POLYTECHNIC UNIVERSITY OF MADRID
E.T.S. DE INGENIERÍA AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS
E.T.S. DE INGENIEROS INFORMÁTICOS

APPLIED PROJECT MACHINE LEARNING

PRE-PROCESSING AND CLUSTERING WORK

Computational Biology

IMPLEMENTED BY

Alberto González Calatayud and Alejandro Bernabeu Duréndez

ACADEMIC COURSE: 2023-2024

Abstract

This study focuses on the exploration and comparison of multiple data pre-processing approaches and clustering techniques in the analysis of a dataset. The main objective of this research is to determine the most effective and optimal configuration for data processing. Throughout the study, various pre-processing strategies have been considered, from data cleaning and normalisation to feature engineering. Different clustering methods and algorithms have also been evaluated in order to identify the strategy that maximises consistency and efficiency in data clustering.
The focus on optimising pre-processing and clustering stems from the importance of these steps in obtaining accurate and meaningful results in data analytics. Determining the best approach not only drives research progress, but also provides a solid path for decision making in a wide range of applications, from agricultural management to efficiency in data organisation and distribution in various sectors.

Keywords Clustering; Evaluation; Inertia; Pre-Processing; Score; Silhouette

1 INTRODUCTION

The data set to be analyzed in this project consists of a series of data pertaining to measurements made on different fruits. These measurements range from width, length, weight to different parameters such as regularity or cleft of the fruit. Therefore, in order to analyze these data, it is necessary to carry out both a preprocessing (Kang and Tian, 2018) and a classification of the dataframe (Rodriguez et al., 2019).

1.1 Pre-processing

The data pre-processing stage is an essential component of any data analytics and machine learning project. It represents the crucial starting point in the model development cycle, where raw data is conditioned for use in analysis and model training. Preprocessing addresses a number of challenges, from cleaning up missing data to transforming variables, ensuring that the data is in an optimal form for processing (Maharana et al., 2022). This phase plays a critical role in the quality and effectiveness of the final results, as poor quality or poorly prepared data can lead to erroneous conclusions or unreliable models. In this paper, the key steps of pre-processing are explored (Tiu et al., 2021), highlighting their importance in data-driven decision making and in building accurate and robust models.

1.2 Clustering

Clustering is a fundamental concept in machine learning that plays a crucial role in organizing and extracting meaningful patterns from complex data sets. It involves the grouping of similar data points into distinct clusters (Ezugwu et al., 2022) based on inherent similarities or characteristics, thereby revealing underlying structures within the data. This unsupervised learning technique holds paramount importance across this paper, diferent clustering algorithms are tested in order to build the best clustering model.

2 METHODS

This study seeks to apply advanced Machine Learning techniques for the segmentation and clustering of data in a specific dataset. To carry out this analysis, various tools and libraries within the Python programming environment were used, taking advantage of the capabilities of Google Colab.

1. **numpy(np):** Version 1.23.5
2. **pandas(pd):** Version 1.5.3.
3. **matplotlib:** Version 3.7.1.
4. **seaborn:** Version 0.12.2.
5. **scipy:** Version 1.11.3.
    - .stats
    - .spatial.distance
    - .cluster.hierarchy
6. **scity-learn (sklearn):** Version 1.2.2.
    - .KMeans
    - .silhoutte_score
    - .silhoutte_samples
    - .StandardScaler
    - .MinMaxScaler

3 RESULTS & DISCUSSION

3.1 Pre-Processing

In the first stage, the data is observed, including the type of data to be worked with and the different variables involved.

	weight	length	width	regularity	cleft
0	1.205	4.60	2.847	5.691	Small
1	1.726	5.978	3.594	4.539	Large
2	1.126	4.516	2.710	5.965	Average
3	1.755	5.791	3.690	5.366	Large
4	1.238	4.666	2.989	6.157	Small

Table 1: *First lines (head 5) of the dataset in which can be observed and analysed the type of data to work with.*

3.1.1 Cleaning and tyding the data

Using the *head* command, one can visualize the data type and variables that comprise the "*fruits.csv*" dataset. It can be seen in the Table 1. Four numerical variables corresponding to "*weight*", *length*, "*width*" and "*regularity*", and one categorical variable "*cleft*" can be observed. This suggests that if the information contained in "*cleft*" is meaningful to the analysis, which seems consistent, some mechanism of "*enconding*" will be needed. Upon initial inspection, it becomes evident that there is a column lacking a title, and this column does not offer any significant information.

3.1.2 Delete uninformative column

This particular column corresponds to the number of rows, and when the dataset is transformed into a pandas dataframe, an additional column is automatically added to represent the row numbers, often named "Unnamed: 0." Consequently, when converting the dataset into a pandas dataframe, an extra column without a header is generated by default to indicate the row numbers.
After a review of the data, the decision was taken to remove the column called “Unnamed: 0” from the dataset. This column does not seem to provide relevant information for the analysis, so it was decided to discard it in order to simplify and clean up the dataset.

