Exploring Genetic Patterns in Cancer Transcriptomes

An Unsupervised Learning Approach

Eloísa Toledo Iglesias

MU Bioinf. i Bioest.
Bioinformática Estadística y Aprendizaje Automático

Supervisor:
Romina Astrid Rebrij

Subject coordinator:
Carles Ventura Royo

16/01/2024

Universitat Oberta de Catalunya

UNIVERSITAT DE BARCELONA

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Exploring Genetic Patterns in Cancer Transcriptomes: An Unsupervised Learning Approach*
Nombre del autor:	*Eloísa Toledo Iglesias*
Nombre del consultor/a:	*Romina Astrid Rebrij*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	*01/2024*
Titulación o programa:	Máster universitario en bioinformática y bioestadística
Área del Trabajo Final:	*Bioinformática Estadística y Aprendizaje Automático*
Idioma del trabajo:	*Inglés*
Palabras clave	*Cancer Transcriptome, unsupervised learning*

Resumen del Trabajo

El cáncer es una enfermedad compleja y heterogénea que representa un importante problema de salud pública mundial debido a su creciente número de víctimas y a la ausencia de tratamientos eficaces. Se ha demostrado que las mutaciones y modificaciones del ARN desempeñan un papel crucial en el desarrollo y la progresión de los tumores. En este contexto, el estudio molecular de la biología del cáncer es de suma importancia, debido a su relevancia en la clasificación y comparación de múltiples tipos y subtipos de cáncer, permitiendo el desarrollo de terapias personalizadas y aumentando el éxito del tratamiento. Sin embargo, aunque las tecnologías RNA-seq, como la secuenciación Illumina Hiseq, han revolucionado la investigación médica, implican el análisis de cantidades extensas de complejos datos. Las técnicas de aprendizaje automático no supervisado pueden ser de gran ayuda para crear nuevas clasificaciones del cáncer, superando las limitaciones de las técnicas tradicionales.

En este trabajo, se probaron diferentes enfoques de reducción de la dimensionalidad, como PCA y UMAP, y varios algoritmos no supervisados, incluidos algoritmos de partición, basados en la densidad, jerárquicos y basados en modelos, con el fin de identificar tipos y/o subtipos de cáncer según su expresión génica. Varios algoritmos, como k-means, PAM, CLARA y algoritmos jerárquicos aglomerativos utilizando la técnica UMAP para la reducción dimensional, demostraron la capacidad de clasificar los datos de expresión génica con un alto grado de precisión formando grupos bien separados. Estos resultados confirman el potencial de estos algoritmos para contribuir a la lucha contra el cáncer.

Abstract

Cancer is a complex and heterogeneous disease that represents a major global public health concern due to its escalating casualty rates and the absence of effective treatments. RNA mutations and modifications have been shown to play a crucial role in the development and progression of tumors. In this context, the molecular study of the cancer biology is of paramount importance, due to its relevance in classifying and comparing multiple cancer types and subtypes, allowing the development of more personalized therapies and increasing the treatment success. However, although RNA-seq technologies, such as Illumina Hiseq sequencing, have revolutionized medical research, they involve the analysis of complex and extensive amounts of data. Unsupervised machine learning techniques can be of unparalleled help in creating novel cancer classifications, surpassing the limitations of traditional techniques.

In the present work, different dimensionality reduction approaches, such as PCA and UMAP, and several unsupervised algorithms, including partitioning, density based, hierarchical and model based algorithms, were tested in order to identify types and/or subtypes of cancer according to their gene expression. Several algorithms, namely, k-means, PAM, CLARA and agglomerative hierarchical algorithms using the UMAP technique for dimensional reduction, demonstrated the ability to classify gene expression data with a high degree of accuracy forming well separated clusters. These results confirm the potential of these algorithms to contribute to the fight against cancer.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Context and justification of the work

According to the 2020 World Cancer Report, cancer ranks as the second most prevalent cause of mortality on a global scale, responsible for approximately 9.6 million fatalities in the year 2018 [1]. Lung, breast, colorectal and prostate cancer are amongst the most common cancer types in the world, all of them being expected to double their incidence by 2070 [1,2]
RNA mutations and modifications have recently proven to have a key role in tumorigenesis, and tumor growth and progression, leading to the appearance of subtypes of cancers [3]. Understanding the underlying biology of cancer and the common and differential characteristics among cancers and cancer subtypes, can lead to the identification of targets for new therapies, more specific treatments for patients, as well as finding biomarkers which could enable a rapid and accurate detection of the type of tumour each individual is afflicted with [1].
However, the analysis of RNA-seq data usually involves high amounts of complex and multidimensional data which can be difficult to extract conclusions from [4]. Thus, machine learning techniques can be employed to carry out a pan-cancer analysis, in order to find relevant mutation or mutation clusters capable of distinguishing between cancer types or potentially unveiling new disease subtypes, using RNA-seq data of cancer patients. Therefore, this tool could represent a significant milestone in improving the prognosis of cancer patients.

1.2. Objectives of the work

- **General objectives:**
  1. Develop an unsupervised machine learning methodology that enables to identify RNA expression patterns compatible with different cancer types and/or subtypes.

- **Specific objectives:**

  1. Compare different techniques of feature selection genes.
  2. Identify the optimal clustering solution for the dataset.
  3. Compare the obtained clustering with the cancer classification found in the literature.

1.3. Impact on sustainability, ethical-social aspects, and diversity

The present work focuses on the development of an unsupervised machine learning model which leads to an accurate clustering of the RNA expression data in cancer patients through a pan-cancer approach. This goal involves the pre-processing and processing of data, where it is important to be aware of the environmental and sustainability impact that this may entail. Hence, although this process was carried out by a personal computer and that the amount of data in

study is not significant compared to the amount of data that is being processed daily worldwide, data management and energy efficiency was taken into consideration at every step of the project. Moreover, one of the main objectives of this project is to establish an efficient clustering method and, consequently, of data processing, which can reduce the time required for data managing, and thus, positively improve the energy impact.

In terms of ethical and social aspects, negative impacts are not directly related to this work. However, it is important to emphasise that any step into automatization or the adoption of new technologies may have an impact on the way data is managed, changing the nature of some current jobs, and therefore, being important that society is ready to adapt to these changes. Furthermore, the technical approach of this work does not have direct implications regarding gender, diversity or human rights, due to the anonymity of the data presented and that only RNA expression is taken into account to.

In this context, it is important to mention that this work is helping to move a step closer to meet up with the UN sustainable Development 2030 Agenda which proposes to reduce the total premature mortality from noncommunicable diseases by one third by 2030 [1]. A good understanding of the cancer biology can help to identify new targets for effective treatments or an early diagnosis, among others. Additionally, concerning the Sustainable Development Goals, this work is contributing mainly to the Goal 3 (good health and well-being), and more indirectly to the Goal 9 (industry, innovation and infrastructure) by providing advanced technology for the biomedical research.

1.4. Approach and methodology

This work used the data from the UCI Machine Learning Repository [5] which was obtained from The Cancer Genome Atlas (TGCA) Pan-cancer analysis project [6], analysed in the paper “Identification of common and dissimilar biomarkers for different cancer types from gene expressions of RNA-sequencing data” by Venkataramana, Lokeswari et al. (2020) [7]. It consists of a collection of RNA-Seq gene expression levels measured by Illumina HiSeq platform, including th data from 801 cancer patients suffering from five different types of cancer, and 20531 genes. In contrast to Venkataramana, Lokeswari et al. (2020), whose main goal is to create a supervised machine learning algorithm to predict cancer classification, this work used clustering models to uncover and identify the connections between RNA expression patterns and cancer type and subtypes among these patients, rather than predicting the specific disease type for each patient. Thus, this work is focused on the development of an unsupervised machine learning model for creating accurate clusters for the RNA expression data.

To carry out the project, the following strategy was followed:

- Identifying the relevant data: since 20531 are considered in this study, it is important to remove the irrelevant and redundant data. For this porpoise, different methods were used to ensure a good selection of the data, such

as evaluating the genes with more variability in their expression and recursive feature elimination (RFE) techniques.

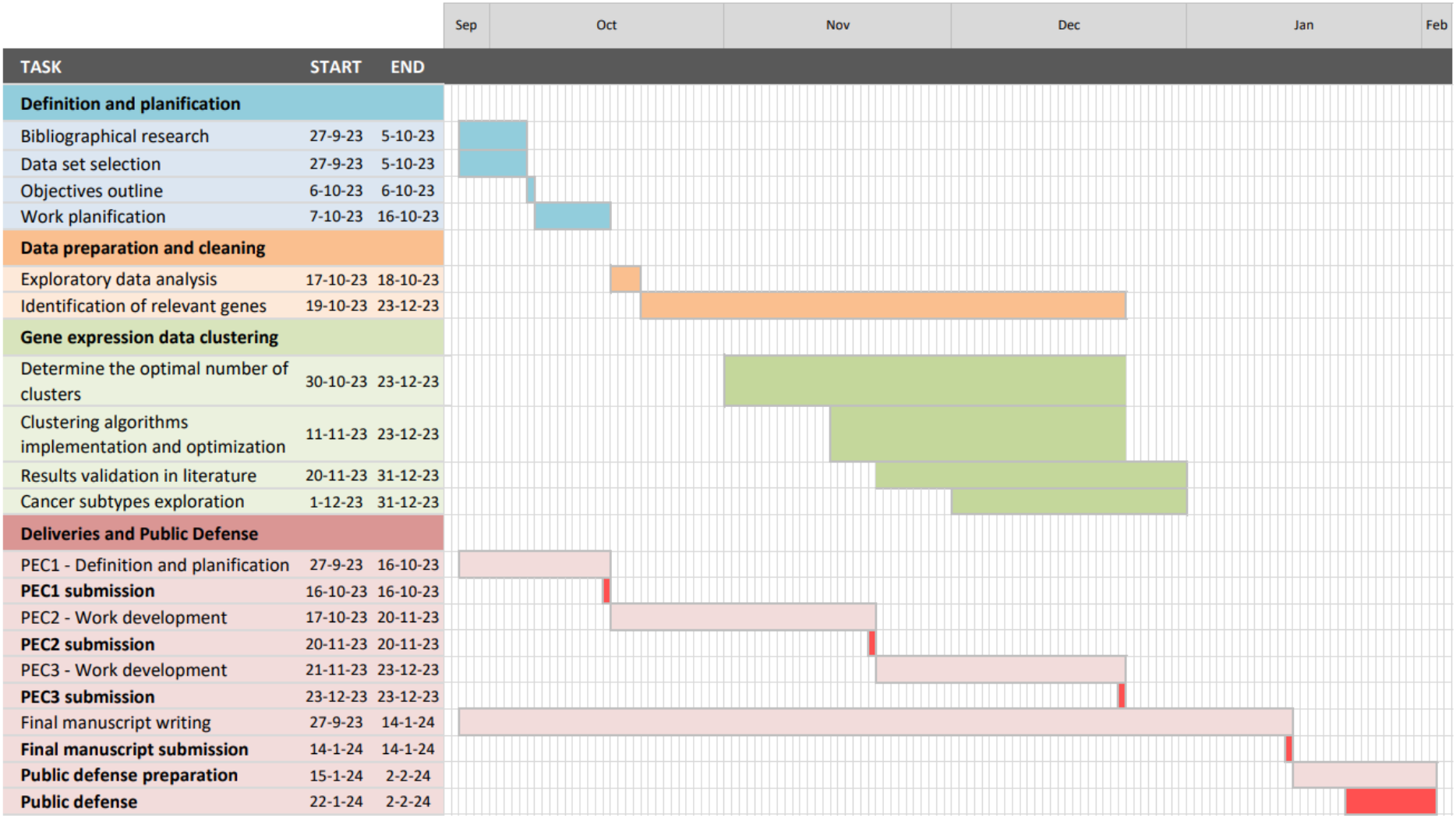
- Determining the optimal number of clusters: different methods were performed to determine the optimal number of clusters for the model, including the elbow method, the silhouette method or the gap statistic method.
- Apply different types of clustering algorithms: depending on the nature of the data, different types of algorithms may be ideal to cluster the data, thus, a number of algorithms were tested to determine the best option, such as k-Means, Partitioning Around Medoids, Clustering for Large Applications, Density-Based Spatial Clustering of Applications with Noise, Hierarchical and Gaussian Mixture algorithms.
- Validating the results with the cancer classifications described in the literature.

1.5. Work planification

The work is divided into four principal tasks, which count with sub-tasks, that together allowed the completion and submission of all the PECs and deliveries. These tasks consist of:

- Task 1. Definition and planification.
    - 1.1. Bibliographical research
    - 1.2. Data set selection
    - 1.3. Objectives outline.
    - 1.4. Work planification
- Task 2. Data preparation and cleaning.
    - 1.1. Exploratory data analysis
    - 1.2. Identification of relevant genes
- Task 3. Gene expression data clustering.
    - 3.1. Determine the optimal number of clusters
    - 3.2. Clustering algorithms implementation and optimization
    - 3.3. Results validation in the available literature
    - 3.4. Exploration of potential cancer subtypes according to the clusters obtained.

The sub-tasks and the schedule for starting, completing and delivering each one of them, are shown in the calendar below.



TASK	START	END
Definition and planification		
Bibliographical research	27-9-23	5-10-23
Data set selection	27-9-23	5-10-23
Objectives outline	6-10-23	6-10-23
Work planification	7-10-23	16-10-23
Data preparation and cleaning		
Exploratory data analysis	17-10-23	18-10-23
Identification of relevant genes	19-10-23	23-12-23
Gene expression data clustering		
Determine the optimal number of clusters	30-10-23	23-12-23
Clustering algorithms implementation and optimization	11-11-23	23-12-23
Results validation in literature	20-11-23	31-12-23
Cancer subtypes exploration	1-12-23	31-12-23
Deliveries and Public Defense		
PEC1 - Definition and planification	27-9-23	16-10-23
PEC1 submission	16-10-23	16-10-23
PEC2 - Work development	17-10-23	20-11-23
PEC2 submission	20-11-23	20-11-23
PEC3 - Work development	21-11-23	23-12-23
PEC3 submission	23-12-23	23-12-23
Final manuscript writing	27-9-23	14-1-24
Final manuscript submission	14-1-24	14-1-24
Public defense preparation	15-1-24	2-2-24
Public defense	22-1-24	2-2-24

Figure 1. Gantt diagram of the work tasks and subtasks.

Hence, the milestones to be achieved in each one of the PECs were:

- PEC1 (due date 16/10/2023): after having selected a line of work and a database, the main and specific objectives of the work outline and a work planification schedule were detailed.
- PEC 2 (due date 20/11/2023): for the first phase of the work development, data exploration was carried out. Furthermore, the preparation and cleaning were started, as well as the determination of the optimal number of clusters, the implementation and optimisation of clustering algorithms will be started.
- PEC 3 (due date 23/12/2023): all the tasks started in the previous PEC were concluded. Moreover, the final clustering algorithm was selected, and the clusters were compared and validated with the data labels reported in the literature. The presence of cancer types and/or subtypes was also assessed.
- Final manuscript (due date 14/01/2024): a final manuscript containing the chapters briefly described in section 1.7, was delivered.
- Public defence (due date 22/01 – 02/02/2024): a PowerPoint presentation and its corresponding speech will be prepared.

During the realisation of this work, there were different risks to consider:

- There was a risk that the selected dataset may contain errors or noisy data, which could impact the accuracy of the results. Selecting the relevant data was an important part of the process to try to mitigate this risk. However, the process of selecting relevant genes also involved potential errors that may affect the data's ability to distinguish mutation patterns in different cancer types and subtypes. Different types of tests were used to select the relevant data in order to ensure a good selection.
- The choice of clustering algorithms is crucial, and selecting suboptimal ones may result in lower-quality outcomes. It's important to ensure that the chosen algorithms are the most suitable for the project's objectives, so different algorithms were tested, comparing their results and selecting the optimal algorithms for this purpose.
- Since a significant part of this project involves result interpretation, there is a risk of errors or misinterpretations of the clusters and their association with the existing literature. Proper validation throughout objective indices found in the literature were carried out to avoid this problem.
- The tasks outlined in this project were suggested to encounter unexpected challenges, potentially causing delays in the work plan. Adjustments to the project's timelines were necessary to accommodate the challenge of selecting the best method of selecting the relevant genes.

1.6. Summary of the obtained products

- Final manuscript: the final manuscript consists of a document containing all the work developed during the Master’s thesis, including:
    - A contextualization of the main topic of this work (cancer disease and how a pan-cancer approach to study the effect on the transcriptome of patients with different types of cancer can help

to understand more about these diseases and potentially find new cancer subtypes).
    - The materials and methods followed to develop the work.
    - The results obtained, their discussion and comparison with the available literature, the conclusions, and future perspectives of the work.
- Final presentation: a final PowerPoint presentation will be made in order to summarize, describe and justify the work developed during this project.

1.7. Brief description of the chapters in the manuscript

- State of art: this chapter will provide an overview of the context in which the work is situated (cancer, pan-cancer, unsupervised machine learning, etc), its significance, and the hypotheses formulated.
- Materials and methods: a description of the materials (software, database) used throughout the work will be detailed in this section, as well as the methods for defining the relevant data, clustering number selection and clustering algorithms, among others.
- Results: the results obtained in the different tasks of this project will be thoroughly presented, described
- Discussion: interpretation of the results and their comparison with the literature.
- Conclusion and future perspectives: this chapter will provide a summary of the key findings obtained, the implications and contributions that these findings may have, an evaluation of the success in the accomplishment of the objectives proposed and outline future research directions.

2. State of art

2.1. Cancer

Cancer is a disease in which abnormal cells divide uncontrollably and potentially spread to other tissues and organs [8]. The growing number of cancer patients represents a major concern for public health worldwide, since it is one of the most common causes of premature mortality, expecting to duplicate its already worrisome incidence by 2070 [1,2].

Cancer can be caused by various factors, such as genetic inheritance (5-10% of cancers) and environmental factors (90-95% of cancers), including chemicals, food, pollutants, radiations, among others. The multistep process by which a stem cell becomes an abnormal cell is called carcinogenesis [8]. This process requires a combination of different mutations involving the activation of protooncogenes (cell cycle related genes with important functions in proliferation regulation) and the inhibition of tumor suppressor genes, leading to a change of the normal balance between apoptosis and proliferation [8,9]. Therefore, the genetic complexity implied in cancer apparition and progression difficult the treatment and management of this multifactorial disorder [9].
Seeking to accelerate the understanding of the molecular basis of cancer, projects such as The Cancer Genome Atlas (TCGA) have emerged. Within this project, the TCGA Pan-Cancer analysis tackles the challenge by examining different genomic samples across a variety of cancers, regardless of their origin (organ or tissue), in order to find similarities between types or subtypes of cancer. The molecular similarities between cancer types and subtypes can help to uncover the underlying biology of less studied cancers when compared to more extensively researched ones, as well as helping to select treatments based on successful cases of similar cancers [10].

2.2. RNA

The transcriptome is the set of all transcripts produced in one or a population of cells in a particular moment, including ribosomal RNAs, messenger RNAs, transfer RNAs and regulatory noncoding RNAs [11]. Transcriptome facilitates the study of disorders in comparison with the use of the genome since the transcriptome is smaller and the impact of its modification or mutations is more likely to have a direct impact on the expression and function [12]. Alternative splicing, gene fusion, RNA editing, or nucleotide variation can have a crucial role in the development of cancer, its expansion, progression and differentiation in subtypes of cancer, promoting its diversity and complicating its management and study [3,12]. Thus, the study of RNA expression is of paramount importance for understanding cancer biology leading to the discovery of new targets for preventing, controlling and curing this disease as well as new effective therapies [1,3].

In this context, RNA-seq is a next generation sequencing (second-generation) technique (NGS) which allows to identify and quantify all the transcripts or a selection of them extracted from a sample [13]. Although RNA and DNA sequencing dates back to the late 70's with the development of Sanger

sequencing, the emergence of NGS in the last decade, has opened up a whole new field of possibilities in science [12–14]. The NGS improves the Sanger method by allowing the simultaneous sequencing of large quantities of sequences instead of performing the process for one sequence at a time, with very high throughput and much lower cost [13,14].
RNA-seq is based on the creation of a cDNA library through the reverse transcription of fragmented RNAs in the samples, facilitated by the action of reverse transcriptases. The preparation of the sequences for the sequencing process involves adding adaptors to the ends of the cDNA chains, which are necessary for binding to the flowcells. Additionally, 3' adenylation of the sequences might be included to prevent them from overlapping. Following the amplification of the sequences by PCR to amplify the signals, the sequences are added to the flowcells where sequencing is performed by incorporating fluorescent dNTPs into the single-stranded cDNAs through complementation. Thanks to the fluorescently labelled nucleotides, high-resolution images are captured, allowing the obtention of the sequence [12,13]. Currently, one of the most widely adopted sequencing systems is Illumina (Illumina HiSeq and MiSeq sequencing), commanding over 70% of the market share. This system employs PCR bridge amplification across the flow cell to generate clusters of replicated DNA fragments. Furthermore, it incorporates fluorescently labelled reversible terminators during sequencing, contributing to cost-effectiveness [13].
However, as mentioned before, the analysis of NGS technology data in general and RNA-seq in particular, often proves challenging due to the complexity and multidimensionality of the data involved [4]. Therefore, bioinformatics tools have been simultaneously developed to cope with the difficult analysis, allowing researchers to draw conclusions from the large masses of data [13].

2.3. Machine learning

Machine learning focuses on developing algorithms and models to enable computers to learn from previous experience. These algorithms allow us to uncover patterns which might be overlooked by human operators. Among the different classes of machine learning, unsupervised machine learning identifies groups within unclassified or non-labelled data, allowing the algorithm to rely its choses only by the characteristics of the data and not on predefined features [15–17]. In the context of cancer research, unsupervised machine learning can be a crucial tool for creating novel classifications, using its ability to reveal hidden types and subtypes of tumors through gene expression which surpasses the limitations of traditional morphological and histopathological classifications [18].

Unsupervised classification or clustering algorithms typically address challenges posed by high dimensional, noisy or incomplete datasets, complicating the task of selecting the most accurate algorithm for each specific situation. Fortunately, a diverse range of clustering algorithms have been developed, facilitating the process of finding a suitable algorithm for datasets by comparing the results and quality of the clusters created [19]. Clusters can be defined as groups of data samples of similar natured, selected by different classification criterion [20]. Depending on how the cluster's structures are formed, clustering algorithms are separated in different categories or techniques, some of them are: partitional, density, hierarchical and model-based [19,21].

2.3.1. Partitioning clustering

Partitioning clustering methods create k cluster, being k a predefined parameter, following an objective function depending on the algorithm used. Each point of the data set must belong to a cluster and clusters must be formed by at least one point [20].

- K-means: each point in the dataset is initially randomly assigned to one of k clusters. Subsequently, centroids for each cluster are calculated as the average of the coordinates of all the points within that cluster. The algorithm then iterates through each data point, reassigning it to the cluster which centroid is closer (typically based on Euclidean distance). After, centroids are recalculated for the updated clusters. This process is repeated until a stable configuration is reached and no further modifications are observed in the points assignment [19,20].
- Partitioning Around Medoids (PAM) or K-medoids: in this case, the iterative technique to improve the clusters remains the same, however, the clusters are formed based on medoids. Medoids are real data points, specifically, the points that have the smallest average distance to the other points within the cluster. Although this algorithm is more computationally expensive than K-means, however, it is less sensitive to outliers or noisy data [20].
- Clustering for Large Applications (CLARA): this algorithm was designed to improve the PAM algorithm, facilitating the clustering for large data sets. The data set is first divided in random samples of 40+2k points. The PAM algorithm is then applied to each sample, identifying the medoids. The best medoids are selected based on the sample with the lowest sum of distances (typically Euclidean distance). These selected medoids are then applied to the whole data set, producing the final clustering results [19,20].

2.3.2. Density-Based clustering

Density-based methods create clusters based on high-density areas in the feature space. These methods consider clusters as groups of high-density objects surrounded by regions of low-density objects [20,22]. They do not require a predefined number of clusters as input parameters and, in contrast to partitioning methods, they can create non-compact and non-spherical clusters [22,23].

An example of these clustering methods is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [20,22,23]. This algorithm is capable of finding clusters with random shapes while isolating noise from the data set under study. To achieve the objective, the algorithm initially scans the dataset, looking for points with a minimum number of neighbours (MinPts) within a selected radius (Eps). If this criterion is reached, the element is considered as the core point. Following this selection, the algorithm starts expanding the clusters by adding those points that continue meeting the density threshold. The iteration process continues until no further points are unclassified. Those points that do not have the required number of MinPts within the marked radius are considered noise [23,24].

2.3.3. Hierarchical clustering

Hierarchical methods organize the points of a data sets into a tree of clusters based on a proximity matrix (pairwise distances between data points) [21,25]. The results are represented in binary trees or dendrograms where the whole data set is shown as the root node and each data point, as leaf nodes. Furthermore, the intermediate nodes allow to visualize which points are closer or far from the one particular point and the height of the branches also gives useful information about the distance between points and between points and clusters [25].

The nested partitions can be performed using two different techniques:

- Agglomerative Hierarchical Clustering: in this case, the algorithm goes from singleton clusters (each data point is considered a separate cluster) to iteratively merge the closest clusters until a cluster including all the data points is reached [19,21,25].
- Divisive Hierarchical Clustering: the clustering is oppositely made to the agglomerative. The algorithm starts from a single cluster that is iteratively divided into smaller clusters until singleton clusters are reached [19,21,25].

2.3.4. Model-based clustering

Model-based clustering assumes that the data set under study is divided into clusters following a specific probability distribution. These algorithms attempt to find the best fit between the data and the mathematical model defining the clusters [20,21,26].

An example of model-based algorithms is the Gaussian Mixture Model. This algorithm assumes that the data set is generated from a mixture of Gaussian distributions, each characterized by different means and covariances [26,27]. Means and covariances are randomly chosen to initialize the algorithm, as well as the weight for each distribution (initially the same for all the clusters). The algorithm calculates the probability of each data point belonging to each cluster based on these parameters. Subsequently, the model's distributions are then updated, assigning each point to the distribution for which it has the highest probability of belonging. The parameters (mean, covariance, and weights) are then recalculated based on the new distributions. This iterative process is repeated until convergence is reached [26].

2.4. Related works

In the recent years, the scientific community has become increasingly aware of the importance of machine learning in the medical research. The rapid development of the new technologies has enabled the exploration of diverse approaches to apply these algorithms into different fields [18,28–30]. Several authors have explored the use of machine learning to classify gene expression data from cancer patients, with the aim of achieving high levels of classification accuracy within reasonable computational times. Both supervised and unsupervised approaches have been investigated, but no consensus has been reached regarding the best method for identifying cancer types and subtypes [31,32].

Among the supervised machine learning algorithms, Support Vector Machines, have shown promising results in terms of classification accuracy for multiclass cancer datasets and in the identification of subtypes of cancers within individual cancer datasets [32–34]. Regarding the unsupervised algorithms, de Souto M.C.P. et al. (2008) compared seven different unsupervised algorithms, including hierarchical models, k-means, Gaussian mixture, spectral clustering and shared nearest neighbour clustering. The study concludes that the k-means algorithm had the best performance across several datasets [28]. Furthermore, Perera M.A.I. et al. (2020) analysed the potential of k-means, hierarchical and PAM algorithms in the identification of types and subtypes of cancers, where k-means was also highlighted [18]. However, only a few studies have been found comparing different unsupervised machine learning methods in cancer gene expression data. Other works have demonstrated the good performance of individual unsupervised algorithms mainly in breast cancer, such as hierarchical algorithms [35] and DBSCAN and k-means algorithms [36] for identifying patterns and clustering gene expression datasets.

Therefore, the potential of unsupervised machine learning algorithms in cancer research needs to be further explored in order to determine the optimal method for classifying this type of data, focusing on the inherent structure of the data and not relying predefined labels. Furthermore, through pan-cancer or multiclass approaches, a deeper understanding of the molecular differences between various cancer types and subtypes can be gained while also testing the algorithms in a broader approach. In this context, this work contributes to this exploration by providing a comparative analysis between unsupervised machine learning algorithms based on several evaluation indices to assess their effectiveness and accuracy in the classification of different types and/or subtypes of cancer, taking a pan-cancer approach.

3. Materials and Methods

This project is based on a random sample of an RNA sequencing dataset originally obtained from The Cancer Genome Atlas (TGCA) Pan-cancer analysis project [6]. This particular sample was obtained from the Machine Learning Library [5]. The dataset encompasses gene expression profiles of patients diagnosed with different cancer types, measured by the Illumina Hiseq platform. It comprises 801 observations (corresponding to each patient) and 20531 genes.

This work was carried out using the R software (version 4.3.2.) within the RStudio environment. The corresponding code, packages and functions employed can be found on the Github link created for this purpose [37].

3.1. Exploratory analysis

The data structure is studied for the dataset under study. The mean, median, extreme values and standard deviation are also calculated for each variable.

3.2. Feature subset selection

Considering that the main objective of this thesis is to find a classification model with an unsupervised approach capable of distinguishing cancer types based on the gene expression data, the standard deviation is used to discard the variables with a standard deviation equal to zero, since no variation between cancer types could be explained by these features.
With the dataset without the variables with zero standard deviation (sd0), three different dimensional reduction approaches were performed.

3.2.1. Principal Component Analysis (PCA) selecting the variables that explain 95% of the total variance

Given that the data frame in the analysis still contained a considerable number of variables compared to the number of samples, it is important to select the most relevant variables for the clustering.
For the purpose mentioned above, PCA based feature extraction is a method which has proven to be a good option for selecting relevant variables in datasets with a larger number of variables compared to the observations [38]. Therefore, this method was carried out in order to reduce the number of dimensions. Afterwards, an index which organizes the variables in order of importance was created. To do this, the importance of each variable was calculated by multiplying the contribution of each variable to each principal component (PC) and the percentage of variance explained for that particular PC. Then, the variables that explain the 95% of the variance were selected to create a new data frame that was used for the incoming steps (PCA95).

3.2.2. PCA selection of the 800 most relevant variables

In order to improve the computational efficiency and stability of the algorithms, the same PCA filtering of the data was performed, but in this case, only the first 800 variables (PCA800) that explained the most variance were chosen.

3.2.3. Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique, but differs from PCA in that it is designed to find non-linear structures in the data and preserves the local and global structure of the data [39,40]. This technique has been shown to be more sensible than PCA on transcriptome datasets [40]. Therefore, UMAP was used to perform an alternative variable or feature selection to PCA, with the aim of selecting the best one.

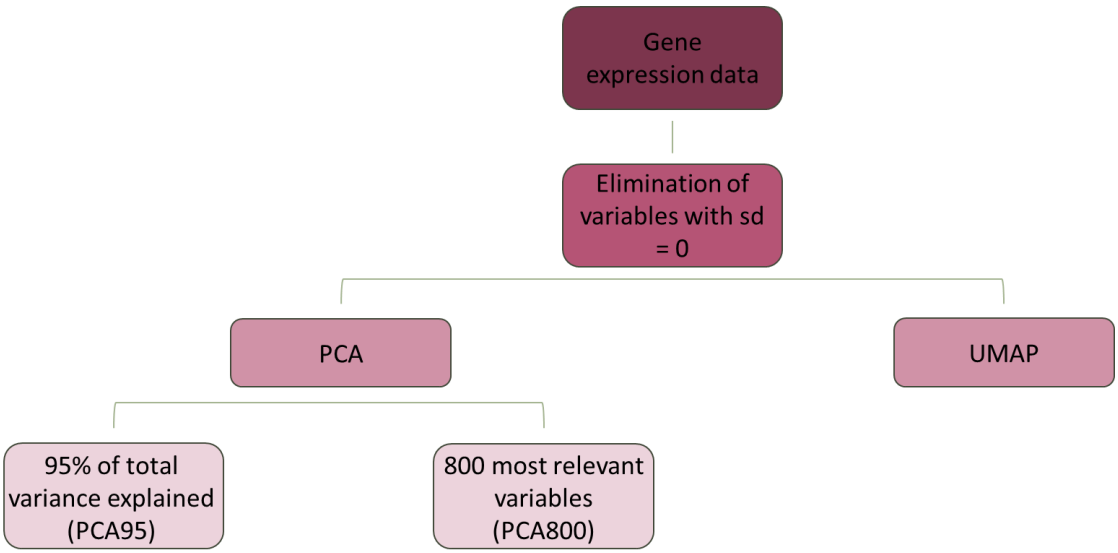


Figure 2. Diagram of the different approaches used to perform dimensionality reduction of the gene expression data.

3.3. Algorithm implementation

Using the PCA95, PCA800 and UMAP datasets created, different types of unsupervised algorithms were applied, in particular, partitioning (k-means, PAM and CLARA), density-based (DBSCAN), hierarchical (agglomerative using complete, average, single and ward, and divisive models) and model based (Gaussian mixture) algorithms. The use of this variety of algorithms brings the possibility of selecting the algorithm that best fits the characteristics of the data due to the different strengths, weaknesses and assumptions that each algorithm has [18].

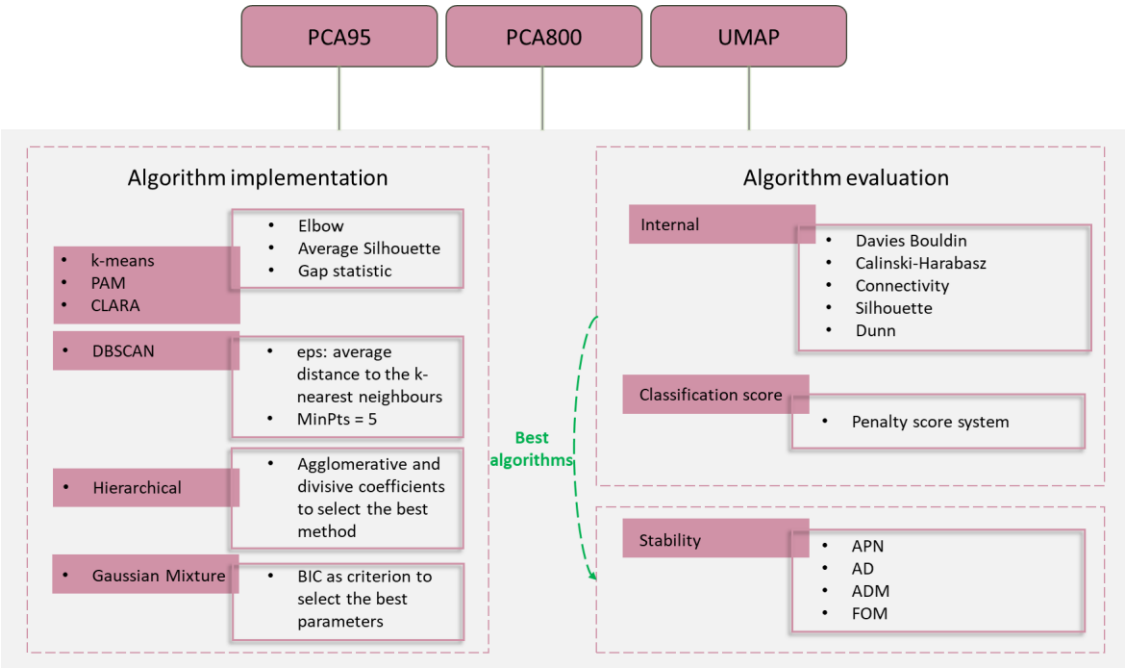


Figure 3. Diagram of the algorithms run on the different datasets obtained during the feature selection and the indices used to evaluate and compare the implemented models.

3.3.1. Partitioning algorithms (k-means, PAM and CLARA)

Before implementing the partitioning algorithms, different methods were used to select the optimal number of clusters.

- Elbow method: this method assesses the intra-cluster variance for several values of k clusters. It is expected that as the number of clusters increases, the variance or heterogeneity within clusters decreases, until all observations are placed within their cluster, so increasing the number of clusters would not improve the model but may lead to overfitting. The optimal number of clusters is chosen by the elbow formed in the line resulting from representing the variance within clusters versus the number of clusters, as shown in Figure 3 [41].

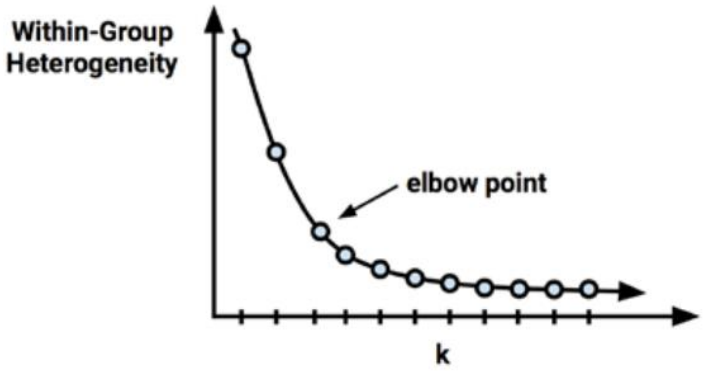


Figure 4. Example of a result plot of the elbow method, showing the internal variance of the clusters versus the number of clusters. The point where an elbow can be seen is indicated.

- Average Silhouette method: this method evaluates how well an object fits into its cluster by assessing the distance between that observation and the rest of the points in its cluster and comparing it to the neighbouring clusters [42]. Therefore, this method calculates the observation's average silhouette for several k values. The optimal number of clusters is determined by the maximum average silhouette calculated [42].
- Gap statistic method: unlike the Elbow and Average Silhouette method, which evaluate global clustering features, the Gap statistic method uses statistical analysis to calculate the optimal number of clusters. This method compares the internal variation within k clusters with the variance expected under a null reference distribution (which is created by randomly sampling the original dataset). The optimal number of clusters is chosen by the number of clusters that maximises the gap statistic [42].

Once the optimal numbers of clusters were selected, the k-means, PAM and CLARA algorithms were implemented. In the cases where different number of clusters were indicated by the three methods, all of them were tested. Furthermore, in some cases, the algorithms were run with a certain number of clusters in order to compare the results with previous ones.

3.3.2. DBSCAN

In this case, it is not necessary to select the number of clusters, but eps (maximum reachability distance) and minPts (minimum number of points required to define a cluster) must be defined beforehand. To select the eps parameter, the average distance to the k-nearest neighbours was calculated for each point and the eps was selected as the approximate point where the curve of the graph shows an elbow [42]. For the minPts, several values were tested with for the number of eps indicated by the k-nearest neighbours plot and the one with the better results was selected.

3.3.3. Hierarchical

A hierarchical approach was also tested. Both agglomerative and divisive methods were implemented and compared using the agglomerative and divisive coefficients. The hierarchical algorithm with the highest coefficient was selected for further analysis [42].
In the case of the agglomerative model, several methods were evaluated:

- Maximum or complete linkage: it calculates the distance between two clusters using the two most distant data points (one from each cluster). The two clusters with the shortest distance are then merged. This method tends to produce compact clusters.
- Minimum or single linkage: it calculates the distance between two clusters using the two closest data points (one from each cluster). The two clusters with the shortest distance are then merged. This method tends to create long clusters.
- Mean or average linkage: it calculates the distance between two clusters using the average distance of all the points between the two clusters. The two clusters with the shortest distance are then merged.

- Ward's minimum variance method: in each iteration of the model, the clusters that when merged minimize less the intra-cluster variance, are combined [42].

Once the approach (agglomerative or divisive) and the method (in the case of the agglomerative approach) is selected, the classified data is divided into k number of clusters. As suggested in the literature, the number of clusters obtained in the elbow method for the k-means algorithm was considered [43,44], except in the case of PCA95, in which because of the results of the elbow method and the background knowledge of the original data, it was decided to use k = 5 to ensure a better classification by the model.

3.3.4. Gaussian mixture

The Gaussian mixture model was also performed. A range from 1 to 9 clusters were tested with different covariance matrix parameterizations. To select the best parameters to carry out the model, the Bayesian Information Criterion is used, looking to maximize its value [42].

3.4. Algorithm evaluation

After running all the described algorithms with the optimal parameters found, it is important to calculate objective measurements of the quality of the clusters since the plots shown in the results do not represent the whole data in most of the cases. An internal evaluation of the different algorithms was therefore carried out. Furthermore, since in this particular case, the real classification of the data is available, a comparison was made between the original labelled data and the algorithms classification.

3.4.1. Internal evaluation

For the six algorithms, the Calinski-Harabasz, Connectivity, Silhouette and Dunn indices were calculated. In addition, the Davies Bouldin index was only calculated for the kmeans, PAM and CLARA algorithms.
The Davies Bouldin and Dunn indices provide an insight into the compactness and separation of clusters, while the Calinski-Harabasz assesses the ratio of inter-cluster variance to intra-cluster variance. The Connectivity measure, using clValid and internal metrics, provides a deeper understanding of how well defined and connected the clusters are. And the Silhouette analysis evaluates the cohesion and separation of clusters for individual data points [45–48].

All the indices calculated for the same algorithm using different parameters were then compared in order to choose the best parameters in each case. Afterwards, the best results for each algorithm were compared in order to classify the different algorithms according to the best internal structure of the clusters.

3.4.2. Classification assessment

In order to better understand the performance of the algorithms, the labels (cancer type) associated with the original data were used to assess the

classification accuracy of each algorithm. The original data counts with data of patients with five different types of cancers labelled as BRCA (breast cancer), COAD (colon cancer), KIRC (kidney cancer), LUAD (lung cancer) and PRAD (prostate cancer).
For this purpose, a function was created in R to be able to know which patients are grouped together and how many of them belong to the same type of cancer according to the original labelled data. The function is explained and displayed in the Github link provided.

This function also allows the identification of the type of cancer to which each cluster is associated. With this information a scoring system was created. All the algorithms start with a score of 801, which is the maximum number of observations to be classified and therefore, the maximum number of instances that can be correctly classified. Given that there is a predominant type of cancer in each cluster created, the observations of other cancer types that are classified in the same cluster of the predominant cancer type are considered as classification errors and treated as a penalty, so the number of errors is subtracted from 801. Those clusters with only one observation of each type of cancer are treated as penalties. Furthermore, the observations not classified by the DBSCAN algorithm are considered as errors. Therefore, the best parameters for each algorithm were selected according to the classification score and then used to compare between algorithms in order to find the best ones.

3.4.3. Stability evaluation

With the aim of finding another way to differentiate and evaluate the algorithms, the stability of the best algorithms found based on the internal and classification evaluations is compared. Also, the computational requirements for performing the algorithms were considered in order to select the best algorithms for this step.
The stability evaluation assesses the robustness and reliability of a clustering classification by comparing the clusters given by the algorithm using the whole data, with the clusters created when a column or feature of the dataset is removed. The columns are deleted one by one [42].

Several stability parameters were measured such as:

- Average Proportion of Non-overlap (APN) which measures the average proportion of observations that change their cluster whether the full data is used or when a column is removed.
- Average Distance (AD): represents the average distance values between the two conditions (whole data and data with removed columns).
- Average Distance between Means (ADM) measures the average distance between the cluster centres in the two conditions (whole data and data with removed columns).
- Figure of Merit (FOM) measures how the removal of a column or variable affects the intra-cluster variance [46].

3.4.4. Comparison of the original data distribution separated within the clusters obtained by the best algorithm

In order to study if it was possible to distinguish any differences between the clusters suggested by the best algorithm, the mean and median of the gene expression of each patient within the different clusters were calculated and the distribution was represented. The calculations were made using original data.

4. Results

4.1. Exploratory analysis

During the exploratory analysis of the dataset, it was possible to analyse its structure confirming that it contains information on 20532 features of 801 cancer patients. Among the variables, one is a character variable, corresponding to names of the samples (patients) and the rest are numeric variables, corresponding to the gene expression count matrix.

In Figure 5 the results of the mean, median and extreme values of all the numeric variables present in the data are displayed. As we can see, the ranges of the four calculated parameters are similar, all of them going approximately from 0 to 15, except for the maximum values which go up to 20. Furthermore, all the variables are believed to be in the same units since they are expression levels of specific genes from cancer patient samples measured by Illumina Hiseq Platform. Therefore, the data were assumed to be on the same scale, so no normalization was carried out prior to feature selection.

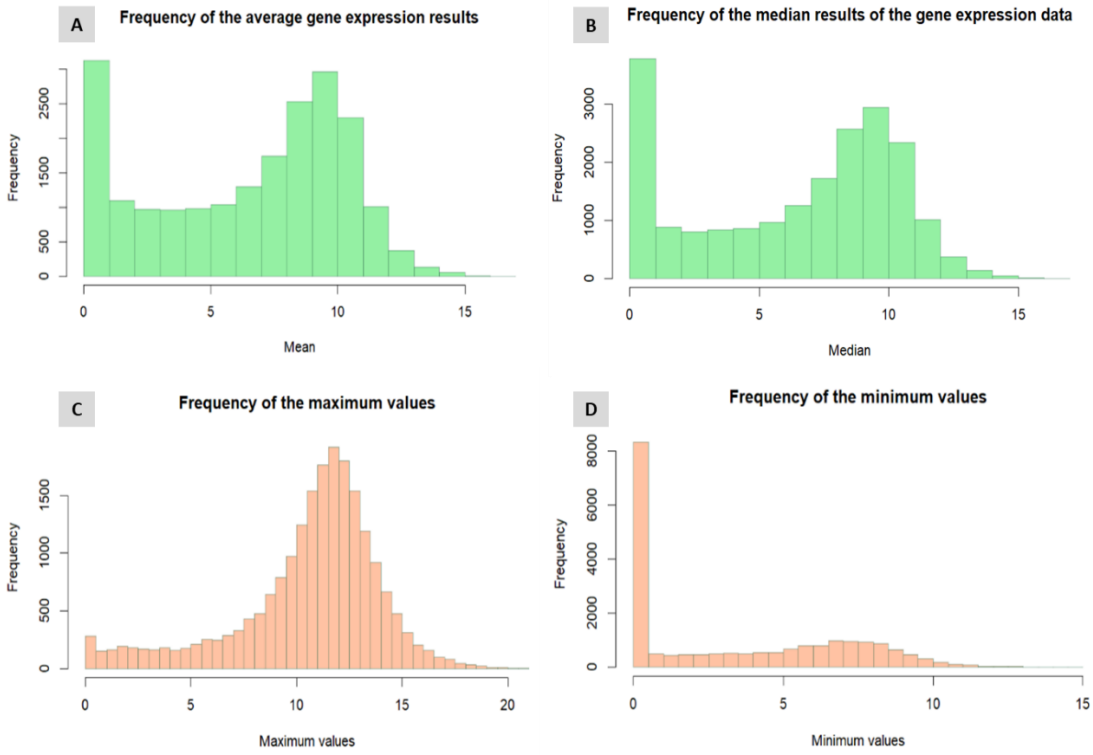


Figure 5. Representation of the frequencies of the mean (plot A), median (plot B), maximum (plot C) and minimum (plot D) values of the features contained in the dataset under study.

Regarding the standard deviation, in Figure 6 it can be seen that there are variables with very low standard deviation (sd), including 267 variables with sd equal to zero.

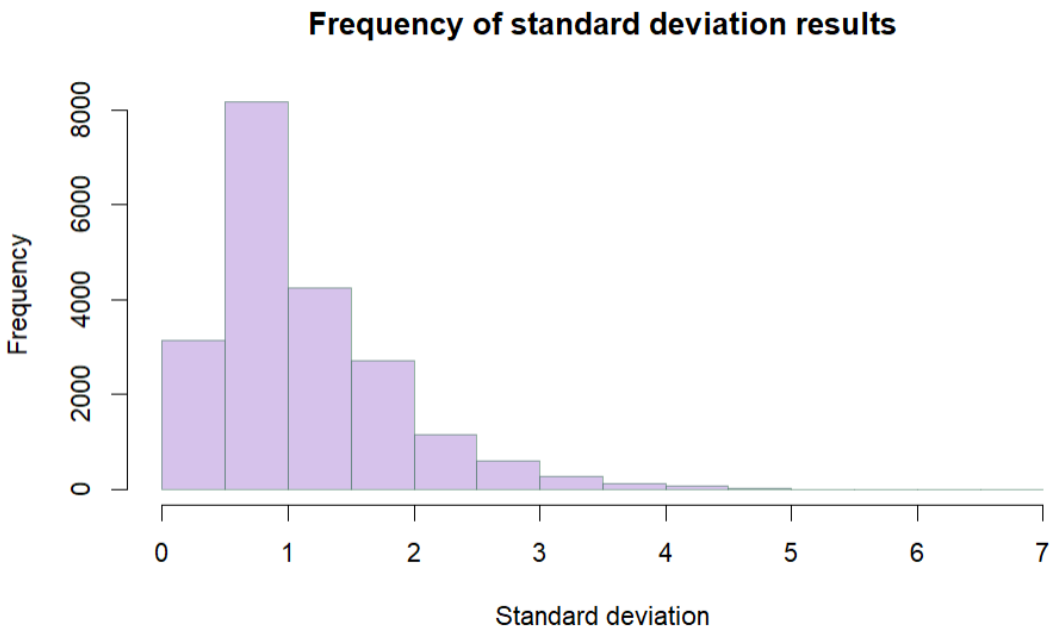


Figure 6. Representation of the frequency of the Standard deviation results obtained for the different features present in the data.

4.2. Feature subset selection

As can be seen in Table 1, after removing the variables with a standard deviation equal to zero, 20264 features remained in the dataset. Furthermore, three different datasets were created after applying the methods and approaches selected for the dimensional reduction of the sd0 data. The PCA95 dataset showed that the 13357 most relevant genes allowed to explain the 95% of the variance. Moreover, the PCA800 dataset with 800 variables was created in order to have fewer variables than observation, but only the 25.18% of the total variance is explained by this dataset. Finally, the UMAP technique allowed to create a two-dimensional dataset.

Table 1. Number of variables of the original data and the created datasets sd0, PCA95, PCA800 and UMAP.

Dataset	Number of variables
Original data	20531
sd0	20264
PCA95	13357
PCA800	800
UMAP	2

4.3. Algorithm implementation

4.3.1. PCA95

4.3.1.1. k-means

As can be seen in Figure 7, all three methods (elbow, average silhouette and gap statistic) give different results. Thus, k-means algorithm was run with k = 3, k = 6 and k = 8.

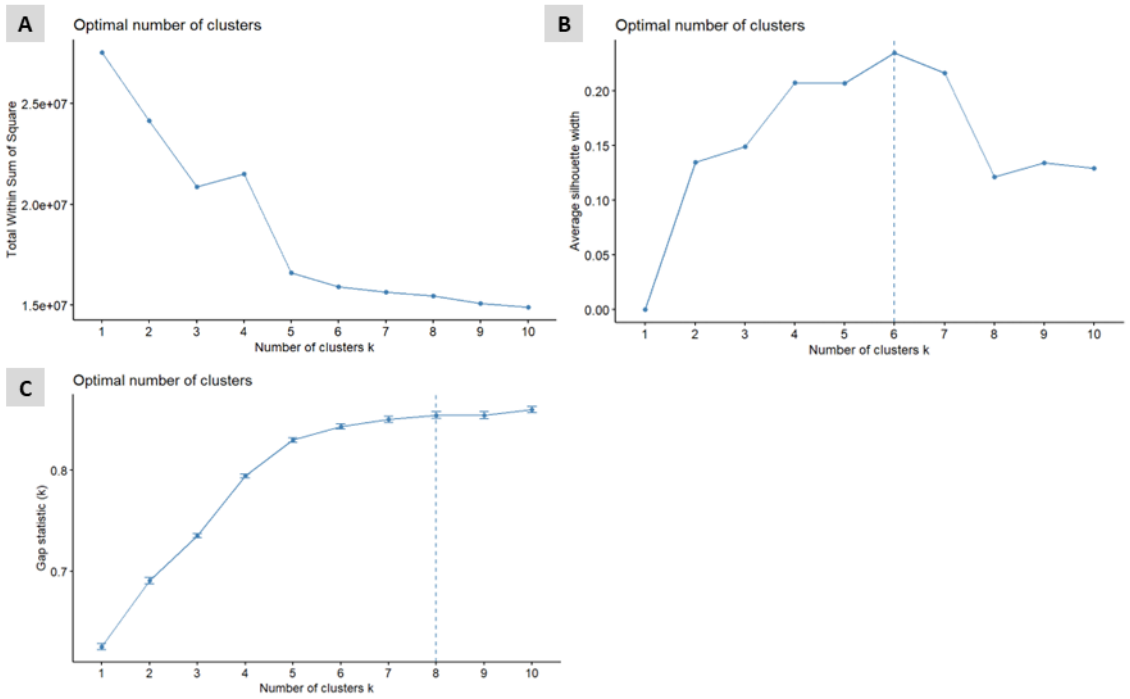


Figure 7. Graphical results of the optimal number of clusters according to the elbow method (plot A), the average silhouette method (plot B) and the gap statistic method (plot C) for the k-means algorithm using the PCA95 dataset.

As can be seen in the k-means algorithm plots in Figure 8, only the first two dimensions are represented. However, the amount of variance that can be explained is by these two dimensions is small since Dimension 1 (Dim1) explains only 12.7% of the variance and the second (Dim2) only 9.4%. Therefore, these graphs might not be a good representation of the performance of the algorithm. Nevertheless, considering only this percentage of explained variance, as can be observed in plots A, B and C of Figure 8, none of the k (k = 3, k = 6 and k = 8) produced separate and well-defined clusters.

It is known that this dataset consists of samples from patients with five different cancers. Therefore, and taking into account the success of Venkataramana L. *et al* (2020) in classifying the data into five groups, the algorithm was re-run with k = 5 to see if any improvements were observed. As can be seen in the plot D of Figure 8, the result is similar to the ones seen before, with all the clusters partially or almost completely overlapping for the explained variance represented. However, as mentioned before, the graph may not be a good representation of the algorithm's performance.

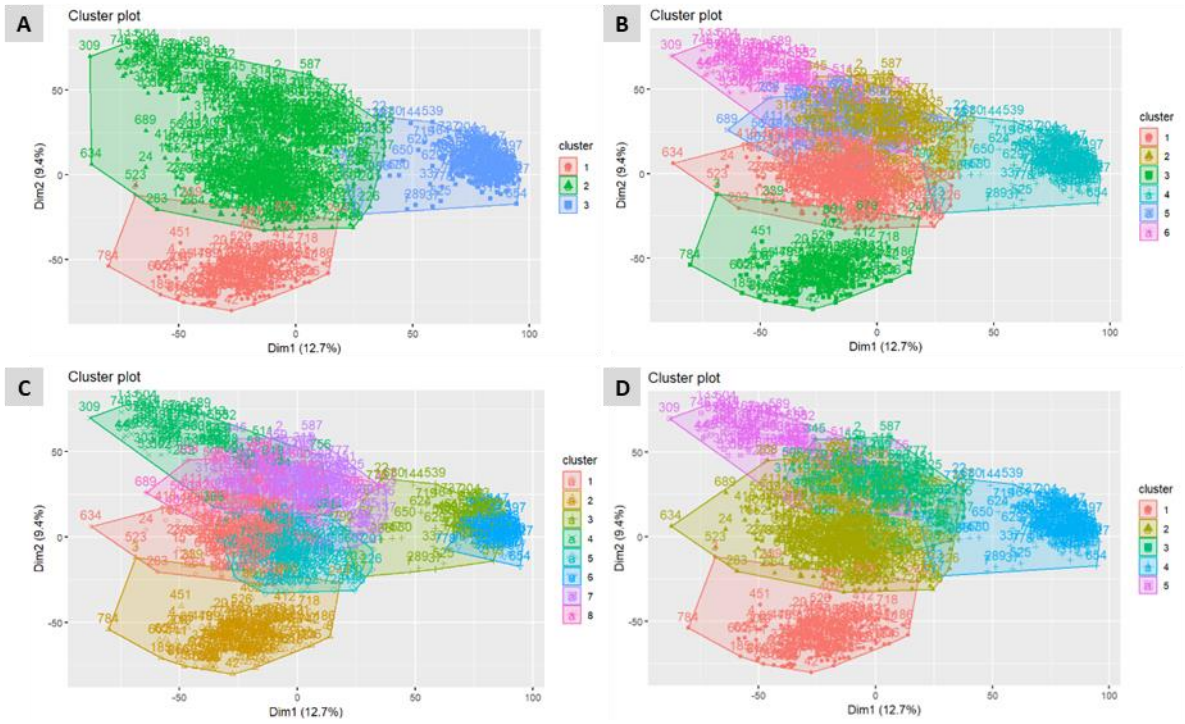


Figure 8. Graphical results of k-means algorithm using PCA95 dataset for k = 3 (plot A), k = 6 (plot B), k = 8 (plot C) and k = 5 (plot D).

4.3.1.2. PAM

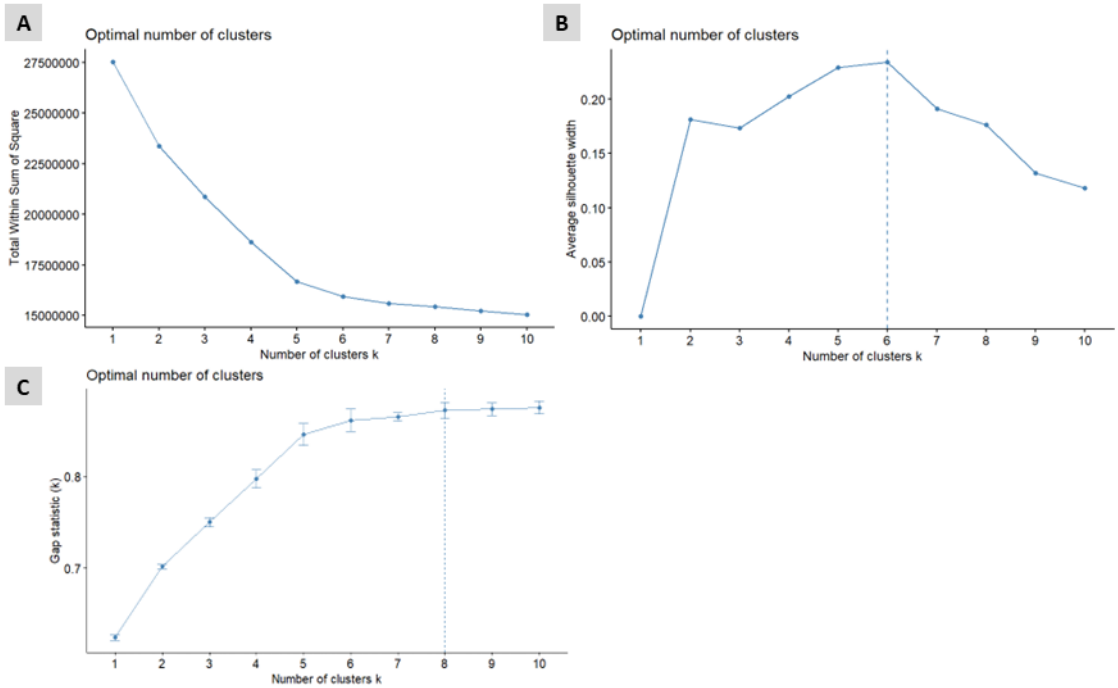


Figure 9. Graphical results of the optimal number of clusters according to the elbow method (plot A), the average silhouette method (plot B) and the gap statistic method (plot C) for the PAM algorithm using the PCA95 dataset.

In this case, the graph of the elbow method (plot A in Figure 9) shows a small elbow in the curve that allows the identification of the optimal number of clusters as k = 5 according to this method. On the contrary, the average silhouette method

(plot B in Figure 9) and the gap statistic method (plot C in Figure 9) identify the same optimal number of clusters as in the k-means algorithm, 6 and 8 clusters, respectively. In addition to the results obtained, the algorithm using k = 3 was also performed for the PAM algorithm in order to be able to compare the results.

As it can be seen in the plot A of Figure 10, the PAM algorithm with k = 3, which was carried out only for comparison with the results of the k-means algorithm, the division is worse performed than in the previous algorithm, since the cluster 1 almost completely overlaps with the cluster 2, whereas in the k-means algorithm the distinction between these two clusters is better defined for the Dim 1 and Dim 2. This was expected once neither of the methods used to calculate the optimal number of algorithms for this algorithm suggested k = 3. Regarding the results obtained for k = 5 and k = 6, they are visually almost identical or similar in the case of k = 8, to those obtained for the k-means algorithm.

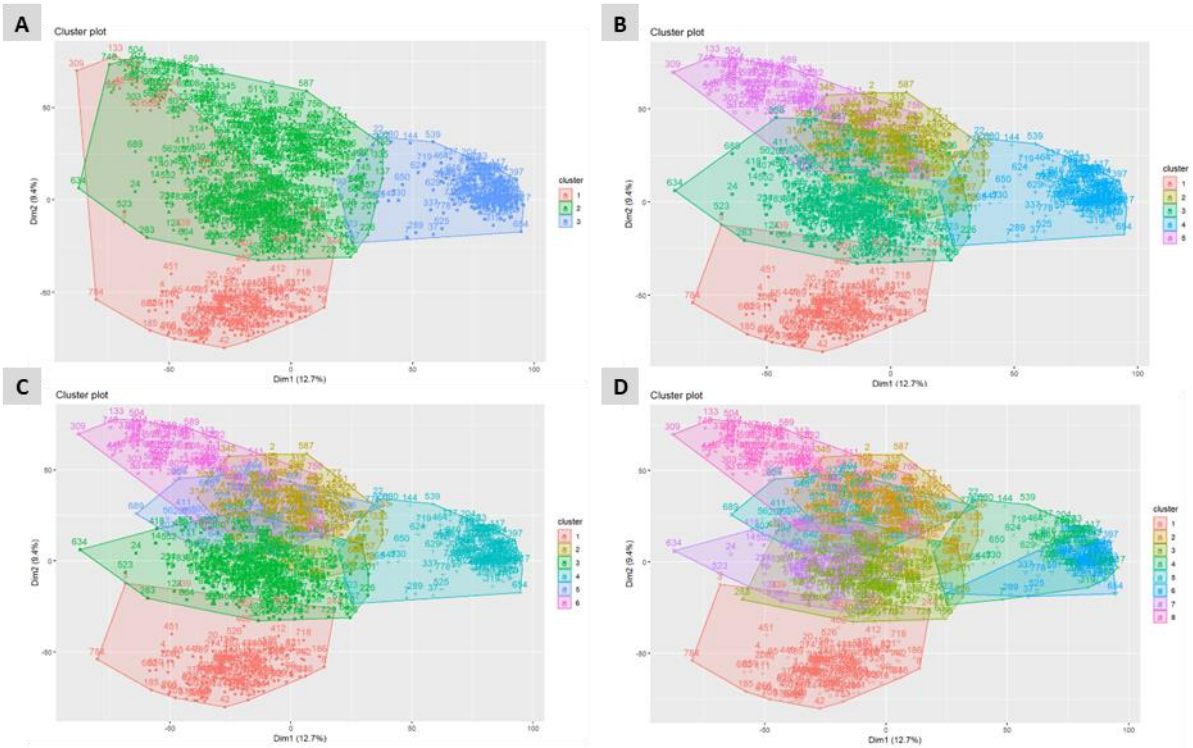


Figure 10. Graphical results of PAM algorithm using PCA95 dataset for k = 3 (plot A), k = 5 (plot B), k = 6 (plot C) and k = 8 (plot D).

4.3.1.3. CLARA

The elbow method graph for the CLARA algorithm (plot A of Figure 11) shows a slight elbow at 5 clusters. Regarding the average silhouette method (plot B of Figure 11), the same result was obtained as in this method of the previous algorithms (k = 6), while in the gap statistic method (plot C of Figure 11), seven clusters are chosen as the optimal number. Considering the results, the algorithm was implemented with the three obtained values of k = 5, k = 6 and k = 7. Also, k = 3 and k = 8 were run.

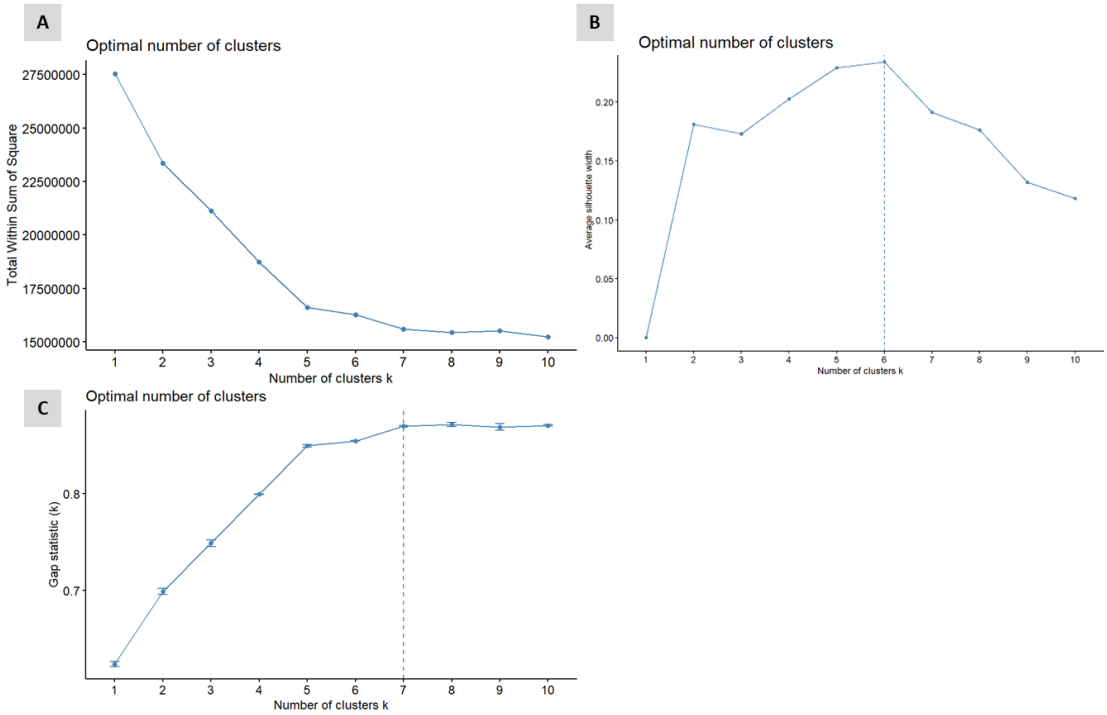


Figure 11. Graphical results of the optimal number of clusters according to the elbow method (plot A), the average silhouette method (plot B) and the gap statistic method (plot C) for the CLARA algorithm using the PCA95 dataset.

Comparing the results obtained by the CLARA and PAM algorithms, the results obtained for k = 3 and k = 5, are the almost visually identical for Dim1 and Dim2. It is worth noting that CLARA is the only algorithm that has so far clearly suggested k = 5 as one of the optimal number of clusters methods (elbow), since although in PAM a small elbow can also be seen in k = 5, is much less pronounced that in the CLARA algorithm. However, the algorithm result is visually identical to the previous ones.

Furthermore, there are differences in the graphs obtained for k = 6 and k = 8 compared to the k-means and PAM results for the first two dimensions. In this case, for k = 6, the cluster 3 reaches higher points, while the cluster 2 is narrower. Moreover, the cluster 6 is almost completely defined within the cluster 2. For k = 8, the cluster proposed in the other two models, which separates the data located on the right of the graph, is no longer present, while in this case, another cluster is observed in the upper centre of the graph. Also, the clusters 3 and 5 are almost identical. Moreover, for k = 7, while the cluster seven (6 in k=6 and 8 in k =8) is smaller than in k = 6 and k = 8, the cluster five which does not appear in k = 6, is defined as in k = 8, and overlaps almost completely with the clusters 2, 6 and 3.

Overall, it is not possible to draw any conclusions since the graphical representations only allow the analysis of the first two dimensions of the data, which in this case, represent a small percentage of the total variance. Hence, a qualitative analysis of the clusters is needed in order to compare these algorithms.

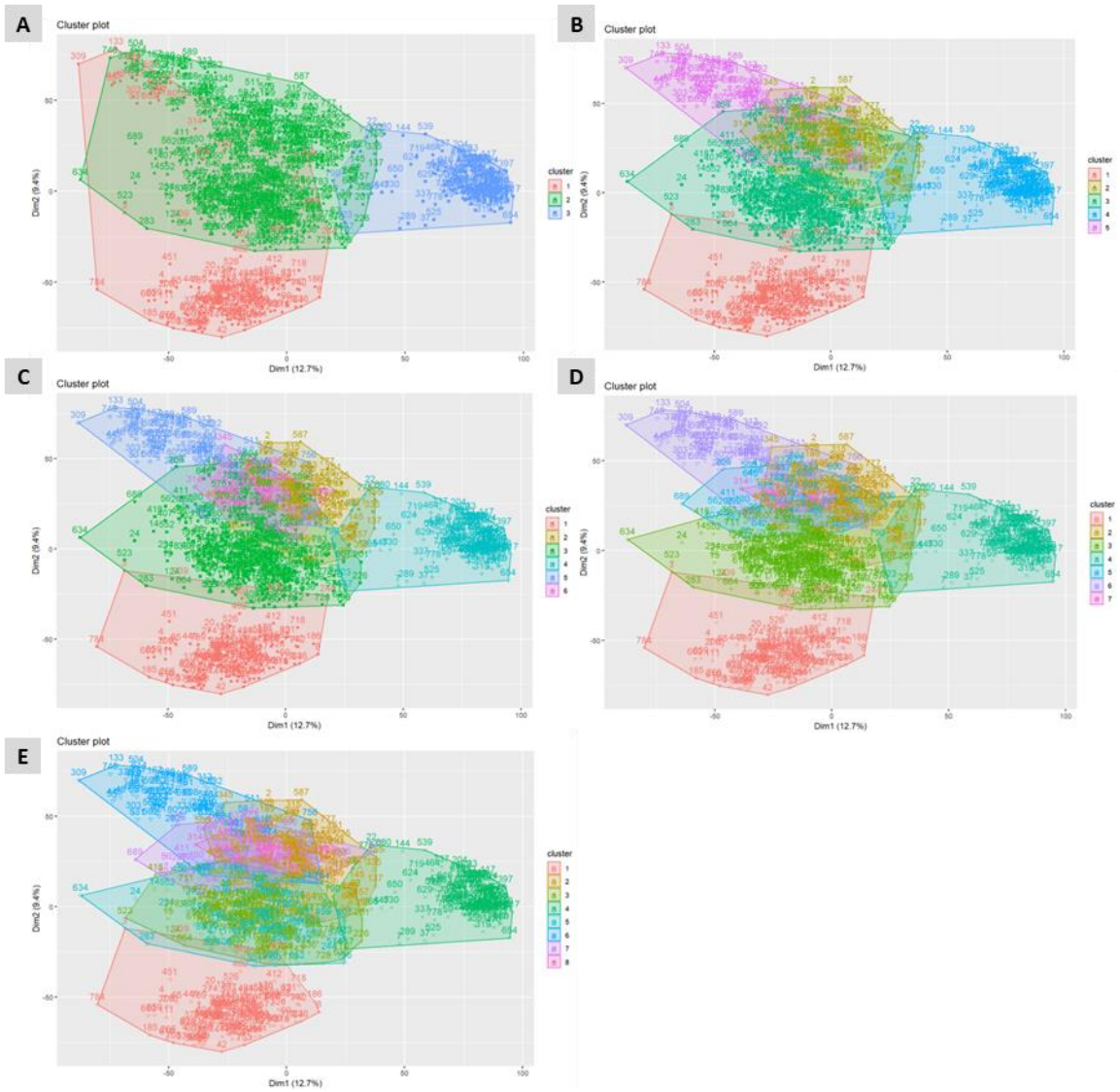


Figure 12. Graphical results of CLARA algorithm using PCA95 dataset for k = 3 (plot A), k = 5 (plot B), k = 6 (plot C), k = 7 (plot D) and k = 8 (plot E).

4.3.1.4. DBSCAN

Having established that eps will be around 200 for minPts = 5 as indicated by Figure 13, different values of eps were tried, in particular, eps = 174, 161 and 171. Furthermore, different values of minPts were also applied, however, all of them worsen the results. The best results are displayed in Figure 14.

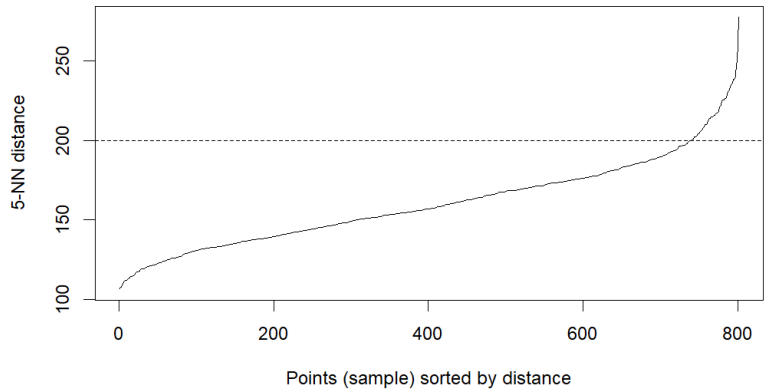


Figure 13. Representation of the average distance to the k-nearest neighbours plot for k = 5, using PCA95 dataset. A discontinuous line is shown around the point where the line forms an elbow, indicating the approximate best eps value.

As shown in Figure 14 and Table 2, three different results from three different eps values are presented. The first case (plot A and eps = 174) is the one that shows the lowest number of values of unclassified points (cluster 0). It shows five clusters, three of which overlap considerably for Dim 1 and Dim 2 which explained 12.7% and 9.4% of the variance, respectively. The second case (plot B and eps = 161) is the one with the highest number of unclassified points. It also shows five clusters, however, they are almost completely separated from each other. It is important to mention that out of 801 points, 354 were not assigned to any cluster. Finally, the third case (plot C and eps = 171), with an intermediate value of eps, also shows an intermediate value of unclassified points, but in this case six clusters are defined. Clusters 1 to 5 are almost the same as in the first two cases, but a sixth cluster is defined within the limits of the cluster four. This is reminiscent of the results obtained with the CLARA algorithm for k = 6. It is important to note that these plots might not reflect the reality of the clusters because they represent only a small part of the variance in the data.
Furthermore, it can be seen that the shape of the clusters is almost the same as those obtained with the previous algorithms. The DBSCAN cluster is capable of designing non-spherical or ovoid clusters [49], however, in this case, we can observe that the shape of the clusters does not change much, suggesting that this algorithm might not bring any advantage. Further evaluation of the DBSCAN clusters is described in section 4.2 in order to provide more rigorous measurements of the cluster quality.

Table 2. Number of observations classified in each cluster by the DBSCAN algorithm with minPts = 5 and eps = 174, eps = 161 and eps = 171, using the PCA95 dataset. The instances that could not be classified by the algorithm are indicated in cluster number 0.

	Clusters						
eps	**0**	**1**	**2**	**3**	**4**	**5**	**6**
174	199	129	200	125	80	68	
161	354	120	126	117	38	46	
171	238	125	177	125	68	61	7

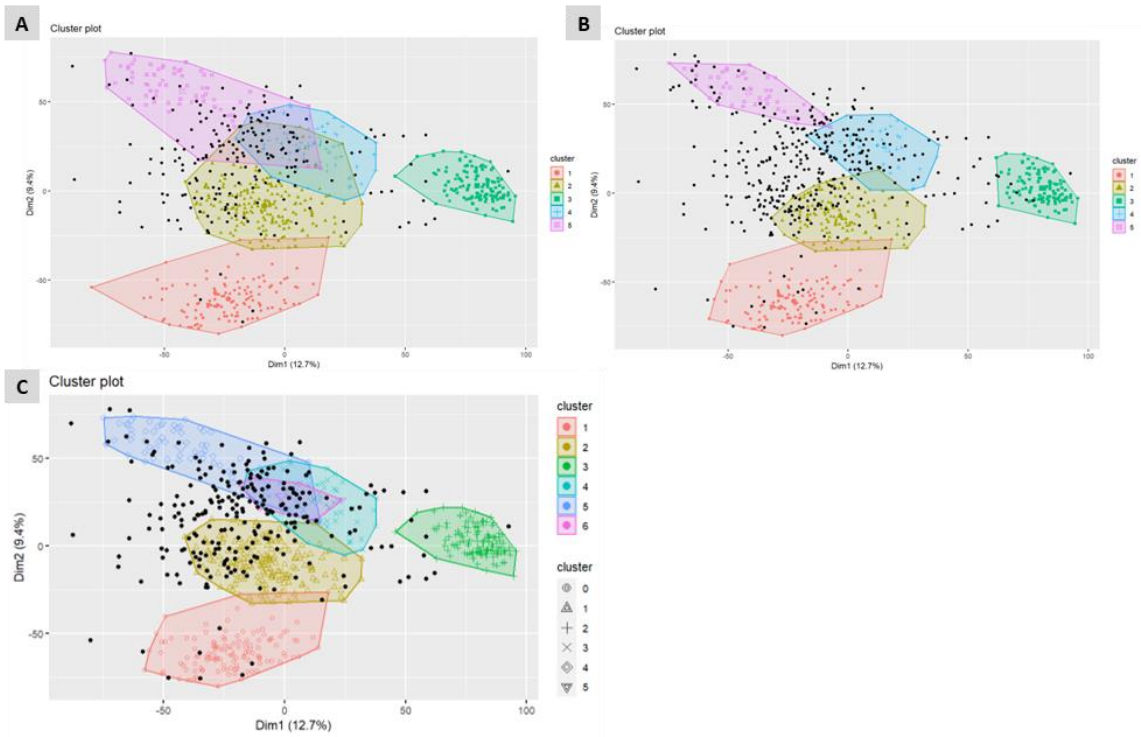


Figure 14. Graphical results of DBSCAN algorithm using PCA95 dataset with minPts = 5 and eps = 174 (plot A), eps = 161 (plot B) and eps = 171 (plot C).

4.3.1.5. Hierarchical

In order to ensure the better classification strategy, and since partitioning and density classification methods have already been applied, the hierarchical method was implemented.

As can be seen in Table 3, the best method was the Ward's method, since it achieved the highest agglomerative coefficient among all the methods performed, being very close to 1. As mentioned before, this method minimizes the intra-cluster variation and maximizes inter-cluster variation, and it was found to be less sensitive to noise than the other agglomerative methods [50]. Thus, this method was chosen for the following steps for the PCA95 data.

Table 3. Agglomerative and divisive coefficients of the different hierarchical algorithms performed using PCA95 dataset.

Method	Average	Single	Complete	Ward's
Agglomerative coefficient	0.6203821	0.5045055	0.4189288	0.9451525
Divisive coefficient	0.6082847			

Afterwards, although the elbow method for the k-means model suggested that the optimal number of clusters was 3, the data classified by the Ward's agglomerative method, was divided into five clusters, based on the background knowledge of the original data in order to try to reach the best performance of this algorithm. In Figure 15, the five clusters created are represented in a dendrogram. It can be seen in the representation that there are two main divisions of clusters, where in the first division only one cluster is separated from the other four, suggesting that this cluster is the one with the most differences and the other

clusters have more similar gene expressions. However, to evaluate the accuracy of the classification, it is necessary to use different indices and parameters that allow to assess the performance of this algorithm.

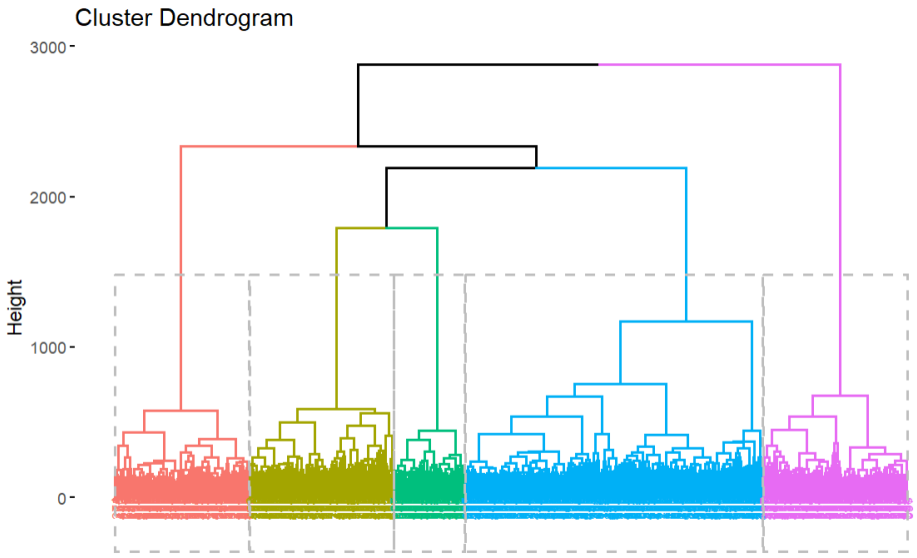


Figure 15. Dendrogram of the classification of the agglomerative hierarchical algorithm using Ward's method on PCA95 data.

4.3.1.6. Gaussian mixture

Finally, the Gaussian mixture model was carried out to evaluate whether the classification of the data perform better when the distribution of the data is used as a criterion. Thanks to the Mclust function used and the subsequent comparison of the BIC values obtained using different parameters, the model was run with nine clusters and VEI (the volumes of the clusters vary, their shapes are equal and their orientation is equal to the coordinate axes) as the defined covariance parameterization.

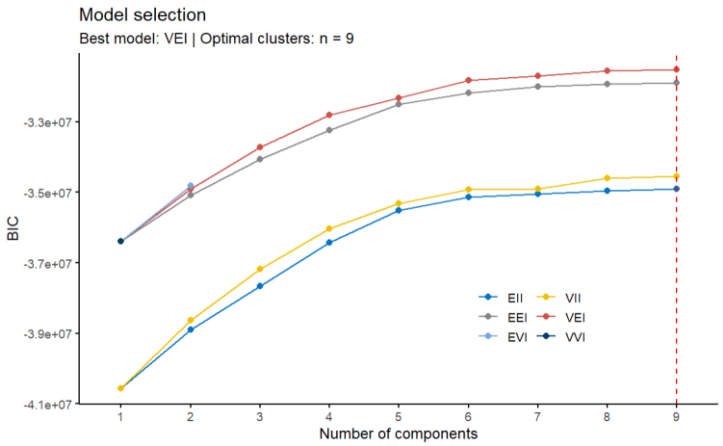


Figure 16. Representation of the BIC values obtained for the different numbers of clusters and the covariance parametrizations tested, particularly, EII (equal volume, equal shape, identical orientation), VII (varying volume, spherical covariance, identical orientation), EEI (equal volume, equal shape, identical orientation), VEI (varying volume, equal shape, identical orientation), EVI (equal volume, varying shape, identical orientation) and VVI (varying volume, varying shape, identical orientation) in the Gaussian mixture model using the PCA95 dataset.

Although only close to 20% of the variance of the data is represented in Figure 17, it appears that the clusters created overlap. Nine clusters were created and, as can be seen, several points corresponding to each clusters are represented outside the boundaries of the outlined clusters, making vague the division between them. Furthermore, in Table 4 it can be seen that the cluster with the highest number of observations is the cluster 4, although it is not the visually the largest cluster. However, in Figure 17 it can be observed that many observations corresponding to cluster 4 are displayed in the surroundings of this cluster, overlapping with the neighbouring clusters. Moreover, it is worth mentioning that all the clusters have more than 50 observations.

Table 4. Number of observations classified in each cluster by the VEI Gaussian Model algorithm with 9 clusters using PCA95 data.

Clusters								
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
51	68	119	215	52	52	93	78	73

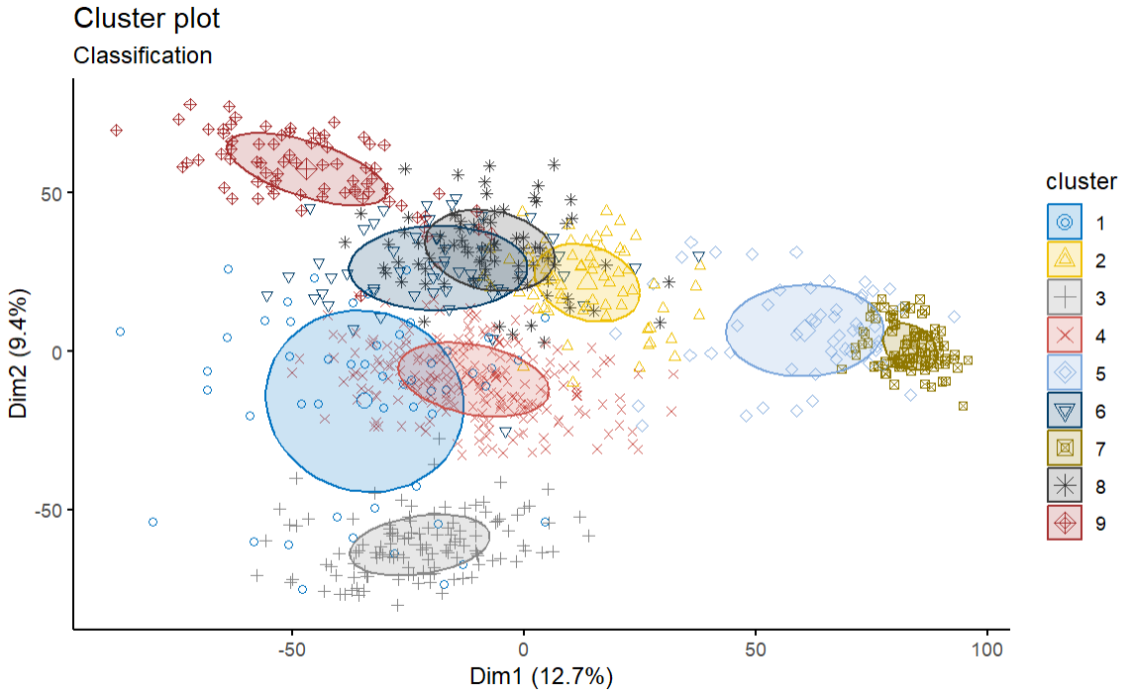


Figure 17. Graphical results of VEI Gaussian mixture algorithm using PCA95 dataset for 9 clusters.

4.3.2. PCA800

4.3.2.1. k-means

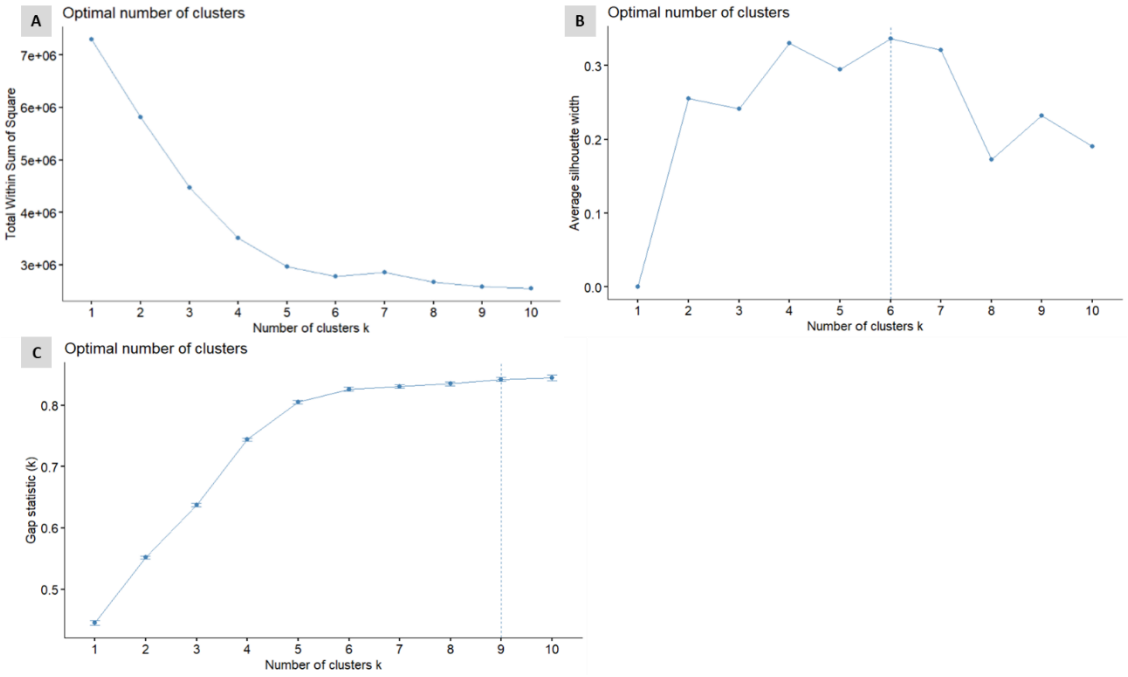


Figure 18. Graphical results of the optimal number of clusters according to the elbow method (plot A), the average silhouette method (plot B) and the gap statistic method (plot C) for the k-means algorithm using the PCA800 dataset.

As can be seen in Figure 18, the optimal number of clusters is 6 according to the elbow and average silhouette methods and 9 to the gap statistic method. However, the algorithm for k = 5 was also performed as it is known that there are 5 cancer types in the original data. It is also important to mention that the optimal number of clusters obtained is different from the PCA95 data, except for k = 6 (average silhouette method).

In this case (Figure 19), the first two dimensions explain approximately the 42.8%, a higher value than in the PCA95, but still a low value to make assumptions about the performance of the classification algorithm based on the graphical representation of the algorithms. Nevertheless, compared to the PCA95 results, the clusters can be seen to be more differentiated, since although some of them might overlap in some points, the boundaries of the majority of the clusters seem to be well defined, except for k = 9. However, it should be noted that the explained variance represented in these plots and in the PCA95 plots is not the same, which may affect the accuracy of the comparison made between the results of the two feature extraction methods.

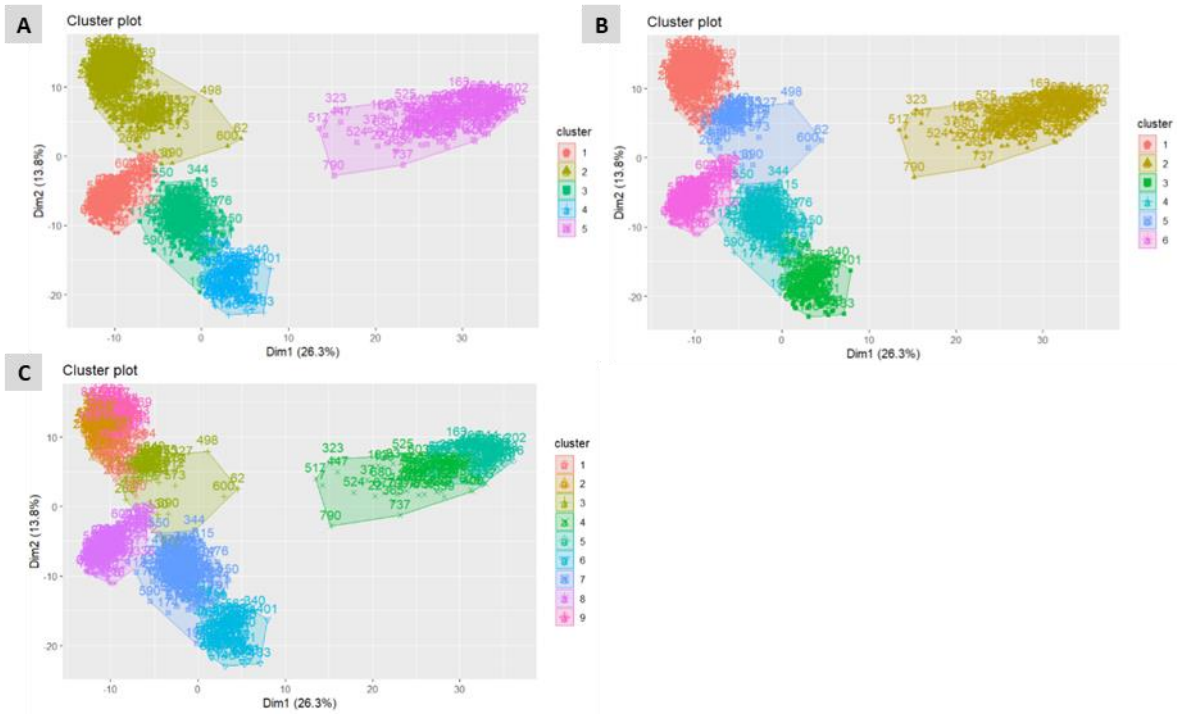


Figure 19. Graphical results of k-means algorithm using PCA800 dataset for k = 5 (plot A), k = 6 (plot B) and k = 9 (plot C).

4.3.2.2. PAM

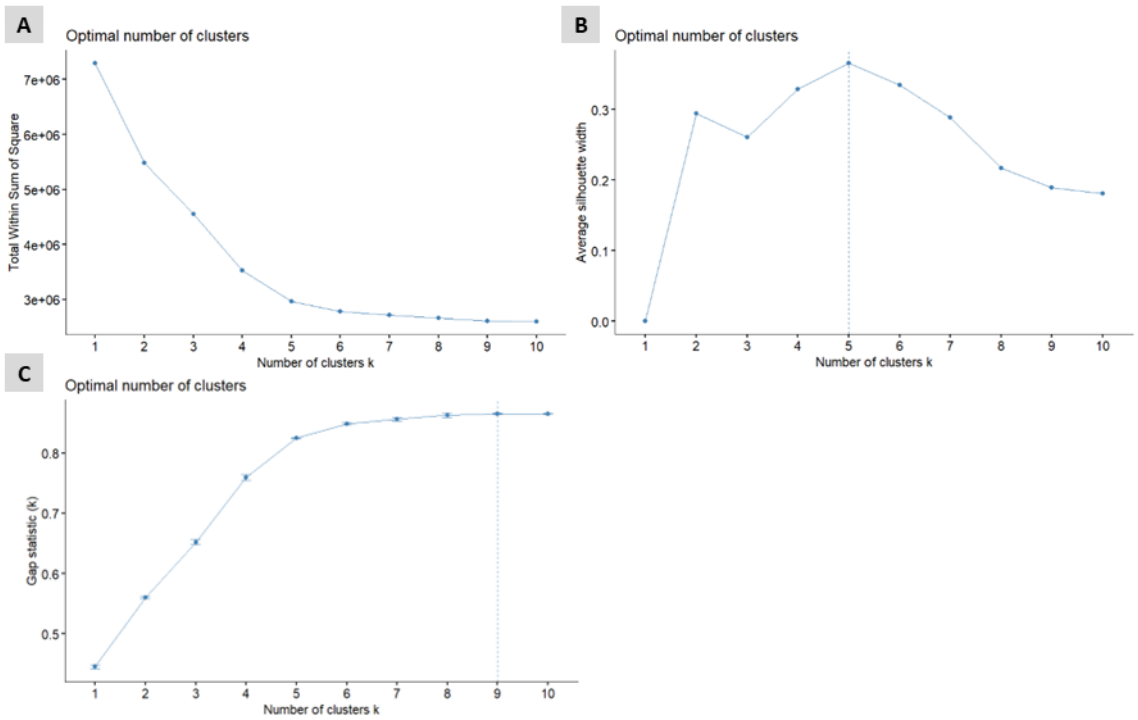


Figure 20. Graphical results of the optimal number of clusters according to the elbow method (plot A), the average silhouette method (plot B) and the gap statistic method (plot C) for the PAM algorithm using the PCA800 dataset.

In Figure 20, the results of the optimal number of clusters are represented. In this case, both the elbow and silhouette methods indicate that the optimal number of clusters is 5, however, the elbow in the elbow method is very small. Furthermore, the gap statistic method again indicates that 9 clusters is the best choice. However, k = 6 was also run in order to compare with the k-means algorithm results.

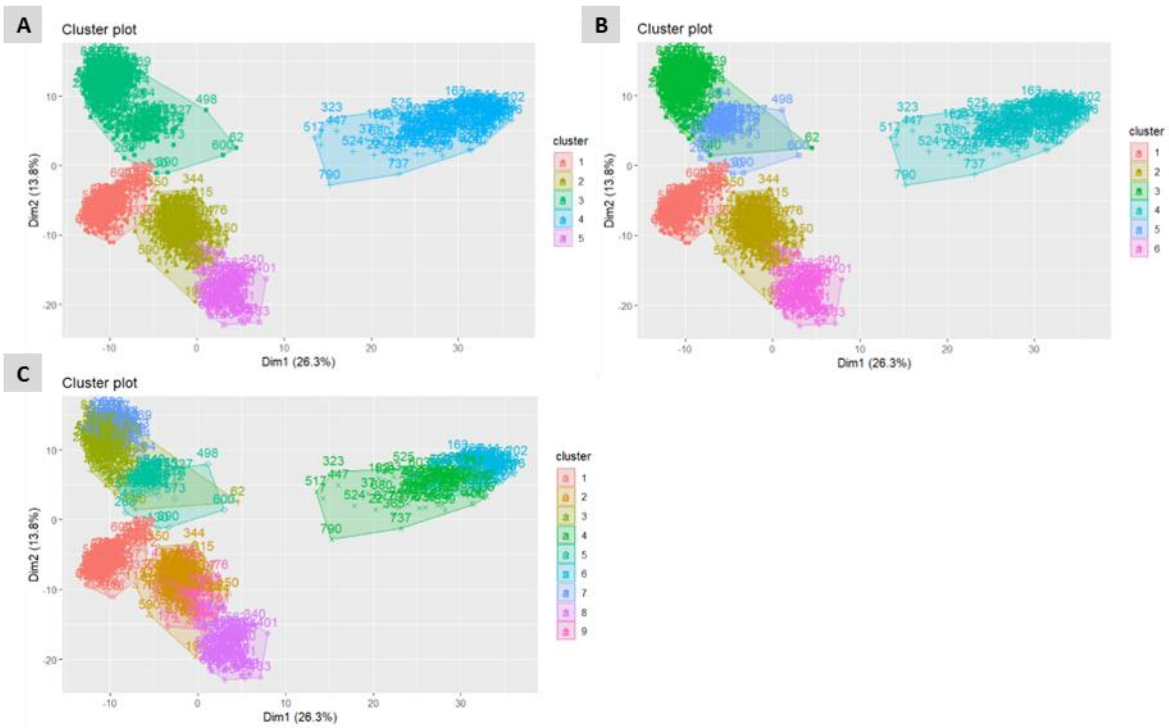


Figure 21. Graphical results of PAM algorithm using PCA800 dataset for k = 5 (plot A), k = 6 (plot B) and k = 9 (plot C).

In the plots of the PAM algorithm implementation for k = 5, k = 6 and k = 9 shown in Figure 21, it is possible to see that there seems to be more overlap between clusters in the case of k = 6 and k = 9. However, for k = 5, the graphical representation for this case and the one of the PCA95 dataset seems to be very similar.
Compared to the results obtained for the PCA95 dataset, only the elbow method for calculating the optimal number of clusters gave the same results (k = 5). Regarding the graphical results, there are obvious differences between the clusters in the two datasets, as the size and shape of the clusters are completely different. However, as mentioned above, the explained variance shown in the plot is different.

4.3.2.3. CLARA

The elbow method for the CLARA algorithm using the PCA800 dataset does not give a clear elbow in the graph, but a soft elbow can be seen again in k = 5. Moreover, the silhouette method also indicates that the optimal number of clusters for this algorithm is 5, while the gap statistic suggests that k = 6 is the best choice. k = 9 was also run in order to compare with the previous results.

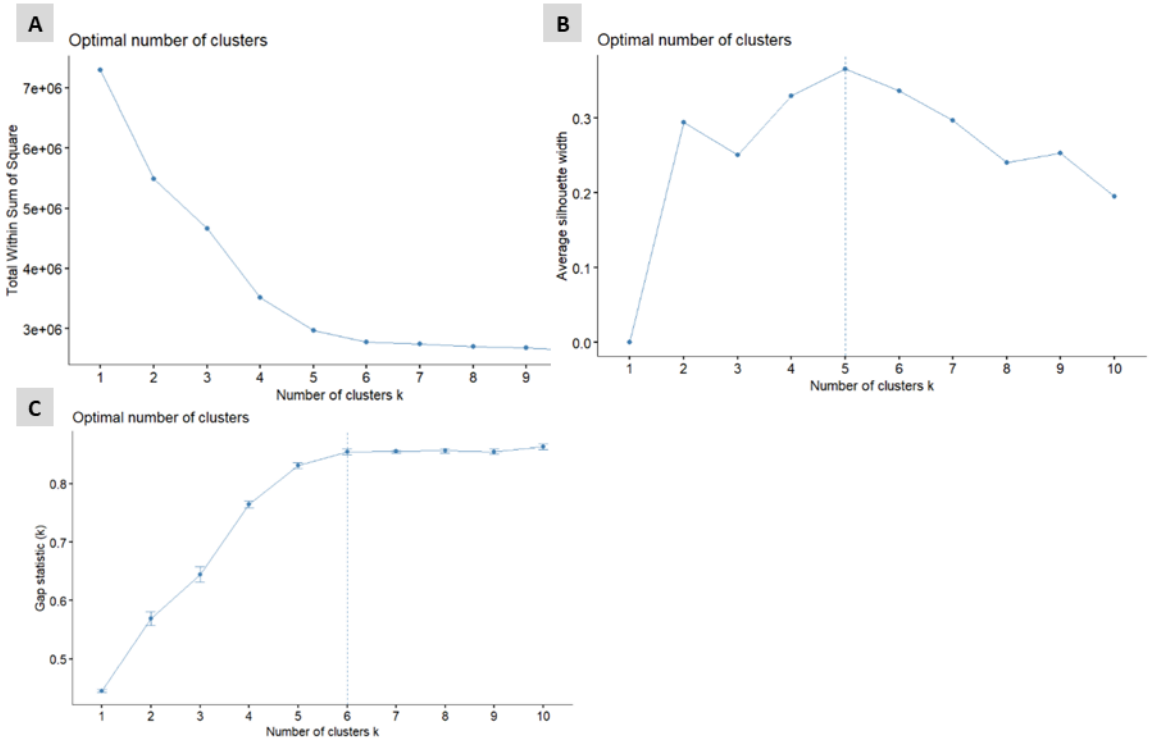


Figure 22. Graphical results of the optimal number of clusters according to the elbow method (plot A), the average silhouette method (plot B) and the gap statistic method (plot C) for the CLARA algorithm using the PCA800 dataset.

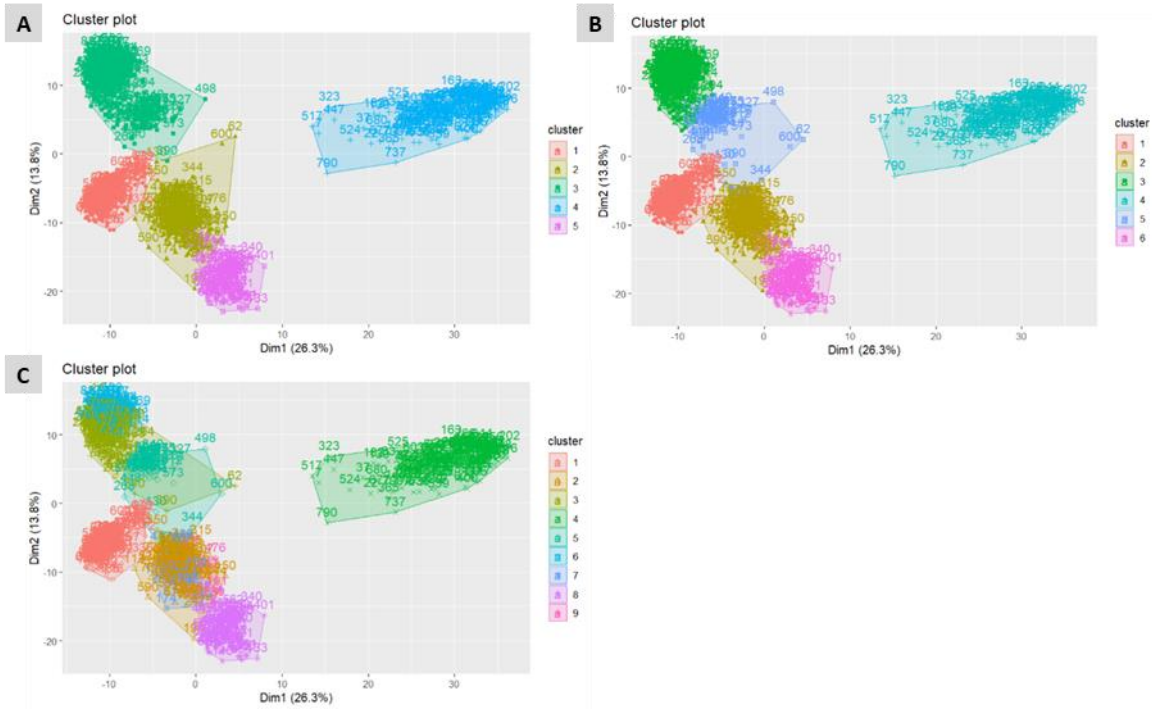


Figure 23. Graphical results of CLARA algorithm using PCA800 dataset for k = 5 (plot A), k = 6 (plot B) and k = 9 (plot C).

The CLARA algorithm for k = 5 and k = 6 seems to give similar results to the k-means algorithm for the same number of clusters respectively, with almost no overlapping clusters. However, in the case of k = 9, for this explained variance, it can be seen that all clusters, except cluster 4, are fully or partially overlapping.

Similarly to PAM, none of the results were alike to those obtained for the optimal number of clusters in the PCA95 dataset, except for k = 5 in the elbow method. Besides that, when comparing the results for k = 5 and k = 6, clear differences can be seen in the graphical representations, not only in the shape of the clusters but also in the in the size. However, these interpretations have to be considered with caution, as the variance explained is different in the two cases.

4.3.2.4. DBSCAN

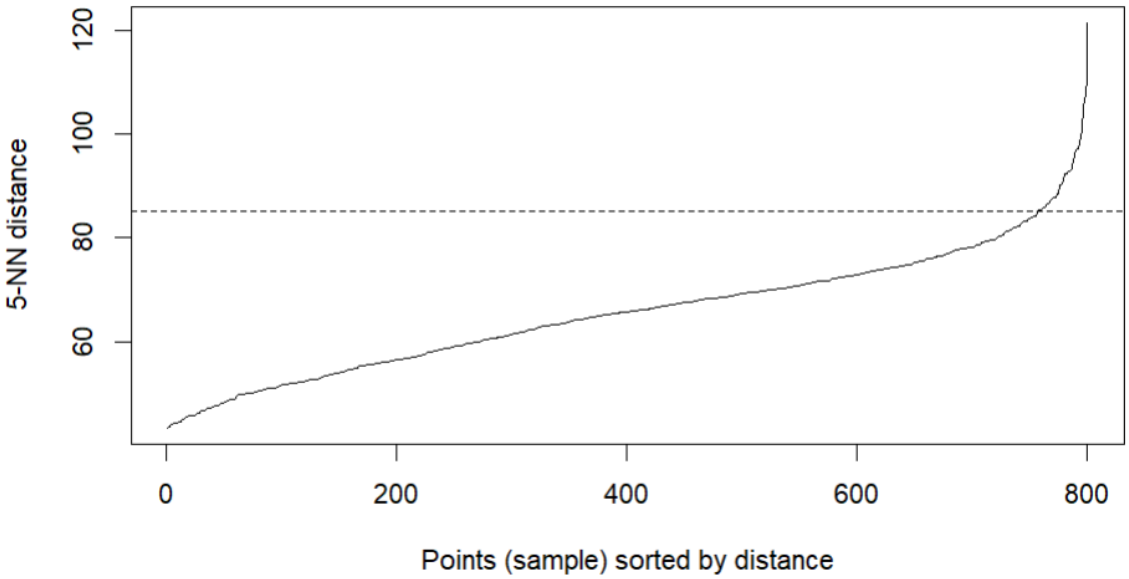


Figure 24. Representation of the average distance to the k-nearest neighbours plot for k = 5, using PCA800 dataset. A discontinuous line is shown around the point where the line forms an elbow, indicating the approximate best eps value.

As can be seen in Figure 25 and Table 5, in this case, three different results for three different values of eps are represented. All of them show five clusters to classify the data, but with different numbers of unclassified data. The first parameter combination was chosen due to the elbow shown in the average distance to the k-nearest neighbours' plot (Figure 24).
Although, the three options seem to be graphically resemblant (Figure 25), there might be differences that are not represented for this amount of explained variance. However, only two clusters appear to be partially overlap. Furthermore, it is also important to consider how the number of unclassified instances differs between the eps values, since in one of the cases it goes up to 71 (eps = 80). Although with eps = 97, this value is much lower than in with the other two eps values, this algorithm is still unable to classify 7 points which, in this context of cancer patients, could still be too high to be able to use this algorithm.
However, it is worth to note that the number of classified points has decreased considerably compared to the results obtained with the PCA95 data, since in PCA95, the number of unclassified points was always higher than 199 out of 801.

Table 5. Number of observations classified in each cluster by the DBSCAN algorithm with minPts = 5 and eps = 85, eps = 80 and eps = 97 using the PCA800 dataset. The instances that could not be classified by the algorithm are indicated in cluster number 0.

	Clusters					
eps	**0**	**1**	**2**	**3**	**4**	**5**
85	36	135	123	289	141	77
80	71	135	110	271	137	77
97	7	136	137	300	143	78

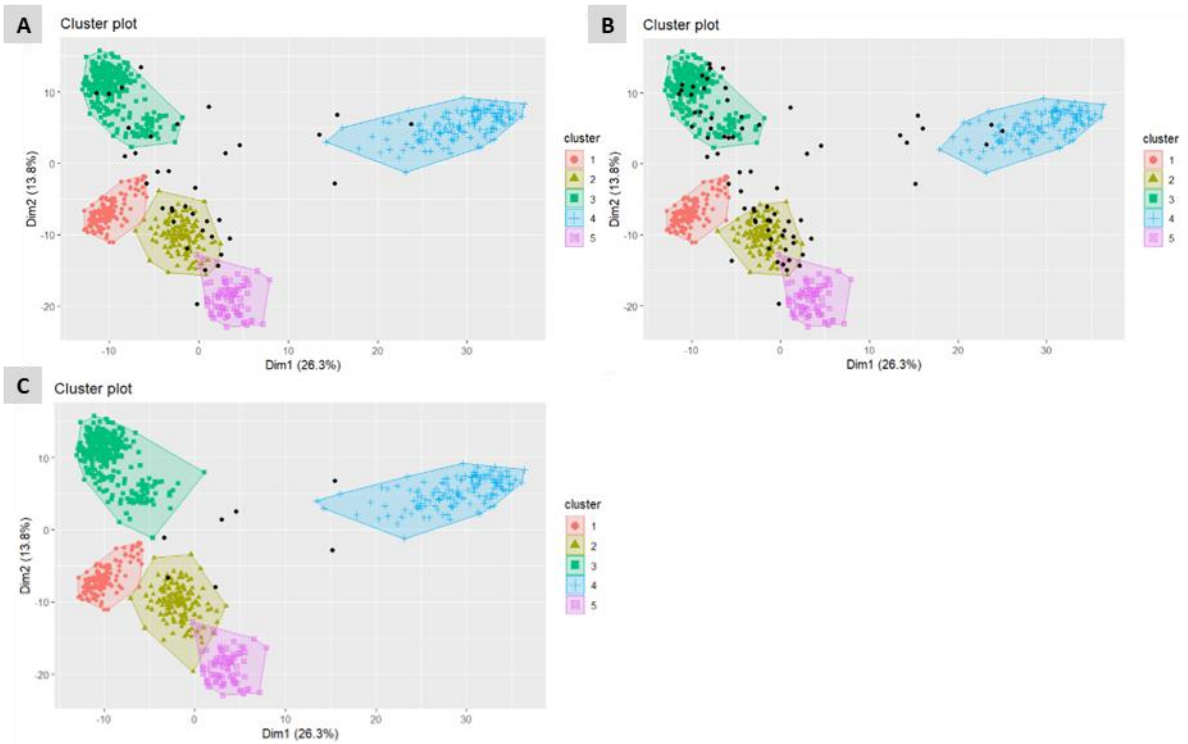


Figure 25. Graphical results of DBSCAN algorithm using PCA800 dataset with minPts = 5 and eps = 85 (plot A), eps = 80 (plot B) and eps = 97 (plot C).

4.3.2.5. Hierarchical

As with the hierarchical algorithm for the PCA95 data, the approach and method with the highest coefficient was the agglomerative with the Ward's method (Table 6). In this case, the classification was made into six different clusters, as suggested by the elbow method for the k-means algorithm, as can be seen in Figure 26. Moreover, it is worth mentioning that the agglomerative coefficient of the PCA800 dataset is higher than for the PCA95 dataset in each method and approach performed for the hierarchical algorithm.

Table 6. Agglomerative and divisive coefficients of the different hierarchical algorithms performed using PCA95 dataset.

Method	Average	Single	Complete	Ward's
Agglomerative coefficient	0.6809188	0.6039179	0.4781062	0.9663623
Divisive coefficient	0.6723683			

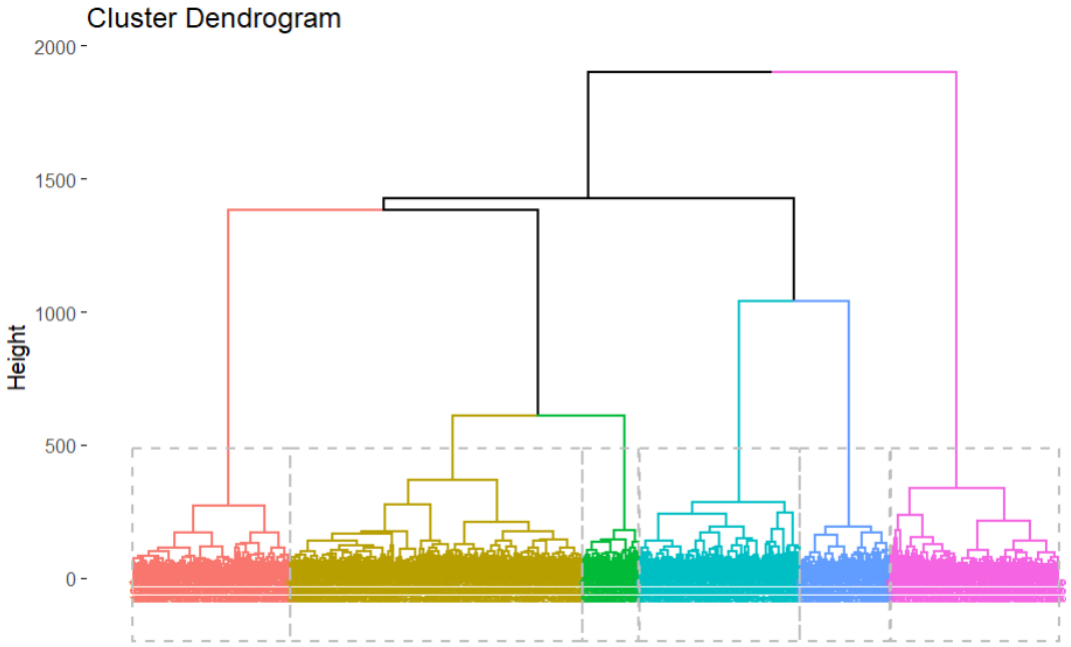


Figure 26. Dendrogram of the classification of the agglomerative hierarchical algorithm using Ward's method on the PCA800 dataset.

4.3.2.6. Gaussian mixture

In this case, as shown in Figure 27, the parameters that showed the best BIC for the Gaussian mixture model are seven clusters and EEI (the volumes of the clusters are equal, their shapes are equal, and their orientation is equal to the coordinate axes) as covariance parametrization.

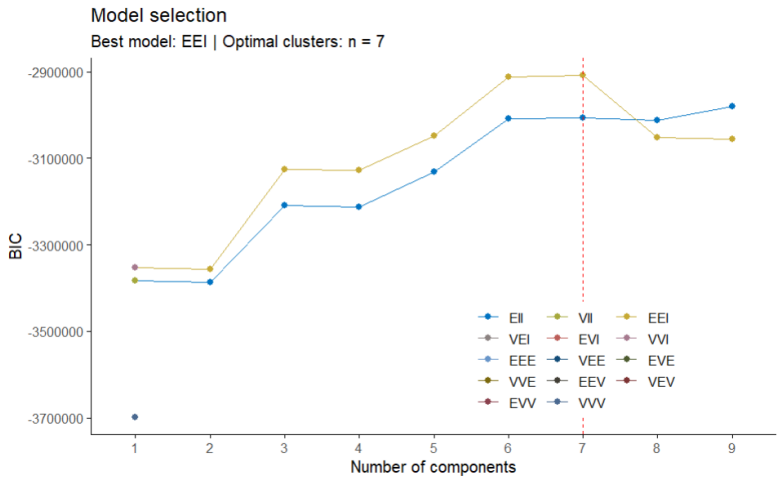


Figure 27. Representation of the BIC values obtained for the different number of clusters and the covariance parametrizations tested, particularly, EII (equal volume, equal shape, identical orientation), VII (varying volume, spherical covariance, identical orientation), EEI (equal volume, equal shape, identical orientation), VEI (varying volume, equal shape, identical orientation), EVI (equal volume, varying shape, identical orientation), VVI (varying volume, varying shape, identical orientation), EEE (equal volume, equal shape, orientation in p-dimensional space), VEE (varying volume, equal shape, p-dimensional space), EVE (equal volume, varying shape, p-dimensional space), VVE (varying volume, varying shape, p-dimensional space), EEV (equal volume, equal varying, varying orientation), VEV (varying volume, equal shape, varying orientation), EVV (equal volume, varying shape, varying orientation) and VVV (varying volume, varying shape, varying orientation) in the Gaussian mixture model using the PCA800 dataset.

In Table 7 it can be seen the distribution of the instances present in the PCA800 dataset throughout the seven clusters considered by the model. Also, the graphical representation of the model is represented in Figure 28. In this context, it is worth mentioning that two of the clusters (clusters 6 and 7 in Table 7) have only one observation within their boundaries, suggesting that this algorithm does not capture well the differences between the different groups of data. Regarding the rest of the clusters, although there are observations that are far from the borders of the clusters, they generally do not overlap, with the exception of the clusters 2 and 3.

In comparison with the result obtained for the PCA95 dataset, it can be noted that two fewer clusters are proposed as the best option for the model, but, as mentioned before, clusters 6 and 7 are only fill with one observation in contraposition with the PCA95 model in which all the clusters have more than 50 observations.

Table 7. Number of observations classified in each cluster by the EEI Gaussian Model algorithm with 7 clusters using the PCA800 dataset.

Clusters						
1	**2**	**3**	**4**	**5**	**6**	**7**
438	57	144	78	82	1	1

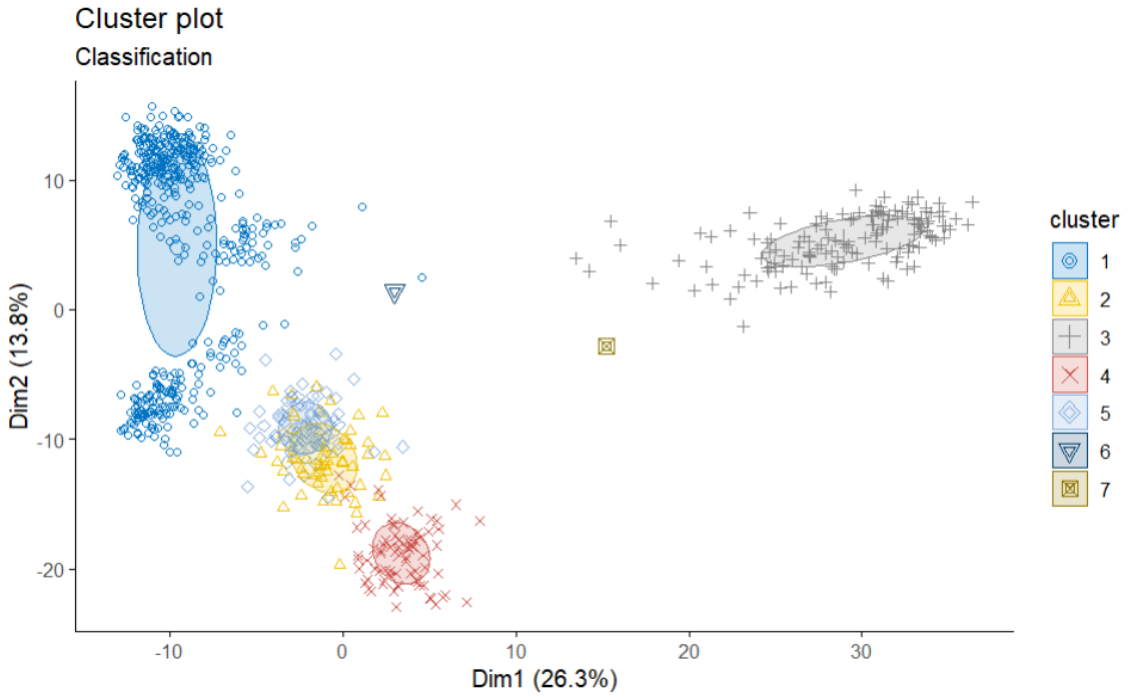


Figure 28. Graphical results of EEI Gaussian mixture algorithm using the PCA800 dataset for 7 clusters.

4.3.3. UMAP

As referred in Table 1, after performing the UMAP technique for feature subset selection, a bidimensional dataset is obtained. In Figure 29, a representation of the distribution of this dataset is shown. It can be seen that the data appears to be grouped into six distinct groups, which are widely separated throughout the space.

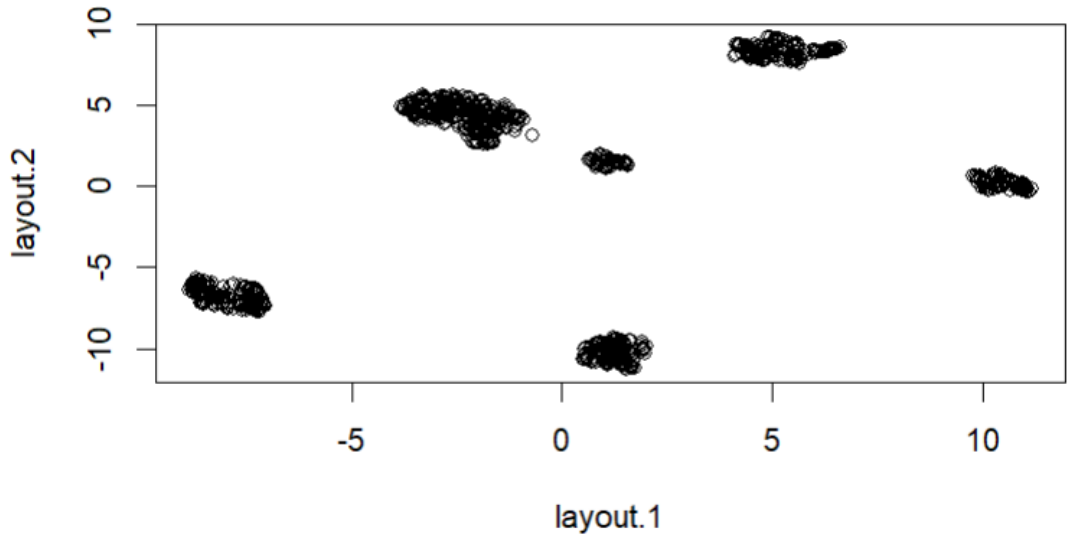


Figure 29. Representation of the distribution of the UMAP dataset.

4.3.3.1. k-means

As can be seen in Figure 30, the elbow method suggests that the best number of clusters is 5 clusters, while the silhouette and gap statistics suggest 7 clusters. Therefore, k-means was run for k = 5 and k = 7. Furthermore, k-means was also carried out with k = 6, as the plot with the UMAP output shows six differentiated groups of data, suggesting that 6 clusters might be a good parameter for the classification task.

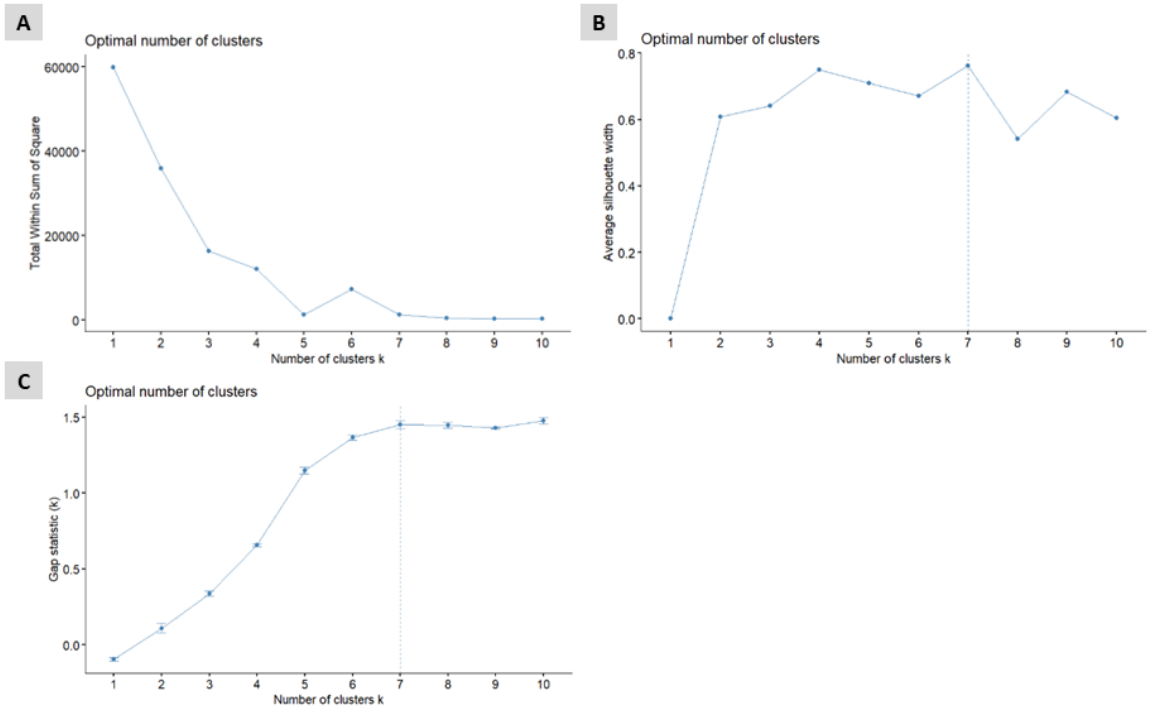


Figure 30. Graphical results of the optimal number of clusters according to the elbow method (plot A), the average silhouette method (plot B) and the gap statistic method (plot C) for the k-means algorithm using the UMAP dataset.

As expected, the number of clusters that better visually matches the groups of the UMAP data distribution, is k = 6 (plot B of Figure 31), with cluster corresponding to one group. This is also important to highlight since the algorithm was able to detect the differences between each group and separate them into different clusters. Although, all the UMAP data is represented in these graphs (Figure 31), it is important to have into account that it is a representation of the dimensional reduction which might not accurately represent the full data variance and information may be lost.

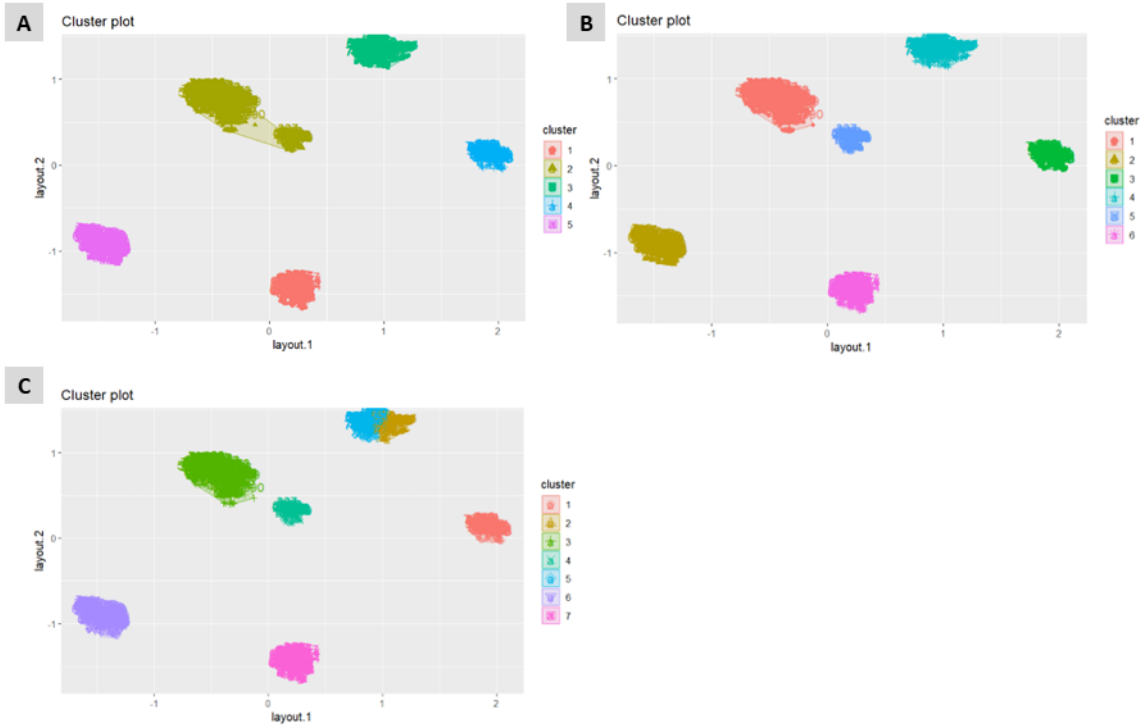


Figure 31. Graphical results of k-means algorithm using UMAP dataset for k = 5 (plot A), k = 6 (plot B), k = 7 (plot C).

4.3.3.2. PAM

As indicated in Figure 32, the optimal number of clusters suggested for this algorithm is k = 5 and k = 8. However, k = 6 and k = 7 were also performed in order to be able to compare the results with the previous algorithm.

In Figure 33, the graphical results of the PAM algorithm implemented on the UMAP dataset for the different number of k clusters is shown. It can be observed that the results for k = 5 and k = 6 (plot A and B) are visually very similar to those of the k-means algorithm. However, when comparing the two algorithms performed for k = 7 (plot C), it can be seen that the group of data that is divided in two to create the seventh cluster is different between the k-means and the PAM algorithm. Furthermore, in this algorithm, the number of clusters that visually better adjust to the groups of data is also k = 6.

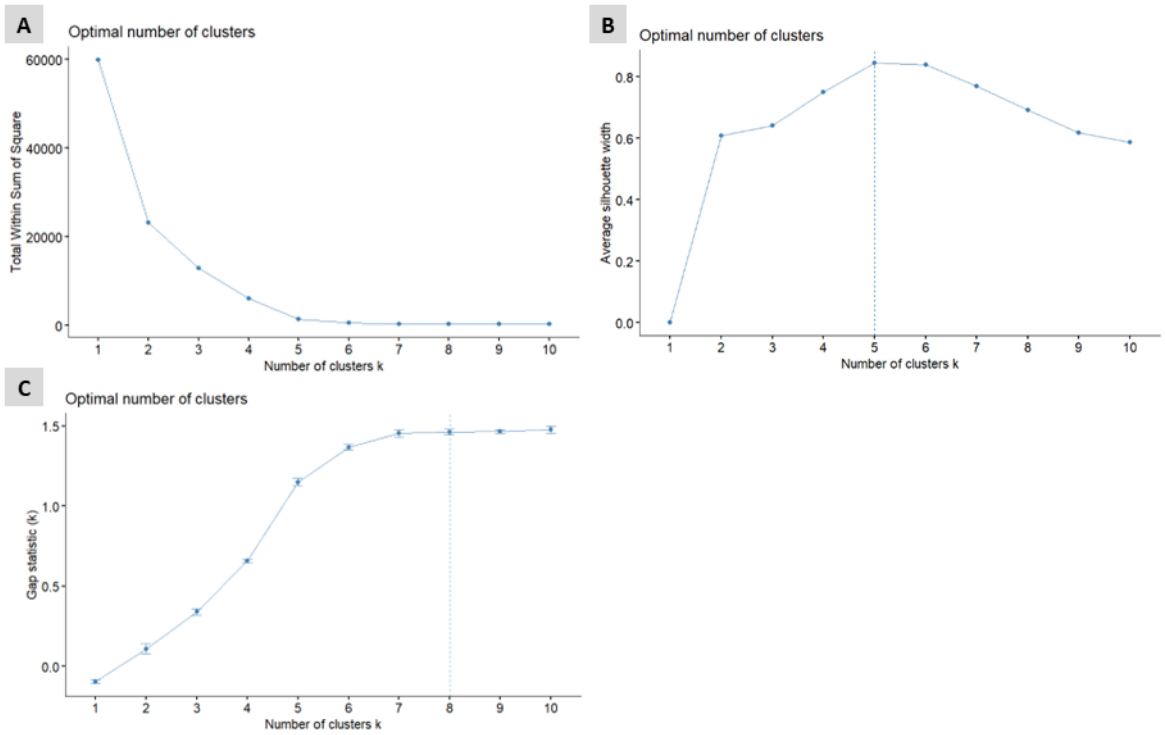


Figure 32. Graphical results of the optimal number of clusters according to the elbow method (plot A), the average silhouette method (plot B) and the gap statistic method (plot C) for the PAM algorithm using the UMAP dataset.

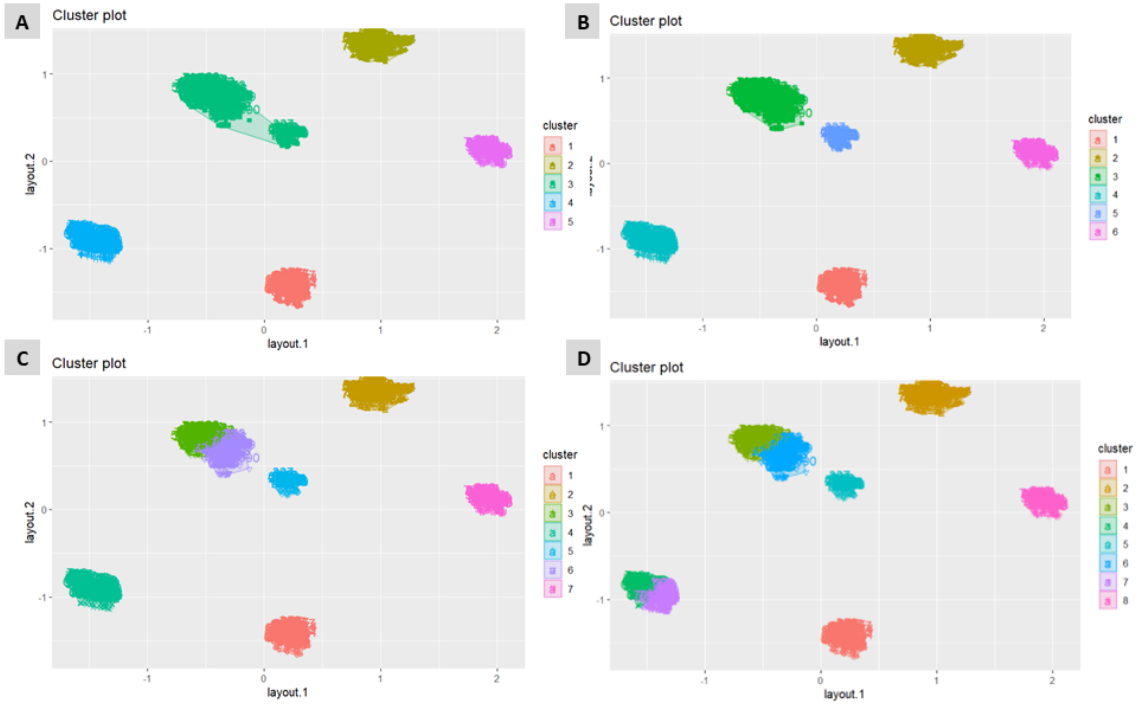


Figure 33. Graphical results of PAM algorithm using PCA95 dataset for k = 5 (plot A), k = 6 (plot B), k = 7 (plot C) and k = 8 (plot D).

4.3.3.3. CLARA

As with the PAM algorithm, the optimal numbers of clusters according to the elbow method, silhouette, and gap statistic are k = 5 and k = 8 (Figure 34).

Therefore, the CLARA algorithm was also performed with k = 6 and k = 7 in order to compare the results between the algorithms.

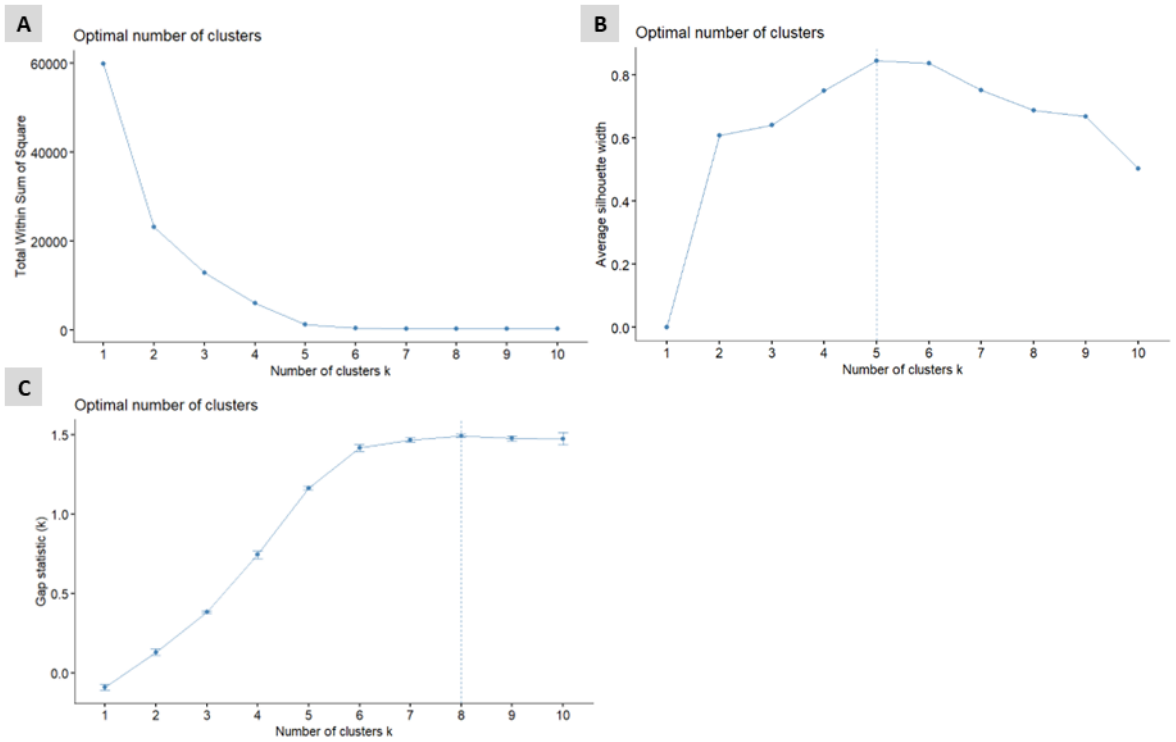


Figure 34. Graphical results of the optimal number of clusters according to the elbow method (plot A), the average silhouette method (plot B) and the gap statistic method (plot C) for the CLARA algorithm using the UMAP dataset.

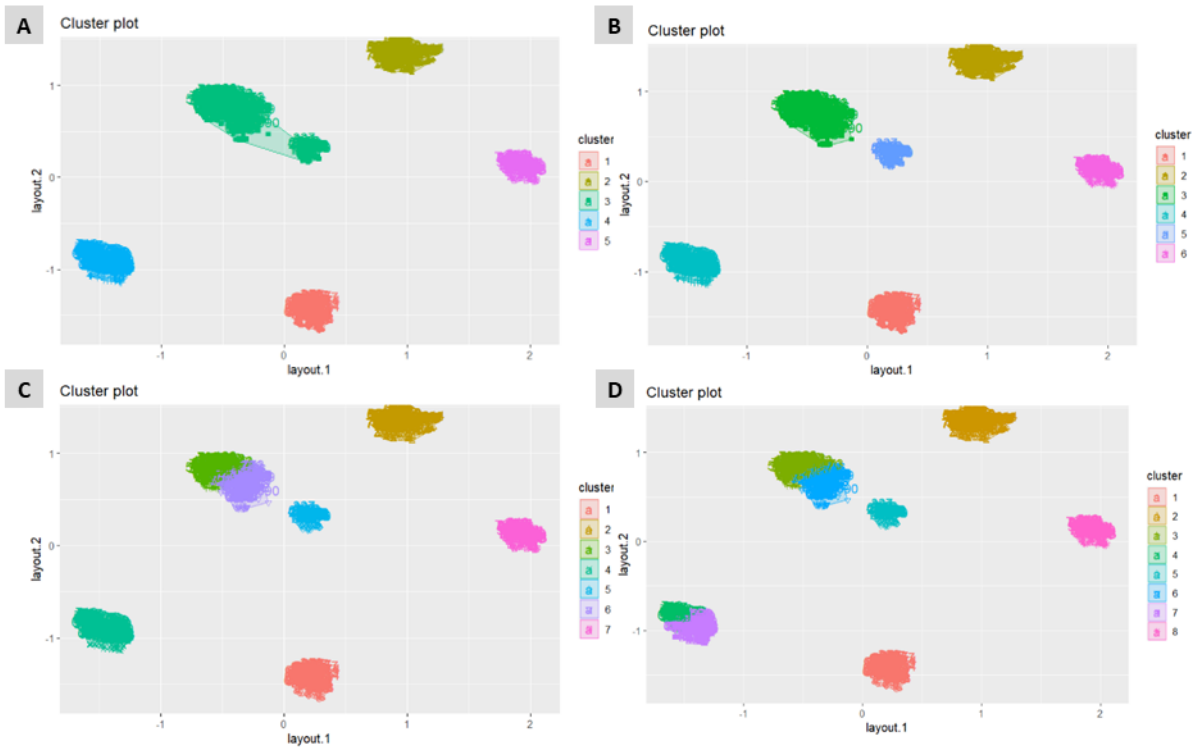


Figure 35. Graphical results of CLARA algorithm using UMAP dataset for k = 5 (plot A), k = 6 (plot B), k = 7 (plot C) and k = 8 (plot D).

As can be seen in Figure 35, the results are visually almost identical to those obtained for the PAM algorithm.

With regard to the results of the partitioning models (k-means, PAM and CLARA) implemented on the UMAP dataset, the number of clusters that the graphical representation indicates as performing a better classification of the data is when k = 6 parameter is used, since, as mentioned before, each one of the clusters corresponds to one of the groups of data displayed in the data space. However, to confirm this hypothesis, several algorithm evaluation were performed (section 4.4).

4.3.3.4. DBSCAN

The elbow of the curve of the k-nearest neighbours plot (Figure 36) shows that the approximate value of the optimal eps is 0.3, so different values around this number were used. Also, minPts was chosen to be 5, since this was the value chosen for the PCA95 dataset, and also coincides with the number of cancers types that are known to be represented in the original data.

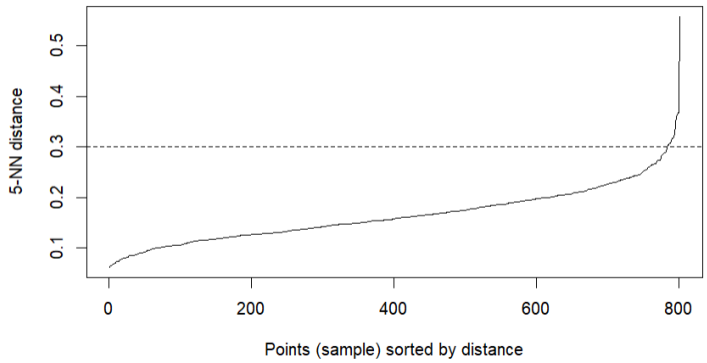


Figure 36. Representation of the average distance to the k-nearest neighbours plot for k = 5, using the UMAP dataset. A discontinuous line is shown around the point where the line forms an elbow, indicating the approximate best eps value.

As can be seen in Table 8 and in the plots in Figure 37, the algorithm with eps = 0.37 is the one that classifies the highest number of points, leaving only two unclassified. Furthermore, this parameter combination shows 6 clusters, which coincides with the data groups that can be visualized in the UMAP data distribution plot (Figure 29). In the case of the other two parameter combinations (eps = 0.3 and eps = 0.25), they classified the data into 7 and 9 clusters, leaving 12 and 44 unclassified observations, respectively. Furthermore, with eps = 0.25, a cluster with only 5 observations is generated, which seems to be a very small number to be consider as a relevant subtype of cancer.

Finally, it is worth noting that in comparison with the results of the DBSCAN algorithm in the PCA95 and PCA800 datasets, the number of unclassified data was reduced to 2 out of 801 in the case of eps = 37, while the minimum reached in the other two datasets were 199 and 7 respectively. Also, it is important to highlight that this algorithm was also able to detect the six main groups of data that can be distinguished in the UMAP data distribution plot.

Table 8. Number of observations classified in each cluster by the DBSCAN algorithm with minPts = 5 and eps = 0.3, eps = 0.25 and eps = 0.37, using the UMAP dataset. The instances that could not be classified by the algorithm are indicated in cluster number 0.

	Clusters									
eps	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0.3	12	135	139	230	145	47	15	78		
0.25	44	5	106	123	221	140	46	15	77	24
0.37	2	136	140	252	146	47	78			

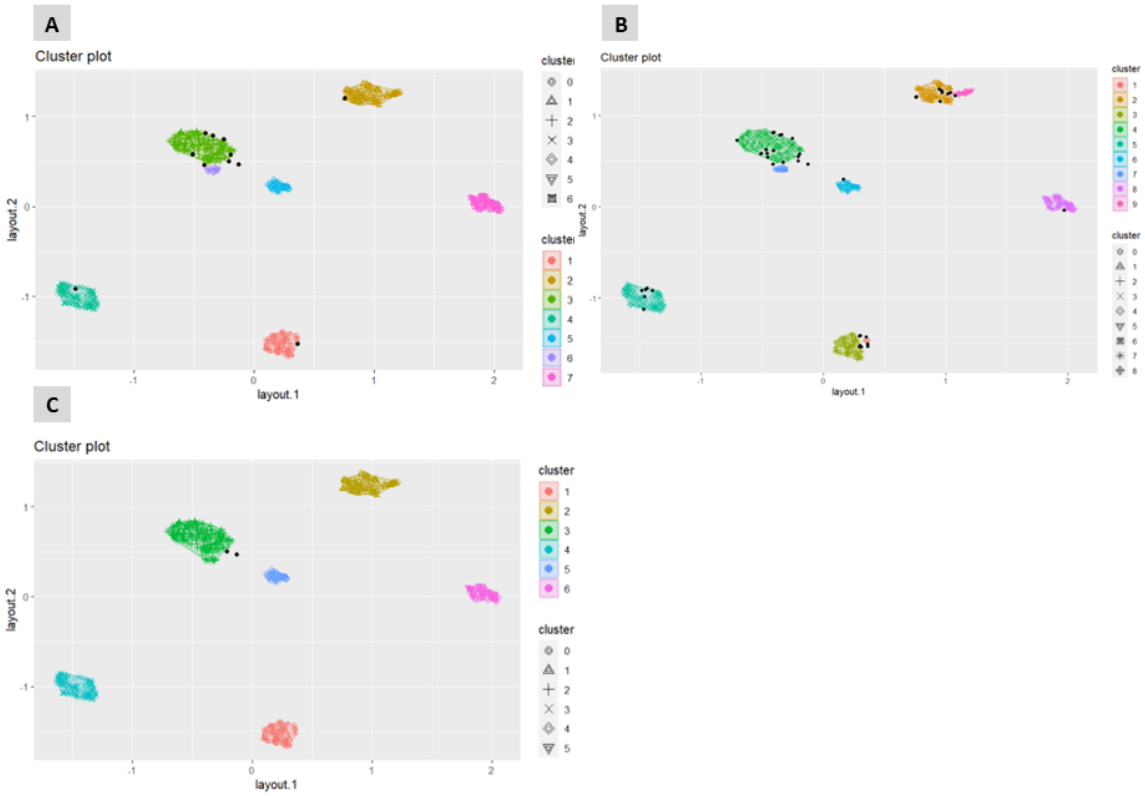


Figure 37. Graphical results of DBSCAN algorithm using the UMAP dataset with minPts = 5 and eps = 0.3 (plot A), eps = 0.25 (plot B) and eps = 0.37 (plot C).

4.3.3.5. Hierarchical

As can be seen in Table 9, the method with the highest coefficient is again the agglomerative approach following the Ward's method. Hence, this approach and method were used for the following analysis. It is important to note that this method was always the best in all cases (PCA95, PCA800 and UMAP).

Table 9. Agglomerative and divisive coefficients of the different hierarchical algorithms performed using the UMAP dataset.

Method	Average	Single	Complete	Ward's
Agglomerative coefficient	0.9962865	0.9950505	0.9934537	0.999701
Divisive coefficient	0.9958642			

Furthermore, to calculate the optimal number of clusters to separate the data, the elbow method for the k-means algorithm was used as a reference. Therefore, the separation was performed using five clusters, as can be seen in Figure 37. In this

case, the first division shown in the dendrogram separates the data into 4 groups, contrary to the PCA95 and PCA800 results where the first division was into 2 groups.

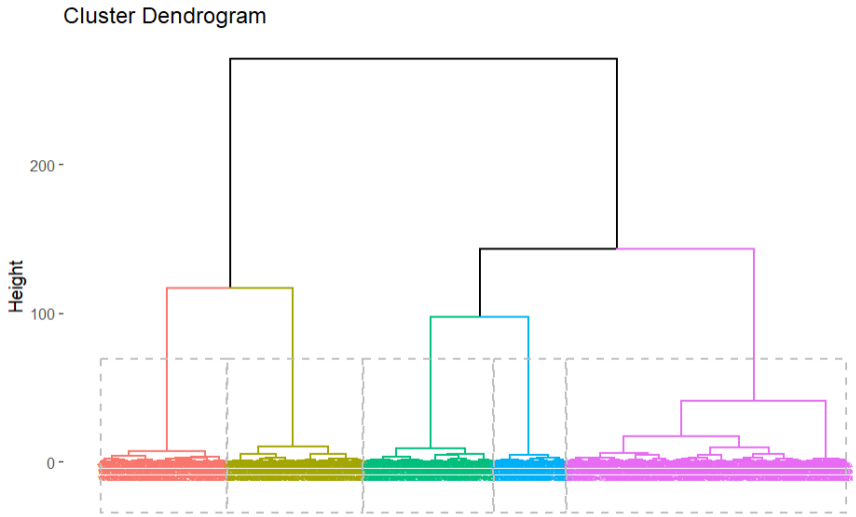


Figure 38. Dendrogram of the classification of the agglomerative hierarchical algorithm using Ward's method on the UMAP dataset.

4.3.3.6. Gaussian mixture

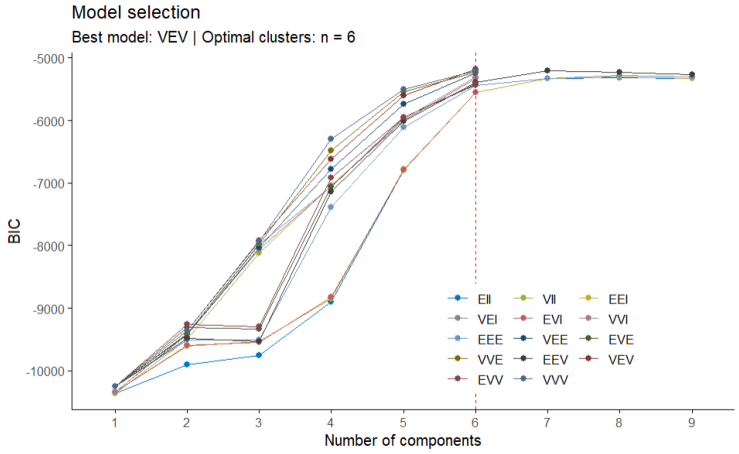


Figure 39. Representation of the BIC values obtained for the different number of clusters and the covariance parametrizations tested, particularly, EII (equal volume, equal shape, identical orientation), VII (varying volume, spherical covariance, identical orientation), EEI (equal volume, equal shape, identical orientation), VEI (varying volume, equal shape, identical orientation), EVI (equal volume, varying shape, identical orientation), VVI (varying volume, varying shape, identical orientation), EEE (equal volume, equal shape, orientation in p-dimensional space), VEE (varying volume, equal shape, p-dimensional space), EVE (equal volume, varying shape, p-dimensional space), VVE (varying volume, varying shape, p-dimensional space), EEV (equal volume, equal varying, varying orientation), VEV (varying volume, equal shape, varying orientation), EVV (equal volume, varying shape, varying orientation) and VVV (varying volume, varying shape, varying orientation) in the Gaussian model using the UMAP dataset.

As shown in Figure 39, the covariance parametrization that shown the highest BIC was VEV (the volumes of the cluster vary, their shapes are equal and their orientation also varies) with 6 clusters, which is the same number of data groups as in the UMAP data distribution representation (Figure 29).

In Table 10, the distribution of the observations into the different clusters created by the model is displayed. As mentioned above, 6 clusters were formed coinciding with the data groups of UMAP data, as shown in Figure 40. In comparison with the Gaussian mixture model representations obtained for the PCA95 and PCA800 datasets, in this case, the clusters seem to be more compact and adjusted to the data distribution.

Table 10. Number of observations classified in each cluster by the VEV Gaussian Model algorithm with 6 clusters using the UMAP dataset.

Clusters					
1	**2**	**3**	**4**	**5**	**6**
136	140	254	146	47	78

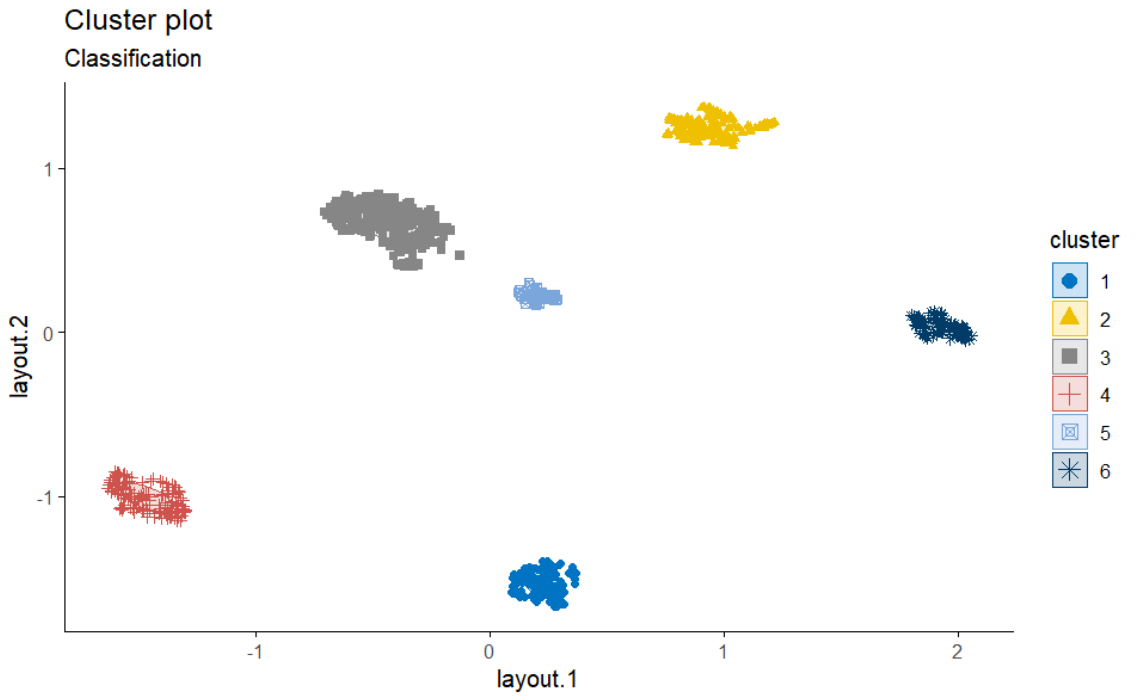


Figure 40. Graphical results of VEV Gaussian mixture algorithm using the UMAP dataset for 6 clusters.

4.4. Algorithm evaluation

After performing all the algorithms presented in the previous sections, it is important to calculate objective measurements of the quality of the clusters, as the graphs shown do not represent the whole data in most of the cases. Thus, an internal evaluation of the different algorithms was carried out (section 4.4.1). Furthermore, since in this particular case, the real classification of the data is available, a comparison between the real classification data and the algorithms' results was performed (section 4.1.2).

4.4.1. Internal evaluation

Table 11. Davies Bouldin index, Calinski-Harabasz index, connectivity, silhouette and Dunn index results for k-means, PAM, CLARA, hierarchical, Gaussian mixture and DBSCAN using the PCA95 dataset. The best results for each algorithm are highlighted in green and the best algorithm for each index is highlighted in bold.

	k	k-means	PAM	CLARA	Hierarchical	Gaussian	DBSCAN	
Davies Bouldin	k = 3	1.8002	1.6878	1.7168				
	k = 5	1.8128	1.5689	**1.5421**				
	k = 6	1.9397	1.6150	1.7908				
	k = 7	/	/	1.6531				
	k = 8	2.5839	1.8783	1.8783				
Calinski-Harabasz	k = 3	133.2549	127.3870	123.8122	/	/	eps = 161	356.6057
	k = 5	131.0755	129.8547	130.7383	130.2399	/	eps = 165	340.6976
	k = 6	116.2632	115.6746	106.9693	/	/	eps = 171	450.3961
	k = 7	/	/	98.1767	/	/	eps = 174	**669.2802**
	k = 8	91.0528	89.5897	87.6659	/	/	/	/
	k = 9	/	/	/	/	76.5445	/	/
Connectivity	k = 3	**13.2992**	43.7369	96.7381	/	/	eps = 161	585.1968
	k = 5	25.2492	36.1155	15.0623	17.1571	/	eps = 165	540.5052
	k = 6	25.5853	21.7171	211.6770	/	/	eps = 171	502.1655
	k = 7	/	/	182.2599	/	/	eps = 174	439.8532
	k = 8	229.6496	243.2925	240.7889	/	/	/	/
	k = 9	/	/	/	/		/	/
Silhouette	k = 3	0.1926	0.1731	0.1678	/	/	eps = 161	0.0914
	k = 5	0.2015	0.2025	0.2306	0.0787	/	eps = 165	0.09355
	k = 6	**0.2348**	0.2338	0.1864	/	/	eps = 171	0.1315
	k = 7	/	/	0.1927	/	/	eps = 174	0.1718
	k = 8	0.1764	0.1911	0.1777	/	/	/	/
	k = 9	/	/	/	/		/	/

	k	k-means	PAM	CLARA	Hierarchical	Gaussian	DBSCAN	
Dunn	k = 3	0.3533	0.3266	0.3421	0.5395	/	eps = 161	**0.6618**
	k = 5	0.4369	0.4226	0.4776	/	/	eps = 165	0.6459
	k = 6	0.4662	0.4535	0.3597	/	/	eps = 171	0.6256
	k = 7	/	/	0.4103	/	/	eps = 174	0.5922
	k = 8	0.3918	0.3335	0.4103	/	/		
	k = 9	/	/	/	/			

Table 12. Davies Bouldin index, Calinski-Harabasz index, connectivity, silhouette and Dunn index results for k-means, PAM, CLARA, hierarchical, Gaussian mixture and DBSCAN using the PCA800 dataset. The best results for each algorithm are highlighted in green and the best algorithm for each index is highlighted in bold.

	k	k-means	PAM	CLARA	Hierarchical	Gaussian	DBSCAN	
Davies Bouldin	k = 5	1.2221	**1.2011**	1.2037				
	k = 6	1.4123	1.3469	1.3638				
	k = 9	2.2048	1.7118	1.6167				
Calinski-Harabasz	k = 5	**290.1842**	**290.1842**	289.6865	/	/	eps = 85	346.2119
	k = 6	258.5821	258.2977	258.2577	257.2931	/	eps = 80	**396.6218**
	k = 7	/	/	/	/	118.5451	eps = 97	304.0451
	k = 9	178.5634	177.3796	171.0974	/	/	/	/
Connectivity	k = 5	50.5036	**3.7123**	**3.7123**	/	/	eps = 85	90.9441
	k = 6	50.5036	12.0948	14.2591	6.6413	/	eps = 80	182.5829
	k = 7	/	/	/	/	58.8198	eps = 97	21.3917
	k = 9	125.6694	240.0373	94.4385	/	/	/	/
Silhouette	k = 5	0.2266	**0.3656**	**0.3656**	/	/	eps = 85	0.3497
	k = 6	0.2988	0.3348	0.3364	0.3520	/	eps = 80	0.3285
	k = 7	/	/	/	/	0.2326	eps = 97	0.3583
	k = 9	0.2374	0.1896	0.3639	/	/	/	/
Dunn	k = 5	0.2798	**0.6493**	**0.6493**	/	/	eps = 85	0.7124
	k = 6	0.3002	0.4620	0.4620	0.6577	/	eps = 80	0.7475
	k = 7	/	/	/	/	0.3744	eps = 97	0.6905
	k = 9	0.3009	0.2904	0.2535	/	/	/	/

Table 13. Davies Bouldin index, Calinski-Harabasz index, connectivity, silhouette and Dunn index results for k-means, PAM, CLARA, hierarchical, Gaussian mixture and DBSCAN using the UMAP dataset. The best results for each algorithm are highlighted in green and the best algorithm for each index is highlighted in bold.

	k	k-means	PAM	CLARA	Hierarchical	Gaussian	DBSCAN	
Davies Bouldin	k = 5	0.2458	0.2523	0.2523				
	k = 6	0.2052	**0.2040**	**0.2040**				
	k = 7	0.4059	0.3784	0.3786				
	k = 8	/	0.5063	0.5087				
Calinski-Harabasz	k = 5	8638.2440	8638.2440	8638.2440	8638.2440	/	eps = 0.3	14876.7000
	k = 6	18010.2800	18010.2800	18010.2800	/	18010.2800	eps = 0.25	18262.1300
	k = 7	16330.9600	21395.3500	21366.0400	/	/	eps = 0.37	17253.4200
	k = 8	/	**21433.4100**	21078.9000	/	/	/	/
Connectivity	k = 5	**0**	**0**	**0**	**0**	/	eps = 0.3	42.6524
	k = 6	**0**	**0**	3.0718	/	**0**	eps = 0.25	120.9389
	k = 7	12.0476	13.7179	21.4313	/	/	eps = 0.37	5.0401
	k = 8	28.5917	35.4635	42.0718	/	/	/	/
Silhouette	k = 5	**0.8452**	**0.8452**	**0.8452**	**0.8452**	/	eps = 0.3	0.7132
	k = 6	0.8404	0.8404	0.8387	/	0.7656	eps = 0.25	0.4755
	k = 7	0.7695	0.7688	0.7530	/	/	eps = 0.37	0.6948
	k = 8	0.7555	0.6931	0.6885	/	/	/	/
Dunn	k = 5	**0.9606**	**0.9606**	**0.9606**	**0.9606**	/	eps = 0.3	0.1147
	k = 6	0.5337	0.5337	0.1493	/	0.0286	eps = 0.25	0.1015
	k = 7	0.0631	0.0330	0.0228	/	/	eps = 0.37	0.7589
	k = 8	0.0474	0.0234	0.0207	/	/	/	/

4.4.1.1. PCA95

The results of all the indices calculated for each algorithm run using the PCA95 dataset are displayed in Table 11. Furthermore, a summary of the best results is displayed in Table 14. It can be seen that there is no algorithm that stands out from the others, since those that obtain the best results in more than one indices (k-means and DBSCAN), they do so with different parameters, either the number of clusters (k-means with k = 3 and k = 5) or eps (DBSCAN with eps = 174 and eps = 161).

Table 14. Summary of the best results for each of the internal evaluation indices (Davies Bouldin index, Calinski-Harabasz index, connectivity, silhouette, and Dunn index) calculated using the PCA95 dataset. The number of clusters and eps value, in the case of DBSCAN, are given.

	k-means	**CLARA**	**DBSCAN**
Davies Bouldin		k = 5 (1.5421)	
Calinski-Harabasz			eps = 174 (669.2802)
Connectivity	k = 3 (13.2992)		
Silhouette	k = 6 (0.2348)		
Dunn			eps = 161 (0.6618)

4.4.1.2. PCA800

In both Table 12 (where the results of the internal evaluation indices for the PCA800 dataset are represented) and 15 (where the best results are shown), it can be observed that the PAM algorithm with k = 5 has the best results for each one of the indices. Although for some indices, there are other algorithms that reach the same value, the PAM with k = 5 is the only one that is highlighted in four out of five indices. Therefore, for PCA800 data the best algorithm in terms of internal validation, is PAM using k = 5.

Table 15. Summary of the best results for each of the internal evaluation indices (Davies Bouldin index, Calinski-Harabasz index, connectivity, silhouette, and Dunn index) calculated using the PCA800 dataset. The number of clusters and eps value, in the case of DBSCAN, are given.

	PAM	**CLARA**	**DBSCAN**
Davies Bouldin	k = 5 (1.2011)		
Calinski-Harabasz			eps = 80 (396.6218)
Connectivity	k = 5 (3.7123)	k = 5 (3.7123)	
Silhouette	k = 5 (0.3656)	k = 5 (0.3656)	
Dunn	k = 5 (0.6493)	k = 5 (0.6493)	

4.4.1.3. UMAP

In Table 13, the results of the internal validation for each algorithm and parameter used for the UMAP dataset are shown. In this case, it can be seen that several algorithms reached the same value for different indices, making it difficult to select the best algorithm in terms of the internal structure.
Furthermore, in Table 16, which is a summary of the best results obtained can be seen, it can be confirmed that several algorithms are highlighted as the best for each parameter. However, it can be observed that all the algorithms with the best indix values were run with k = 5 or k = 6, with the exception of Calinki-Harabasz which shows the best value with PAM k = 8.

Table 16. Summary of the best results for each of the internal evaluation indices (Davies Bouldin index, Calinski-Harabasz index, connectivity, silhouette, and Dunn index) calculated using the UAMP dataset. The number of clusters are given.

	k-means	**PAM**	**CLARA**	**Hierarchical**	**Gaussian**
Davies Bouldin		k = 6 (0.2040)	k = 6 (0.2040)		
Calinski-Harabasz		k = 8 (21433.41)			
Connectivity	k = 5 and k = 6 (0)	k = 5 and k = 6 (0)	k = 5 (0)	k = 5 and k = 6 (0)	k = 6 (0)
Silhouette	k = 5 (0.8452)	k = 5 (0.8452)	k = 5 (0.8452)	k = 5 (0.8452)	
Dunn	k = 5 (0.9606)	k = 5 (0.9606)	k = 5 (0.9606)	k = 5 (0.9606)	

4.4.1.4. Comparison between datasets

Therefore, comparing the results of all the indices obtained for the PCA95, PCA800 and UMAP datasets, it can be seen that, except for Davies Bouldin, which is highest in the PCA800 dataset (Table 12) with 1.2010 for the PAM algorithm with k = 5, the rest of the indices are highest in the algorithms using the UMAP dataset. Although no particular algorithm or parameter can be chosen for the UMAP dataset as the best in terms of internal evaluation, since, as mentioned above, the same value is obtained in more than one algorithm (Table 13), the UMAP technique seems to have provided a simpler data layout that allows the algorithms to create compact and well-defined clusters among the data. In addition, it was noted during the code run that UMAP was much less computationally demanding, providing results in a much shorter time than the rest of the datasets obtained by PCA.

4.4.2. Classification assessment

In order to assess the classification accuracy of the different models implemented, a scoring system was developed (described in section 3.4.2). This score system allowed to rank the algorithm according to the number of observations that were successfully classified into groups where the majority of the instances correspond to the same cancer type as that particular observation. Furthermore, this system permits the evaluation of algorithms with different numbers of clusters, since it does not take into account the number of clusters and only consider as errors those observations that are located in a cluster where the predominant type of cancer is different. In this way, this system does not penalize an algorithm that divides the data into more than five clusters, which is the number of groups in the labelled original data, allowing not to discard algorithms able to detect subtypes of cancers.

Table 17. Classification score of all the algorithms performed using the PCA95, PCA800 and UMAP datasets.

	PCA95		**PCA800**		**UMAP**	
Algorithm	Parameters	Score	Parameters	Score	Parameters	Score
k-means	k = 3	581	k = 5	798	k = 5	**800**
	k = 5	**797**	k = 6	797	k = 6	**800**
	k = 6	796	k = 9	797	k = 7	**800**
	k = 8	**797**				
PAM	k = 3	220	k = 5	798	k = 5	**800**
	k = 5	790	k = 6	798	k = 6	**800**
	k = 6	**797**	k = 9	798	k = 7	**800**
	k = 8	**797**			k = 8	**800**
CLARA	k = 3	581	k = 5	**800**	k = 5	**800**
	k = 5	796	k = 6	796	k = 6	**800**
	k = 6	793	k = 9	791	k = 7	**800**
	k = 7	795			k = 8	**800**
	k = 8	796				
DBSCAN	eps = 175	602	eps = 85	765	eps = 0.3	798
	eps = 161	447	eps = 80	730	eps = 0.25	787
	eps = 171	563	eps = 97	794	eps = 0.37	755
	eps = 165	490				
Hierarchical	k = 5	792	k = 6	799	k = 5	**800**
Gaussian mixture	k = 9	775	k = 7	663	k = 6	**800**

The output number of matches with the labels of the original data for each algorithm are available in the supplementary materials, along with the graphical representation of the classification carried out by the remaining algorithms which are not presented in the following pages [37]. In Table 17, the results of the penalty score system are presented. It can be observed that in the case of the PCA95 dataset, four algorithms reached the same maximum punctuation (797 out of 801), particularly, k-means with k = 5 and k = 797 and PAM with k = 6 and k = 8. In the case of k-means, it can be seen in Figure 41, that for the algorithm with five clusters, all the BRCA patients are classified in the same cluster as well as all the PRAD patients. However, one KIRC and two LUAD are located in the

cluster corresponding to the BRCA patients. Thus, this algorithm had four observations considered as errors. Regarding the same algorithm but run with eight clusters, the BRCA patients are divided into three different clusters, two of them accounting for 124 patients and the other one with 52 patients. The KIRC patients are also divided into two different clusters, one with 84 patients and the second with 62 patients. Furthermore, 77 out of 78 COAD patients were classified in the same cluster, but the remaining one was classified in the cluster corresponding to the LUAD patients, and at the same time, all the LUAD patients were placed into the same cluster, except for three who were positioned in the BRCA cluster with 52 patients. Finally, all the PRAD patients were classified into the same cluster. Therefore, this algorithm also showed four misclassified observations.

In the case of the PAM algorithm with k = 6, the Figure 41 shows that the extra cluster observed is given by the division of the BRCA patients cluster into two clusters containing 252 and 48 patients. Furthermore, all the KIRC were classified into the same clusters and all the PRAD patients were also placed into only one cluster. However, one KIRC patient and one LUAD patient were classified into the BRCA cluster of 252 patients, while two other LUAD patients were placed in the BRCA cluster of 48 patients. Concerning the PAM algorithm with k = 8, the BRCA patients were again divided into three clusters but with a different quantity of patients per cluster compared to the k-means with k = 8 (with clusters of 162, 90 and 48 patients). Furthermore, the KIRC patients were also divided into two clusters of 84 and 62 patients. Moreover, four patients were also misclassified in this case, specifically, one COAD and one LUAD patients were classified into the BRCA cluster with 162 patients and two LUAD patients into the BRCA cluster with 162 patients. It is worth noting that in both k-means and PAM algorithms with k = 8, the same clusters were divided into the same number of clusters, although they contained different numbers of patients per cluster.

It is important to mention that, as can be seen in Figure 41, all algorithms were able to create clusters where homogeneity of cancer types predominated, i.e. we did not find clusters that have the same approximate number of patients with different cancers or suggesting that the division of the data into clusters was random.

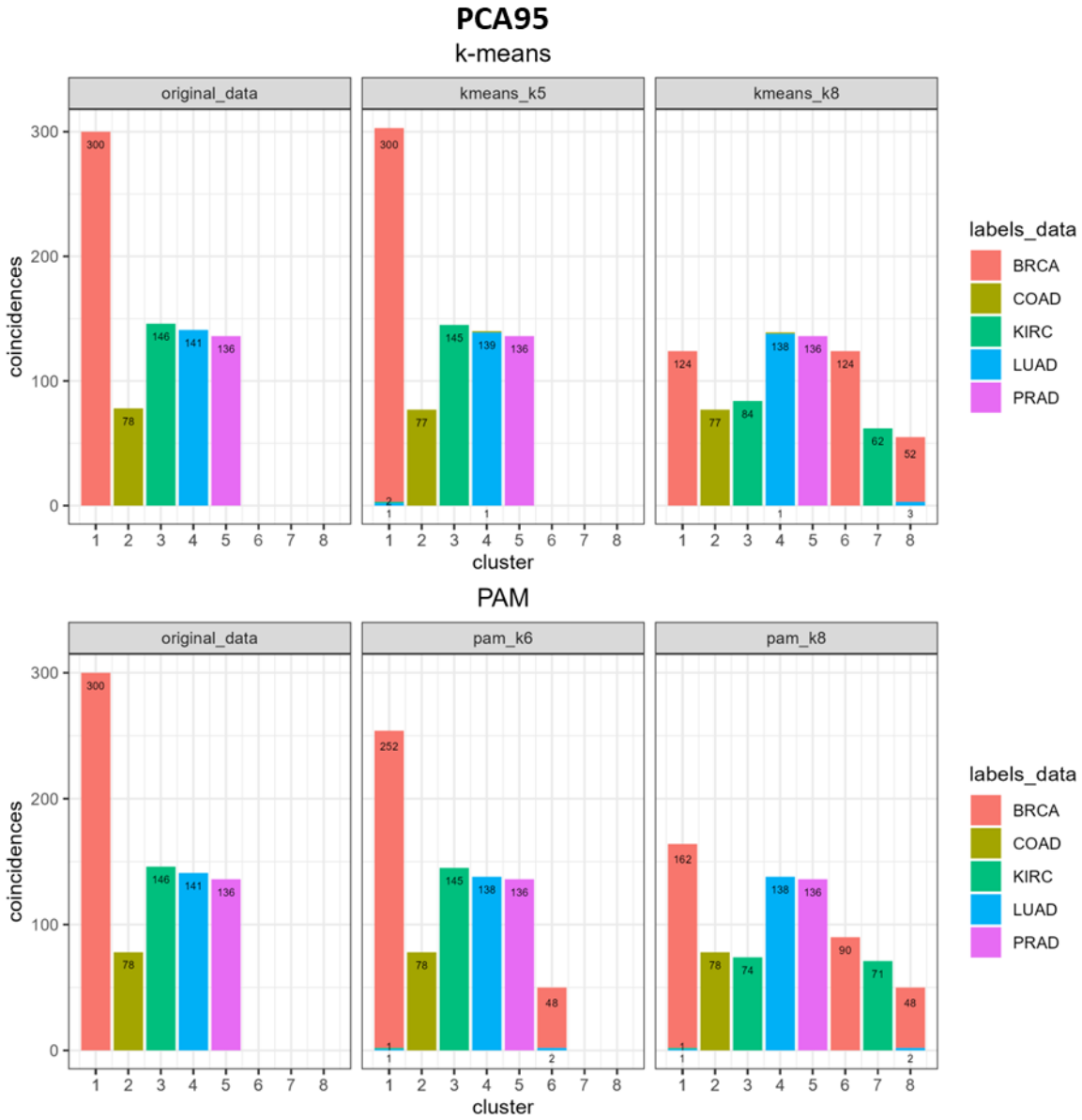


Figure 41. Comparison between the classification of the original data and the classification of the algorithms performed using the PCA95 dataset that obtained the best classification scores, specifically, k-means algorithm with k = 5 and k = 8, and PAM algorithm with k = 6 and k = 8.

Regarding the PCA800 data, as can be seen in Table 17, the best classification score was reached by the CLARA algorithm with five clusters, achieving a score of 800 points. As represented in Figure 42, all BRCA, COAD, LUAD and PRAD were classified as in the original data. However, in the case of the LUAD patients, although 145 were located into the same cluster, one of them was placed into the same cluster as the LUAD patients. Therefore, this algorithm has only one classification error, having a higher classification score than the algorithms implemented using the PCA95 dataset.

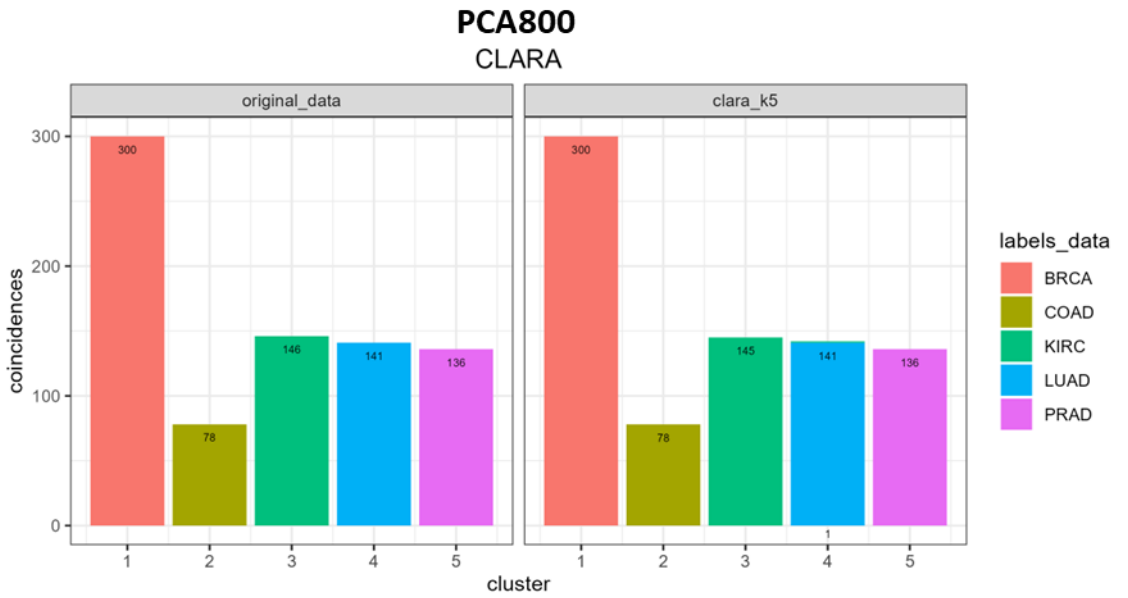


Figure 42. Comparison between the classification of the original data and the classification of the algorithm performed using the PCA800 dataset that obtained the best classification scores, specifically, CLARA algorithm with k = 5.

Regarding the UMAP dataset, as can be seen in Table 17, all of the algorithms tested had the same score, reaching 800 points, except for the DBSCAN algorithms which have lower scores. Therefore, they all obtained the same score as the CLARA k = 5 algorithm using the PCA800 dataset. It is important to mention that all the algorithms implemented using the UMAP dataset with k = 5, specifically, k-means (Figure 43), PAM (Figure 44), CLARA (Figure 44) and agglomerative hierarchical with Ward's method (Figure 45) algorithms, showed the same classification results. These algorithms classify all the observations as in the original labelled data, with the exception of one observation corresponding to a LUAD patient, which is classified in the same cluster as the BRCA patients. Furthermore, the same classification pattern is observed for the algorithms implemented with k = 6, in particular, k-means (Figure 43), PAM (Figure 44), CLARA (Figure 44) and Gaussian model mixture. In these cases, the BRCA patients are divided into two clusters, one of 254 patients and the other with 46 patients. Furthermore, they have in common with the algorithms with k = 5 that one LUAD patient is misclassified, in this case, also grouped with BRCA patients, specifically, into the second BRCA cluster.

In the case of the algorithms performed with k = 7 and k = 8, the results vary between algorithms, however they all share that one LUAD patient is classified into a BRCA cluster, as in the previous algorithms. In the case of the k-means algorithm with k = 7, the BRCA patients are still placed into two different clusters (one with 254 observations and the other with 45, plus the misclassified LUAD patient) and also, the LUAD patients are divided into two clusters with 87 and 53 instances. However, in the case of PAM and CLARA with k = 7, the only cancer type that it is separated into various clusters is BRCA, being divided into three clusters in both algorithms, while each of the remaining cancer types are classified into one cluster per type. Furthermore, in PAM and CLARA with k = 8, the BRCA and KIRC observations are again separated into three and two clusters, respectively, while the rest of the types of cancer patients are placed into a cluster per cancer type. Therefore, it seems that the way that the patients

are divided into clusters is similar between algorithms and it depends on the number of clusters. Also, it can be noted that the group of patients that appeared to be separated before the others is BRCA, which is also the group with a larger group of patients.

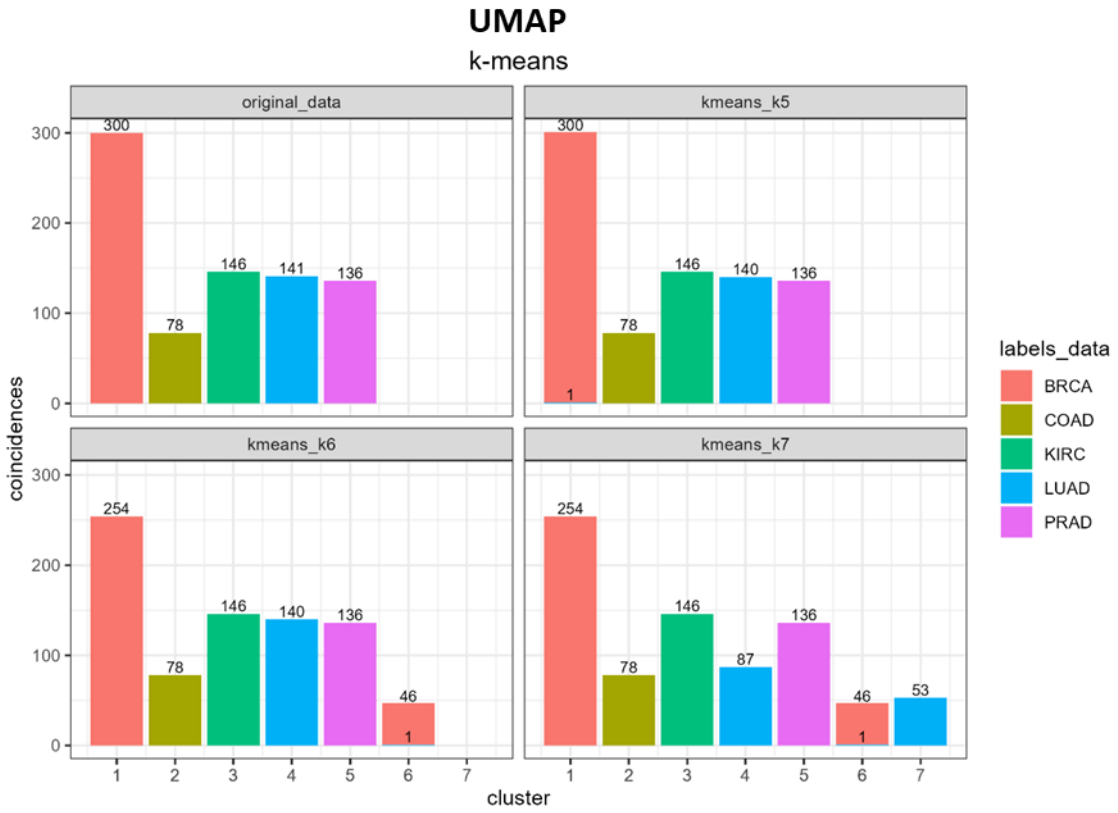


Figure 43. Comparison between the classification of the original data and the classification of the k-means algorithms performed using the UMAP dataset that obtained the best classification scores, specifically, k-means with k = 5, k = 6 and k = 7.

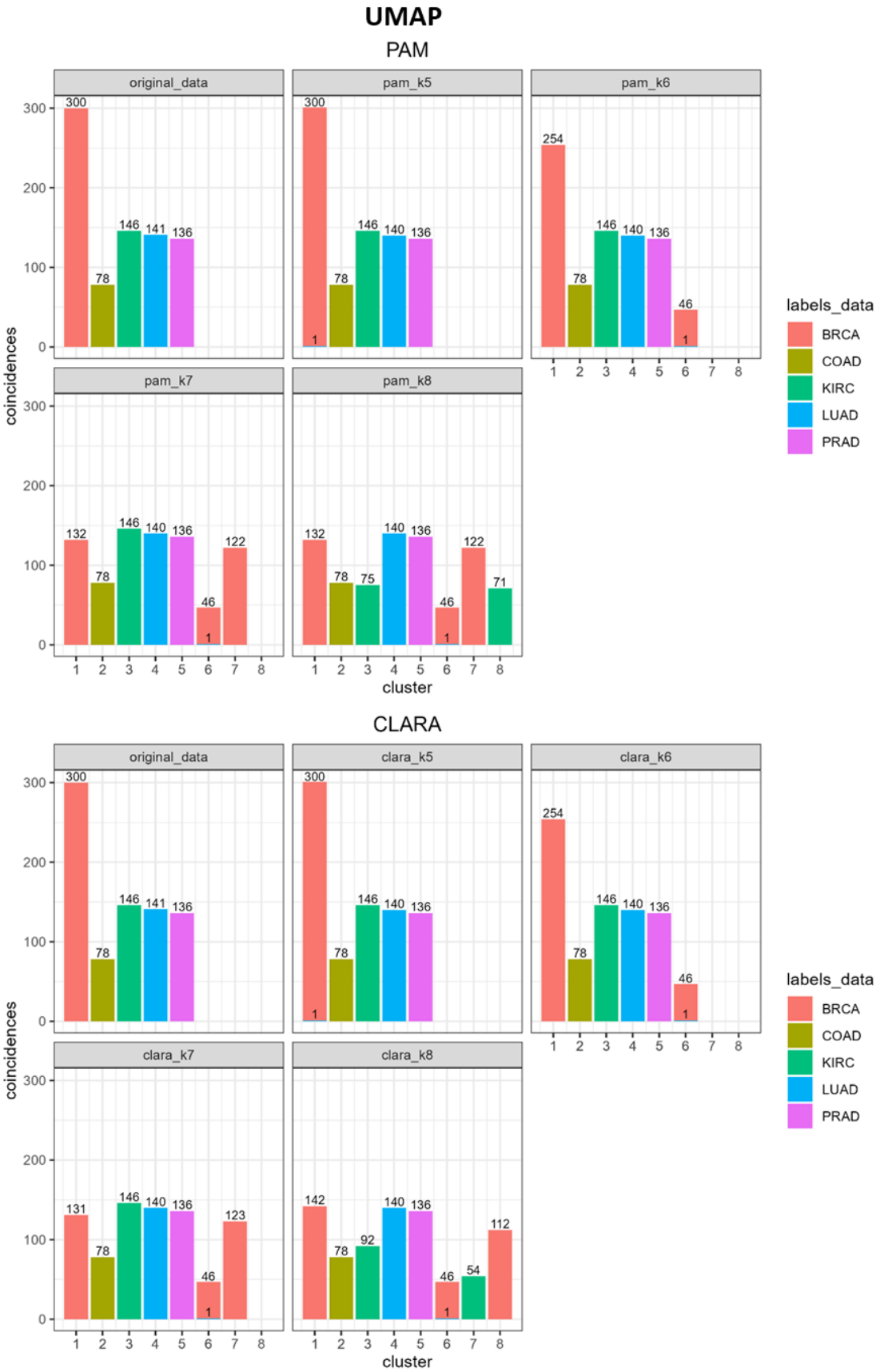


Figure 44. Comparison between the classification of the original data and the classification of the PAM and CLARA algorithms performed using the UMAP dataset that obtained the best classification scores, specifically, both PAM and CLARA algorithms with k = 5, k = 6, k = 7 and k = 8.

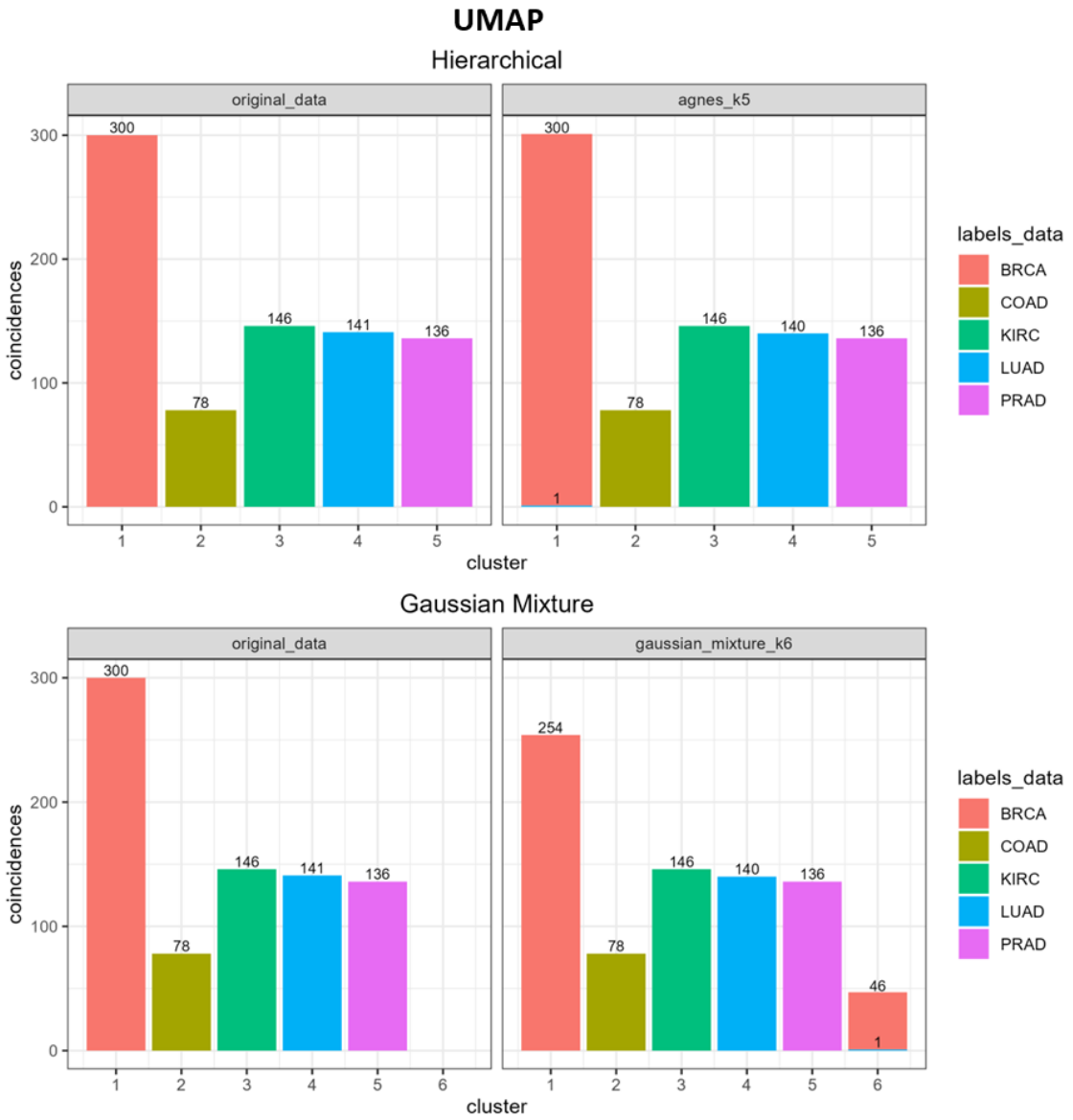


Figure 45. Comparison between the classification of the original data and the classification of the hierarchical and Gaussian mixture algorithms performed using the UMAP dataset that obtained the best classification scores, specifically, hierarchical with k = 5 and Gaussian mixture with k = 6.

4.4.2.1. Comparison between datasets

Regarding the classification accuracy evaluation of the three dimensionally reduced data (PCA95, PCA800 and UMAP), it was found that the PCA95 dataset is the one with lower scores, since both the PCA800 and UMAP datasets reached 800 out of 801 correctly classified observations in at least one of the algorithms. Furthermore, it is worth mentioning that using the UMAP technique to reduce the dimensionality, several of the algorithms reached 800 correctly classified observations for different parameters, but, in the case of the PCA800 dataset only CLARA with k = 5 had this score value. Besides that, several of the algorithms that had the best classification score using the UMAP dataset, also had the best values in different internal evaluation indices, pointing them as the best

algorithms tested, namely, k-means, PAM, CLARA and the agglomerative hierarchical algorithm with k = 5.

4.4.3. Stability evaluation

Considering the previous results (internal validation and classification score), the stability of the best algorithm was tested with the aim of differentiating the best algorithms in order to select one. The UMAP results were chosen for this section because this technique is less computationally intensive than the PCA800 and the classification score was the same as for UMAP. Thus, considering that k-means, CLARA, PAM and hierarchical with k = 5 had the best results in the internal validation, being the only algorithms that had the best score in three different indices, and also, had 800 points in the classification score, their stability was tested.
As can be seen in Table 19, all the algorithms show the same exact values for all the stability parameters tested. Therefore, the previous results and the stability measurements suggest that these algorithms have the same performance in clustering the gene expression data of cancer type of patients.

Table 18. Stability results of the best algorithms in terms of internal evaluation and score classification, specifically, k-means, PAM, CLARA and hierarchical algorithms with k = 5, using the UMAP dataset.

	k = 5			
Algorithm	**APN**	**AD**	**ADM**	**FOM**
k-means	0.0990	2.1467	1.2738	2.2151
PAM	0.0990	2.1467	1.2738	2.2151
CLARA	0.0990	2.1467	1.2738	2.2151
Hierarchical	0.0990	2.1467	1.2738	2.2151

4.4.4. Comparison of the original data distribution separated within the clusters obtained by the best algorithm

Since it was not possible to differentiate between the four best algorithms using the internal evaluation, classification accuracy and the stability, the k-means algorithm was chosen for this step, as it is more frequently highlighted in the literature as the most suitable algorithm for this type of data [18,28]. Furthermore, since the four algorithms had exactly the same performance and the same classification result, the result of the k-means could be extrapolated to the others.

In this context, Figure 46 shows the distribution of each the mean and the median of each cluster is represented. As can be seen in the figure, all the clusters appear to have very similar distributions, suggesting that there are no significant differences between them. The only cluster whose interquartile range is further apart from the rest is cluster 4. However, looking at the scale of the figure, the difference between this cluster and the others is also very small.

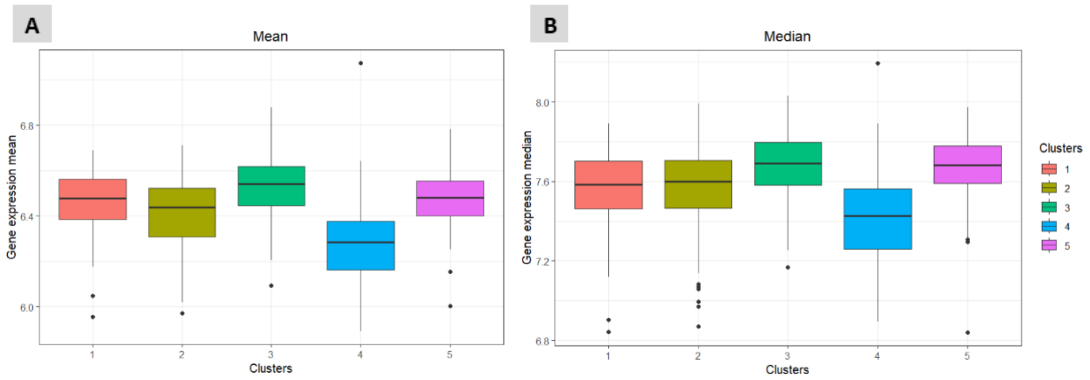


Figure 46. Representation of the distribution of the mean (A) and the median (B) of the original data gene expression of the patients within each cluster obtained by the algorithms k-means with five clusters.

5. Discussion

Cancer is a multifactorial disease that causes millions of deaths each year and is expected to double in incidence over the next fifty years [1,2]. The complexity and diversity of cancer complicate the study its biology, delays the discovery of effective cures and treatments, as well as complicates the identification and classification of the different cancer types and subtypes [1,9]. In response to these challenges, several projects have emerged with the aim of finding ways to accelerate the cancer research [10]. NGS in combination with machine learning has proven to be of great relevance to the cancer molecular study and comprehension, particularly, in the study of its transcriptome [13]. Among machine learning techniques, unsupervised machine learning has gained attention because of its ability to find hidden associations between molecular disorders based on the gene expression characteristics rather than predefined labels [15–18]. In this context, an RNA sequencing dataset containing the gene expression information of cancer patients with different types of cancer was employed in order to identify an unsupervised machine learning model capable of recognizing cancer types and/or subtypes within the data. For this porpoise, several dimensionality reduction approaches and unsupervised algorithms were tested.

During this work, it was demonstrated that unsupervised machine learning algorithms are capable of finding underlying similarities in gene expression within each originally labelled type of cancer, as several algorithms successfully distinguished and separated the cancer types present in the original labelled data with a high percentage of accuracy. Specifically, k-means, PAM, CLARA (partitioning models) and agglomerative hierarchical algorithms with k = 5, using the UMAP technique outcome, were able to correctly classify 800 out of 801 observations (99.875% of accuracy), while also generating compact and well-defined clusters (internal evaluation).
However, despite the efforts to select a single method to perform this clustering task, none of the evaluation methods carried out could differentiate between these algorithms. In order to choose between k-means, PAM, CLARA and hierarchical algorithms, it would be necessary to test different datasets in order to determine which algorithm is better adapted to general cases, being able to perform a successful classification in any dataset in which it is implemented. It is crucial not only to be able to classify this particular dataset but also, to ensure a good performance when the algorithm is applied to diverse data sets. Additionally, by applying the algorithms in different datasets, it would be possible to determine whether the recurrent misclassification of one LUAD patient by these algorithms is an isolated error or if they are unable to correctly detect the differences between certain types of cancer, particularly between BRCA and LUAD cancers. Although the algorithms make only 1 error out of 801 observations, in the context of cancer, this error can be fatal. Therefore, it is essential to verify if this error could be repeated on a larger scale, whether it occurs recurrently or if, on the contrary, this particular patient exhibited characteristics that were out of the ordinary within that particular type of cancer, leading the algorithm to reach an incorrect conclusion.
Nevertheless, it is worth noting that the fact that different algorithms produce the same results suggests that these algorithms do not depend heavily on the

intrinsic characteristics of each model, and also the existence of patterns in the data that allow the algorithms to classify the observations correctly, demonstrating once again that gene expression data are a key tool in the study of cancer.

Furthermore, this study found that the UMAP technique was generally more efficient than PCA approaches. The UMAP technique produces a bidimensional matrix of data that allowed the algorithms under study to reach a conclusion from the data faster than the PCA outcomes, implying that its use is less computationally expensive. Furthermore, the UMAP dataset had the best results in both the internal evaluation and the classification accuracy assessment, consistently outperforming PCA in several evaluation indices for almost all the algorithm tested. These observations were expected since, according to Dorrity M.W. et al. (2020), the recently developed UMAP technique is more sensitive in performing the dimensional reduction of data while preserving the patterns and structure than PCA, when applied to gene expression data [40]. Furthermore, Yang Y. et al (2021) support this notion by comparing various dimensionality reduction methods, including PCA and UMAP, across 71 large transcriptomic datasets. Their study concludes that UMAP had the superior performance, allowing the creation of clusters with biological and clinical significance [51]. Moreover, it has been proven that UMAP technique provides faster run times and higher reproducibility and accuracy than PCA, further supporting the findings of this work [52]. Regarding the two approaches carried out with the PCA technique, it is important to highlight that although the variables within the PCA800 dataset only explained the 25% of the total variance, this dataset generally had better results in terms of internal structure and precision score than the PCA500 dataset, in which the 95% of the total variance is explained, containing a much larger number of variables. This suggests that the majority of the variables in the original dataset do not provide useful information about the differences between cancer types or subtypes, and only a small percentage of the genes contribute to the identification of those differences. In line with these observations, as mentioned in section 4.4.4, after analysing the distribution of each cluster from the k-means algorithm in the original data, it was found that all the clusters appeared to have the same distribution. However, while executing the algorithms, distinct differences were observed among the clusters, which appeared well-defined and clearly separated. Therefore, this fact again suggests that only a small subset of variables among the 20531 variables present in the original data, possesses the ability to distinguish the patients in their cancer types, while the majority of variables might obscure the differences between patients. This underscores the importance of applying dimensionality reduction techniques. Furthermore, reducing the number of dimensions allows to move a step closer to the discovery of biomarkers that could allow the study of the differences between patients and the identification of cancer types directly from the original data, by focusing on only the most relevant genes from the beginning of the analyses.

Moreover, it is worth mentioning that there were algorithms that also accurately classified the data with similar internal validation values and equally high score classifications but placed the data into six clusters. However, the algorithms that were selected as the best ones using five clusters instead of six, were chosen because they performed best on more evaluation indices.

Nevertheless, it is important to emphasize that when the distribution of the UMAP dataset was plotted, six groups of data could be seen. Furthermore, the sixth group of data was noticeably smaller than the others, being represented close to the largest one, which might imply that this small group is more similar to the larger group than to the rest of the clusters. Also, given the information that the original labelled data provides, it seems likely that the largest group corresponds to the BRCA patients since it is the most represented cancer type in the original dataset. Considering the algorithms that showed better classification scores and that separate the data into 6 clusters, all of them divide the BRCA cluster into two clusters, one with 254 patients and the other with 46 (plus one misclassified LUAD patient). Therefore, this suggests that the smaller group shown in the plot is formed by patients suffering from a cancer subtype of BRCA cancer, but it is important to perform further analysis to confirm this hypothesis. In this regard, the separation of this cluster of BRCA patients into two, could be due to the fact that the number of BRCA patients in the data is higher than in the rest of cancer types and therefore, being a larger group, it is expected to have a higher variance among the data, making it more likely to be separated into different clusters. Thus, it is crucial to analyse the characteristics of the members of the extra cluster and to study in detail the molecular differences from the rest of the subjects in the main BRCA cluster before concluding that a cancer subtype had been found. To do so, it is necessary to identify differentially expressed genes between the two clusters, using a combination of differential expression analysis techniques and the metadata available in TGCA files [6] and to perform gene set enrichment analysis, such as Gene Ontology enrichment analysis and pathway analysis, in order to understand the biological processes underlying the observed differences in gene expression [53,54]. In addition, the analysis of more BRCA expression data using the same algorithms could allow confirming or rejecting this hypothesis.

6. Conclusions and future perspectives

The present work allowed the successful classification of gene expression data from cancer patients into different cancer types, achieving a high percentage of accuracy into well-defined and separated clusters. However, determining a single unsupervised algorithm as the optimal method proved challenging since several algorithms achieved the same performance, namely, k-means, PAM, CLARA (partitioning methods) and agglomerative hierarchical algorithms, all of them implemented for five clusters. Moreover, although further analyses are required to confirm the hypothesis, the models implemented suggest the presence of a subtype of BRCA cancer within the data under study. Furthermore, it was possible to verify that among the dimensionality reduction methods, UMAP technique was the one that gave the best results after the implementation of the algorithms as well as implying faster running times.
This study highlights the importance of the unsupervised machine learning algorithms, which, in combination with NGS techniques are capable of accelerating the unravelling of the underlying mysteries of the cancer biology and helping to find novel cancer classifications that facilitate the discovery of effective treatments.

Regarding the impact on sustainability, ethical-social aspects, and diversity no significant impacts were predicted during the design stage of the project and no unexpected impacts occurred during its development.

In terms of future work, in order to determine the robustness of the models and their performance in different datasets with similar types of data, it is necessary to implement them across a variety of datasets. Performing these models across different datasets, will allow to confirm whether the classification errors shown in each selected algorithm are due to the presence of an outlier in the data or due to the inability of the model to perform an accurate classification of certain patients.
In addition, to verify not only the presence of a BRCA subtype in the dataset but also, the ability of the algorithm to detect it, further analyses will be needed.
Performing differential expression analysis together with gene set enrichment analysis will allow a detailed exploration of the differences between the two clusters generated by the best algorithms using six clusters. This deeper exploration would help to study the functions and pathways associated with the most differentially expressed genes of these groups of patients and provide insights into the features that may differentiate the two clusters. Consequently, this approach will contribute to a more thorough understanding of potential subtypes within the BRCA cancer type.

7. Glossary

AD	Average Distance
ADM	Average Distance between Means
APN	Average Proportion of Non-overlap
BRCA	breast cancer
COAD	colon cancer
cDNA	Complementary Deoxyribonucleic Acid
CLARA	Clustering for Large Applications
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
Dim	Dimension
dNTPs	Deoxynucleotide Triphosphates
FOM	Figure of Merit
KIRC	kidney cancer
LUAD	lung cancer
NGS	Next Generation Sequencing
MinPts	Minimum number of neighbours
PAM	Partitioning Around Medoids
PC	Principal Component
PCA	Principal Component Analysis
PCA95	PCA outcome selecting the variables explaining the 95% of the variance
PCA800	PCA outcome selecting the 800 most relevant variables in terms of explained variance
PCR	Polymerase Chain Reaction
PRAD	prostate cancer
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
sd	Standard deviation
sd0	dataset without the variables with null standard deviation
TCGA	The Cancer Genome Atlas
UN	United Nations
UMAP	Uniform Manifold Approximation and Projection

8. Bibliography

1. World health organization: regional office for Europe *World Cancer Report: Cancer Research for Cancer Development.*; Wild, C.P., Weiderpass, E., Stewart, B.W., Eds.; IARC, 2020; ISBN 9789283204473.
2. Soerjomataram, I.; Bray, F. Planning for Tomorrow: Global Cancer Incidence and the Role of Prevention 2020–2070. *Nat Rev Clin Oncol* **2021**, *18*, 663–672, doi:10.1038/s41571-021-00514-z.
3. Goodall, G.J.; Wickramasinghe, V.O. RNA in Cancer. *Nat Rev Cancer* **2021**, *21*, 22–36, doi:10.1038/s41568-020-00306-0.
4. Osama, S.; Shaban, H.; Ali, A.A. Gene Reduction and Machine Learning Algorithms for Cancer Classification Based on Microarray Gene Expression Data: A Comprehensive Review. *Expert Syst Appl* **2023**, *213*, doi:10.1016/j.eswa.2022.118946.
5. Fiorini, S. Gene Expression Cancer RNA-Seq Available online: https://archive.ics.uci.edu/dataset/401/gene+expression+cancer+rna+seq (accessed on 11 October 2023).
6. TCGA_Pancancer Available online: https://www.synapse.org/#!Synapse:syn300013/files/ (accessed on 11 October 2023).
7. Venkataramana, L.; Jacob, S.G.; Saraswathi, S.; Venkata Vara Prasad, D. Identification of Common and Dissimilar Biomarkers for Different Cancer Types from Gene Expressions of RNA-Sequencing Data. *Gene Rep* **2020**, *19*, doi:10.1016/j.genrep.2020.100654.
8. Cole, L.; Kramer, P.R. Human Cancers and Carcinogenesis. In *Human Physiology, Biochemistry and Basic Medicine*; Elsevier, 2016; pp. 197–200.
9. Tab, M.L.; Tan, H.K.; Muhammad, T.S.T. Apoptosis and Cancer. In *Cancer Immunology*; Rezaei, N., Ed.; Springer: Berlin, Heidelberg, 2015; pp. 209–242 ISBN ISBN 978-3-662-44005-6.
10. Cline, M.S.; Craft, B.; Swatloski, T.; Goldman, M.; Ma, S.; Haussler, D.; Zhu, J. Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser. *Sci Rep* **2013**, *3*, doi:10.1038/srep02652.
11. Schneider, M. V.; Orchard, S. Omics Technologies, Data and Bioinformatics Principles. In *Bioinformatics for Omics Data. Methods and Protocols*; Mayer, B., Ed.; Humana Press, 2011; Vol. 719, pp. 3–30 ISBN ISBN 978-1-61779-026-3.
12. Feng, H.; Qin, Z.; Zhang, X. Opportunities and Methods for Studying Alternative Splicing in Cancer with RNA-Seq. *Cancer Lett* **2013**, *340*, 179–191, doi:10.1016/j.canlet.2012.11.010.
13. Kulski, J.K. Next-Generation Sequencing — An Overview of the History, Tools, and “Omic” Applications. In *Next Generation Sequencing - Advances, Applications and Challenges*; InTech, 2016.
14. McCombie, W.R.; McPherson, J.D.; Mardis, E.R. Next-Generation Sequencing Technologies. *Cold Spring Harb Perspect Med* **2019**, *9*, doi:10.1101/cshperspect.a036798.
15. Handelman, G.S.; Kok, H.K.; Chandra, R. V.; Razavi, A.H.; Lee, M.J.; Asadi, H. EDoctor: Machine Learning and the Future of Medicine. *J Intern Med* **2018**, *284*, 603–619, doi:10.1111/joim.12822.

16. Goecks, J.; Jalili, V.; Heiser, L.M.; Gray, J.W. How Machine Learning Will Transform Biomedicine. *Cell* **2020**, *181*, 92–101, doi:10.1016/j.cell.2020.03.022.
17. Koteluk, O.; Wartecki, A.; Mazurek, S.; Kołodziejczak, I.; Mackiewicz, A. How Do Machines Learn? Artificial Intelligence as a New Era in Medicine. *J Pers Med* **2021**, *11*, 1–22, doi:10.3390/jpm11010032.
18. Perera, M.A.I.; Wijesinghe, C.R.; Weerasinghe, A.R. Analysis of Expression Data Using Unsupervised Techniques. In Proceedings of the 20th International Conference on Advances in ICT for Emerging Regions, ICTer 2020 - Proceedings; Institute of Electrical and Electronics Engineers Inc., November 4 2020; pp. 119–124.
19. Rodriguez, M.Z.; Comin, C.H.; Casanova, D.; Bruno, O.M.; Amancio, D.R.; Costa, L. da F.; Rodrigues, F.A. Clustering Algorithms: A Comparative Approach. *PLoS One* **2019**, *14*, doi:10.1371/journal.pone.0210236.
20. Saket J, S.; Pandya, S. An Overview of Partitioning Algorithms in Clustering Techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* **2016**, *5*, 1943–1946.
21. Aghabozorgi, S.; Seyed Shirkhorshidi, A.; Ying Wah, T. Time-Series Clustering - A Decade Review. *Inf Syst* **2015**, *53*, 16–38, doi:10.1016/j.is.2015.04.007.
22. Campello, R.J.G.B.; Kröger, P.; Sander, J.; Zimek, A. Density-Based Clustering. *Wiley Interdiscip Rev Data Min Knowl Discov* **2020**, *10*, doi:10.1002/widm.1343.
23. Luchi, D.; Loureiros Rodrigues, A.; Miguel Varejão, F. Sampling Approaches for Applying DBSCAN to Large Datasets. *Pattern Recognit Lett* **2019**, *117*, 90–96, doi:10.1016/j.patrec.2018.12.010.
24. Ali, T.; Asghar, S.; Sajid, N.A. Critical Analysis of DBSCAN Variations. *2010 International Conference on Information and Emerging Technologies* **2010**, 1–6, doi:10.1109/ICIET.2010.5625720.
25. Xu, R.; Wunsch, D.C. Clustering Algorithms in Biomedical Research: A Review. *IEEE Rev Biomed Eng* **2010**, *3*, 120–154, doi:10.1109/RBME.2010.2083647.
26. Crespo-Roces, D.; Méndez-Jiménez, I.; Salcedo-Sanz, S.; Cárdenas-Montes, M. Generalized Probability Distribution Mixture Model for Clustering. In *Hybrid Artificial Intelligent Systems*; de Cos Juez, F.J., Villar, J.R., de la Cal, E.A., Herrero, Á., Quintián, H., Sáez, J.A., Corchado, E., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2018; pp. 251–263 ISBN 978-3-319-92638-4.
27. Hastie, T.; Tibshirani, R.; Friedman, J. Unsupervised Learning. In *The Elements of Statistical Learning*; Springer: New York, NY, 2009; pp. 485–585.
28. de Souto, M.C.P.; Costa, I.G.; de Araujo, D.S.A.; Ludermir, T.B.; Schliep, A. Clustering Cancer Gene Expression Data: A Comparative Study. *BMC Bioinformatics* **2008**, *9*, doi:10.1186/1471-2105-9-497.
29. Kononenko, I. Machine Learning for Medical Diagnosis: History, State of the Art and Perspective. *Artif Intell Med* **2001**, *23*, 89–109, doi:https://doi.org/10.1016/S0933-3657(01)00077-X.
30. Petretta, M. Applications of Machine Learning in Medicine. *Biomed J Sci Tech Res* **2019**, *20*, doi:10.26717/bjstr.2019.20.003503.

31. Mazlan, A.U.; Sahabudin, N.A. binti; Remli, M.A.; Ismail, N.S.N.; Mohamad, M.S.; Warif, N.B.A. Supervised and Unsupervised Machine Learning for Cancer Classification: Recent Development. In Proceedings of the 2021 IEEE International Conference on Automatic Control and Intelligent Systems, I2CACIS 2021 - Proceedings; Institute of Electrical and Electronics Engineers Inc., June 26 2021; pp. 392–395.
32. Mazlan, A.U.; Sahabudin, N.A.; Remli, M.A.; Ismail, N.S.N.; Mohamad, M.S.; Nies, H.W.; Warif, N.B.A. A Review on Recent Progress in Machine Learning and Deep Learning Methods for Cancer Classification on Gene Expression Data. *Processes* **2021**, *9*, doi:10.3390/pr9081466.
33. Piao, Y.; Piao, M.; Ryu, K.H. Multiclass Cancer Classification Using a Feature Subset-Based Ensemble from MicroRNA Expression Profiles. *Comput Biol Med* **2017**, *80*, 39–44, doi:10.1016/j.compbiomed.2016.11.008.
34. Best, M.G.; Sol, N.; Kooi, I.; Tannous, J.; Westerman, B.A.; Rustenburg, F.; Schellen, P.; Verschueren, H.; Post, E.; Koster, J.; et al. RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. *Cancer Cell* **2015**, *28*, 666–676, doi:10.1016/j.ccell.2015.09.018.
35. Hamoudi, R.; Bettayeb, M.; Alsaafim, A.; Hachim, M.; Nassir, Q.; Nassif, A.B. Identifying Patterns of Breast Cancer Genetic Signatures Using Unsupervised Machine Learning. *IEEE International Conference on Imaging Systems and Techniques (IST)* **2019**, 1–6, doi:10.1109/IST48021.2019.9010510.
36. Agapito, G.; Milano, M.; Cannataro, M. A Python Clustering Analysis Protocol of Genes Expression Data Sets. *Genes (Basel)* **2022**, *13*, doi:10.3390/genes13101839.
37. Etoledoi/TFM_EloisaToledoIglesias: Code and Supplementary Materials Available online: https://github.com/etoledoi/TFM_EloisaToledoIglesias (accessed on 13 January 2024).
38. Fujisawa, K.; Shimo, M.; Taguchi, Y.H.; Ikematsu, S.; Miyata, R. PCA-Based Unsupervised Feature Extraction for Gene Expression Analysis of COVID-19 Patients. *Sci Rep* **2021**, *11*, doi:10.1038/s41598-021-95698-w.
39. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**, doi:https://doi.org/10.48550/arXiv.1802.03426 Focus to learn more.
40. Dorrity, M.W.; Saunders, L.M.; Queitsch, C.; Fields, S.; Trapnell, C. Dimensionality Reduction by UMAP to Visualize Physical and Genetic Interactions. *Nat Commun* **2020**, *11*, doi:10.1038/s41467-020-15351-4.
41. Lantz, B. Finding Groups of Data – Clustering with k-Means. In *Machine Learning with R*; Nair, A., Jones, J., D'Souza, N., Shah, R.C., Lobo, A., Priya, S., Eds.; Packt Publishing, 2015; pp. 285–310 ISBN 9781784393908.
42. Kassambara, A. *Practical Guide To Cluster Analysis in R. Unsupervised Machine Learning*; 1st ed.; STHDA, 2017;
43. Weiqiang, C. RPubs - Datacamp R - Unsupervised Learning in R Chapter 2 (Hierarchical Clustering) Available online: https://rpubs.com/AIventurer/unsupervised_learning_ch2 (accessed on 18 December 2023).

44. Aggregation Methods in R. – Data Science Portfolio Available online: https://www.alldatascience.com/clustering/aggregation-methods-in-r/ (accessed on 18 December 2023).
45. Ekemeyong Awong, L.E.; Zielinska, T. Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification. *Sensors* **2023**, *23*, 7925, doi:10.3390/s23187925.
46. Brock, G.; Pihur, V.; Datta, S.; Datta, S. *ClValid , an R Package for Cluster Validation*; 2008;
47. Bolshakova, N.; Azuaje, F. Cluster Validation Techniques for Genome Expression Data. *Signal Processing* **2003**, *83*, 825–833, doi:10.1016/S0165-1684(02)00475-9.
48. Abbas T., S. 5 Major Evaluation Metrics for Clustering Available online: https://medium.com/@SyedAbbasT/5-major-evaluation-metrics-for-clustering-74ea8b301e68 (accessed on 20 November 2023).
49. Davidson, B. Cluster Analysis Available online: https://bookdown.org/brittany_davidson1993/bookdown-demo/ (accessed on 12 November 2023).
50. Eszergár-Kiss, D.; Caesar, B. Definition of User Groups Applying Ward's Method. *Transportation Research Procedia* **2017**, *22*, 25–34, doi:10.1016/j.trpro.2017.03.004.
51. Yang, Y.; Sun, H.; Zhang, Y.; Zhang, T.; Gong, J.; Wei, Y.; Duan, Y.G.; Shu, M.; Yang, Y.; Wu, D.; et al. Dimensionality Reduction by UMAP Reinforces Sample Heterogeneity Analysis in Bulk Transcriptomic Data. *Cell Rep* **2021**, *36*, doi:10.1016/j.celrep.2021.109442.
52. Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.A.; Kwok, I.W.H.; Ng, L.G.; Ginhoux, F.; Newell, E.W. Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat Biotechnol* **2019**, *37*, 38–47, doi:10.1038/nbt.4314.
53. Jha, A.; Quesnel-Vallières, M.; Wang, D.; Thomas-Tikhonenko, A.; Lynch, K.W.; Barash, Y. Identifying Common Transcriptome Signatures of Cancer by Interpreting Deep Learning Models. *Genome Biol* **2022**, *23*, doi:10.1186/s13059-022-02681-3.
54. Liu, Y.R.; Jiang, Y.Z.; Xu, X.E.; Yu, K. Da; Jin, X.; Hu, X.; Zuo, W.J.; Hao, S.; Wu, J.; Liu, G.Y.; et al. Comprehensive Transcriptome Analysis Identifies Novel Molecular Subtypes and Subtype-Specific RNAs of Triple-Negative Breast Cancer. *Breast Cancer Research* **2016**, *18*, doi:10.1186/s13058-016-0690-8.