Departamento de Computação/UFS

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

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

Este estudo apresenta uma abordagem híbrida que combina técnicas de clusterização com modelos de linguagem em larga escala (LLMs) para classificação semântica de dados não supervisionados. Utilizando o *Breast Cancer Dataset*, foram aplicados três algoritmos de clusterização (K-Means, Gaussian Mixture Model e Agglomerative Clustering) para segmentar os dados clínicos em grupos não rotulados. Em seguida, os modelos GPT-4o-mini e GPT-4o da OpenAI foram utilizados para atribuição semântica aos grupos formados, classificando-os como malignos ou benignos. A performance da abordagem foi avaliada utilizando as métricas Adjusted Rand Index (ARI) e Normalized Mutual Information (NMI), comparando os agrupamentos com os rótulos reais do conjunto de dados. Os resultados indicam que as LLMs são capazes de interpretar contextualmente os agrupamentos gerados, fornecendo rótulos coerentes mesmo quando os algoritmos de clusterização apresentam baixa precisão de separação segundo as métricas convencionais, evidenciando a capacidade dos modelos de linguagem em inferir corretamente significado para os clusters gerados com eficácia considerável, mesmo em cenários de segmentação imperfeita.
Palavras-chave: Clusterização, Modelos de Linguagem, Aprendizado Não Supervisionado.

1. Introdução

Atualmente, o crescimento exponencial de dados gerados em uma sociedade cada vez mais automatizada trouxe a necessidade de uma análise de dados mais dinâmica e adaptável. Ao lidar com números grandes de dados, o Aprendizado de Máquina (Machine Learning) combinado com Modelos de Linguagem (LLMs - Large Language Models) formam uma ferramenta poderosa para auxiliar no processamento e interpretação de informações.

A clusterização, uma técnica de aprendizado não supervisionado, desempenha um papel crucial ao agrupar dados semelhantes sem a necessidade de rótulos prévios [1]. No entanto, por basear-se em medidas de similaridade entre os dados, a clusterização tradicional possui limitações, principalmente quanto à interpretação e à classificação semântica dos grupos formados. Isso expõe a necessidade de técnicas complementares que possam agregar significado aos grupos obtidos.

Neste contexto, o objetivo deste trabalho é aplicar uma LLM para classificar os dados agrupados não supervisionados e posteriormente avaliar o desempenho da abordagem. O estudo foi feito através da utilização de três técnicas conhecidas de clusterização: a K-means, Gaussian Mixture Model (GMM) e Agglomerative Clustering, aplicadas a um conjunto de dados de

exames médicos com o objetivo de auxiliar na detecção de câncer de mama, dividindo-o em dois grupos distintos. Após a divisão, a LLM escolhida foi utilizada para atribuir rótulos semânticos aos grupos, classificando-os em Benigno ou Maligno. Os resultados obtidos foram comparados com os rótulos esperados com métricas de avaliação, fornecendo uma análise do desempenho da abordagem proposta e da capacidade da LLM em identificar padrões e classificar informações atribuindo significados contextuais aos agrupamentos.

Dessa forma, a metodologia proposta demonstra sua versatilidade ao permitir a aplicação do modelo em outros conjuntos de dados com outras características e especificações. A utilização de LLMs com o seu potencial de análise semântica e contextualização, revela uma ferramenta fundamental para promover a dinamicidade e adaptabilidade dos modelos ao lidar com diferentes tipos de informações e domínios.

2. FUNDAMENTAÇÃO TEÓRICA

Ao longo do trabalho diversos conceitos e fundamentos da Inteligência Artificial e da Análise de Dados foram utilizados como base para definição do escopo e para a implementação da abordagem proposta. Nesta seção, serão apresentados brevemente os principais conceitos utilizados, entre eles: Aprendizado de Máquina, com o foco no aprendizado não supervisionado, identificação de entidades e clusterização de dados; e os Modelos de linguagem em Grande Escala, as LLMs .

2.1 APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina (Machine Learning - ML) é um ramo da Inteligência Artificial voltado para o desenvolvimento de sistemas capazes de aprender a partir de dados e tomar decisões ou fazer previsões com base em padrões identificados [2]. Essa capacidade de aprendizado sem instruções explícitas torna o ML uma ferramenta poderosa para lidar com grandes volumes de dados e resolver problemas complexos nas mais diversas áreas.

Os algoritmos de Aprendizado de Máquina podem ser utilizados em diferentes contextos a depender do tipo de informação disponível nos dados, permitindo desde análises preditivas até descobertas de padrões ocultos. Dentre as abordagens de ML pode-se identificar três principais categorias: o aprendizado por reforço, em que o modelo aprende através da tentativa e erro, recebendo recompensas ou penalidades por ações de um agente em um ambiente; o aprendizado supervisionado, no qual o modelo é treinado com dados rotulados aprendendo a mapear entradas para saídas conhecidas; e o aprendizado não supervisionado, que lida com dados não rotulados e busca identificar estruturas ocultas como agrupamentos ou relações entre variáveis, sendo o foco deste trabalho através da aplicação de técnicas de clusterização [2].