Upon reviewing the data descriptions of the different variables, it is evident that all columns contain missing values. In addition, an anomalous value in the variable “*weight*” of 99 is likely to be present and will be checked.

3.1.3 Outlier

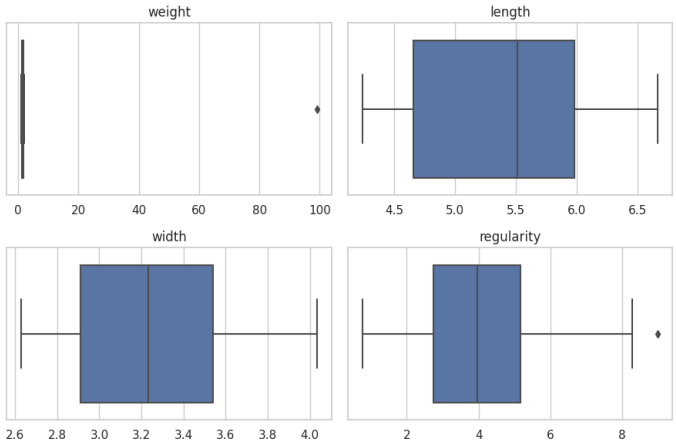

Figure 1: *Representation of the distribution of the data of the numerical variables by boxplot. An out-of-bounds value can be observed for the variable “weight”.*

As mentioned above, the anomalous data corresponding to the "weight" variable is reviewed and it is decided to replace it with the average of the other values in the column, since, when compared with these, everything seems to indicate that it is an error.

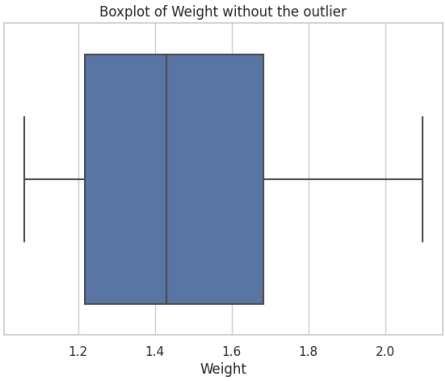

Figure 2: *Representation of the distribution of the data of the variable “weight” after the imputation of the outlier by the mean of the values corresponding to its column.*

3.1.4 Missing Data Analysis

This is essential for understanding the completeness of the data in the set and for transparent communication of results. The presence of missing data can affect the quality of subsequent analyses and models and may lead to incorrect or biased interpretations.

In addition, by identifying missing values it can be decided whether to remove the feature or to keep them by imputation.

The generated code was responsible for calculating the amount of missing data for each column and expressing this result both in absolute values and as a percentage of the total data.

	Total	Missing values %
length(cm)	24	13,33
width(cm)	17	9,44
weight(kg)	12	6,67
regularity	10	5, 55

Table 2: *Table containing the % percentage of missing data in each numeric column of the dataset.*

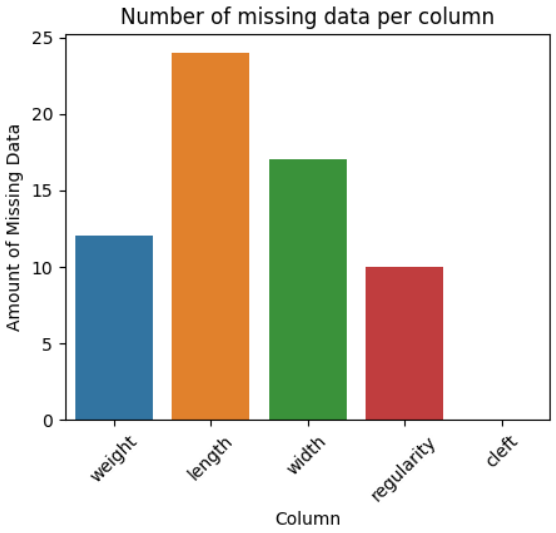

Figure 3: *Graphical representation by bar chart showing the number of missing values for each variable in the set.*

Additionally, a bar chart showing the amount of missing data in each column can be seen in *Figure3*. This graph provides a visual representation of the columns with missing data and their magnitude. This allows for a quick and effective identification of the distribution of missing data so that decisions can be made for their treatment.

Rows with 0 missing values	122
Rows with 1 missing values	53
Rows with 2 missing values	5
Rows with 3 missing values	0
Rows with 4 missing values	0

Table 3: *Count of rows with missing values*

Furthermore, as can be seen in the Table 3, the possibility of having rows with all missing values, i.e. rows with no values for any of the variables, was checked. In this case there are no rows containing more than two missing values so no rows have been removed from the data.

