Departamento de Computação/UFS

# Semantic Clustering and Classification of Unsupervised Data Using Large Language Models (LLMs)

Clusterização e Classificação Semântica de Dados Não Supervisionados utilizando Modelos de Linguagem (LLMs).

Heitor Saulo Dantas Santos [1]; Itor Carlos Souza Queiroz [1]; Lanna Luara Novaes Silva [1]; Lavínia Louise Rosa Santos [1]; Rômulo Menezes De Santana [1]

[1] *Departamento de Computação (DCOMP)*
*Universidade Federal de Sergipe (UFS)*
*Av. Marechal Rondon, s/n– Jardim Rosa Elze– CEP 49100-000*
*São Cristóvão – SE– Brazil*

heitor.santos@dcomp.ufs.br, itor_carlos@academico.ufs.br, lannaluara@academico.ufs.br, lavinialouise@academico.ufs.br, rmsantana@dcomp.ufs.br

This study presents a hybrid approach that combines clustering techniques with large language models (LLMs) for the semantic classification of unsupervised data. Utilizing the Breast Cancer Dataset, three clustering algorithms (K-Means, Gaussian Mixture Model, and Agglomerative Clustering) were applied to segment the clinical data into unlabeled groups. Subsequently, the GPT-4o-mini and GPT-4o models from OpenAI were used for semantic assignment to the formed groups, classifying them as malignant or benign. The performance of the approach was evaluated using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) metrics, comparing the clusters with the real labels of the dataset. The results indicate that LLMs are capable of contextually interpreting the generated clusters, providing coherent labels even when the clustering algorithms show low separation accuracy according to conventional metrics, highlighting the ability of language models to correctly infer meaning for the generated clusters with considerable effectiveness, even in scenarios of imperfect segmentation.
**Keywords:** Clustering, Language Models, Unsupervised Learning.

## 1. Introduction

Currently, the exponential growth of data generated in an increasingly automated society has brought about the need for more dynamic and adaptable data analysis. When dealing with large volumes of data, Machine Learning (ML) combined with Large Language Models (LLMs) forms a powerful tool to assist in the processing and interpretation of information.

Clustering, an unsupervised learning technique, plays a crucial role by grouping similar data without the need for prior labels [1]. However, because it relies on similarity measures between data points, traditional clustering has limitations, especially regarding the interpretation and semantic classification of the formed groups. This highlights the need for complementary techniques that can add meaning to the obtained groups.

In this context, the objective of this work is to apply an LLM to classify the unsupervised clustered data and subsequently evaluate the performance of the approach. The study was conducted using three well-known clustering techniques: K-means, Gaussian Mixture Model (GMM), and Agglomerative Clustering, applied to a medical examination dataset with the aim of assisting in breast cancer detection, dividing it into two distinct groups. After the division, the chosen LLM was used to assign semantic labels to the groups, classifying them as Benign or

Malignant. The results obtained were compared with the expected labels using evaluation metrics, providing an analysis of the proposed approach's performance and the LLM's ability to identify patterns and classify information by assigning contextual meanings to the groupings.

Thus, the proposed methodology demonstrates its versatility by allowing the model to be applied to other datasets with different characteristics and specifications. The use of LLMs with their potential for semantic analysis and contextualization reveals a fundamental tool for promoting the dynamism and adaptability of models when dealing with different types of information and domains.

## 2. THEORETICAL FRAMEWORK

Throughout this work, various concepts and fundamentals of Artificial Intelligence and Data Analysis were used as a basis for defining the scope and for the implementation of the proposed approach. In this section, the main concepts used will be briefly presented, including: Machine Learning, with a focus on unsupervised learning, entity recognition, and data clustering; and Large Language Models, LLMs.

### 2.1 MACHINE LEARNING

Machine Learning (ML) is a branch of Artificial Intelligence focused on developing systems capable of learning from data and making decisions or predictions based on identified patterns [2]. This ability to learn without explicit instructions makes ML a powerful tool for handling large volumes of data and solving complex problems in various fields.

Machine Learning algorithms can be used in different contexts depending on the type of information available in the data, enabling everything from predictive analyses to the discovery of hidden patterns. Among the ML approaches, three main categories can be identified: reinforcement learning, where the model learns through trial and error, receiving rewards or penalties for an agent's actions in an environment; supervised learning, in which the model is trained with labeled data, learning to map inputs to known outputs; and unsupervised learning, which deals with unlabeled data and seeks to identify hidden structures such as groupings or relationships between variables, being the focus of this work through the application of clustering techniques [2].

#### *2.1.1 SUPERVISED AND UNSUPERVISED LEARNING*

As previously introduced, supervised learning is a Machine Learning technique in which the model is trained based on a labeled dataset, meaning the data provided for training already has a known output. The objective of this technique is to build a model that, once trained, has learned a function that maps inputs to the desired outputs so that, when it receives new data, it can correctly predict the corresponding labels. This type of approach is commonly used for classification and regression tasks [3].