2.1.1 APRENDIZADO SUPERVISIONADO E NÃO SUPERVISIONADO.

Como já introduzido anteriormente, o aprendizado supervisionado é uma técnica de Aprendizado de Máquina na qual o modelo é treinado com base em um conjunto de dados rotulados, ou seja, os dados fornecidos para treinamento já possuem uma saída conhecida. O objetivo dessa técnica é construir um modelo que, após treinado, tenha aprendido uma função que mapeie as entradas para as saídas desejadas para que, ao receber novos dados, consiga prever corretamente os rótulos correspondentes a eles. Esse tipo de abordagem é comumente utilizada para tarefas de classificação e regressão [3].

Por outro lado, o aprendizado não supervisionado opera sobre dados não rotulados, isto é, aqueles que não possuem saídas ou categorias predefinidas, por conta disso o modelo deve ser capaz de identificar por conta própria padrões ou estruturas ocultas inerentes aos dados durante o treinamento. Dessa forma, em vez de prever uma saída conhecida para novos exemplos, o seu objetivo é organizar ou agrupar os dados de maneira a revelar características compartilhadas que não estavam explicitamente declaradas, sendo útil principalmente quando não se conhece

previamente a categorização dos dados ou quando se busca descobrir novas formas de organização dentro do conjunto analisado [3].

No contexto deste trabalho, o aprendizado não supervisionado foi adotado por sua capacidade única de analisar grandes volumes de dados sem a necessidade de um conhecimento prévio sobre sua categorização. Ao aplicar algoritmos de clusterização como K-means, Gaussian Mixture Model e Agglomerative Clustering, é possível agrupar exames de pacientes com base em suas características, revelando possíveis padrões associados a diferentes diagnósticos. No entanto, uma limitação significativa dessa abordagem é que, embora possa identificar agrupamentos naturais dos dados, não atribui significado semântico a esses grupos. Diante disso, buscando superar essa questão, essa etapa antecede a utilização de uma LLM para realizar a classificação dos dados agrupados, complementando a análise com uma camada semântica capaz de atribuir rótulos significativos aos grupos identificados.

2.1.2 Clusterização e Identificação de entidades

Clusterização é o processo de dividir um conjunto de dados em clusters de modo que os elementos dentro de um mesmo cluster apresentem alta similaridade entre si, enquanto elementos de clusters distintos sejam o mais diferentes possível. A identificação de entidades através de técnicas de clusterização representa um dos pilares fundamentais do aprendizado não supervisionado, permitindo a descoberta de padrões e estruturas ocultas em conjuntos de dados não rotulados [1]. Essas técnicas, como explicado anteriormente, operam sob o princípio básico de agrupar os elementos semelhantes e separar os distintos sem qualquer conhecimento prévio sobre suas categorias ou rótulos. No entanto, a natureza não supervisionada desses métodos introduz desafios significativos no que diz respeito à avaliação da qualidade dos clusters formados.

Embora a clusterização seja eficaz na descoberta de padrões, sua avaliação representa um grande desafio. Isso ocorre porque, na ausência de *ground truth*, isto é, uma verdade de referência com rótulos conhecidos, torna-se difícil medir objetivamente a qualidade dos agrupamentos [4]. Uma alternativa a ausência de *ground truth* é o uso de métricas internas, como a distância intra-cluster, que mede a coesão entre os elementos de um mesmo cluster, e a distância inter-cluster, que mede a separação entre diferentes clusters. Tais métricas são úteis, mas não garantem que os grupos formados tenham significado prático ou coerência semântica [5].

Outra forma de se avaliar a qualidade dos clusters é através de métricas externas quando se dispõe de algum conhecimento prévio sobre os dados, isto é, quando há *ground truth*. Algumas dessas métricas são a Rand Index (RI) que mede a similaridade entre uma clusterização obtida e a verdadeira classificação, comparando como pares de pontos são agrupados ou separados em ambas; a Adjusted Rand Index (ARI), versão normalizada do RI que leva em consideração a possível ocorrência de concordâncias aleatórias e a Normalized Mutual Information (NMI), a qual mede a quantidade de informação compartilhada entre os clusters e os rótulos verdadeiros, também normalizando o valor para permitir comparações entre diferentes configurações [5].

Neste estudo, dispõe-se dos rótulos verdadeiros dos dados do dataset utilizado, correspondentes à classificação médica definitiva dos tumores dos exames como malignos ou benignos. Essas informações, embora disponíveis, são completamente excluídas do processo de clusterização, preservando integralmente a natureza não supervisionada do experimento. No entanto, a existência desses rótulos serve para permitir uma avaliação objetiva da qualidade dos clusters formados através de métricas externas como as citadas anteriormente. Para avaliar a clusterização realizada neste trabalho as métricas de avaliação escolhidas foram a ARI e a NMI, que avaliam a similaridade entre os clusters gerados e a classificação esperada, ajustando os resultados para o acaso e considerando diferentes estruturas de agrupamento. Além disso, realiza-se também uma verificação valor a valor dos dados agrupados com o objetivo de identificar quantitativamente a porcentagem de acerto dos clusters gerados com relação aos grupos esperados.