In summary, these steps are essential to ensure data quality, make informed decisions on how to address missing data, and ensure that subsequent analyses and modelling are based on robust and reliable information.

3.1.5 Renaming Column Headings

In addition, before addressing the missing data, the column headings in the dataset were modified. The original headings did not contain information on the units of measurement of the characteristics, which made it difficult to interpret the data.

Before	After
length	length(cm)
width	width(cm)
weight	weight(kg)
regularity	regularity

Table 4: *Table containing the new names of the columns*

3.1.6 Imputation of Missing Data with Averaging

Following the assessment of the presence of missing values and their distribution, it has been decided to perform an imputation process with the mean of the existing values in the respective columns.
As can be seen in Figures 2 and Figure 3, the percentage or number of missing values in any of the columns is not significantly high (no more than 10%), with the exception of the variable "*Lenght*", where it reaches 13%. However, this value is not high enough to justify the elimination of this column, given the limited number of columns or variables available. In addition, the variable "*Lenght*" has a significant importance in the characteristics of the fruits and is expected to provide a high informative value to the learning algorithm.

For this purpose, a study of possible imputation strategies has been carried out, including substitution by the mean value, substitution by 0, substitution by the maximum value or some value of the percentiles.

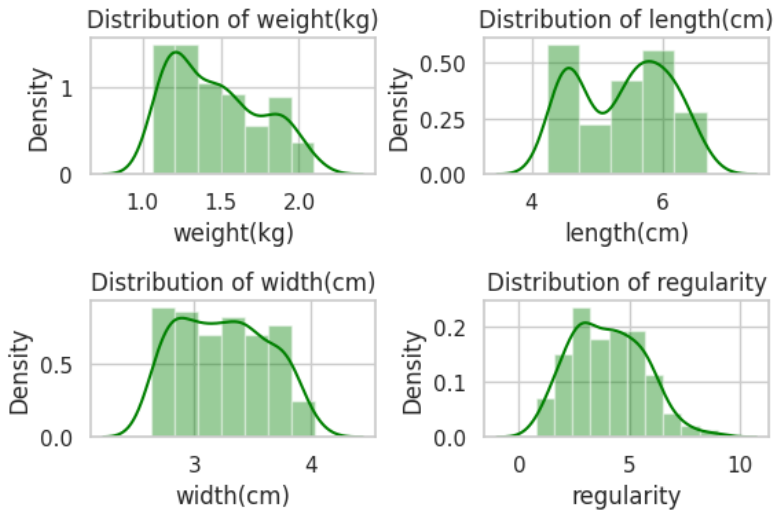

Figure 4: *Graphical representation by distplot with kde of the distribution of the numerical variables with the aim of studying them in order to choose a better imputation strategy.*

It has been decided to impute the missing values using the mean value of each variable since, as can be seen in the Figure 4, these variables maintain to a greater or lesser extent a normal distribution so that imputation with the mean value is an appropriate strategy. In this context, the mean is considered a suitable value to replace missing values.

The choice of the mean is based on its ability to preserve the central tendency of the data, which helps to maintain the statistical consistency of the data set.

3.1.7 Alternatives to the use of averaging

Furthermore, as these are fruit characteristics, it is assumed that missing values should not be replaced by zero or other fixed or random values.

- Replacement by zero: This strategy could introduce significant biases in the analysis, as assigning a null value to fruit characteristics could distort the interpretation of the data as characteristics such as "Lenght", "width" and "weight" cannot have a value of 0. For example, if "Lenght" were replaced by zero, information on fruit length would be lost, which could be critical in further analysis.

- Replacement by a fixed value other than the mean: Using a fixed value (e.g. the minimum or maximum of the column) can generate distortions similar to the zero strategy. In the case of fruit length, if replaced by the maximum value, the variability of this characteristic would be lost.

- Substitution by random values: Random value assignment was another strategy considered since, as mentioned above, missing values are not very abundant, but still the mean was chosen since the values follow a normal distribution and substitution by random values can be problematic, as it introduces artificial variability in the data and hinders the reproducibility of the results.

In summary, the choice of the mean as an imputation strategy is based on its ability to maintain statistical consistency of the data and preserve valuable information on fruit characteristics. The alternative strategies mentioned above may introduce biases and reduce the usefulness of the data in subsequent analyses.

3.1.8 Ordinal Enconding of Categorical Variables

In the preprocessing of the data, an ordinal encoding has been implemented in the "*cleft*" column by creating a dictionary that assigns categories to ordinal numeric values. This approach has been adopted to represent cleft size information on an ordinal scale, assigning numerical values sequentially to the categories "*Very Small*", "*Small*", "*Average*", "*Large*", and "*Very Large*", respectively.