On the other hand, unsupervised learning operates on unlabeled data, that is, data that does not have predefined outputs or categories. Because of this, the model must be able to independently identify patterns or hidden structures inherent in the data during training. Thus, instead of predicting a known output for new examples, its goal is to organize or group the data in a way that reveals shared characteristics that were not explicitly stated, being particularly useful when the categorization of the data is not known beforehand or when seeking to discover new ways of organizing the analyzed dataset [3].

In the context of this work, unsupervised learning was adopted for its unique ability to analyze large volumes of data without the need for prior knowledge about its categorization. By applying clustering algorithms such as K-means, Gaussian Mixture Model, and Agglomerative, it is possible to group patient exams based on their characteristics, revealing potential patterns

associated with different diagnoses. However, a significant limitation of this approach is that, while it can identify natural groupings in the data, it does not assign semantic meaning to these groups. In light of this, seeking to overcome this issue, this step precedes the use of an LLM to perform the classification of the clustered data, complementing the analysis with a semantic layer capable of assigning meaningful labels to the identified groups.

### *2.1.2 Clustering and Entity Identification*

Clustering is the process of dividing a dataset into clusters such that elements within the same cluster exhibit high similarity to each other, while elements from different clusters are as different as possible. Entity identification through clustering techniques represents one of the fundamental pillars of unsupervised learning, enabling the discovery of patterns and hidden structures in unlabeled datasets [1]. These techniques, as explained earlier, operate under the basic principle of grouping similar elements and separating distinct ones without any prior knowledge about their categories or labels. However, the unsupervised nature of these methods introduces significant challenges regarding the evaluation of the quality of the formed clusters.

Although clustering is effective in discovering patterns, its evaluation poses a significant challenge. This is because, in the absence of ground truth, that is, a reference truth with known labels, it becomes difficult to objectively measure the quality of the groupings [4]. An alternative to the absence of ground truth is the use of internal metrics, such as intra-cluster distance, which measures the cohesion between the elements of the same cluster, and inter-cluster distance, which measures the separation between different clusters. Such metrics are useful, but they do not guarantee that the formed groups have practical significance or semantic coherence [5].

Another way to evaluate the quality of clusters is through external metrics when some prior knowledge about the data is available, that is, when there is ground truth. Some of these metrics are the Rand Index (RI), which measures the similarity between an obtained clustering and the true classification by comparing how pairs of points are grouped or separated in both; the Adjusted Rand Index (ARI), a normalized version of the RI that takes into account the possible occurrence of random agreements; and the Normalized Mutual Information (NMI), which measures the amount of information shared between the clusters and the true labels, also normalizing the value to allow comparisons between different configurations [5].

In this study, the true labels of the data in the used dataset are available, corresponding to the definitive medical classification of the tumors in the exams as malignant or benign. This information, although available, is completely excluded from the clustering process, fully preserving the unsupervised nature of the experiment. However, the existence of these labels allows for an objective evaluation of the quality of the formed clusters through external metrics such as those mentioned above. To evaluate the clustering performed in this work, the chosen evaluation metrics were ARI and NMI, which assess the similarity between the generated clusters and the expected classification, adjusting the results for chance and considering different grouping structures. In addition, a value-by-value verification of the clustered data is also performed to quantitatively identify the percentage of accuracy of the generated clusters in relation to the expected groups.

However, even though this study uses the real labels to evaluate the quality of the formed groupings, its main objective is centered on simulating scenarios in which the correct output of the data is not available, which is common in real unsupervised learning tasks. In this context, the application of a language model becomes essential, as it allows for the assignment of semantic meaning to the clusters generated from unlabeled data. In this way, the proposed

model seeks to circumvent the absence of prior information about the data categories by using the interpretative capacity of the LLM as a complementary mechanism to traditional clustering.

### 2.2 Language Models

Large Language Models (LLMs) are Deep Learning Network models trained on vast amounts of text and billions of parameters to process, generate, and identify patterns in natural language [6]. These models have stood out for their potential in various Natural Language Processing tasks such as text generation, summarization, translation, and entity identification (classification). Currently, the use of LLM-based applications is numerous and already part of the daily lives of citizens worldwide.

The use of LLMs with zero-shot and few-shot learning approaches demonstrates the power of these models for data classification tasks without the need for traditional supervised training [7]. This shows that the models can perform classification tasks solely through a textual description via prompt, with or without examples. This potential of LLMs is directly related to the Transformer architecture, present in models like GPT and BERT and proposed by Vaswani et al. (2017) [8]. The use of the Self-Attention mechanism with Multi-Head Attention allows the model to capture contextual relationships between words in a sequence in a parallel and simultaneous manner, which enables the construction of useful representations for data grouping and semantic labeling. Furthermore, through Encoder-Decoder mechanisms and Positional Encoding, models that use the Transformer architecture learn dense and contextualized semantic representations of text. In this way, models like GPT (Generative Pre-trained Transformers), used in this work, prove to be an effective instrument for the semantic classification of unsupervised data.

