THE APPLICATION OF CLUSTERING TECHNIQUES IN GLASS CLASSIFICATION

Student ID: S3957034

Student Name: Truong Hong Van

Email: s3957034@rmit.edu.vn

Affiliations: RMIT University.

Date of Report: 29/05/2024

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honour code by typing "Yes": *Yes*.

Table of Contents

Executive summary

This research aimed to examine the methods to identify the type of glass pieces according to their refractive index and elemental proportions. This holds significant importance in forensic science, where glass fragments found in crime sciences can provide critical evidence. The data was first inspected to detect anomalies and key linear relationships among the features. Subsequently, we applied KMeans and DBSCAN clustering algorithms to uncover patterns among seven glass types. Parameter tuning methods such as the Elbow method and k-distance graph were applied to enhance the efficiency of the two models. Regarding the result, KMeans demonstrated better performance regarding accuracy and purity scores, making it more suitable for the glass identification task. It is recommended that the KMeans model should be adopted in similar material clustering cases, with careful adjustment of parameters and integration of domain knowledge.

Introduction

In forensic science, glass is one of the most popular physical evidence types encountered at crime scenes [1]. Therefore, glass analysis plays a crucial role in criminal investigation. The glass fragments at a crime scene can help identify potential causes of crime. For instance, bottle glass may be present in an assault case, and window glass may be evidence of a burglary. Moreover, pieces of glass can be found on a person's clothing, and the crime suspects can be identified if there is a fragment with a similar root to the glass evidence found on their clothes.

The goal of glass analysis is to determine the origin of fragments, achieved by examining the physical properties (color, thickness, refractive index) and chemical composition [2]. However, the accuracy of glass classification through sole physical properties has been questioned, because modern glass types recently have had a similar range of refractive indices [3]. This raises two challenges: First, the similarity of the appearance of the glass might be misleading, resulting in false identification. Secondly, it might be challenging to identify a new source of glass. Thus, the elemental analysis of glass is believed to help identify vital chemical elements and categorize glass types with high accuracy. With this approach, this research will center around the research question:

How can we identify the type of glass piece according to its refractive index and elemental proportions?

To address this problem, we will apply machine learning to model the chemical elements, using K-means and DBSCAN clustering algorithms. Based on an element's weight percentage in the corresponding oxide and refractive index, the models are expected to perform glass identification with acceptable precision.

Methodology

Description of dataset

The Glass Identification dataset was obtained from the open-source Machine Learning Repository to support the study [4]. The data consists of 214 samples representing the fragments extracted from 7 types of glass. Every fragment has ten attributes categorized into three groups:

- Refractive index (RI): This property measures the rate at which light bends as it traverses through glass [5], expressed in continuous values.

- Elemental proportions: The dataset involves 8 elements that are assessed based on the weight percentage in corresponding oxides. The specific attributes are detailed as follows (Figure 1).

Attribute	Description	Data type	Unit measurement
Na	Sodium	Continuous	Weight percent of Sodium oxide (Na2O)
Mg	Magnesium	Continuous	Weight percent of Magnesium oxide (MgO)
Al	Aluminum	Continuous	Weight percent of Aluminum oxide (Al2O3)
Si	Silicon	Continuous	Weight percent of Silicon oxide (SiO2)
K	Potassium	Continuous	Weight percent of Potassium oxide (K2O)
Ca	Calcium	Continuous	Weight percent of Calcium oxide (CaO)
Ba	Barium	Continuous	Weight percent of Barium oxide (BaO)
Fe	Iron	Continuous	Weight percent of Iron oxide (Fe2O3)

Figure 1. Dataset description

- Type of glass: This categorical attribute indicates distinct glass types. A fragment is categorized into a type of glass given its composition, which is also the primary focus of this research. There are seven categories, namely building windows (float and non-float), vehicle windows (float and non-float), containers, tableware, and headlamps in order from 1 to 7.

Data preparation

Firstly, the glass dataset is loaded into the IPython environment as a data frame, featuring 11 features as columns and 214 observations as rows. Concerning the dataset overview:

The attributes are all numeric types, with the ID columns as the index and the float columns as the chemical element proportions. Although the Type column is of an integer type, it has seven unique values. Thus, this column can serve as a categorical column.

- Each column contains 214 instances, suggesting no missing values in the dataset.
- The data types are consistent object data types, suggesting no data type entry error.

Moreover, the observation reveals varying scales in element distributions, which might need to be standardized before the model development. Following that, we primarily inspect the impossible values and outliers.