Before	After
Very small	1
Small	2
Average	3
Large	4
Very Large	5

Table 5: *Table containing the new names of the columns*

The choice to use ordinal encoding rather than One Hot Encoding for this specific variable is based on the intrinsic nature of the ordinal relationship between the "*cleft*" categories. Ordinal encoding assigns values that reflect the inherent order of the categories, capturing the hierarchical relationship between them. This approach is particularly appropriate when there is significant ordering between categories, such as in the case of "*cleft*", where "*Very Small*" is inherently smaller than "*Small*", and so on.
As for One Hot Encoding, it transforms categorical variables into binary vectors, creating a binary column for each category so that in this type of data that presents an ordinal relationship, such as in the case of "*cleft*" ("*Very Small*" to "*Very Large*"), One Hot Encoding would not capture the natural hierarchy between the categories, resulting in a redundant and dimensionally larger representation, which is not efficient.

3.1.9 Normalization

Since Ordinal Encoding has already been applied in the "cleft" column, meaning that all features are numerical, in this context, it appears that the features are on a consistent scale and would not pose a specific problem related to the scale of the features when applying a machine learning model. However, it is always advisable to evaluate the scale and distribution of the features and consider how they might influence the model before making a final decision.

Two approaches to data processing, standardisation (Standardisation or Z-score normalisation) and normalisation (Min-Max scaling), have been considered in order to prepare the data for further analysis. After careful analysis, normalisation has been chosen as the most appropriate method.

Z-Score Normalization: Adjusts features to have a mean of 0 and a standard deviation of 1, which may be useful especially when features have different scales and distributions, and may be relevant in distance-based algorithms, K-Nearest Neighbors and support vector machines (SVMs).

Normalisation (Min-Max Scaling): Normalisation adjusts features to be in the range [0, 1], which can be beneficial when all features are required to share the same scale and not be affected by outliers. This could be really relevant in algorithms such as K-Means.

Therefore, it has been chosen to perform a normalisation by adjusting the feature values to be in a specific range [0-1] in the case of the Min-Max scale. The rationale behind the normalisation of these specific columns lies in several important considerations such as better comparability, better performance and interpretation and visualisation.

3.2 Hierarchical Clustering

Hierarchical clustering is a fundamental technique in Machine Learning. Unlike other clustering methods this technique does not require prior specification of the number of clusters (Ezugwu et al., 2022). So in this case, it is applied to the data in order to explore and understand it's structure.

3.2.1 Vary Linkage

For this purpose, Vary Linkage is used to determine the distance or similarity between clusters at each step of the dendogram construction. There are several linkage methods commonly used in hierarchical clustering. Each has its own rules for calculating the distance/similarity between clusters so the final groupings are different.

Considering the context and structure of the data, several linkage methods have been tested and the results compared. In order to obtain an optimal final grouping of the data. The methods compared were as follows:

1. Single Linkage: The distance between two clusters is defined as the shortest distance between any pair of points belonging to the two respective clusters. This tends to produce elongated clusters and can be sensitive to outliers.

2. Complete Linkage: Calculates the distance between two clusters as the longest distance between any pair of points belonging to the two clusters. This method tends to produce compact and balanced clusters.

3. Average Linkage: Calculates the distance between two clusters as the average distance of all pairs of points between the two clusters. This approach tends to produce more balanced clusters in size.

3.2.2 Distance Metrics

Secondly, it has been necessary to determine the distance metrics. These are mathematical functions used to quantify the relationship between variables according to their attributes. The choice of these metrics also depends on the nature of the data as well as the objectives of the analysis. The Euclidean distance has been used in this case as it is the one that best suits the data.

- **Euclidean Distance**: This is the best known and most widely used distance metric. It measures the linear distance between two points in Euclidean space.

- **Manhattan distance**: This distance measures the sum of the absolute distances along each dimension. It is so named because it can be seen as the distance to get from one point to another in a city where it is only possible to move horizontally and vertically.

- **Cosine distance**: This distance measures the directional similarity between two vectors. This distance is 0 if the vectors are perpendicular and 1 if their direction is identical. Additionally, the smaller the cosine distance, the more similar the vectors are.

3.2.3 Dendograms Plot and Analysis

Once all the linkage possibilities and distance metrics have been defined, it proceeds with the comparison of the groupings generated by each option.

To do so, it is necessary to generate a dendrogram for each of the possibilities. In the context of hierarchical clustering, dendrograms are fundamental visual representations that allow an intuitive visualisation of how data are related to each other at different levels of granularity. They also facilitate the identification of patterns and relationships between data, as they can reveal the existence of natural clusters or subgroups in the data.