Another ally of the presented proposal and LLMs is the use of the Prompt Engineering technique [9], which consists of a textual and systematic elaboration of the instructions passed to the model to guide its behavior. A well-formulated prompt is extremely important for the LLM to correctly interpret its function and task. When interacting with an LLM, it is important that the instructions are clear, objective, specific, and adjusted according to the need and nature of the task. The combination of a structured and correctly elaborated prompt with the application of the LLM results in a powerful tool.

## 3. Methodology

As presented in the previous topics, the approach used clustering techniques such as K-Means and Gaussian Mixture Model (GMM) applied to a patient dataset for breast cancer detection. After clustering, the gpt-4o-mini model and, in some cases, OpenAI's gpt-4o were applied for the semantic classification of the identified labels. Subsequently, the obtained results were analyzed and compared with the expected classification. Furthermore, as a way to evaluate the results, the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) metrics were employed, with the aim of presenting concrete values of what was obtained and compared.

The tests and experiments performed were done using a machine with the follow specifications 16 GB Intel Core i5-9300H Windows 10 and the libraries used were: pandas, for analyzing and manipulating data in tables; openai, for accessing the OpenAI API; KMeans, for the clustering algorithm; LabelEncoder, for transforming categorical labels into numbers; GaussianMixture, Gaussian mixture model for clustering; AgglomerativeClustering, for the clustering algorithm; adjusted_rand_score, a metric to evaluate the similarity between groupings; normalized_mutual_info_score, a metric that measures the normalized mutual information between groupings. In addition to auxiliary libraries such as json, numpy, os, and matplotlib.pyplot.

It is important to highlight that the dataset was chosen to exemplify the approach and to analyze performance, but the methodology of clustering and LLM application is not restricted to

the studied data; the nature of the LLM allows its application to other data domain sets. Next, a brief overview of each technique and the necessary tools in the approach will be presented.

## 3.1 Clustering Techniques

### *3.1.1 K-means*

In this work, the K-Means clustering algorithm was employed to segment the data related to breast cancer diagnosis, with the aim of automatically identifying groupings of patients with similar characteristics, without the use of supervised labels. K-Means was chosen for its simplicity, computational efficiency, and good performance in tasks where the data tends to organize into well-defined spherical clusters. Considering that the dataset in question presents numerical metrics derived from imaging exams (for example, radius_mean, smoothness_mean, among others), K-Means proved appropriate for finding underlying patterns among patients, grouping them based on their geometric similarities in the attribute space.

The K-Means algorithm works iteratively as follows [10]:

- Initialization: Randomly chooses a number k of centroids (in this case, k=2).
- Assignment: Each data point is assigned to the nearest centroid based on the Euclidean distance.
- Update: The centroids are recalculated as the mean of 1 the points assigned to them.
- Convergence: The assignment and update process is repeated until the centroids no longer change significantly between iterations, or until a maximum number of iterations is reached.

The algorithm is sensitive to the scale of the data and the presence of collinearities, so it is necessary to preprocess the data to avoid highly correlated features. The following section details another clustering technique.

### *3.1.2 Gaussian Mixture Model (GMM)*

The Gaussian Mixture Model [11] (GMM) clustering algorithm was also employed with the objective of identifying patient groupings based on their clinical and imaging characteristics, without the use of supervised labels, that is, without using the column that identifies the outcome of that instance in the dataset.

The GMM algorithm was chosen for its flexibility and sophistication in the statistical modeling of data. In contrast to algorithms like K-Means, which impose rigid restrictions on the shape of its clusters (spherical and equidistant), GMM allows for the representation of each cluster as a multivariate Gaussian distribution, possessing its own mean and covariance [12]. This characteristic makes this algorithm extremely efficient when the data exhibits heterogeneous variances or overlaps between the possible groups.

As the dataset used contains numerical attributes derived from medical examinations — such as radius_mean, texture_mean, and compactness_mean — which may exhibit different scales and correlations, GMM proved to be a suitable choice for capturing more subtle patterns of grouping.

The GMM algorithm operates using a probabilistic mixture model and employs the Expectation-Maximization (EM) algorithm to adjust its parameters [13]. Its operation occurs in iterative cycles and consists of two steps:

- E-step (Expectation): Calculates the probability of each sample belonging to each of the possible Gaussian distributions present, based on the current parameters, such as mean and covariance.
- M-step (Maximization): Updates the parameters of the Gaussian distributions present in order to maximize the likelihood of the observed data.

### 3.1.3 Agglomerative clustering

Another clustering technique employed was Agglomerative Clustering [14]. Unlike other techniques, data manipulation begins by considering each data point as its own individual cluster. Based on the similarities between the clusters, they are grouped together until the defined number of groups is reached, always following the similarity between the data. In this way, Agglomerative uses a bottom-up approach [15]. The technique has three main steps:

1. **Initialization:** Each data point is an individual cluster.
2. **Iterative Grouping:** The algorithm measures the distance between pairs of clusters and merges the closest ones. To define this proximity, the algorithm uses linkage methods, such as single linkage, complete linkage, average linkage, and ward linkage. At each step, the number of clusters reduces by one, and this process continues until the desired number of clusters remains (or another stopping condition is met).
3. **Hierarchy Formation:** During the algorithm's process, a tree of the grouping is created, called a dendrogram. The tree shows how the clusters were joined and aggregated over time.

## 3.2 Use of OpenAI

In this work, the OpenAI API was used to assist in the interpretation of the results obtained through unsupervised clustering algorithms in the context of breast cancer classification. After applying the Gaussian Mixture Model (GMM) and K-Means models to the Breast Cancer DataSet, the data was grouped into clusters, making it necessary to interpret the nature of each obtained grouping. That is, it is necessary to associate each cluster with its probable classification – malignant or benign.

As the adopted clustering process is unsupervised, the original labels of the dataset – which indicate the correct classification of each instance – are not used during the formation of the clusters. This makes the interpretation step essential to evaluate whether the clustering results are coherent with respect to the real information present in the adopted dataset.

To verify the coherence of the clustering results with the real data of the dataset, the **gpt-4o-mini** model, provided by OpenAI, was adopted. The choice of this model is due to the balance between performance and computational cost, added to the fact that it can efficiently understand and interpret large volumes of structured data in JSON format; the availability and access to the model within the limitations for running the tests also influenced the decision. Furthermore, its competence in semantic analysis allows for a more detailed interpretation of the data characteristics, which contributes to a more accurate association of the generated clusters with their probable classification – malignant or benign.

In cases where **gpt-4o-mini** did not interpret the data satisfactorily, specific tests were performed with the **gpt-4o model**, chosen for its superior performance, although with a higher usage cost.

## 3.3 Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI)

The quality of clustering is an important challenge in the evaluation of unsupervised learning problems, especially when there is no explicit categorization of the data. Since, in this work, the true labels are known, it was possible to use external metrics to measure how much the formed groupings correspond to the expected classification. For this purpose, the Adjusted Rand Index and the Normalized Mutual Information were chosen due to their robustness, reliability, and widespread use in scientific literature for clustering evaluation.

The Adjusted Rand Index is a metric used to measure the degree of similarity between two groupings, correcting the calculated value to eliminate agreement that could occur by chance. Its value can range from -1 to 1, where -1 means complete disagreement, 0 is what would be expected from a random assignment, and 1 indicates perfect agreement. For this reason, the ARI is a suitable metric for evaluating different clustering results, as the

normalization allows for comparison regardless of the number of clusters or the data distribution [16].

On the other hand, Normalized Mutual Information measures the amount of mutual information shared between the generated clusters and the real labels, being normalized to produce values between 0 and 1. Values close to 0 indicate low correspondence between the groupings and the expected classes, while values close to 1 suggest high similarity. One of the main advantages of NMI is that it does not depend on the order of the group labels, which makes it suitable for evaluating groupings whose labels do not have a predefined order [17].

These metrics were applied after the execution of the clustering algorithms, using the real labels only for evaluation purposes. From them, it was possible to quantify the accuracy of the groupings in relation to the expected classes, allowing us to verify the effectiveness of the clustering before the intervention of the language model for the semantic classification of the data.

### 3.4 Dataset Description

The dataset selected for the tests in this work is the Breast Cancer Dataset, available on the Kaggle platform via the following link: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset. The main objective of this dataset is to aid in the detection of breast cancer based on characteristics observed and extracted from patient examinations. The dataset presents information that makes it possible to build prediction models for classifying cancer cases as malignant — represented by the letter M — or benign — represented by the letter B.

General Information:

- Number of rows: 569
- Number of columns: 32
- Target column: **diagnosis**

Characteristics present in the DataSet:

- Identifier: column used for identification in the dataset. In this dataset, this column is named 'id'.
- Diagnosis: column used to label whether the cancer in question is malignant or benign.
- Descriptive attributes: 30 variables extracted from breast examination images. These variables describe characteristics such as texture, perimeter, area, smoothness, among others.

Thus, the DataSet has 30 numerical attributes derived from these measurements, in addition to the **'id'** column and the **'diagnosis'** output variable. This dataset is widely used in supervised machine learning tasks, especially in binary classification problems.

## 4. Experiments

In this section, the step of the proposed implementation and what was necessary in each stage will be presented. The obtained results will be presented in detail in the next topic of Results and Discussions.

### 4.1 Dataset Processing

As previously informed, the chosen dataset was the Breast Cancer Dataset. For manipulation and pre-processing of the data, some actions were necessary.

1. Transformation of non-numeric columns into labels using LabelEncoder: As it is necessary to remove non-numeric columns, the diagnosis column, previously B for benign and M for malignant, now the reference is 0 for benign and 1 for malignant.

2. Removal of the identification column (“id”) for not presenting relevance in the study of the data;
3. Elimination of highly correlated columns (correlation greater than 0.87) in order to reduce redundancy in the data and avoid bias in the clustering;
4. Removal of the diagnosis column (“diagnosis”) for clustering;

### 4.2 Application of clustering techniques

For the application of clustering techniques, binary classification was fixed (n_components = 2/n_clusters = 2) of the clusters, as the diagnosis can be Benign or Malignant. For each clustering technique, a graph was created showing the division and arrangement of the created clusters. The visualization of the graphs will be presented together with the results. The evaluation of the clusters was validated by means of the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) metrics, which quantify the similarity between the groupings produced by the techniques and the real diagnosis labels - only used for evaluation, and not for training.

### 4.3 OpenAI Application

After the binary division of the data, the next step was to make a request call to OpenAI so that the chosen model could classify the set of obtained data. The gpt-4o-mini and gpt-4o models were used; in some cases where the first did not present satisfactory results, the second was used due to its superior performance. The elaboration of the prompt followed this structure: informs possible classifications; passes each cluster separately; passes the context of the data; and defines a specified format for the return of the call. The prompt in Portuguese used in both models follows:

```
f"""
A seguir, apresento a lista das possíveis classificações:
Maligno
Benigno
Além disso, para cada grupo, disponibilizarei os dados no formato JSON
conforme o exemplo abaixo:
Grupo 1: {cluster_0_json[0:1600]}
Grupo 2: {cluster_1_json[0:1600]}
O contexto desses dados são para a classificação de câncer de mama, onde
cada grupo representa um conjunto de características de pacientes que
fizeram o exame para investigar um possível câncer.
Sua tarefa é analisar as características presentes em cada JSON (dados de
cada grupo) e, com base nessas informações e na lista de classificações
fornecida, atribuir a cada grupo a classificação que melhor o representa,
certifique-se de considerar todas as características presentes em cada
json, e classificar todos os grupos fornecidos. Para cada grupo, por
favor, inclua:
A classificação escolhida.
Uma breve justificativa explicando a relação entre as características do
grupo e a classificação atribuída.
Por favor, apresente os resultados da seguinte forma no formato JSON para
eu converter diretamente a resposta para um arquivo JSON em python:
    "classificação":
        "cluster_0": label_0,
        "cluster_1": label_1,
        ...
    "justificativa":
        "cluster_0": "justificativa_0",
        "cluster_1": "justificativa_1",
        ...
"""
```

### 4.3 Application of Evaluation Metrics

After the application of the LLM, there was a visual verification of the results, and after this verification, the ARI and NMI metrics were applied. The metrics compared the original "diagnosis" column with the classification made by the LLM. In this way, it was possible to quantify and compare the obtained results.

## 5. Results and Discussion

In this section, all the obtained results will be presented, along with the metrics and graphs. Each subtopic represents the results of a clustering technique.

### 5.1 K-Means Clustering Results

The following graph shows the distribution made by K-Means in the dataset. The yellow color represents cluster 1, and the purple color represents cluster 0.

