Student Behavioral Patterns in Online Learning Platform

CECS 456: Machine Learning
Professor Fardadi
Spring 2023

Group #13

Jason Jistiripol
Ryan Gieg
Ryo Fujimura
Jovanni Garcia

Joshua Gherman
Chase Aufmann
Matthew Kriesel
Carlos Bardello

https://github.com/JoshuaGherman/CECS-456_BioMachine

Abstract

This report covers the effectiveness of clustering algorithms when covering a dataset of student's grades. The results of the study have been compared against multiple graphs using silhouette and elbow methodology. The conclusion reached is that there is a positive correlation between a student studying and achieving a higher final grade. The extensibility of this project allows for the same methodology and techniques to be used effectively against other datasets.

Introduction

The following report will cover the topic of student behavioral patterns or more simply the correlation between students who attend class and their grades. The way we have organized the data is to split the data by using k-means clustering algorithms. These algorithms include the elbow and silhouette method. We will also be testing a Hierarchical clustering algorithm, HDBSCAN clustering.

Related Work

Using unsupervised clustering techniques against a dataset of student information is far from unique. There are articles written covering all types of clustering algorithms and against datasets that include more information about the students such as ethnicity and gender. Through our search, we have not found a recent research article using the same dataset as we are. We are hoping that our perspective and deconstruction of the dataset will provide our report with the uniqueness needed.

Dataset and Features

The data reviewed comes from a machine learning repository from UCI and the dataset is called Educational Process Mining (EPM): A Learning Analytics Data Set Data Set. The dataset covers a group of 115 first year students that are undergraduate Engineering majors at the University of Genoa. The dataset displayed when students attended class, what grade they achieved for each session, and what their final grade was.

Bar Graph:

The first graph displayed is a bar graph showing the total number of students and their attendance for each day. As seen in figure 6.1, a large portion of students attended class for each class session, but they are still only 40% of the total class of students. This information helped us gain an idea to the work ethic of the student body.

Correlation Matrix:

The correlation matrix used shows a strong connection to attendance and the final grade of each student. In order to enhance readability, we have decided to combine the dataset's values for final grades one and two to create a BestFinalScore column and we also added a row displaying Attendance to easily acknowledge attendance frequency for each day. The figure 6.2 shows the correlation matrix along with a heat map. The heat map allows for a viewer to discern important information as strong correlations are given a lighter orange color. There is a positive correlation to attendance and final grade.

Box Graph:

Performing exploratory data analysis (EDA) shows the results of student's grades as shown in figure 6.3. The box graph shows a box with the top and bottom reflecting the upper and lower mean of student grades respectively. The orange line shows the actual mean of the total dataset.

Distribution Graph:

Figures 7.1-7.7 show the distribution graphs for the dataset. By showing the distribution graphs for each column, it is observable that the dataset is skewed. The skewing of the dataset is mostly due to the fact that the dataset showed students who showed up on the first day and those students never showed up again as well as students who attended the first session weren't given a grade for that day. Figures 7.1 and 7.2 show the distribution of grades for the final grade and the graph is negatively skewed as most students do not have a value registered in those columns. The closest the dataset gets to a bell curve is for the last graded session,Figure 7.7, but that is still skewed by lack of student attendance.

Variance Graph:

In order to properly implement the full scope of data, a variance graph is needed to show how PCA affects the dataset. Figure 9.1 shows the variance of the dataset. It is shown that at least 5 components are needed in order to achieve 80% variance capture. We want to achieve this variance capture in order to reduce dimensionality of the dataset.

Methods

The following methods are going to be used with a simplified dataset of the full-unaltered dataset. The simple dataset relies on the two created columns from the dataset, Attendance and BestFinalScores. The full dataset uses PCA with a parameter of 5 components to achieve a dataset that both has reduced dimensionality and high variability with the data.

Elbow Method

The Elbow Method displays a range of k with each k given a WCSS (Within-Cluster Sum of Square) value that is evaluated depending upon centroid values. We employed the elbow method by utilizing the KMeans library. In Figure 8.1, we used a range of 1 to 15 clusters for the algorithm in order to determine the value of each cluster, we had the KMeans function run 10 times with different centroid seeds and analyzed the graph to find the k we should use in the clustering algorithm

Silhouette Method

The Silhouette method evaluates the clusters based on the average silhouette score for a range of k values. The higher the silhouette score that the results displayed indicated that the cluster of k amount is the most effective cluster number. We had set the range of k to be the same as our Elbow Method along with the fact that we used 10 different centroid seeds to get the values. Figure 8.2 showcases the average silhouette scores graphed out linearly. When read as an elbow graph, the graph shows the optimal cluster amount for the given dataset. Figure 8.3 shows the silhouette graphs for each cluster, K. The graph is read by uniformity, and size of the clusters. Because of dataset is negatively skewed due to the poor attendance of students, the silhouette method's values are significantly lower when looking at the full scope of the dataset. Even then, we can still extrapolate an optimal amount of clusters.

K-means Clustering

K-means clustering is a clustering algorithm that uses k clusters, which we determined using the elbow and silhouette method, and averages the clusters around the area to classify the clusters as like variables. Figure 8.4 is created using the optimal amount of clusters found through the elbow method and silhouette method. The two-dimensionality of the graph is due to the fact that the graph is relying on the two sets of values, Attendance and BestFinalScores. The black dots show the centroids of each cluster group.

HDBSCANS

Hierarchical Density-based Spatial Clustering of Applications with Noise (hdbscan) is, as the name implies, a hierarchical clustering algorithm that uses the density of the different clusters to group the different clusters. The algorithm clusters all the dense data points, then finds the most stable clusters using hierarchy. This algorithm differs from k-means as it is hierarchical and is less susceptible to noise, because k-means can find a local solution, hdbscan is better at ignoring noise. Figure 8.7 showcases a dendrogram of the hierarchical clustering algorithm. The graph shows multiple waterfall clusters that are differentiated by color and size. Figure 8.8 showcases the dendrogram with the full scope of the dataset.

Results

The results gained through exploratory data analysis reveals that there is a positive correlation to students attending class and having a good final grade. There are outliers found that skew the graphs to showing a lower mean than what should be represented. The best method for showcasing the information would be the Correlation Matrix. The accuracy of the graph is able to clearly represent the scores of the students that would be lost when translating the data to a bar or box graph.

The results from the elbow method have consistently shown that the optimal amount of clusters for the dataset would be 3 clusters. The full dataset only skewed the graph slightly to potentially consider 4 clusters as a possible optimal result. This is the weakness of the elbow graph. An elbow graph is read visually, which makes it easy to understand the results quickly, but it lacks the depth to show exactly how optimal clusters k is.The elbow graph found in Figure 8.5 uses the full dataset within the scope of PCA. The new results achieved can be seen as 2-5 clusters being optimal.

The silhouette method is more accurate than the elbow method when reading the graphs as its purpose is to show whether clusters k is optimal. The results gained through this methodology shows that an optimal amount of clusters is 4.

```
For n_clusters = 2, silhouette score is 0.4567931432449556)
For n_clusters = 3, silhouette score is 0.4194692968302277)
For n_clusters = 4, silhouette score is 0.3685094960638865)
For n_clusters = 5, silhouette score is 0.3973564190747631)
For n_clusters = 6, silhouette score is 0.4002858782938913)
For n_clusters = 7, silhouette score is 0.4159565964326881)
```

After performing K-Means clustering, figure 8.3, with 3 clusters, which we determined was the optimal k with the elbow and silhouette method, we can see the clusters form around grade general brackets, higher scoring students, middle scoring students and low scoring students all being clustered in their own clusters. Students with better behavior, students that attended more sessions, being clustered with other students with higher grades.

The hdbscan algorithm works differently finding how many clusters to use based on the density of the clusters, it determined that 9 clusters were the most stable. Figure 9.1 shows the results of the clustering algorithm, which clustered the data points similarly to k-means showing students that attended more sessions having a better grade, and was grouped by common grade letters, the light blue at the top representing students who got an A, the darker blue a B. The dendrogram shows other possible groups of clusters based on the hierarchy it found.

Conclusion

In conclusion, we performed exploratory data analysis and we were able to find information about our dataset that helped us use machine learning to help determine a correlation between the behavior of students and test scores using two different clustering algorithms. Utilizing the elbow and silhouette method we found an optimal k value and then performed k means clustering. The data is skewed slightly from outliers, with some students performing well despite not attending as many sessions, however this did not affect the clustering. The hdbscan was fairly straightforward as the library and algorithm does most of the work.

Acknowledgements

We thank Professor Fardadi and her TA’s , Rahul Deo Vishwakarma and Carol Eunice Gudumotu, for their assistance and effort in teaching this course.

Contributions

Chase Aufmann: Presentation

Ryo Fujimura: Code

Ryan Gieg, 018301580: Code

Jason Jitsiripol, 027462849: Code

Joshua Gherman, 017614062: Report

Matthew Kriesel, 027735680: Report

Jovanni Garcia, 026718365: Presentation Slides

References

https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+%28EPM%29%3A+A+Learning+Analytics+Data+Set#

Appendix

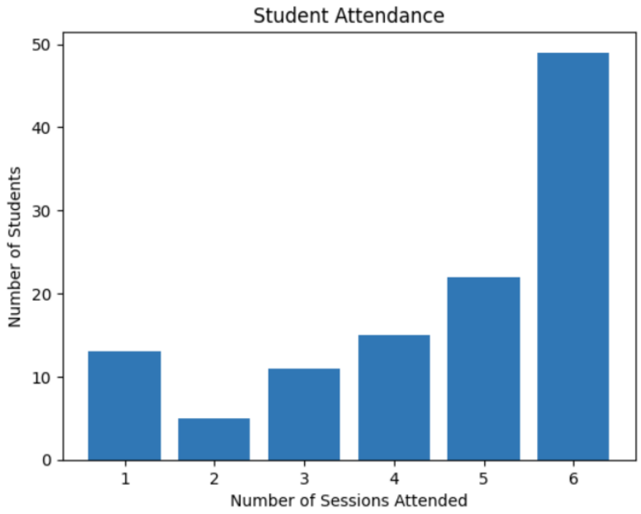

Figure 6.1

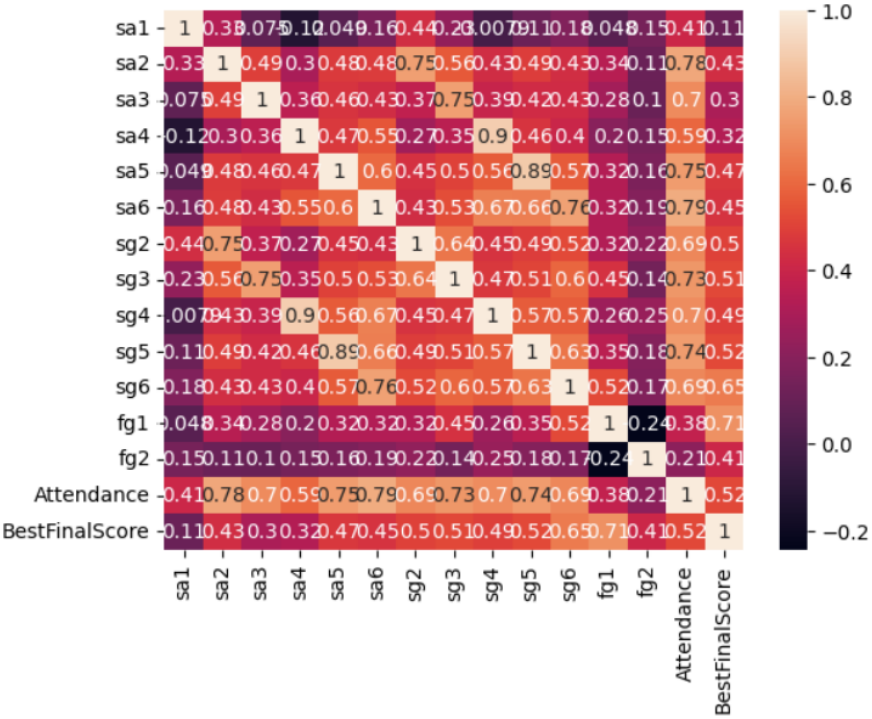

Figure 6.2

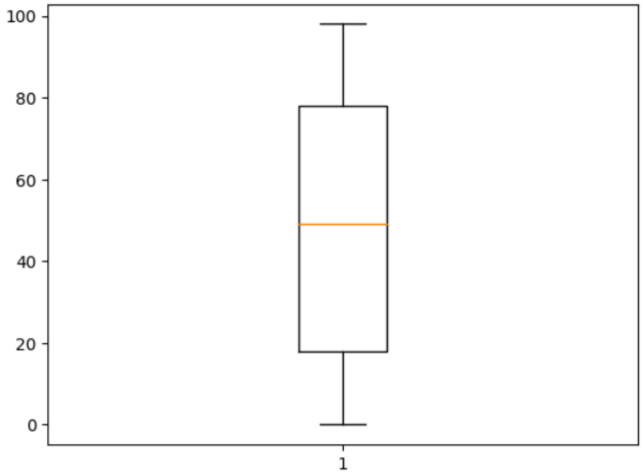

Figure 6.3

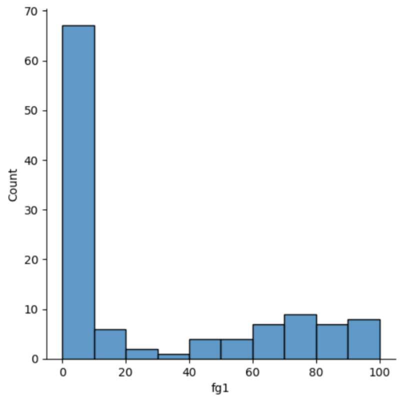

Figure 7.1

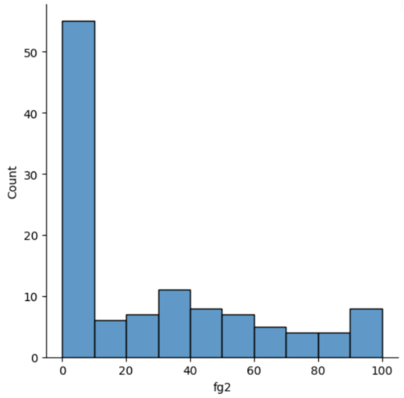

Figure 7.2