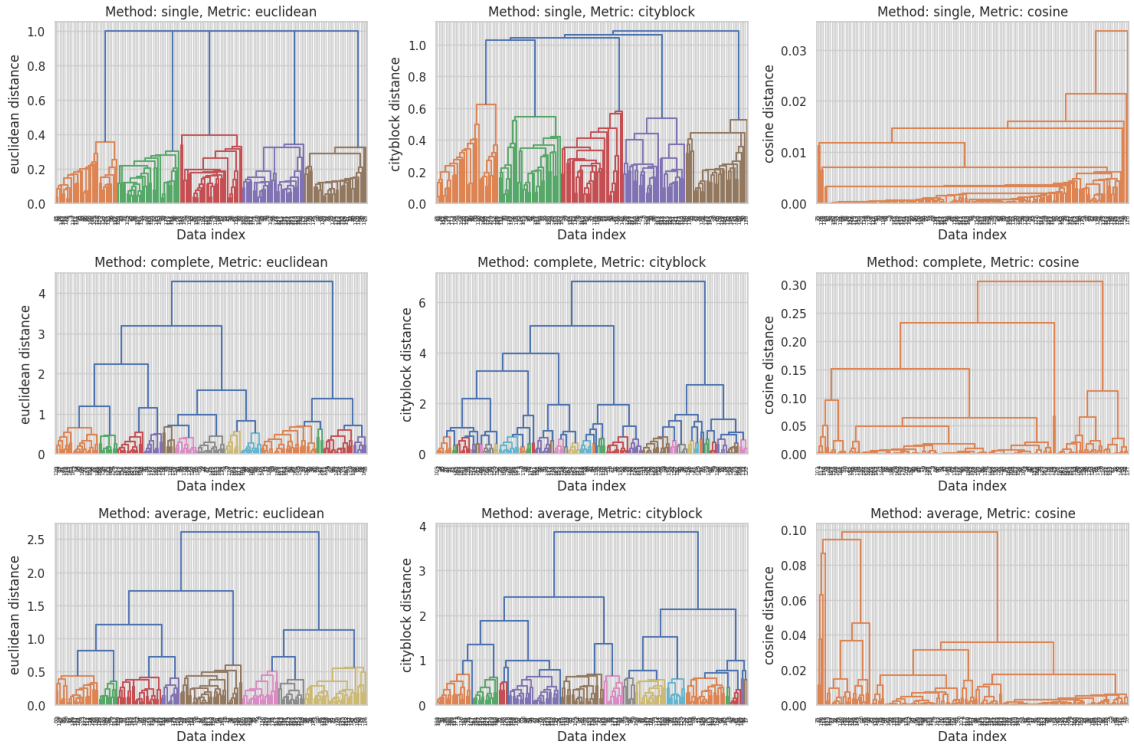

Figure 5: *Comparison of the different dendograms defined by different parameters*

By examining the dendrogram, one can ascertain the optimal number of clusters based on the desired height or levels at which the dendrogram is to be cut.

With all this, the dendrogram is used to compare and evaluate different clustering results. In these nine dendograms, Euclidean distance, Manhattan and cosine have been used. In addition, a different linkage method is used for each of the distance metrics. In order to evaluate which of these combinations generates a better clustering of the data, the cophenetic correlation coefficient is used. This coefficient assesses how closely a dendogram reflects the original distances between data points. In this case, it compares the heights with the original distances between the data points before clustering. That is, it measures how well the distances in the dendogram reflect the original distances in the feature space.

Parameters	Score
Single euclidean	0.6067
Single cityblock	0.6467
Single cosine	0.5279
Complete euclidean	0.7302
Complete cityblock	0.7858
Complete cosine	0.6460
Average euclidean	0.7750
Average cityblock	0.8000
Average cosine	0.7656

Table 6: *Cophenetic Coefficient for each parameters combination*

Best Parameters: (single, Manhattan)

With the information obtained it is possible to answer the following questions:

How does the linkage or the distance influence the result?
When selecting the linkage method, it is essential to consider the desired results and the nature of the data. In the case of the 'complete' method, it tends to generate more compact and well-separated clusters, promoting the formation of defined groups. In contrast, the 'single' method may result in more elongated or chain-like clusters, which could result in more extended and less compact clustering. On the other hand, the choice of distance metric plays a crucial role in cluster formation by defining how similarity or dissimilarity between data points is measured. The distance metric establishes the fundamental concept of proximity, influencing how clusters are configured and how data are grouped based on their characteristics. It is important to select the distance metric that best fits the inherent structure of the data and the specific objectives of the analysis.

In this specific case, the 'single' linkage method has been chosen because of its ability to generate elongated and well-defined clusters, which is consistent with the type of data it is dealing with. In addition, the distance metric 'cityblock' has been chosen because it adequately fits the nature of the data, where distance in terms of city blocks (Manhattan) can effectively capture the spatial and proximity relationships between data points. This choice aligns with the objective of obtaining clusters that accurately reflect the relevant similarities and differences in the dataset.

What seems to be a reasonable clustering?
Reasonable clustering is achieved when elements within a cluster are close to each other in terms of cityblock distance. The choice of the cityblock metric is justified by the ability of this measure to capture the spatial and proximity relationships between data points. In a reasonable clustering, clusters should be well separated and have no significant overlaps. This ensures that the boundaries between groups are distinct and that each element is clearly assigned to a specific cluster.

In addition, the number of clusters selected should be sufficient to capture the key structures and patterns of the data without generating excessive fragmentation or overly general clustering. In other words, the choice of the optimal number of clusters must balance the ability to accurately represent the complexities of the data without over-subdividing or under-clustering it. A reasonable clustering is one that effectively reveals the intrinsic relationships in the data, providing meaningful insights and a clear understanding of the underlying structure.

What are the clusters in that clustering? After examining the resulting dendogram (method='single', metric='cityblock') it is necessary to decide where to cut to obtain the desired number of clusters.

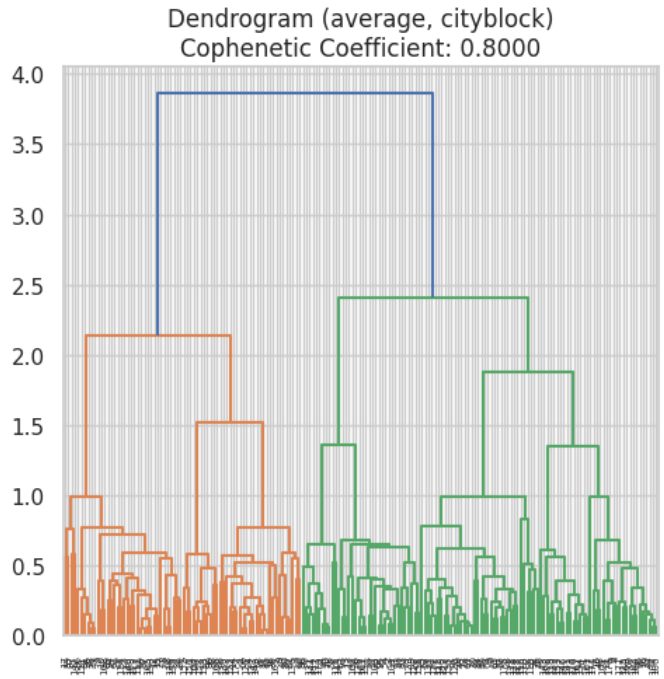

Figure 6: *Dendogram generated with the selected parameters according to the Cophenetic Coefficient.*

In this case, the number of clusters generated by this selected linkage method is 2. These clusters represent groups of data with similarities based on the Manhattan distance. Where a distance of 2.5 means that one cluster is separated from another by a distance of 2.5 units in two-dimensional space.

Once the dendogram has been visualised and 2 clusters have been clearly identified, it is possible to assign the data to these clusters. This is done using the fcluster function of the scipy.cluster.hierarchy library. Which determines the number of clusters into which you want to divide your data.

The criterion used to assign the elements to the clusters is "*maxclust*". This assigns the elements to the clusters in such a way that the maximum number of clusters equals k. Once the data has been assigned, a new column is created containing the cluster assignment for each element of the data, so that each element in "*clusters*" represents to which cluster the corresponding element in the DataFrame belongs.

3.3 Partitional Clustering

Partitional clustering is a grouping technique that divides a dataset into clusters, where each element belongs to exactly one cluster. It is used to identify patterns of similarity between data points, facilitating analysis and decision-making based on the underlying structure of the data.

3.3.1 K-means

The k-means algorithm has been used to perform a clustering analysis on the dataset. Various configurations have been evaluated, covering a range of number of clusters. Inertia, which measures the sum of the squared distances between points and their respective centroids, was used as a metric to discern the optimal number of cluster partitioning.

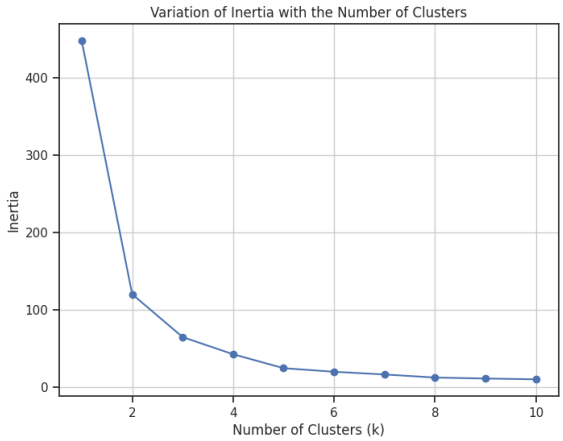

Figure 7: *Inertia variation with the Number of Clusters*

Based on the results obtained from partitional clustering using the k-means method, a substantial change in inertia is observed, depicted in Figure 7. The inertia decreases significantly until reaching an inflection point at k=2. This point suggests a possible natural structure in the data, indicating that splitting the set into two clusters has a considerable impact on reducing intra-cluster variance.
As mentioned above, a different number of clusters have been tested and in all cases a very significant decrease in inertia (decrease in slope) has been observed at k=2, suggesting that this is the most appropriate number and that adding more clusters does not provide a substantial improvement in the reduction of intra-cluster dispersion.

3.3.2 Silhoutte Scores Analysis

The Silhouette coefficient is a metric to evaluate the quality of the clusters obtained from the data set. This metric provides information on how far apart the different groups are separated according to the similarity of the points belonging to these clusters.

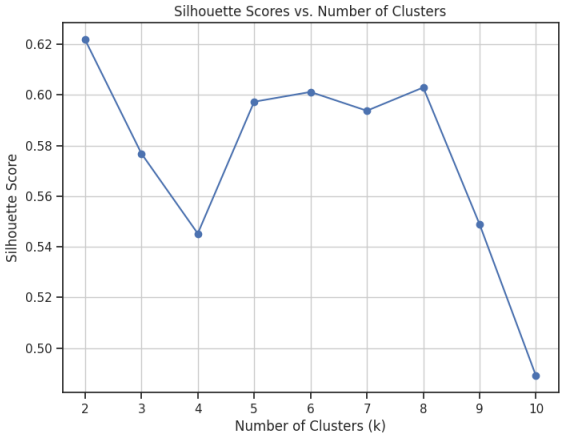

Figure 8: *Silhoutte Score for each cluster number.*

Additionally, when analysing the Figure 8, it stands out that the silhouette index, which measures intra-cluster cohesion and inter-cluster separation, reaches its maximum value also at k=2 (Silhoutte Score = 0.62). This finding reinforces the evidence for an optimal partitioning into two clusters, as the silhouette score indicates a good separation between clusters, supporting the choice of this specific number of clusters.

These results are consistent with previous observation on hierarchical clustering using average Manhattan, where it was suggested that two clusters were an appropriate choice. The convergence of results from different methods supports the robustness of the conclusion that two clusters provide a meaningful representation of the underlying structure of the data.

3.4 Alternatives

In previous sections, the coefficient of coephnet has been used to evaluate which of all the different combinations of parameters to generate the hierarchical clustering was the best. However, the Silhoutte Score allows us to evaluate the quality of the clustering, it is possible to generate these results for each of the linkage and distance combinations. The results are as follows:

Parameters	Score
Single euclidean	0.5242267577412213
Single cityblock	0.5940575734349529
Single cosine	-0.12122942283934854
Complete euclidean	0
Complete cityblock	0
Complete cosine	0.17324822998011044
Average euclidean	0
Average cityblock	0
Average cosine	0.21357091858091115

Table 7: *Silhoutte Score Evaluation*

Best Parameters: (single, cityblock)

In this case, the result is different from the validation by cophenetic correlation. In this case the combination of a 'single' linkage and a 'cityblock' distance metric generates a higher Silhoutte Score. Therefore, it is worth studying which would be the number of clusters that would best group the data according to these metrics.

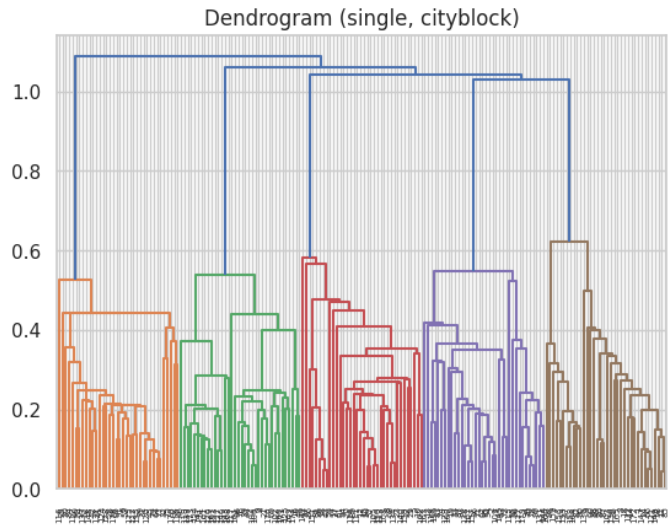

Figure 9: *Dendogram with parameters selected according to the silhouette socre (method=single, metric=manhattan)*

As seen in Figure 9, the parameters 'simple' and 'cityblock' shows that the best classification is into 5 clusters. Where a distance of 0.7 separates a cluster from another in a two-dimensional space.
In addition, the comparison can continue in partitional clustering. Since now the previous graph generated in which the Silhoutte Score is calculated for each number of clusters varies from the previous one.

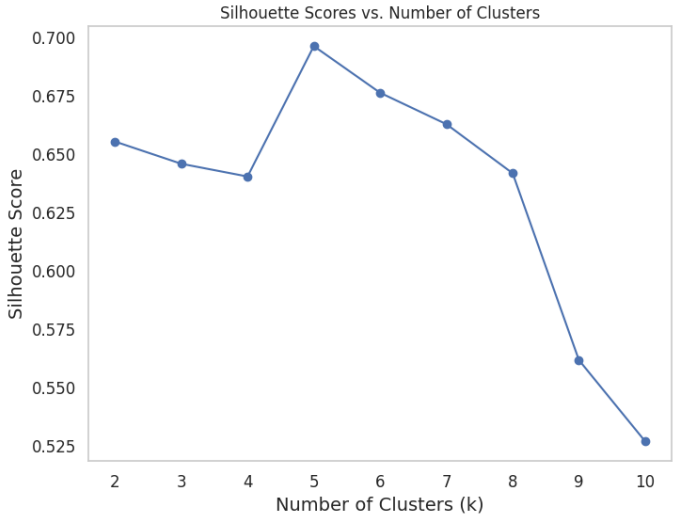

Figure 10: *Silhoutte Score variation according to the number of clusters*

When analyzing the Figure 10, it can be seen that the scores for each number of groupings have varied. There is a large increase in the value at k=5, as expected after analyzing the dendogram with this ratio of parameters.

3.5 Definitive Cluster Algorithm

In order to compare the two classification algorithms seen in the previous sections, visual representations of the grouped data are generated, so that it is possible to finally choose which one is the best suited.

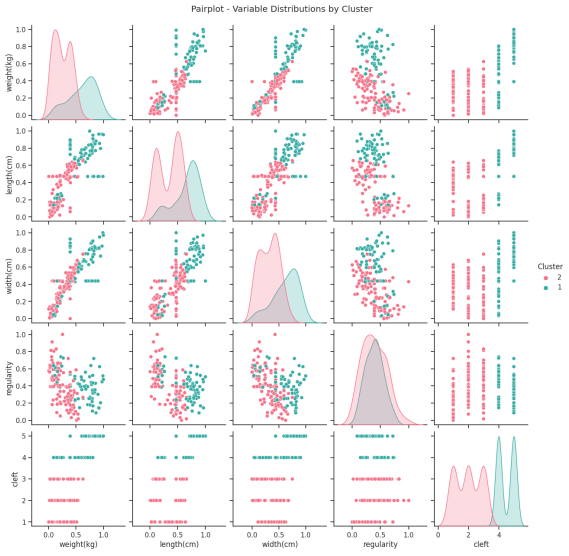

Figure 11: *Pairplot with the distribution of clusters formed (average, Manhattan)*

According to these graphical representations, it is extracted that the classification of Figure 11 is the better of the two. This corresponds to a clustering algorithm in which a linkage metric 'average' and a distance metric 'Manhattan' are selected, which generates two clusters that group the data with a Silhoutte Score of 0.62.

4 CONCLUSIONS

In the development of this study, various clustering strategies and algorithms were examined and evaluated after carrying out a pre-processing of the data.
The results obtained indicate that, depending on the nature of the data, the optimal way to cluster and classify lies in the application of the average link metric and a Manhattan distance metric. By employing these metrics, a clustering into two clusters with a silhouette score of 0.62 is achieved. This value indicates an adequate separation of the data.

- The clustering analysis applied to the fruit dataset, using a combination of hierarchical clustering followed by partitional clustering using k-means and inertia assessment, supported by silhouette analysis, has consistently reaffirmed the results obtained.
- The strategy implemented in the data pre-processing has been carried out considering the most relevant features in accordance with the nature of the data. However, exploring other pre-processing strategies (imputation and normalization) could be effective depending on the desired outcome.

This comprehensive approach has provided a robust validation of the clustering structure identified in the fruit dataset, highlighting the effectiveness of the methodology used in this study.

References

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. https://doi.org/10.1016/j.engappai.2022.104743 (cit. on pp. 1, 8).

Kang, M., & Tian, J. (2018, August). Machine learning: Data pre-processing. https://doi.org/10.1002/9781119515326.ch5 (cit. on p. 1).

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020 (cit. on p. 1).

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., da F. Costa, L., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach (H. A. Kestler, Ed.). *PLOS ONE*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236 (cit. on p. 1).

Tiu, E. S. K., Huang, Y. F., Ng, J. L., AlDahoul, N., Ahmed, A. N., & Elshafie, A. (2021). An evaluation of various data pre-processing techniques with machine learning models for water level prediction. *Natural Hazards*, *110*(1), 121–153. https://doi.org/10.1007/s11069-021-04939-8 (cit. on p. 1).