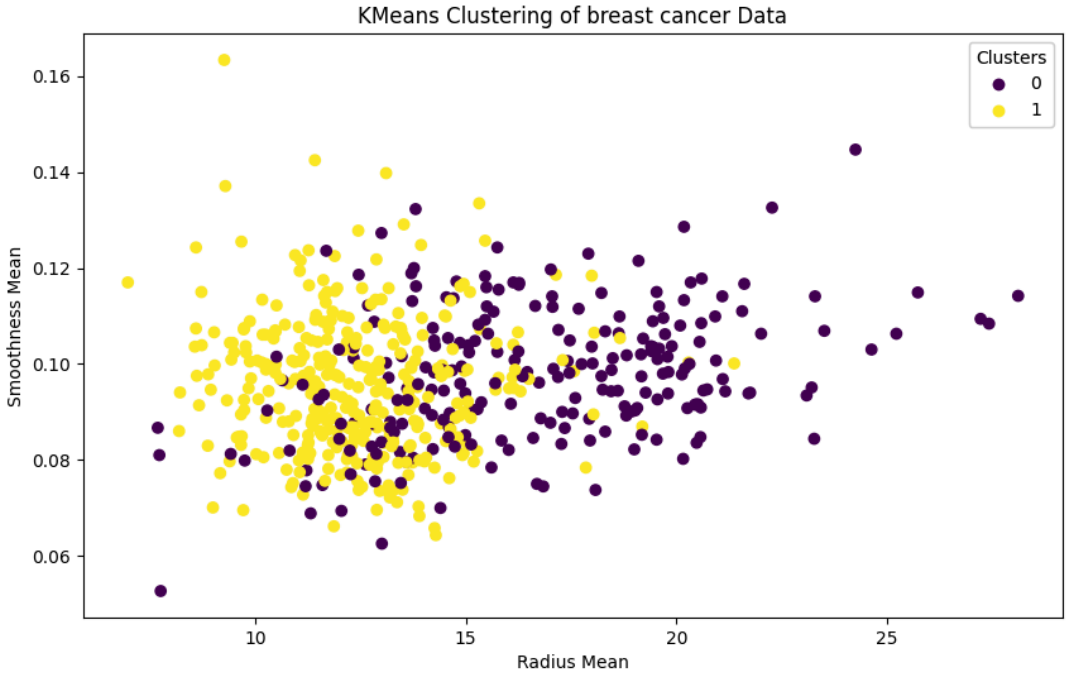


Figure 1: KMeans Distribution Graph

The results obtained by K-Means were:

**KMeans (ari_score) breast cancer: 44%**
**KMeans (nmi_score) breast cancer: 34%**
**KMeans (similaridade entre os clusters e os valores do dataset) breast cancer: 16.70%**

In K-Means, the LLM failed to correctly classify the groups, possibly as a consequence of the low precision in the group separation. According to the Adjusted Rand Index (ARI), the value was approximately 44%, indicating a low agreement between the groupings and the actual diagnoses. The Normalized Mutual Information (NMI) value was about 34%, reinforcing a low informational quality of the segmentation performed. Even though more than 86% of the data is classified correctly, the separation of the groups is not as good as that of the GMM, which may have influenced the final classification. Therefore, the K-means clustering was tested with the gpt-4o model for a better analysis of the performance.

**Kmeans with 4o model (similarity between the clusters and the dataset values) breast cancer: 83.30%**

With OpenAI's superior model, the classification was improved and it was possible to correctly identify which group the data belonged to, but the performance was still lower than the GMM due to the classification of the clusters themselves.

### 5.2 GMM Clustering Results

The following graph shows the distribution made by GMM in the dataset. The yellow color represents cluster 1, and the purple color represents cluster 0.

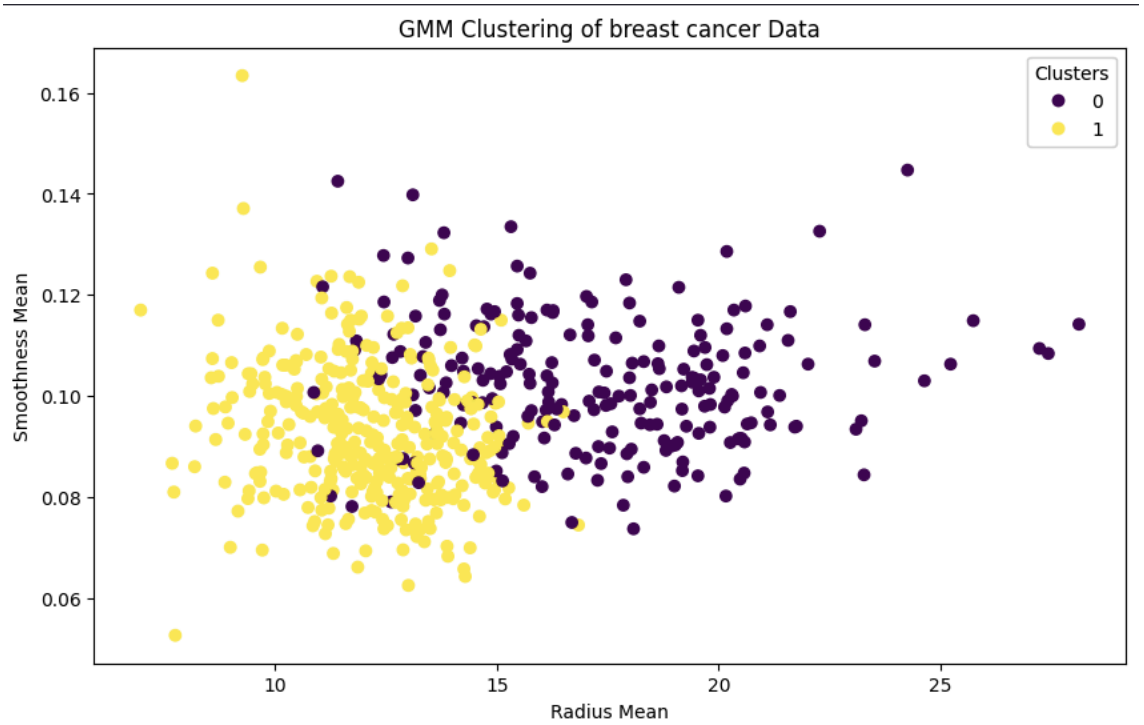


Figure 2: GMM Distribution Graph

The results obtained by GMM were:

**GMM (ari_score) breast cancer: 77%**
**GMM (nmi_score) breast cancer: 66%**
**GMM (similarity between the clusters and the dataset values) breast cancer: 94.02%**

The GMM algorithm had the best performance among the algorithms tested for this dataset. According to the ARI, the GMM was evaluated at approximately 77%, indicating that there is good agreement between the clusters generated by the GMM and the actual diagnoses derived from the analyzed dataset. According to the NMI, it was inferred at approximately 66%, which suggests that the obtained segmentation successfully captured a significant part of the informational structure of the data.

The results obtained using the GMM show that it is capable of performing useful data segmentations, groupings coherent with the clinical patterns obtained in the data associated with breast cancer, even without using the classification labels. The previous results demonstrate that the GMM has enormous potential as a tool for performing data groupings in various contexts. Just as with the use of K-Means, data preprocessing proved essential to ensure the quality of the grouping. A more detailed and in-depth analysis is necessary to identify the benefits and strengths of each algorithm, as this will be a decisive factor for good performance in the final classification.

### 5.3 Agglomerative Clustering Results

The following graph shows the distribution made by Agglomerative Clustering in the dataset. The purple color represents cluster 0, and the yellow color represents cluster 1.

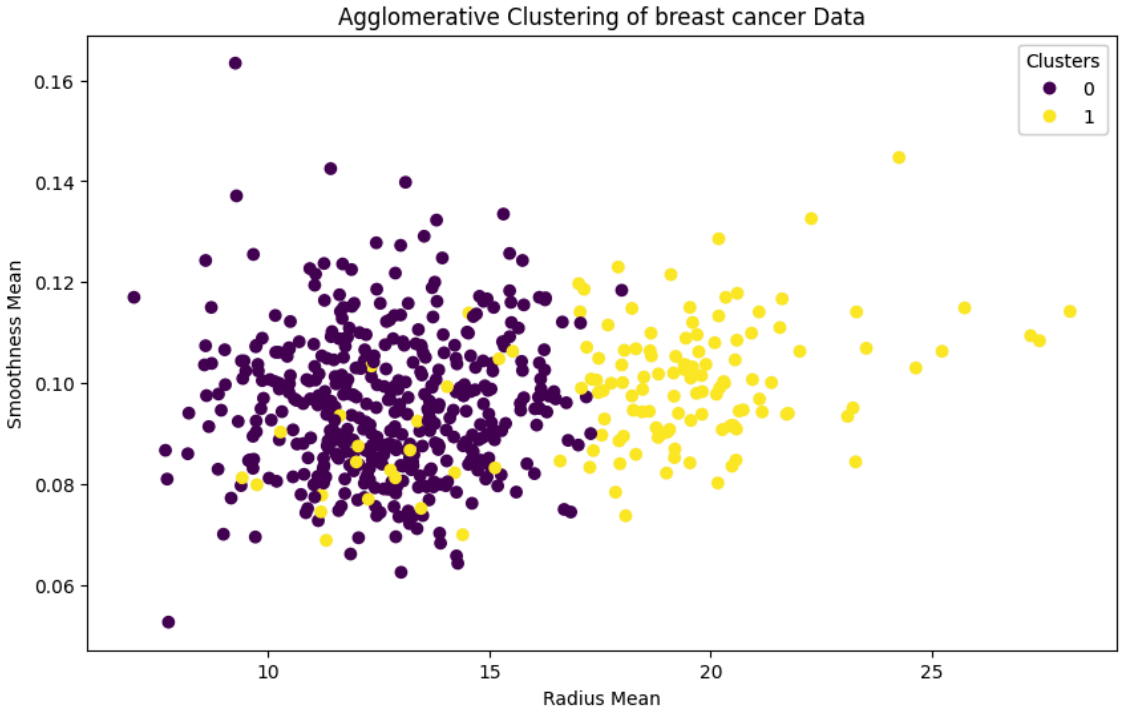


Figure 3: Agglomerative Clustering Distribution Graph

The results obtained by Agglomerative Clustering were:

**Agglomerative (ari_score) breast cancer: 35%**
**Agglomerative (nmi_score) breast cancer: 27%**
**Agglomerative (similarity between the clusters and the dataset values) breast cancer: 80.14%**

In Agglomerative Clustering, the LLM was able to correctly classify the groups. Even though the classification indices were worse, the division pattern presented by the algorithm was closer to the division of the original dataset. This may have possibly influenced the LLM's interpretation in making decisions based on the presented data.

## 6. Conclusion

In summary, the present study explored the application of unsupervised clustering techniques, addressing algorithms such as K-Means, Gaussian Mixture Model (GMM), and Agglomerative Clustering for the segmentation of a breast cancer diagnosis dataset. The main purpose was to identify essential groupings in the data without the use of supervised labels, simulating scenarios where prior knowledge about the classification of the samples is limited. Nevertheless, the initial labeling of the database was used as ground truth. The use of a language model, specifically GPT, proved promising for assigning semantic meaning to these groupings, overcoming one of the inherent limitations of traditional clustering methods.

The experimentation showed that this approach can be very powerful mainly because it:

- Tests the domain knowledge of the LLM: The LLM is capable of correlating statistics (such as mean radius) with clinical knowledge (malignant tumors are larger);
- Reveals the utility of Clusters: Unsupervised clusters can be driven by noise and may not necessarily align with domain knowledge;
- Educational: Highlights challenges in unsupervised learning (for example, clusters may reflect age or measurement bias, and not malignancy).

The final results show that, with a more detailed analysis of the dataset, it is possible to achieve better results, as it became clear that different algorithms lead to extremely different outcomes, and understanding the context in which this data is embedded can greatly help to improve the performance of this classification allied with LLMs.

## 7. References


[1] TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to Data Mining**. [s.l.: s.n.], 2014. **Cap. 8 – Cluster Analysis: Basic Concepts and Algorithms.** Disponível em: https://www.ceom.ou.edu/media/docs/upload/Pang-Ning_Tan_Michael_Steinbach_Vipin_Kumar_-_Introduction_to_Data_Mining-Pe_NRDK4fi.pdf.

[2] IBM. Aprendizado de máquina. Disponível em: <https://www.ibm.com/br-pt/think/topics/machine-learning>.

[3] **Aprendizado supervisionado versus não supervisionado – Diferença entre algoritmos de machine learning** – AWS. Disponível em: <https://aws.amazon.com/pt/compare/the-difference-between-machine-learning-supervised-and-unsupervised/>.

[4] IBM. **Ground Truth in Machine Learning**. Disponível em: <https://www.ibm.com/think/topics/ground-truth>.

[5] ALURA. **Métricas de avaliação para clusterização.** Disponível em: <https://www.alura.com.br/artigos/metricas-de-avaliacao-para-clusterizacao?srsltid=AfmBOoqcS6l3DCWqMfIzwOdvQwG9AEhedhit7ghITkz12x3tjv5G2YqY>.

[6] IBM. **Grandes modelos de linguagem.** Ibm.com. Disponível em: <https://www.ibm.com/br-pt/think/topics/large-language-models>.

[7] BROWN, Tom B. et al. **Language Models are Few-Shot Learners**. arXiv preprint arXiv:2005.14165, 2020. Disponível em: https://arxiv.org/pdf/2005.14165.pdf.

[8] VASWANI, Ashish et al. **Attention is All You Need.** In: **Advances in Neural Information Processing Systems**, v. 30, 2017. Disponível em: https://arxiv.org/pdf/1706.03762.pdf.

[9] LIU, Peng et al. **Pre-train Prompting: A Survey.** arXiv preprint arXiv:2107.13586, 2021. Disponível em: https://arxiv.org/pdf/2107.13586.pdf.

[10] **KMeans.** scikit-learn. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>. Acesso em: 8 abr. 2025.

[11] **Gaussian Mixture Model Explained** | Built In. Built In. Disponível em: <https://builtin.com/articles/gaussian-mixture-model>. Acesso em: 6 abr. 2025.

[12] **2.1. Gaussian mixture models**. scikit-learn. Disponível em: <https://scikit-learn.org/stable/modules/mixture.html>. Acesso em: 8 abr. 2025.

[13] VIPULGANDHI. **Gaussian Mixture Models Clustering** - Explained. Kaggle.com. Disponível:<https://www.kaggle.com/code/vipulgandhi/gaussian-mixture-models-clustering-explained>. Acesso em: 8 abr. 2025.

[14] **AgglomerativeClustering.** scikit-learn. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>.

[15] DENYS GOLOTIUK. **What is Agglomerative clustering and how to use it with Python Scikit-learn**. Medium. Disponível em: <https://medium.com/datadenys/what-is-agglomerative-clustering-and-how-to-use-it-with-python-scikit-learn-7e127ddb148c>.

[16] Adjusted Rand Index (ARI) - OECD.AI. Disponível em: <https://oecd.ai/en/catalogue/metrics/adjusted-rand-index-ari>.

[17] sklearn.metrics.normalized_mutual_info_score. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html>.