Outliers

The inspection of outliers is achieved via the interquartile range, which indicates that values outside the Q3 – Q3 range are outliers. Magnesium is the only element without outliers, whereas the remaining elements contain an excessive number of them. Furthermore, the box plot pattern failed to apply to the distribution of Ba, so we cannot decide whether this element has outliers. To mitigate the effect of the outliers on the learning algorithm, we recognize the necessity to eliminate the observations with more than two outliers from the dataset.

Impossible values

- For elemental proportion, the impossible values can be detected by two criteria:
- The weight percentage of the oxides should range between 0 and 100 since glass must comprise at least two elements.
- The sum of oxide weight percentage should not exceed 100% [6], although it could be less than 100% due to impurities in the composition. Moreover, a minor deviation of 100 is acceptable.

Upon examination, all attributes in the dataset fall into the range of 0 and 100. However, there are 26 observations with the sum of weight percentage larger than 100, with minor variations from 0.01 to 0.1. These variations can be explained by errors in number rounding and measurements. Hence, the values are acceptable, with the belief that it would introduce little error in the modelling stage.

Regarding the refractive index, the general glass refractive index cannot be negative, and the observations meet this condition.

Data exploration

Single feature exploration

We initially construct descriptive statistics to gain an overall understanding of the inspected attributes. This statistical approach aims to delve into following aspects of a single feature:

- Understanding measures of central tendency, from which we can identify the main and supplementary elements in glass composition.
- Assessing the general variations of an element. Assessing the general variations of an element. The coefficient of variation is considered an effective way to assess and compare the variability of the dataset with different means [7].
- Detecting any anomalies in the distribution pattern based on the percentile, min, and max values.

To further investigate the variability of each attribute, we utilize histograms as the second method. This type of visualization helps us observe the shape of data and identify the concentrated value range of the attributes. By that, we can also notice the similarities and differences in the attribute patterns.

1. **Primary and supplementary elements**

It is evident that Silicon (Si) serves as a predominant element in glass composition, with the average percentage of 72.65%. Sodium (Na) and Calcium (Ca) are the subsequent components in glass, accounting for 13.41% and 8.96% respectively, which is one-seventh of the Si content. Conversely, the presence of other elements is relatively minor, typically below 3% percent on average. Among the supplementary elements, Mg comprises 2.68% of the content in a fragment, while Al has an average content of 1.44%. K and Ba represent 0.50% and 0.18% of the content respectively, with Fe being the least prevalent element at 0.06%.

2. **Variability and shape of data**

Regarding the refractive index, the average of 1.52 and a standard deviation of around 0 suggest minimal variability. Also, the distribution of RI is relatively symmetric. We can infer that the values of the refractive index are consistent throughout the data.

In terms of the elements, they can be categorized into two groups depending on the coefficient of variation. In the low variability group, Mg exhibits the highest volatility at 39.74, followed by Al and Ca. Na and Si shows the least volatility at 4.76 and 0.8. Notably, the distribution of Na, Al, Si are relatively symmetric, while the distribution is left-skewed for Mg and right-skewed for Ca. We can deduce that Na, Si, and Al tend to have a more constant distribution among fragments, except for Mg and Ca.

Conversely, the high group with high variability indicates greater content fluctuation among fragments. The element with the highest variability is Ba, with a coefficient of 362.53. Subsequently, Fe and K are also highly variable with 163.86 and 56.62 coefficients respectively. In terms of the shape, Ba and Fe share a similar pattern of right skewness, while K is left-skewed with a small number of outliers of 1 weight percent.

3. Relation to the target

Following the elemental distribution, we examine the distribution of each element in glass categories. This approach aims to identify the potential indicators associated with each glass type during the clustering process. Concerning the low variability group of elements, Na, Al, and Si share a similar pattern, with a higher concentration in glass types 6 and 7 (Figure 2). The proportions of these elements are averagely consistent in glasses 1, 2, and 3, although there exists a significant difference in the scale. Conversely, the Ca element records the highest content in glass type 5, making it a plausible predictor for this category. Regarding the presence of Mg, its concentration is significantly high in glass types 1,2, and 3, while being minimal in glass type 7.

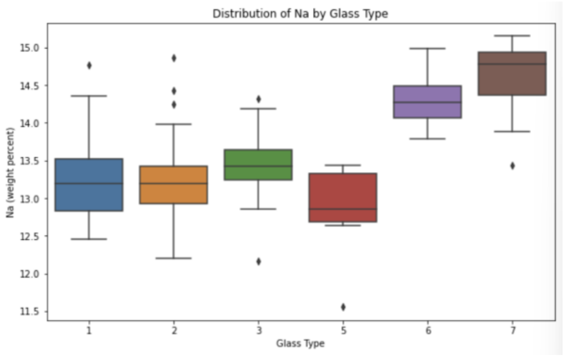


Figure 2. Distribution of Na by glass type

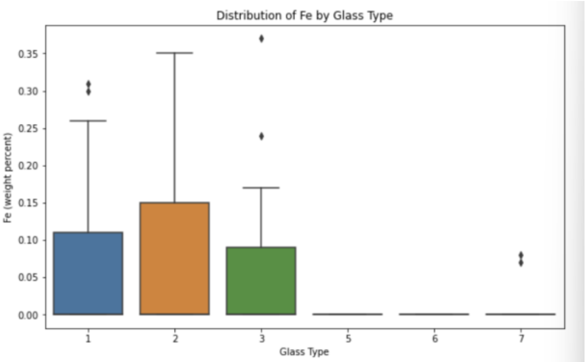


Figure 3. Distribution of Fe by glass type

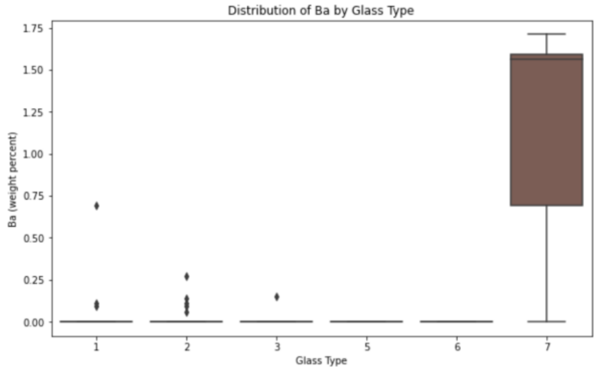


Figure 4. Distribution of Ba by glass type

Concerning the high variability group, the distribution indicates the absence of K in glass type 6 and 7, which is similar to the Fe content (Figure 3). Conversely, glass type 7 exclusively exhibits a notable Ba concentration (Figure 4). This suggests the possibility of predicting a glass type characterized by high content of Ba, or the lack of Fe and K.

Relationship between features

By exploring the relationship between attributes, we can identify key features for the modelling process and eliminate redundant ones to optimize the performance. To achieve this objective, a correlation matrix is constructed, from which the potential correlations can be observed by the rate. The strongest positive correlation is at 0.76, while the strongest negative correlation is recorded at -

0.67. By that, we will examine the pairs with remarkable correlation rates, with associated hypothesis.

1. RI vs Ca

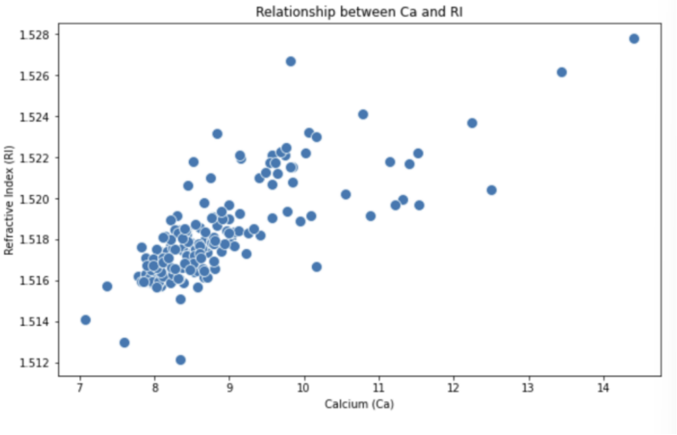


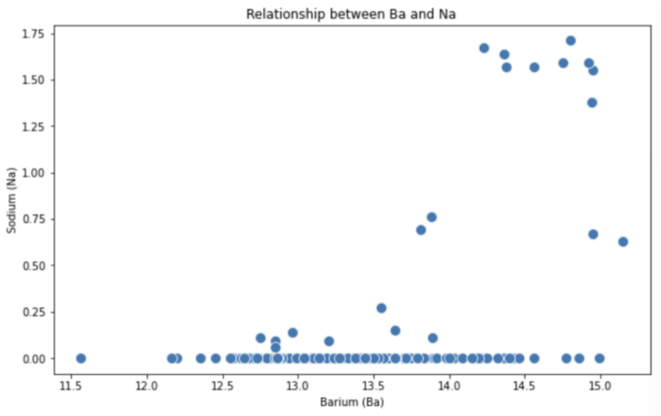
Figure 5. Relationship between Ca and RI

Hypothesis: Higher Ca results in higher RI.

As the content in Ca increases, the rate of refractive index has the tendency to rise as well, with a relatively clear positive linear pattern. By that, we can consider eliminating one element in the feature selection stage, so as to simplify the dataset dimension.

2. Ba vs Na, Ba vs Al

Figure 6. Relationship between Ba and Na

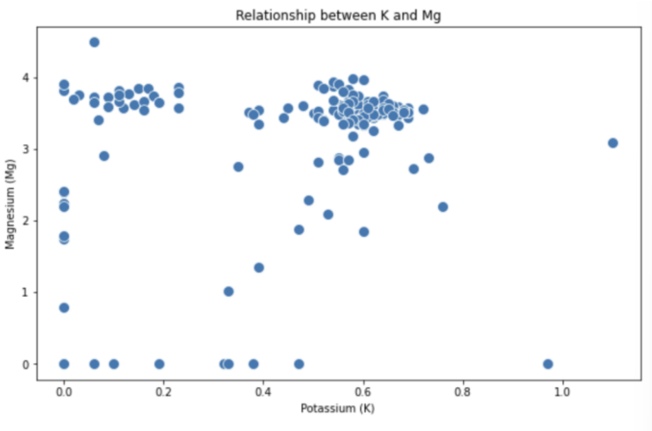


Hypothesis: Ba has a positive linear relationship with Na.

Hypothesis: Higher level of Ba is associated with higher level of Al.

As the Ba proportion increases, the Sodium level primary concentrates at 0 percent, regardless of the fact that there exists a small number of samples with high Ba and Na content. A similar pattern is also observed for the pair Ba – Al, when the Ba content remains at 0 percent while the proportion of Al rises. Thus, we cannot conclude that these pairs of attributes have a linear relationship.

3. K vs Mg



Hypothesis: High level of K is related to high level of Mg.

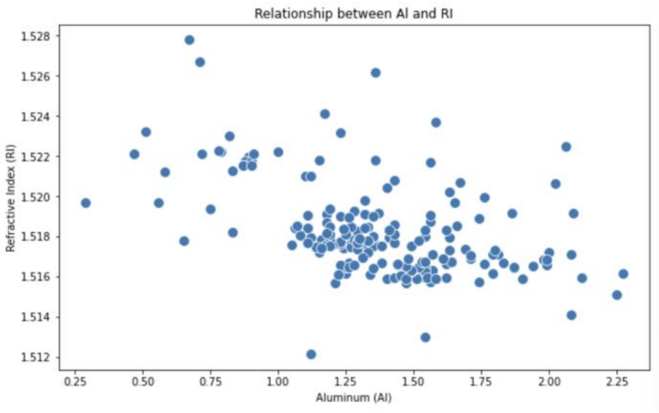
In Figure 7, the data points representing K and Mg concentrate in different clusters, with no particular direction. Thus, there is no relationship between K and Mg in the dataset. However, the distinct cluster distribution could be beneficial in visualization during the clustering process.

Figure 7. Relationship between K and Mg

4. RI vs Si, RI vs Al

Hypothesis: Lower level of Si results in higher rate of RI.

Hypothesis: Low content of Al is associated with high rate of RI.

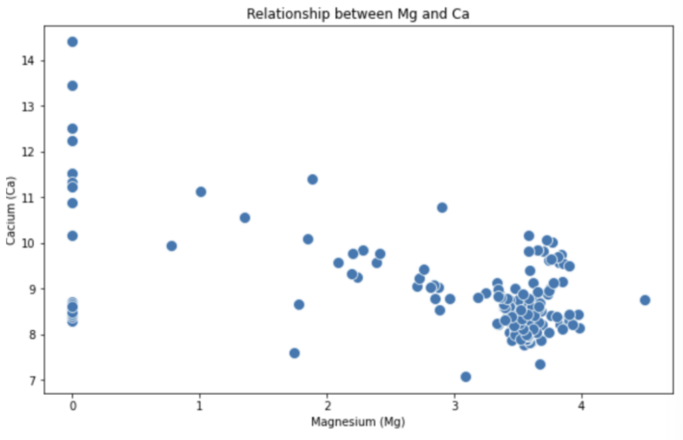


Observing Figure 8, we can see that as the Si content decreases, there tends to be more data points with higher RI values. Conversely, the RI values appear to generally fall when the Si concentration rises. The pair RI – Al also has a similar pattern, with much clearer negative linear direction in the distribution of data points. Hence, we can conclude that RI has a negative linear relationship with Al and Si.

Figure 8. Relationship between Al and RI

5. Ba vs Mg, Mg and Ca

Figure 9. Relationship between Mg and Ca



Hypothesis: Ba and Mg have a strong relationship.

Hypothesis: Higher level of Mg is related to lower level of Ca.

The concentration of Ba remains constant with the increase of Mg, and there are some anomalies with consistent Mg at 0 percent as Ba content increases. Conversely, the pair Mg and Ca exhibit a relatively negative relationship; however, the data points tend to be clustered rather than being distributed with a particular direction. Therefore, we reject two hypotheses made for Ba – Mg and Mg – Ca.

6. Mg vs Al

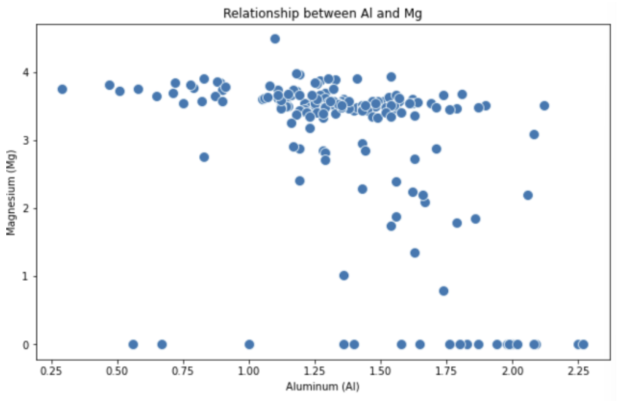


Figure 10. Relationship between Al and Mg

Hypothesis: High content in Al is associated with low content in Mg.

It can be seen from Figure … that the data points represent the inspected pair scatter around, and the content of one element tend to remain constant in response to the rise of the other. This concludes that there is a lack of relationship between Mg and Al.

7. K vs Na

Hypothesis: K has a negative linear relationship with Na.

Although the correlation rate of this pair of attributes is relatively high, the scatter plot shows that the data points concentrate in the center with an excessive number of anomalies in the distribution. Thus, there is not enough evidence to conclude that K has a relationship with Na.

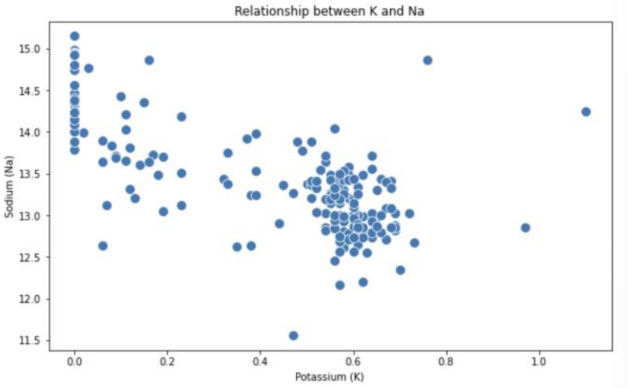


Figure 11. Relationship between K and Na

In conclusion, the refractive index presents a linear relationship with Ca, Si and Al. Conversely, there seems to be a lack of relation among other pairs of attributes inspected.

Data modelling

Concerning the feature selection, we decided to eliminate the refractive index feature, with the belief that it provides a lack of information for the model. The distribution of refractive index is relatively consistent amongst glass types, and the values of refractive index can be represented by the distribution of Ca. Hence, in the model training stage, there are elemental features involved: Na, Mg, Al, Si, K, Ca, Ba, and Fe.

Next, we also recognize the importance of feature scaling, by transforming the data with various magnitudes into a common scale. This approach is believed to reduce the impact of extreme values and outliers, while enhancing the model performance. In particular, the standardization scaling is applied, which transforms the features to have the mean of 0 and standard deviation of 1. Assuming the data is normally distributed, the score will be calculated via the following formula [8]:

$$Z = \frac{x - u}{s}$$

For the model selection, clustering techniques with KMeans and DBSCAN will be applied, with a view to identifying patterns of glass types hidden in the data. To evaluate the performance of the models, two methods are applied:

- Confusion matrix: Since we indeed have predefined labels for comparison, taking advantage of the confusion matrix helps us compare the number of samples in the predicted clusters in relation to the true target labels.
- Accuracy score: To compute the accuracy, each cluster is assigned to the class that is most frequent in the cluster. Then, these most frequent classes will be mapped to the cluster assignment. For example, if cluster 0 has the highest number of samples recorded for target 7, the samples with the cluster assignment as 0 will be predicted as type 7. Finally, the *accuracy_score()* function, taken from the sklearn.metrics module will be utilized to compute the accuracy of the algorithm. This approach allows us to identify if the predicted labels match the true labels of the data.
- Purity score: This metric evaluates the homogeneity of clusters in terms of the most frequent true label in each cluster [9]. The purity value can range from 0 to 1, indicating bad to optimized performance.

KMeans

KMeans algorithm has one primary parameter, which is k (number of clusters). In the initial step, we define k=6 as a fair number of centroids, since we learned that the dataset provided has 6 pre-defined glass types. Thus, the number of clusters set as 6 might reflect effectively the target labels in the case that the model achieves the best performance with high accuracy.

To identify the optimal number of clusters, the Elbow method technique is employed. In particular, the sum of squared distance between each data point in a cluster is calculated, with the k ranging from 2 to 10 [10]. Visualizing the within-cluster sum of square (WCSS) list (Figure 12), the WCSS values fall rapidly before the number of clusters reaches 5. Notably, there is another elbow of k=3, which can be explained by the presence of outliers.

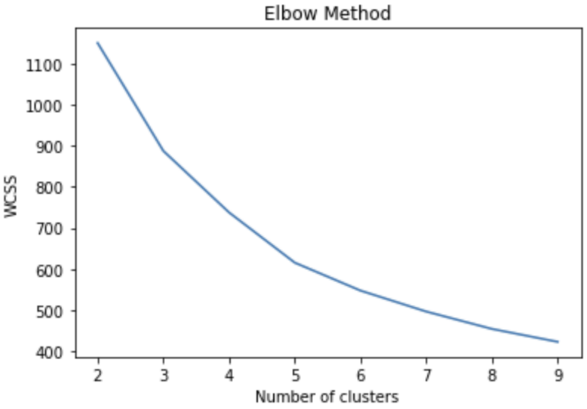


Figure 12. Elbow method

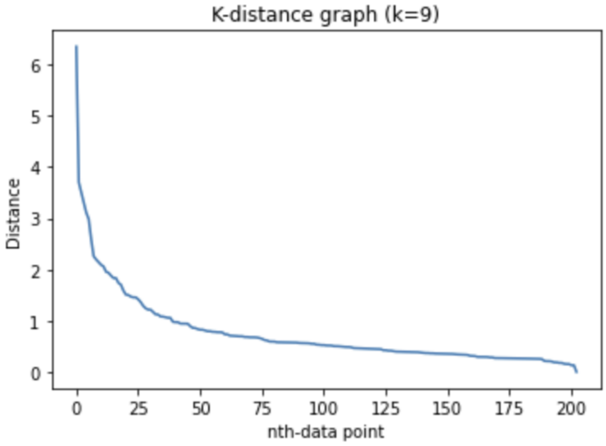


Figure 13. K-distance graph (k=9)

DBSCAN

There are two parameters required for the DBSCAN algorithm, which are the MinPts (min_samples) and Epsilon. MinPts, in this case, is the minimum number of observations, representing as data points in a cluster. The MinPts should be greater than or equal to the dimensionality of the dataset [11], with the optimal value of MinPts being twice the data dimensions [12]. As suggested, we will set the MinPts = 9 and MinPts = 16 respectively for two models and compare their performance.

In terms of the Epsilon, this value could be observed in the average k-distances graph suggested in [13]. The research suggested an intuitive technique to calculate the average distance between the data point and its k nearest neighbor, where k = MinPts. Then, the distances will be plotted in a graph, and the optimal Epsilon value can be identified as the first valley in the graph. With this tuning method, the optimum Epsilon value, corresponding MinPts = 9 and MinPts = 16, is 1.5 for both (Figure 13).

Results and Discussion

KMeans

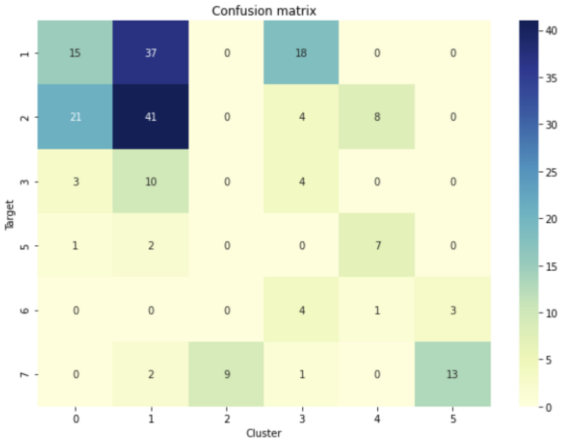


Figure 14. Confusion matrix (k=6)

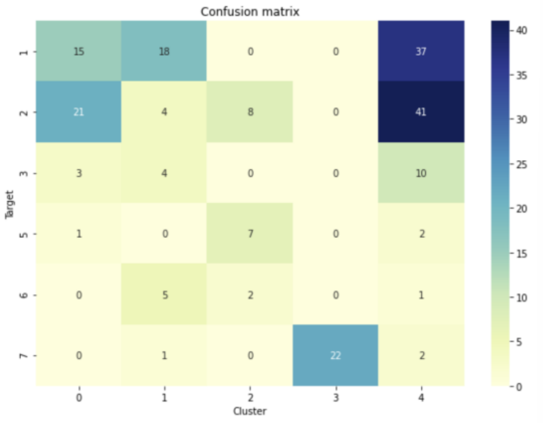


Figure 15. Confusion matrix (k=5)

Figure 14 demonstrates the relationship between the clustering results and the true observation labels, with the number of clusters as 6. Each true label contains a cell with the highest number of observations. Thus, we can infer that the cluster containing that cell might represent the corresponding true label. Based on that notion, each cluster should only represent one target label, creating a diagonal shape for the result. We can see that cluster 5 represents well for target 7, as for cluster 3 with target 6 and cluster 4 for target 5. However, target labels 1,2 and 3 are conflated, when their highest number of observations all belong to cluster 1. This conflation indicates an error in clustering; therefore, the algorithm is not optimized. The accuracy score, which compares the true label against the clustering labels, is 0.54, indicating that only 50% of the observations are correctly identified.

For the optimal k=5 (Figure 15), the result appears to be similar to 6 clusters, where target labels 1,2, and 3 are still conflated in a single cluster. Also, it can be observed that the lower number of clusters could result in the samples being merged into a cluster. The accuracy score calculated for k=5 is also 0.54, but the purity score increases to 0.60.

In general, the KMeans algorithm does not perform well in predicting target label 1, 2, and 3.

DBSCAN

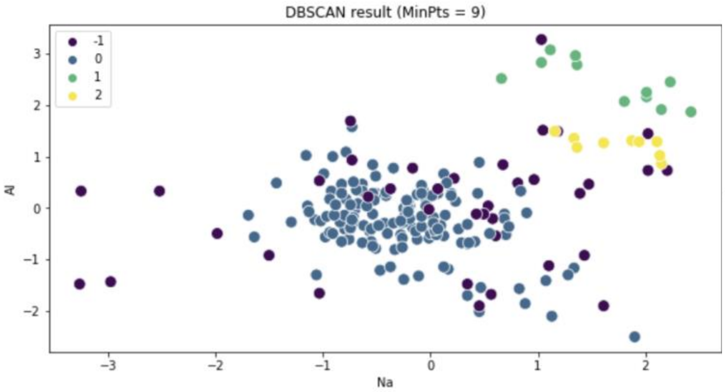


Figure 16. DBSCAN result (MinPts=9)

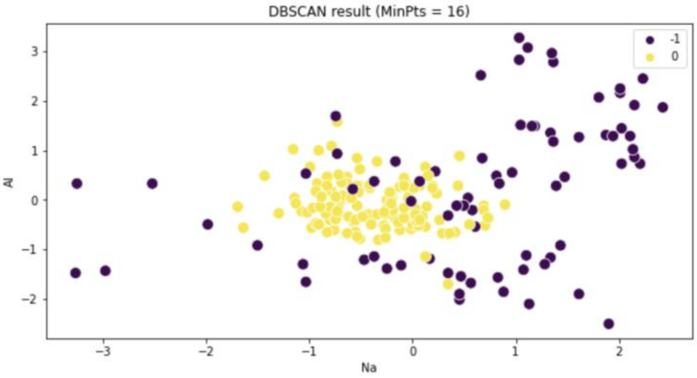


Figure 17. DBSCAN result (MinPts=16)

In terms of the DBSCAN algorithm, with MinPts = 9, there are 3 clusters formed, except for the label -1 indicated the noises. It can be seen in Figure … that the algorithm separates clusters 1 and 2 well;

however, cluster 0 is mixed with the noise. Regarding the target label, both clusters 1 and 2 represent the glass type 7 with 11 and 9 observations respectively, while glass types 1, 2, and 3 are still conflated in cluster 0. These results indicate that the algorithm might not be optimized for identifying the glass types.