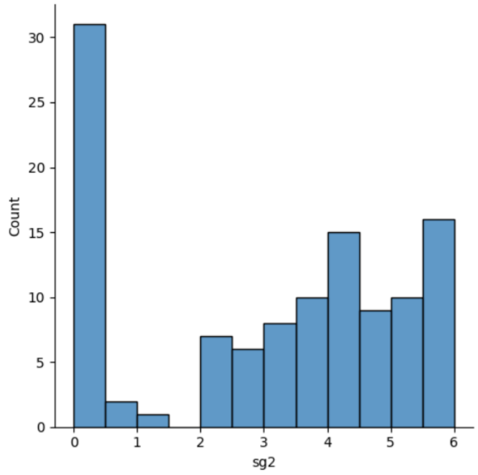

Figure 7.3

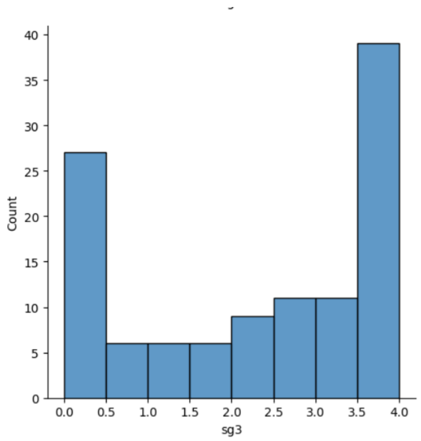

Figure 7.4

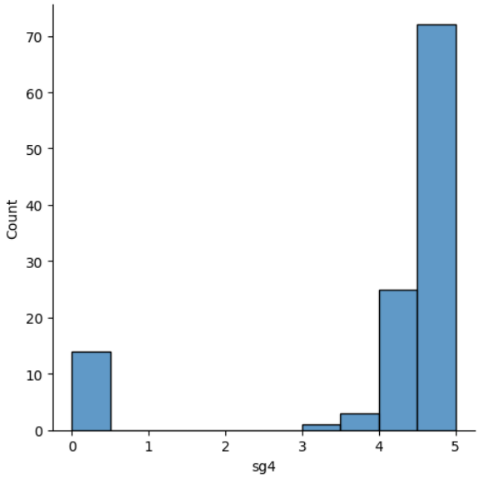

Figure 7.5

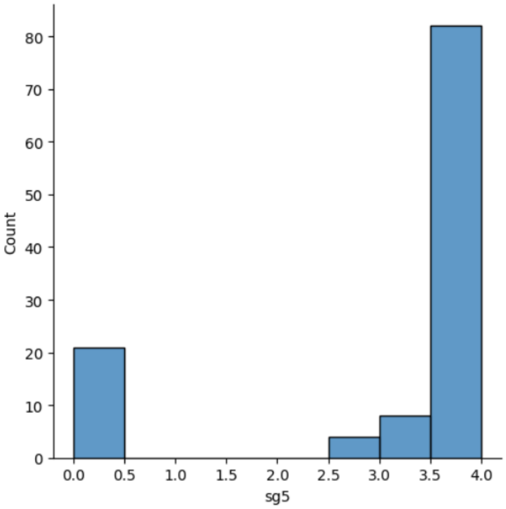

Figure 7.6

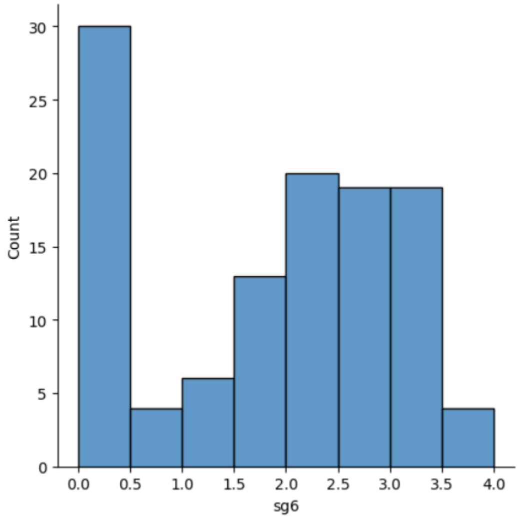

Figure 7.7

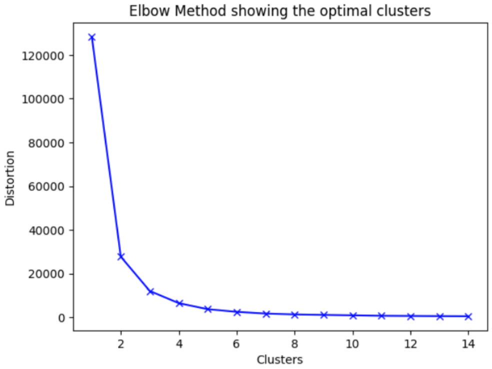

Figure 8.1

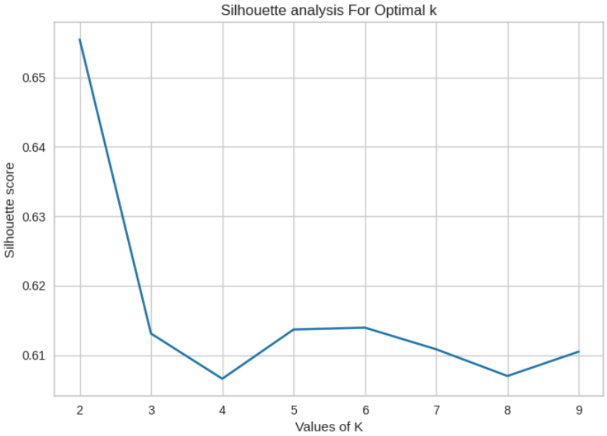

Figure 8.2

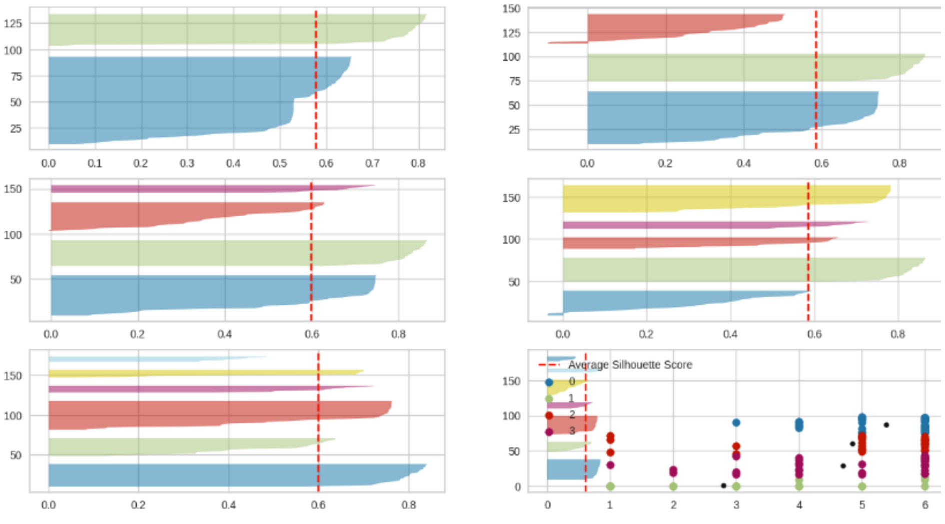

Figure 8.3

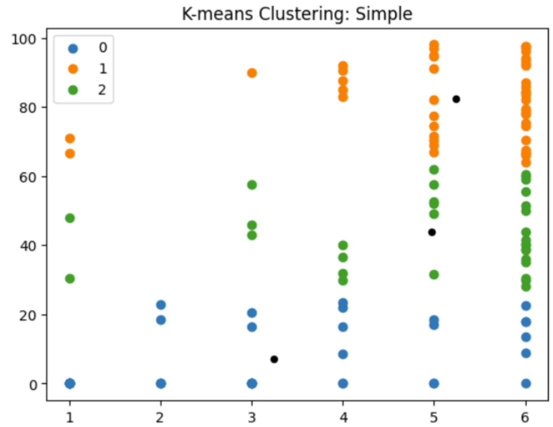

Figure 8.4

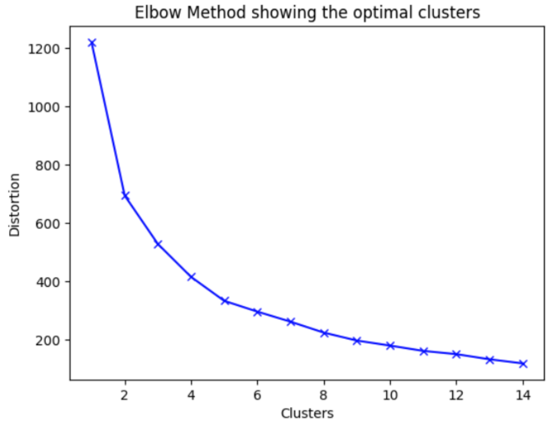

Figure 8.5

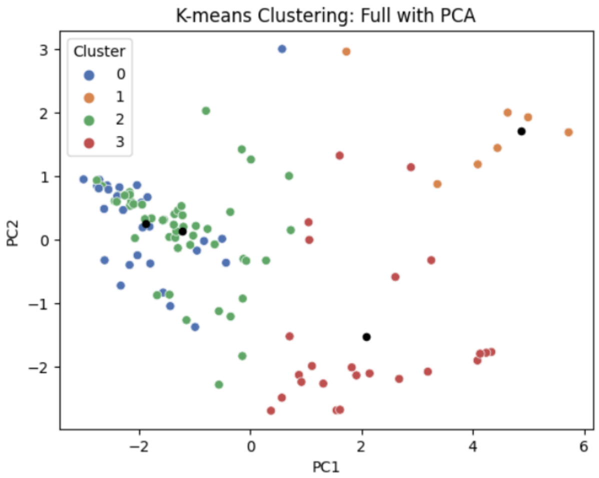

Figure 8.6

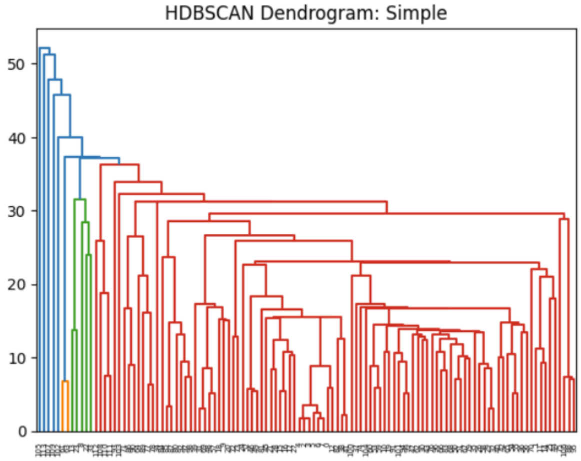

Figure 8.7

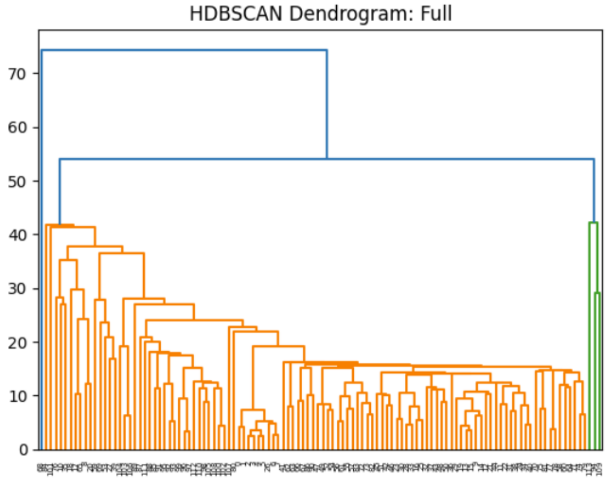

Figure 8.8

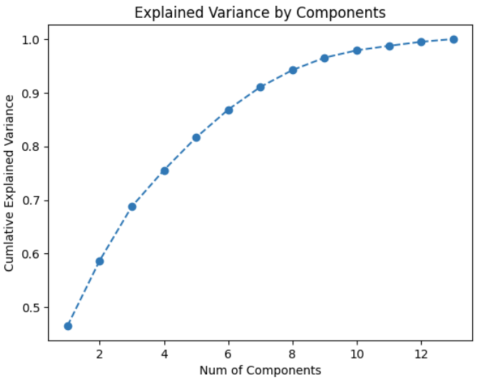
Explained Variance by Components
Cumlative Explained Variance
1.0
0.9
0.8
0.7
0.6
0.5
2
4
6
8
10
12
Num of Components

Figure 9.1

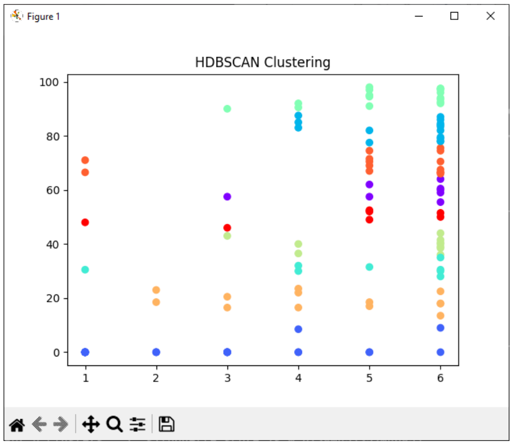
Figure 1
HDBSCAN Clustering
100
80
60
40
20
0
1
2
3
4
5
6