Entretanto, ainda que este estudo utilize os rótulos reais para avaliar a qualidade dos agrupamentos formados, seu objetivo principal está centrado na simulação de cenários em que não se dispõe da saída correta dos dados, o que é comum em tarefas reais de aprendizado não supervisionado. Nesse contexto, a aplicação de um modelo de linguagem torna-se essencial, pois permite atribuir significado semântico aos clusters gerados a partir de dados não rotulados. Dessa forma, o modelo proposto busca contornar a ausência de informação prévia sobre as categorias dos dados, utilizando a capacidade interpretativa da LLM como um mecanismo complementar à clusterização tradicional.

2.2 Modelos de Linguagem

Os Modelos de Linguagem de Grande Escala (Large Language Models - LLM) são modelos de Redes Neurais Profundas (Deep Learning Network) treinadas em grande quantidade de textos e bilhões de parâmetros para processamento, geração e identificar padrões de linguagem natural [6]. Esses modelos têm se destacado pelo seu potencial em diversas tarefas de Processamento de Linguagem Natural como geração de textos, sumarização, tradução e identificação de entidades (classificação). Atualmente, o uso de aplicações baseadas em LLM são inúmeras e já fazem parte do dia a dia dos cidadãos de todo o mundo.

A utilização de LLMs com as abordagens zero-shot e few-shot learning mostram o poder desses modelos para tarefas de classificação de dados sem a necessidade de treinamento supervisionado tradicional[7]. Isso mostra que os modelos podem realizar tarefas de classificação apenas através de uma descrição textual via prompt, com ou sem exemplos. Esse potencial da LLM está diretamente relacionado à arquitetura Transformer, presente em modelos como GPT e BERT e proposta por Vaswani et al. (2017) [8]. O uso do mecanismo de Self-Attention com o Multi-Head Attention permite ao modelo capturar relações contextuais entre as palavras de uma sequência de forma paralela e simultânea, o que permite construir representações úteis para o agrupamento de dados e rotulagem semântica. Além disso, por meio de mecanismos de Encoder-Decoder e codificação posicional (Positional Encoding), os modelos que usam arquitetura Transformer aprendem representações semânticas densas e contextualizadas do texto. Dessa forma, modelo como o GPT (Generative Pre-trained Transformers), utilizado neste trabalho, mostra-se como um instrumento eficaz para classificação semântica de dados não supervisionados.

Outro aliado da proposta apresentada e das LLMs, é a utilização da técnica de Engenharia de Prompt (Prompt Engineering)[9], que consiste em uma elaboração textual e sistemática das instruções passadas ao modelo para guiar seu comportamento. Um prompt bem formulado é de extrema importância para que a LLM interprete corretamente sua função e tarefa. Ao interagir com um modelo de linguagem é importante que as instruções sejam claras, objetivas, específicas e ajustadas conforme a necessidade e natureza da tarefa. A combinação de um prompt estruturado e elaborado corretamente com a aplicação da LLM, resulta em uma grande ferramenta.

3. Metodologia

Como apresentado nos tópicos anteriores, a abordagem utilizada fez-se uso de técnicas de clusterização como K-Means e Gaussian Mixture Model (GMM) aplicadas a um conjunto de dados de pacientes para detecção de câncer de mama. Após a clusterização, o modelo gpt-4o-mini e, em alguns casos, o gpt-4o da OpenAI foi aplicado para a classificação semântica dos rótulos identificados. Posteriormente, os resultados obtidos foram analisados e comparados com a classificação esperada. Ademais, como forma de avaliação dos resultados foram empregadas as métricas Adjusted Rand Index (ARI) e Normalized Mutual Information (NMI), com objetivo de apresentar valores concretos do que foi obtido e comparado.

Os testes e experimentos realizados foram feitos pela máquina com as seguintes especificações 16 GB Intel Core i5-9300H Windows 10 e as bibliotecas utilizadas foram:

pandas, para análise e manipulação de dados em tabelas; openai, para o acesso à API da OpenAI; KMeans, para o algoritmo de clusterização; LabelEncoder, transformação de rótulos categóricos em números; GaussianMixture, modelo de mistura de gaussianas para clusterização; AgglomerativeClustering, para o algoritmo de clusterização; adjusted_rand_score, métrica para avaliar a similaridade entre agrupamentos; normalized_mutual_info_score, métrica que mede a informação mútua normalizada entre agrupamentos. Além de bibliotecas auxiliares como a json, numpy, os, e matplotlib.pyplot.

Importante destacar que o dataset foi escolhido para exemplificação da abordagem e para análise do desempenho, mas a metodologia de clusterização e aplicação da LLM não se restringe aos dados estudados; a natureza da LLM permite sua aplicação em outros conjuntos de domínio de dados. A seguir será apresentado brevemente sobre cada técnica e ferramentas necessárias na abordagem.

3.1 Técnicas de Clusterização

3.1.1 K-means

Neste trabalho foi empregado o algoritmo de clusterização K-Means para segmentar os dados referentes ao diagnóstico de câncer de mama, com o objetivo de identificar automaticamente agrupamentos de pacientes com características semelhantes, sem o uso de rótulos supervisionados. O K-Means foi escolhido por sua simplicidade, eficiência computacional e bom desempenho em tarefas onde os dados tendem a se organizar em agrupamentos esféricos bem definidos. Considerando que o conjunto de dados em questão apresenta métricas numéricas derivadas de exames de imagem (por exemplo, radius_mean, smoothness_mean, entre outras), o K-Means mostrou-se apropriado para encontrar padrões subjacentes entre os pacientes, agrupando-os com base em suas similaridades geométricas no espaço de atributos.

