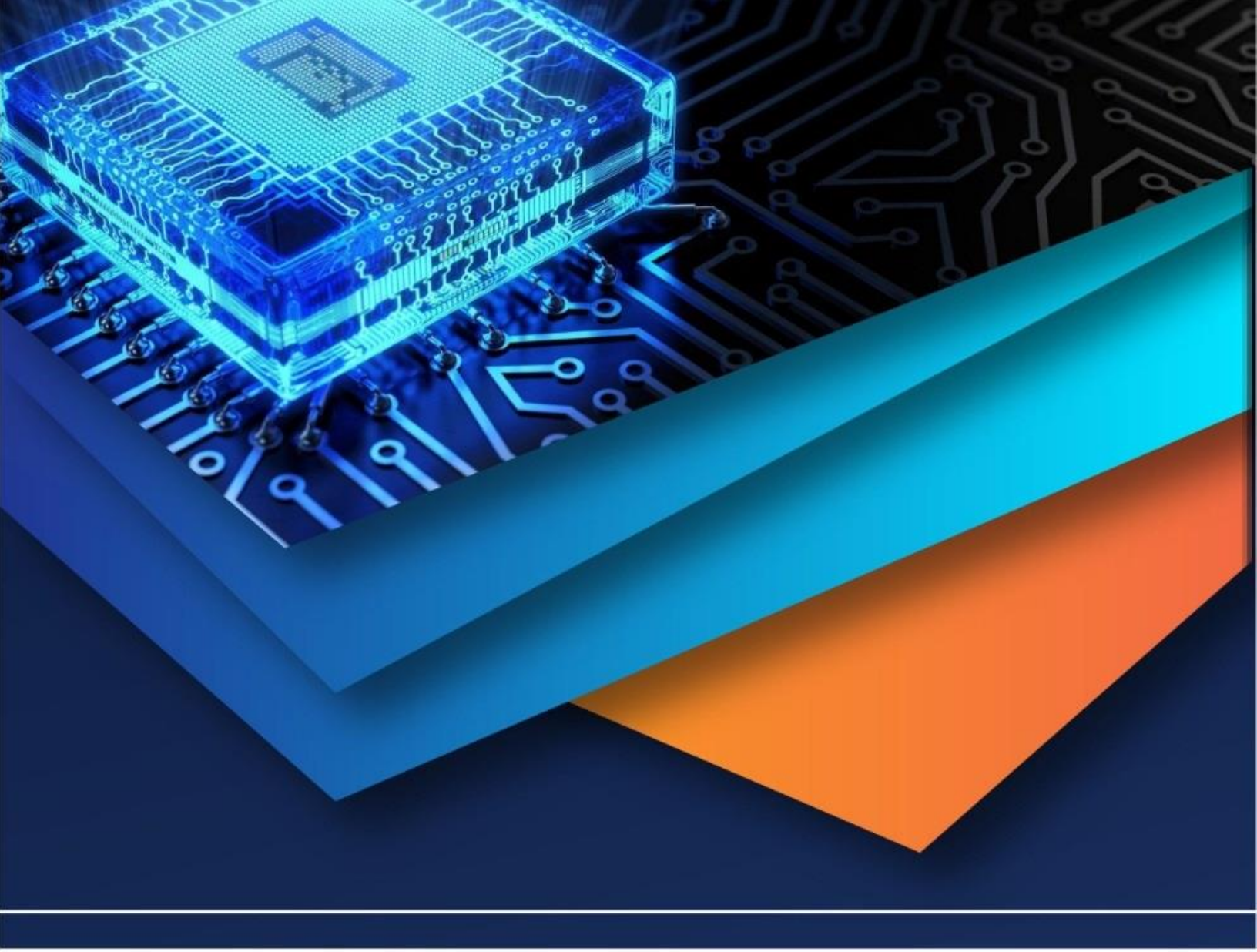
G. Hence, as a conclusion we can say that Agglomerative Clustering is preferable out of the three and then DBSCAN and later Kmeans.

VIII. FUTURE SCOPE

The relative evaluation can be drawn with more number of datasets of different sizes and different dimensionality and a more precise evaluation can be drawn. As in this project three factors were used for evaluation, an evaluation with more number of factors can be drawn and which might lead to more variety of unknown details. A comparision of more number of algorithms is definitely possible as in this case it was three algorithms.

REFERENCES

[1] https://www.geeksforgeeks.org/clustering-in-machine-learning/
[2] https://scikit-learn.org/stable
[3] https://en.wikipedia.org/wiki/Cluster_analysis



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY