LAB 1: CLUSTERING

ALEJANDRO SILVA RODRÍGUEZ
MARTA CUEVAS RODRÍGUEZ

APRENDIZAGE COMPUTACIONAL

UNIVERSIDAD DE MÁLAGA

OCTUBRE 2024

INDEX

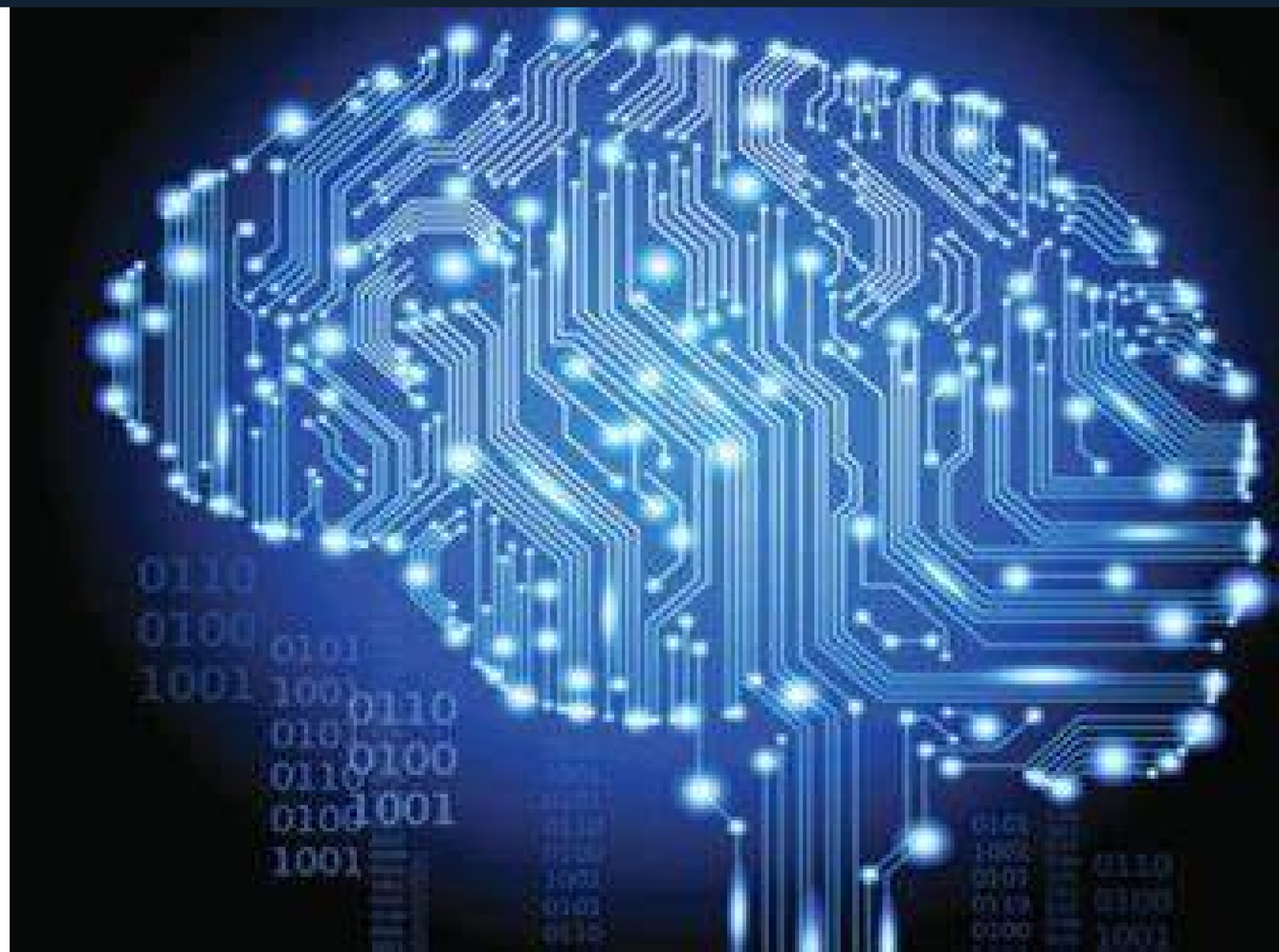
1– INTRODUCTION

INRODUCTION

- Yeast sporulation is a key biological process, crucial for cell differentiation and survival.

- Gene clustering based on expression profiles helps to understand groupings of similar gene expression patterns during sporulation.

- In this work, we use k-Means clustering to analyze yeast sporulation data.



2 – OBJETIVES

OBJETIVES

EVALUATE THE PERFORMANCE OF K-MEANS IN CLUSTERING YEAST GENE EXPRESSION DATA.

USE DIFERENT INTERNAL VALIDATON METHODS TO DETERMINE THE NUMBER OF CLUSTERS

COMPARE THE RESULTS WITH OTHER CLUSTERING METHODS, PARTICULARLY FROM DATTA AND DATTA (2003).

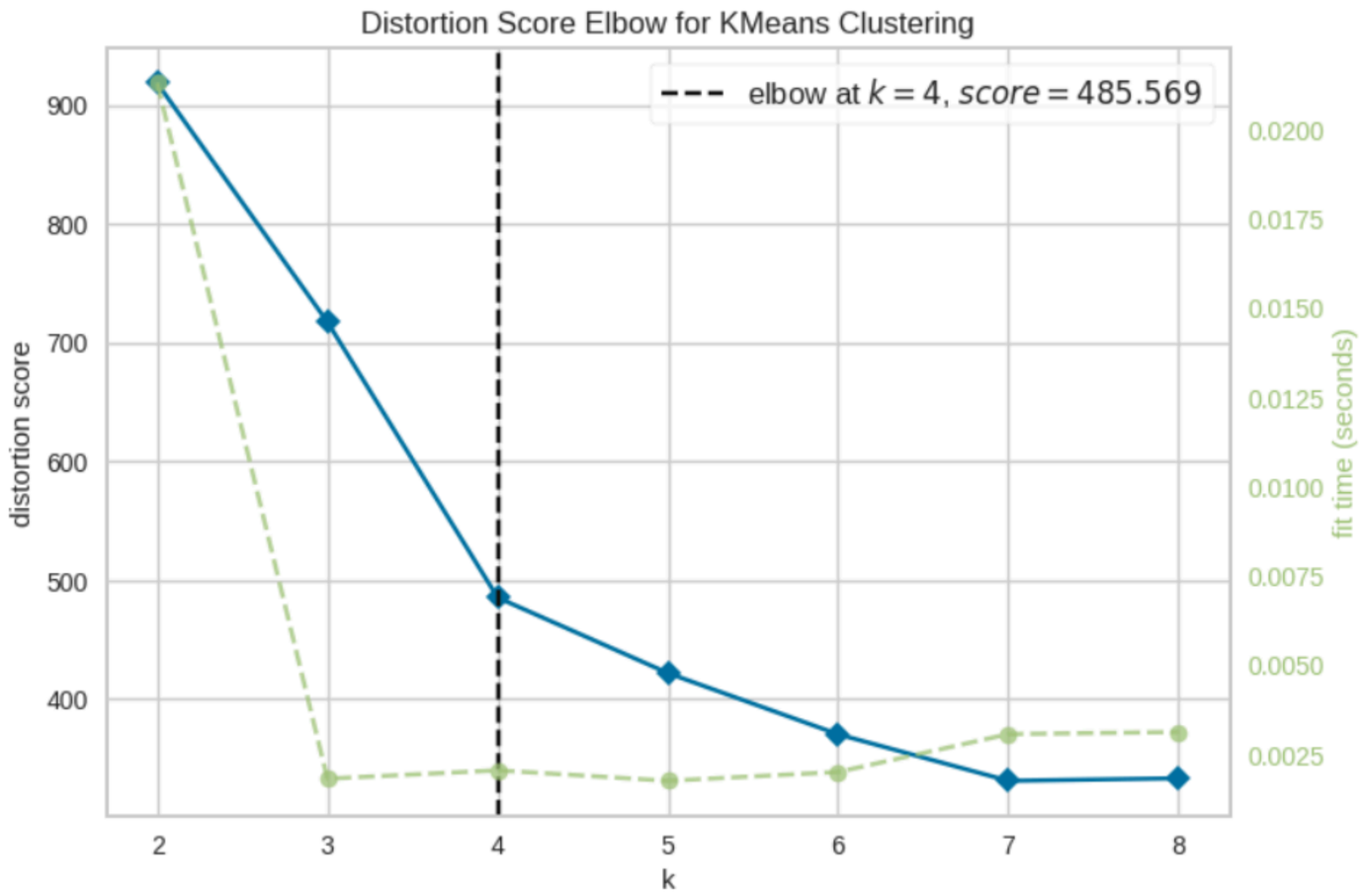
3 – METHODOLOGY AND RESULTS

DATA MANIPULATION

- IMPORT THE DATA TO THE WORKSPACE
- DELETE THE LAST ROWS THAT CONTAINS THE MEAN AND STANDARD DEVIATION
- SEPARATE THE GENES NAME COLUMN AND THE REST OF THE DATA
- NORMALIZE THE DATA USING Z-SCORE

ELBOW METHOD

Now we employ the Elbow Method to identify the optimal number of clusters, k, for the k-Means algorithm.
The Elbow Method is a well-established technique used to determine the appropriate number of clusters by
evaluating the within-cluster sum of squares (also known as inertia) for a range of cluster values.



PCA AND MDS

Principal Component Analysis (PCA)

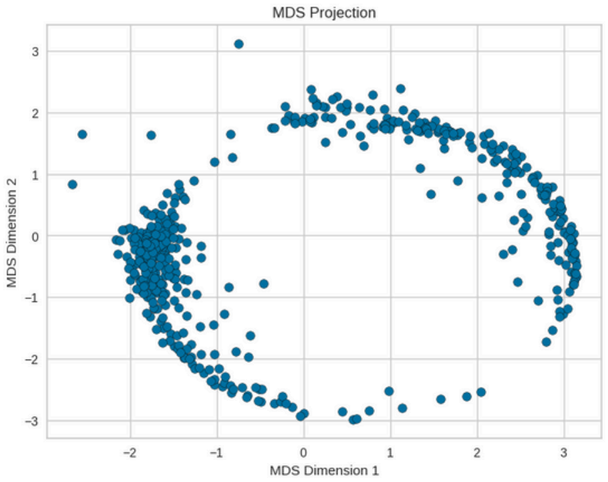
- Reduces high-dimensional gene expression data to 2D.
- Captures significant variance by identifying principal components.

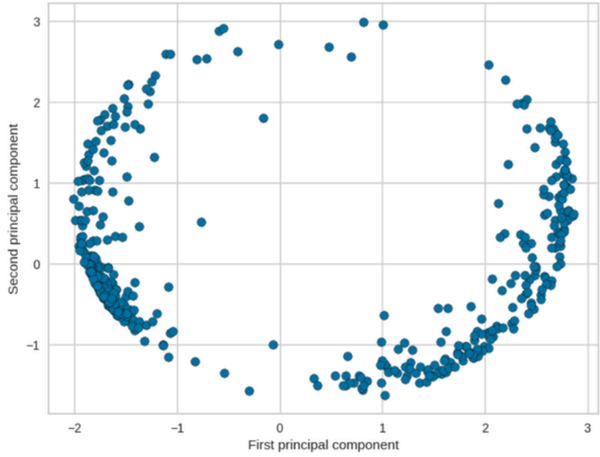
Multidimensional Scaling (MDS)

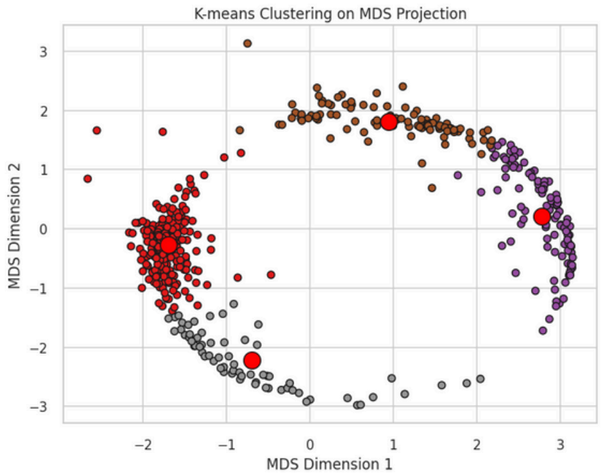
- Aims to preserve data structure based on dissimilarities.
- Minimizes differences between original and reduced distances, useful for clustering.

Visualization

- PCA and MDS provide scatter plots that reveal potential clusters and relationships in gene expression during sporulation.
- K-means clustering identifies k=4 clusters based on MDS-reduced dimensions.

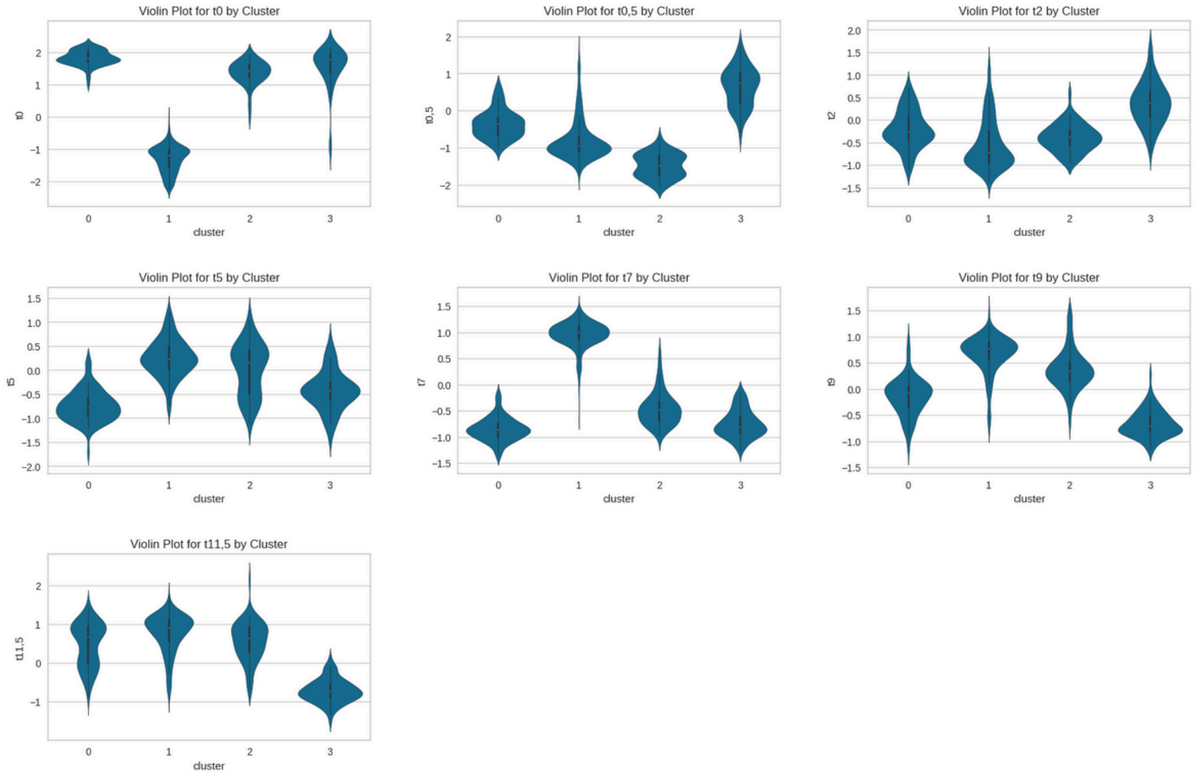






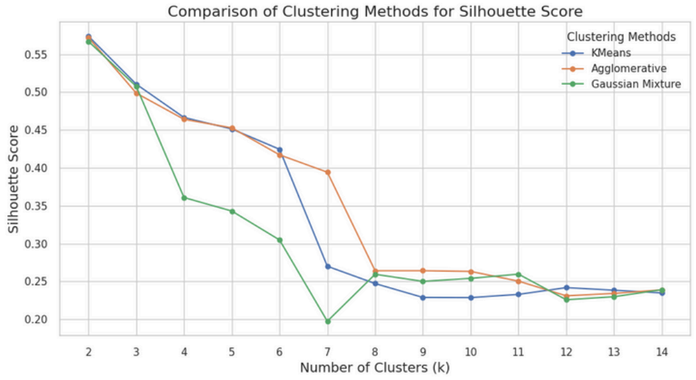
VIOLIN PLOTS

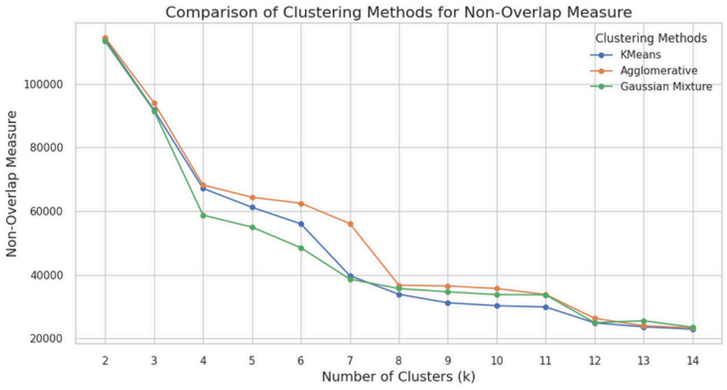
Violin plots effectively illustrate the distribution of each variable, enabling us to observe variations in
central tendency and spread across clusters. This visualization is particularly valuable for interpreting how
the characteristics of gene expression profiles differ in relation to the identified clusters over time.

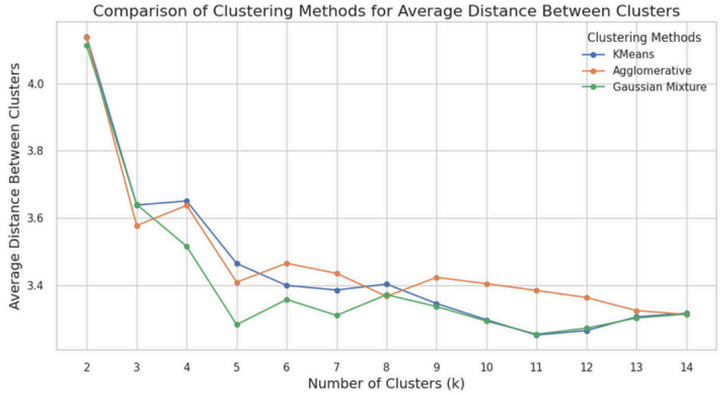


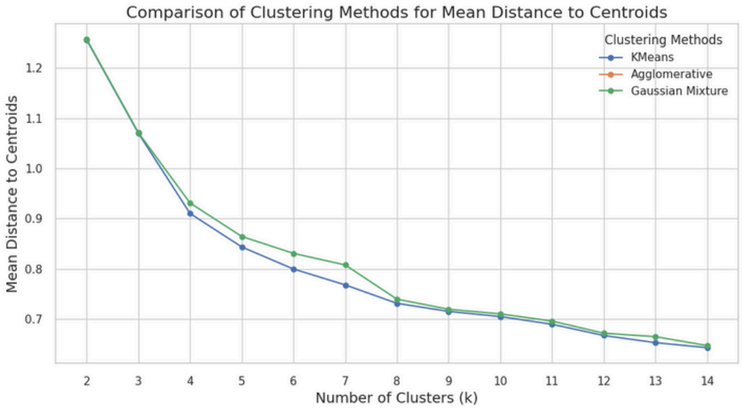
CLUSTERING COMPARISON

In this section, we compare various clustering methods (k-means, agglomerative clustering, and Gaussian mixture models) along with different numbers of clusters. This comparison employs multiple validation techniques to provide a comprehensive understanding of how well these clustering methods perform on the dataset.



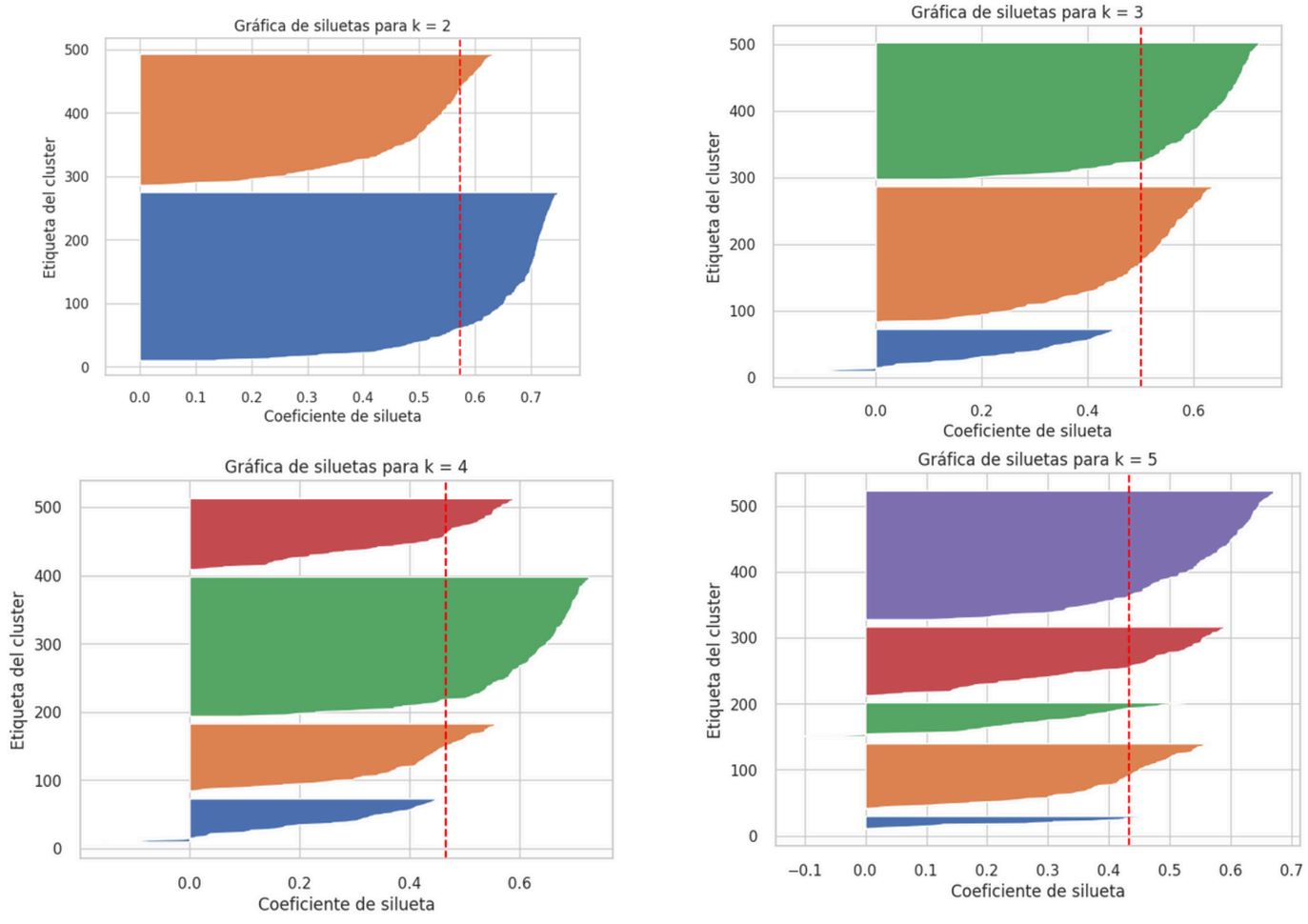






SILHOUETTES ANALYSIS

- Objective: Apply the silhouette method to evaluate clustering results due to inconsistencies with literature findings.
- Evaluation Criteria: This method measures:
  - Cohesion: Closeness of data points within the same cluster.
  - Separation: Distinction between different clusters.
- Goal: Gain insights into cluster separation and validate our results against those in the original study.



3 – CONCLUSIONS

CONCLUSION

OPTIMAL K VALUE: OUR ANALYSIS INDICATES THAT THE OPTIMAL VALUE OF K IS K=2, CONTRASTING WITH PREVIOUS STUDIES THAT SUGGEST K=7.

POTENTIAL REASONS FOR DISCREPANCY: DIFFERENCES MAY STEM FROM VARIATIONS IN DATASETS, EXPERIMENTAL CONDITIONS, AND RELIANCE ON EXTERNAL DATA IN OLDER STUDIES, WHICH MAY NOT REFLECT CURRENT CLUSTERING METHODOLOGIES.

SUPPORTING EVIDENCE: OUR FINDINGS ARE REINFORCED BY:

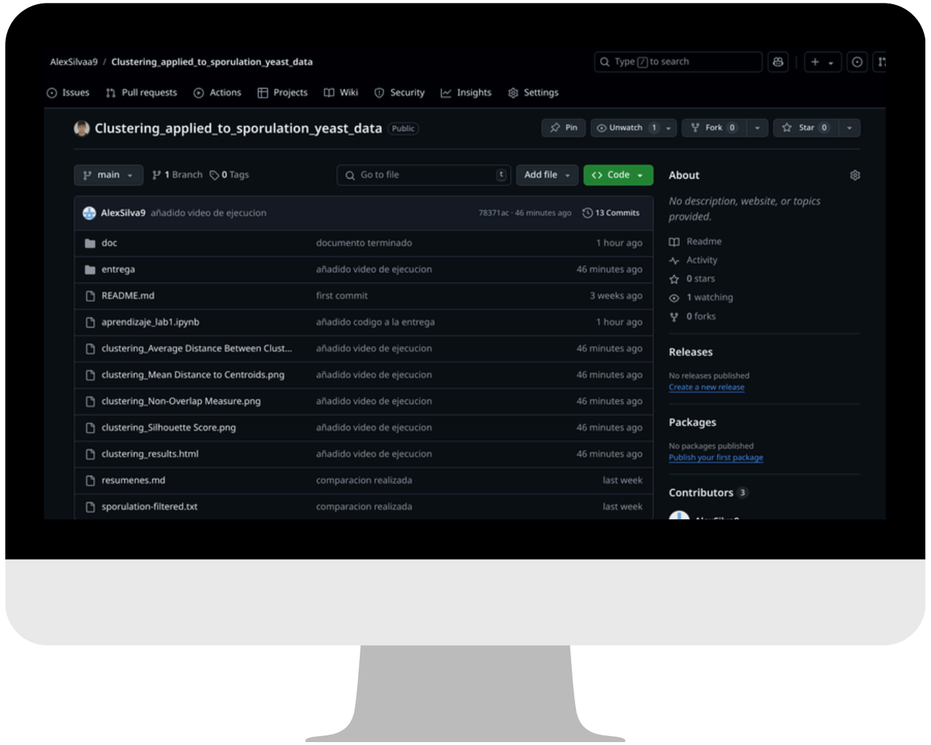
- PRINCIPAL COMPONENT ANALYSIS (PCA) AND MULTIDIMENSIONAL SCALING (MDS) SHOWING A NATURAL GROUPING AROUND TWO CLUSTERS.
- CLUSTERING QUALITY METRICS, INCLUDING SILHOUETTE ANALYSIS, CONSISTENTLY SUPPORTING K=2.

3 – REPOSITORY ACCESS

REPOSITORY ACCESS

- ALL ADDITIONAL INFORMATION, INCLUDING SOURCE CODE AND FULL DOCUMENTATION, IS AVAILABLE IN THE GITHUB REPOSITORY:

https://github.com/AlexSilvaa9/Clustering_applied_to_sporulation_yeast_data



THANK YOU