O algoritmo K-Means funciona iterativamente da seguinte forma [10]:

- Inicialização: Escolhe aleatoriamente um número k de centroides.
- Atribuição: Cada ponto de dado é atribuído ao centroide mais próximo com base na distância Euclidiana.
- Atualização: Os centroides são recalculados como a média dos pontos que lhes foram atribuídos.
- Convergência: O processo de atribuição e atualização é repetido até que os centroides não mudem significativamente entre iterações, ou até que um número máximo de iterações seja atingido.

O algoritmo é sensível à escala dos dados e à presença de colinearidades, por isso faz-se necessário um pré-processamento dos dados, para evitar dados altamente correlacionados. A seguir, detalhes sobre outra técnica de clusterização.

3.1.2 Gaussian Mixture Model (GMM)

O algoritmo de clusterização Gaussian Mixture Model [11] (GMM) também foi empregado com o objetivo de identificar agrupamentos de pacientes com base nas suas características clínicas e de imagens, sem a utilização de rótulos supervisionados, ou seja, sem a utilização da coluna que identifica o resultado daquela instância no conjunto de dados.

O algoritmo GMM foi escolhido visando flexibilidade e sofisticação na modelagem estatística de dados. Em contraponto com algoritmos como K-Means, que impõem restrições rígidas no formato dos seus agrupamentos (esféricos e equidistantes), o GMM permite a representação de cada cluster como sendo uma distribuição Gaussiana multivariada, possuindo média e covariância própria [12]. Tal característica torna esse algoritmo extremamente eficiente quando os dados apresentam variâncias heterogêneas ou sobreposições entre os possíveis grupos.

Como o conjunto de dados utilizado contém atributos numéricos derivados de exames médicos — como radius_mean, texture_mean e compactness_mean — que podem apresentar

diferentes escalas e correlações, o GMM demonstrou ser uma escolha adequada para capturar padrões mais sutis de agrupamento.

O algoritmo GMM opera utilizando um modelo de mistura probabilístico e utiliza o algoritmo Expectation-Maximization (EM) para realizar ajustes nos seus parâmetros [13]. O seu funcionamento ocorre em ciclos iterativos e é composto por duas etapas:

- Etapa E (Expectation): Calcula qual a probabilidade de cada amostra pertencer a cada uma das possíveis distribuições gaussianas presentes, com base nos parâmetros atuais, tais como média e covariância.
- Etapa M (Maximization): Atualiza os parâmetros das distribuições gaussianas presentes de forma a maximizar a verossimilhança nos dados observados.

3.1.3 Agglomerative clustering

Outra técnica de clusterização empregada foi a Agglomerative Clustering [14]. Ao contrário de outras técnicas, a manipulação de dados inicia-se levando em consideração que cada ponto de dados possui seu próprio cluster individual. A partir das semelhanças entre os clusters, eles vão sendo agrupados até a quantidade de grupos definidos seguindo sempre a similaridade entre os dados. Dessa forma, o Agglomerative utiliza-se de uma abordagem bottom-up [15]. A técnica possui três etapas principais:

1. **Inicialização:** Cada ponto de dados, um cluster individual.
2. **Agrupamento Iterativo:** O algoritmo mede a distância entre os pares de clusters e une os mais próximos. Para definir essa proximidade o algoritmo utiliza de linkage methods, como single linkage, complete linkage, average linkage e ward linkage. A cada passo, a quantidade de cluster reduz em uma unidade, e esse processo continua até que reste o número de clusters desejados (ou outra condição de parada)
3. **Formação da hierarquia:** Durante o processo do algoritmo, uma árvore do agrupamento é criada, chamada de dendrograma. A árvore mostra como os clusters foram sendo unidos e agregados ao longo do tempo.

3.2 Uso da OpenAI

Neste trabalho foi utilizado a API da OpenAI para auxiliar na etapa de interpretação dos resultados obtidos através dos algoritmos de clusterização não supervisionada, no contexto de classificação de câncer de mama. Após a aplicação dos modelos Gaussian Mixture Model (GMM) e K-Means ao dataset *Breast Cancer DataSet*, os dados foram agrupados em clusters, tornando-se necessária a interpretação da natureza de cada agrupamento obtido. Ou seja, é necessário associar cada cluster à sua provável classificação - maligno ou benigno.

Como o processo de clusterização adotado é não supervisionado, os rótulos originais do conjunto de dados - que indicam a classificação correta de cada instância - não são utilizados durante a formação dos clusters. Isso torna a etapa de interpretação essencial para avaliar se os resultados da clusterização são coerentes com relação às informações reais presentes no conjunto de dados adotado.