Observing Figure … which demonstrates the clustering result for MinPts=16, all the data points are merged into one cluster, except for the noise. By that, we recognize the importance of the MinPts, for the higher minimum samples raises the risk of mixing the clusters together.

Comparison

Model	Parameters	Accuracy score	Purity score
KMeans	k=6	0.54	0.55
KMeans	K=5	0.54	0.60
DBSCAN	MinPts=9, eps=1.5	0.44	0.51
DBSCAN	MinPts=16, eps=1.5	0.27	0.39

Figure 18. Evaluation metrics

We compare the two algorithms with different parameters in terms of accuracy score and purity score (Figure 18). In the case of glass type identification, it can be observed that the KMeans models achieve a higher accuracy score than the DBSCAN algorithms, with approximately 54% of the observation correctly classified. The DBSCAN algorithms, given the specified parameters, perform below average in this case, with only below 50% of the correctly labelled samples.

In terms of the purity score, KMeans models appear to perform slightly better than the DBSCAN. Particularly, 55% and 60% of purity is observed for the KMeans model, while the least purity score is recorded for the DBSCAN.

Overall, the clustering results indicate that the KMeans algorithm is more efficient in identifying the glass types; therefore, it is recommended that the KMeans models should be applied to similar cases. However, it is noted that the recommendation is based on the algorithm performance for the given parameters only, which does not mean that KMeans is better than DBSCAN under all circumstances.

Conclusion

In this study, we aimed to divide the fragments into distinct types of glass, based on their refractive index and elemental properties, using KMeans and DBSCAN algorithms. The results suggest that the KMeans algorithm performs relatively better for clustering glass types, although it encounters challenges in correctly identifying the types of glass. While KMeans might be more suitable for similar clustering tasks in other glass datasets and materials, however, this approach needs to be refined for practical implementation in glass evidence analysis. In future research, we could focus on optimizing the parameters of the clustering algorithms and incorporating domain knowledge into the clustering process to improve the accuracy of the models.

References

[1] J. M. Curran. “The Statistical Interpretation of Forensic Glass Evidence.” *International Statistical Review / Revue Internationale de Statistique*, vol. 71, no. 3, pp. 497–520, 2003. [Online]. Available: https://www.jstor.org/stable/1403825

[2] University of Central Florida. “Glass – National Center for Forensic Science.” (accessed May 29, 2024). [Online] University of Central Florida. Available: https://ncfs.ucf.edu/research/chemical-evidence/glass/

[3] J. A. Buscaglia, “Elemental analysis of small glass fragments in forensic science,” *Analytica Chimica Acta*, vol. 288, no. 1–2, pp. 17–24, Mar. 1994, doi: https://doi.org/10.1016/0003-2670(94)85112-3.

[4] B. German, 1987. “Glass Identification”. UCI Machine Learning Repository. [Online]. Available: https://doi.org/10.24432/C5WW2P .

[5] Britannica. “Refractive Index.” (accessed May 29, 2024). [Online] Encyclopædia Britannica. Available: https://www.britannica.com/science/refractive-index

[6] GeeksForGeeks. “Percent By Weight Formula.” (accessed May 29, 2024). [Online] GeeksforGeeks. Available: https://www.geeksforgeeks.org/percent-by-weight-formula/.

[7] Advanced Research Computing - Statistical methods and analytics. “FAQ: What is the coefficient of variation?.” (accessed May 29, 2024). [Online]. Advanced Research Computing - Statistical Methods and Analytics. Available: https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-what-is-the-coefficient-of-variation/.

[8] Scikit-Learn, “sklearn.preprocessing.StandardScaler — scikit-learn 0.21.2 documentation.” (accessed May 29, 2024). [Online]. Scikit-learn.org. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[9] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008.

[10] T. Mullin. “DBSCAN Parameter Estimation Using Python.” Medium. Accessed: May 29, 2024. [Online]. Available: https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd.

[11] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, “DBSCAN Revisited, Revisited,” *ACM Transactions on Database Systems*, vol. 42, no. 3, pp. 1–21, Aug. 2017, doi: https://doi.org/10.1145/3068335.

[12] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, “Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 169–194, 1998, doi: https://doi.org/10.1023/a:1009745219419.

[13] N. Rahmah, I. S. Sitanggang. (2016). Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra. Presented at IOP Conf. Series: Earth and Environmental Science. [Online]. Available: https://iopscience.iop.org/article/10.1088/1755-1315/31/1/012012