Para realizar a verificação da coerência dos resultados obtidos com a clusterização com os dados reais do conjunto de dados, foi adotado o modelo **gpt-4o-mini**, disponibilizado pela OpenAI. A escolha desse modelo se deve ao equilíbrio entre desempenho e custo computacional, isso somado ao fato do mesmo conseguir compreender e interpretar de forma eficiente grandes volumes de dados estruturados no formato JSON; também houve a questão da disponibilidade e do acesso ao modelo dentro das limitações para execução dos testes que influenciou na decisão. Ademais, sua competência em análise semântica permite uma interpretação mais detalhada das características dos dados, o que contribui para uma precisão na associação dos clusters gerados com sua provável classificação - maligno ou benigno.

Em casos em que o **gpt-4o-mini** não interpretou os dados de forma satisfatória, foram realizados testes pontuais com o modelo **gpt-4o**, escolhido por sua performance superior, embora com um custo de utilização mais elevado.

3.3 Adjusted Rand Index (ARI) e Normalized Mutual Information (NMI)

A qualidade da clusterização é um desafio importante na avaliação de problemas de aprendizado não supervisionado, principalmente quando não há uma categorização explícita dos dados. Uma vez que, neste trabalho, os rótulos verdadeiros são conhecidos, foi possível utilizar métricas externas para medir o quanto os agrupamentos formados correspondem à classificação esperada. Para tal finalidade, optou-se pela utilização do Adjusted Rand Index e da Normalized Mutual Information devido à sua robustez, confiabilidade e ampla utilização na literatura científica para avaliação de clusterização.

O Adjusted Rand Index é uma métrica usada para medir o grau de semelhança entre dois agrupamentos, corrigindo o valor calculado de maneira a eliminar a concordância que poderia ocorrer ao acaso. Seu valor pode variar de -1 a 1, onde -1 significa completa discordância, 0 é o que seria esperado de uma atribuição aleatória e 1 indica concordância perfeita. Por essa razão, o ARI é uma métrica adequada para a avaliação de diferentes resultados de clusterização, uma vez que a normalização permite a comparação independentemente do número de clusters ou da distribuição dos dados [16].

Por outro lado, a Normalized Mutual Information mede a quantidade de informação mútua compartilhada entre os clusters gerados e os rótulos reais, sendo normalizada para produzir valores entre 0 e 1. Valores próximos de 0 indicam baixa correspondência entre os agrupamentos e as classes esperadas, enquanto valores próximos de 1 sugerem alta similaridade. Uma das principais vantagens da NMI é que ela não depende da ordem dos rótulos dos grupos, o que a torna adequada para avaliar agrupamentos cujos rótulos não possuem uma ordem predefinida [17].

Essas métricas foram aplicadas após a execução dos algoritmos de clusterização, utilizando os rótulos reais apenas para fins de avaliação. A partir delas, foi possível quantificar a precisão dos agrupamentos em relação às classes esperadas, permitindo verificar a eficácia da clusterização antes da intervenção do modelo de linguagem para a classificação semântica dos dados.

3.4 Descrição do Dataset

O conjunto de dados selecionado para os testes deste trabalho é o *Breast Cancer Dataset*, disponível na plataforma Kaggle por meio do seguinte link: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset. Esse DataSet tem como objetivo principal auxiliar na detecção de câncer de mama com base em características observadas e extraídas em exames realizados nos pacientes. O conjunto de dados apresenta informações que possibilitam a construção de modelos de previsão para classificação dos casos de câncer como **malignos** — representados pela letra M — ou **benignos** — representados pela letra B.

Informações gerais:

- Quantidade de linhas: 569
- Quantidade de colunas: 32
- Coluna objetivo: **diagnosis**

Características presentes no DataSet:

- Identificador: coluna utilizada para identificação no dataset. Neste dataset, essa coluna é nomeada como **id.**
- Diagnóstico: coluna utilizada para rotular se o câncer em questão é maligno e benigno.
- Atributos descritivos: 30 variáveis extraídas de imagens de exames de mama. Tais variáveis descrevem características como textura, perímetro, área, suavidade, entre outros.

Dessa forma, o DataSet possui 30 atributos numéricos derivados dessas medições, além da coluna **id** e da variável de saída **diagnosis**. Este conjunto de dados é amplamente utilizado em tarefas de aprendizado de máquina supervisionado, especialmente em problemas de classificação binária.

4. Experimentos

Nesta seção será apresentado o passo da implementação proposta e o que foi necessário em cada etapa. Os resultados obtidos serão apresentados detalhadamente no próximo tópico de Resultados e Discussões.

4.1 Processamento do Dataset

Como informado anteriormente, o *dataset* escolhido foi o *Breast Cancer Dataset.* Para manipulação e pré processamento dos dados algumas ações foram necessárias.

1. Transformação de colunas não numéricas em labels utilizando a LabelEncoder: Como é necessário retirar colunas não numéricas, a coluna diagnosis antes B, para benigno e M para maligno, agora a referência é 0, para benigno e 1 para maligno.
2. Remoção da coluna de identificação (“id”) por não apresentar relevância no estudo dos dados;
3. Eliminação de colunas altamente correlacionadas (correlação maior que 0.87), a fim de reduzir a redundância nos dados e evitar viés no agrupamento;
4. Remoção da coluna de diagnóstico (“diagnosis”) para clusterização;

4.2 Aplicação das técnicas de clusterização

Para aplicação das técnicas de clusterização, foi fixado classificação binária (n_components = 2/n_clusters = 2) dos clusters, pois o diagnóstico pode ser Benigno ou Maligno. Para cada técnica de clusterização, um gráfico foi criado mostrando a divisão e disposição dos clusters criados. A visualização dos gráficos será apresentada juntamente com os resultados. A avaliação dos clusters foi validada por meio das métricas Adjusted Rand Index (ARI) e Normalized Mutual Information (NMI), as quais quantificam a similaridade entre os agrupamentos produzidos pelas técnicas e os rótulos reais de diagnóstico - apenas usados para avaliação, e não para treinamento.

4.3 Aplicação da OpenAI

Após a divisão binária dos dados, o próximo passo foi realizar uma chamada de requisição para o OpenAI, para que o modelo escolhido classificasse o conjunto dos dados obtidos. Foram utilizados os modelos **gpt-4o-mini** e o **gpt-4o**; em alguns casos em que o primeiro não apresentou resultados satisfatórios, o segundo foi utilizado por sua performance superior. A elaboração do prompt seguiu esta estrutura: informa possíveis classificações; repassa cada cluster separadamente; repassa o contexto dos dados; e define um formato especificado para o retorno da chamada. Segue o prompt utilizado em ambos os modelos:

```
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 Aplicação das métricas de Avaliação

Após a aplicação da LLM, houve uma verificação visual dos resultados e após essa verificação as métricas de ARI e NMI foram aplicadas. As métricas comparavam a coluna "diagnosis" original com a classificação feita pela LLM. Dessa forma, foi possível quantificar e comparar os resultados obtidos.

5. Resultados e Discussão

Nesta seção serão apresentados todos os resultados obtidos, juntamente com as métricas e gráficos. Cada subtópico representa os resultados de uma técnica de clusterização.

5.1 Resultados da clusterização realizada pelo K-Means

Segue o gráfico mostrando a distribuição feita pelo K-Means no conjunto de dados. A cor roxa representa o cluster 0 e a cor amarela representa o cluster 1.

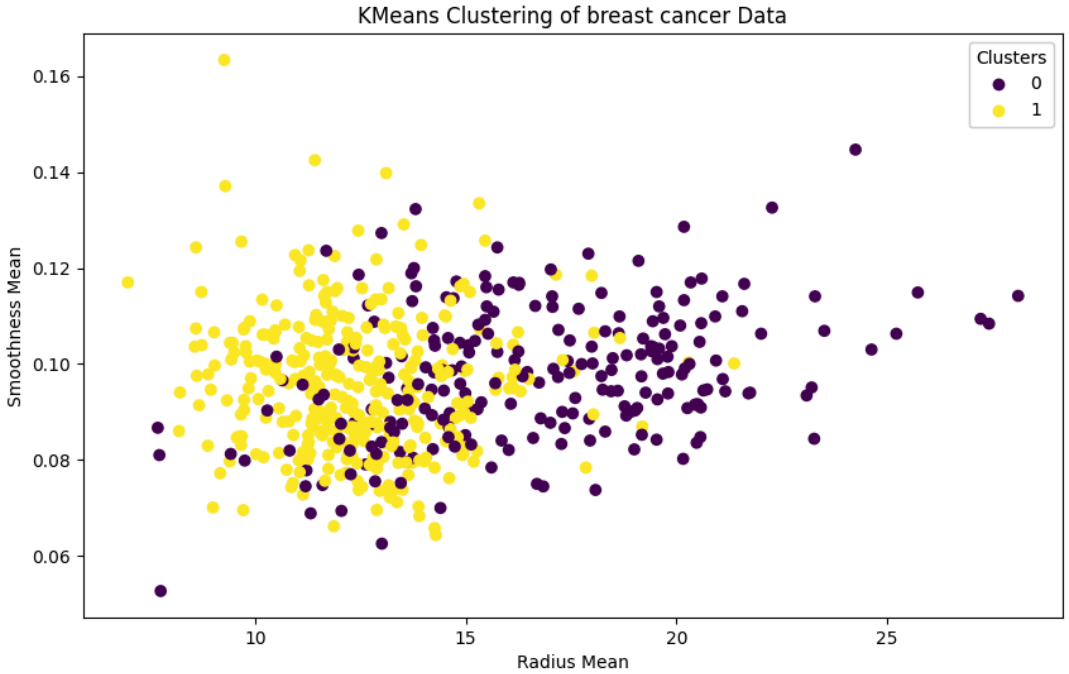

Figura 1: Gráfico de distribuição do KMeans

Os resultados obtidos pelo K-Means foram:

```
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%
```

No K-Means a LLM falhou em classificar corretamente os grupos, possivelmente, em consequência da baixa precisão da separação dos grupos. Pelo Adjusted Rand Index (ARI) o valor foi aproximadamente 44%, indicando uma baixa concordância entre os agrupamentos e os diagnósticos reais. Pelo Normalized Mutual Information (NMI) o valor foi cerca de 34%, reforçando uma baixa qualidade informacional da segmentação realizada. Mesmo que mais de 86% dos dados estejam classificados corretamente, a separação dos grupos não é tão boa quanto a do GMM, o que pode ter influenciado na classificação final. Dessa forma, foi testada a clusterização realizada pelo K-means com o modelo gpt-4o para melhor análise do desempenho.

```
Kmeans com o modelo 4o (similaridade entre os clusters e os valores do
dataset) breast cancer: 83.30%
```

Com o modelo superior da OpenAI a classificação foi melhorada e foi possível identificar corretamente a qual grupo os dados pertenciam, mas o desempenho ainda assim foi menor que o GMM, devido à classificação dos *clusters* em si.

5.2 Resultados da clusterização realizada pelo GMM

Segue o gráfico mostrando a distribuição feita pelo GMM no conjunto de dados. A cor roxa representa o cluster 0 e a cor amarela representa o cluster 1.

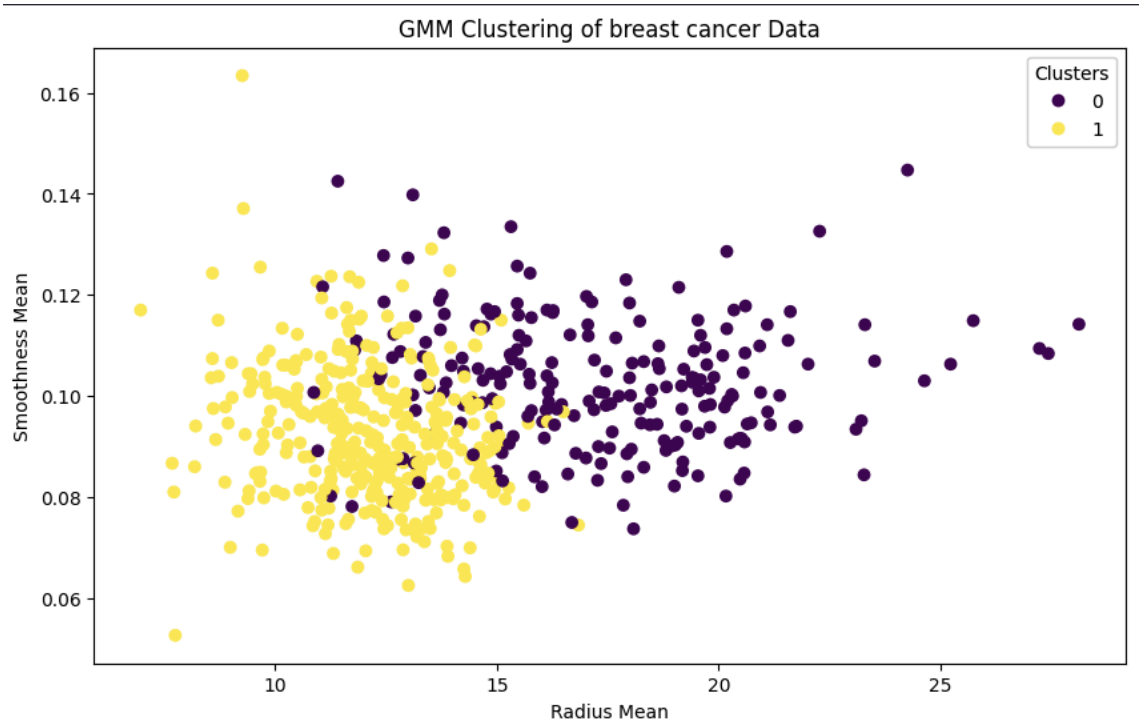

Figura 2: Gráfico de distribuição do GMM

Os resultados obtidos pelo GMM foram:

```
GMM (ari_score) breast cancer: 77%
GMM (nmi_score) breast cancer: 66%
```

GMM (similaridade entre os clusters e os valores do dataset) breast cancer: 94.02%

O algoritmo GMM teve o melhor desempenho entre os algoritmos testados para esse *dataset* utilizado. Pelo ARI, o GMM foi avaliado aproximadamente em 77%, indicando que há uma boa concordância entre os clusters gerados pelo GMM e o diagnósticos reais derivados do conjunto de dados analisados. Pelo NMI, foi inferido em aproximadamente 66%, o que sugere que a segmentação obtida teve sucesso em capturar parte significativa da estrutura informacional dos dados.

Os resultados obtidos utilizando do GMM evidenciam que o mesmo é capaz de realizar segmentações úteis de dados, agrupamentos coerentes com os padrões clínicos obtidos nos dados associados ao câncer de mama, mesmo sem a utilização dos rótulos de classificação. Os resultados anteriores demonstram que o GMM tem enorme potencial como ferramenta para realizar agrupamentos de dados em diversos contextos. Tal qual no uso do K-Means, o pré-processamento de dados se mostrou essencial para garantir a qualidade do agrupamento. Uma análise mais detalhada e profunda se mostra necessária para identificar os benefícios e forças de cada algoritmo, já que será um fator decisivo para um bom desempenho da classificação final.

5.3 Resultados da clusterização realizada pelo Agglomerative Clustering

Segue o gráfico mostrando a distribuição feita pelo Agglomerative Clustering no conjunto de dados. A cor roxa representa o cluster 0 e a cor amarela representa o cluster 1

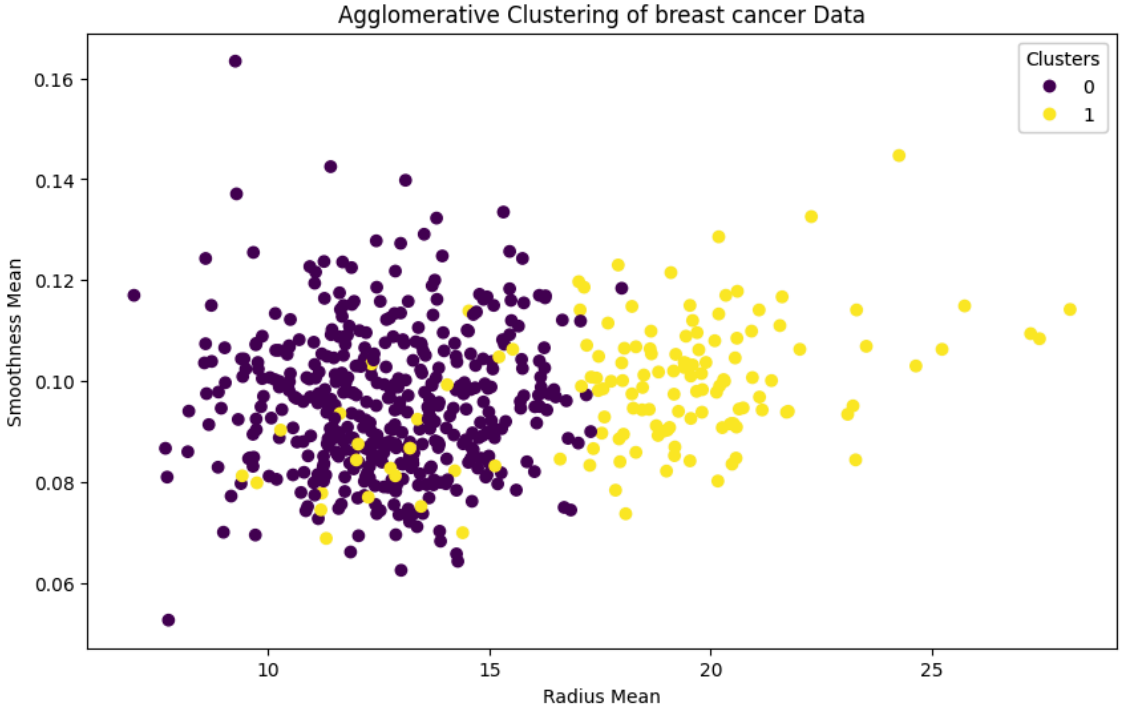

Figura 3: Gráfico de distribuição do Agglomerative Clustering

Os resultados obtidos pelo Agglomerative foram:

Agglomerative (ari_score) breast cancer: 35%
Agglomerative (nmi_score) breast cancer: 27%
Agglomerative (similaridade entre os clusters e os valores do dataset) breast cancer: 80.14%

No Agglomerative Clustering, a LLM conseguiu classificar corretamente os grupos. Por mais que os índices de classificação fossem piores, o padrão de divisão apresentado pelo

algoritmo foi mais próximo com a divisão do *dataset* original. Possivelmente, isso pode ter influenciado na interpretação da LLM na tomada de decisão a partir dos dados apresentados.

6. Conclusão

Em suma, o presente estudo explorou a aplicação de técnicas de *clustering* não supervisionado abordando algoritmos como K-Means, Gaussian Mixture Model (GMM) e Agglomerative Clustering para a segmentação de um conjunto de dados de diagnóstico de câncer de mama. A principal finalidade foi identificar agrupamentos essenciais nos dados sem o recurso de rótulos supervisionados, simulando cenários onde o conhecimento prévio sobre a classificação das amostras é limitado. Ainda assim, a rotulação inicial da base de dados foi utilizada como *ground truth*. A utilização de um modelo de linguagem, especificamente o GPT, demonstrou ser promissora para atribuir significado semântico a esses agrupamentos, superando uma das limitações inerentes aos métodos de clusterização tradicionais.

A experimentação mostrou que essa abordagem pode ser muito poderosa principalmente por:

- Testar o conhecimento de domínio da LLM: A LLM é capaz de correlacionar estatísticas (como média do raio) com conhecimento clínico (tumores malignos são maiores);
- Revela a utilidade de *Clusters*: Os *clusters* não supervisionados podem ser impulsionados por ruído e não necessariamente se alinharem com conhecimento do domínio;
- Educacional: Destaca desafios no aprendizado não supervisionado (por exemplo, os clusters podem refletir idade ou viés de medição, e não malignidade).

Os resultados finais mostram que, com uma análise mais detalhada do dataset, é possível chegar em resultados melhores, já que ficou claro que diferentes algoritmos levam a resultados extremamente diferentes, e entender o contexto no qual esses dados estão inseridos pode ajudar muito para aprimorar o desempenho dessa classificação aliada às LLMs.

7. Referências Bibliográficas

[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>.