Revista Brasileira de Ensino de Física, vol. 47, e20240443 (2025)
www.scielo.br/rbef
DOI: https://doi.org/10.1590/1806-9126-RBEF-2024-0443

Introdução a métodos de aprendizado de máquina não supervisionados através de um experimento simples de medidas de densidades de sólidos

Introducing unsupervised machine learning through a simple experiment on measuring the density of solids

Daniel N. Fachieri[1], Alexandre A. P. Suaide*[1]

[1]Universidade de São Paulo, Instituto de Física, 05508-090, São Paulo, SP, Brasil.

Recebido em 04 de dezembro de 2024. Revisado em 17 de março de 2025. Aceito em 22 de abril de 2025.

Este artigo apresenta uma abordagem didática para introduzir métodos de aprendizado de máquina não supervisionado utilizando um experimento prático de medição de densidades de sólidos plásticos em laboratório de física. No experimento, diferentes tipos de plásticos são analisados quanto à sua densidade, utilizando um procedimento experimental simples de medida de massa e volume. Os dados obtidos são então processados com algoritmos de clusterização, como K-means, DBSCAN e mistura gaussiana, para classificar os materiais com base em suas propriedades. A metodologia visa integrar conceitos de física experimental e aprendizado de máquina, proporcionando aos alunos uma experiência interdisciplinar que une análise de dados e processos de medidas. Os resultados mostram como a aplicação prática facilita a compreensão dos algoritmos e estimula o pensamento crítico sobre a análise de dados.
Palavras-chave: Física experimental, aprendizado de máquina, clusterização.

This paper proposes a pedagogical framework for introducing unsupervised machine learning techniques through a hands-on experiment conducted in a physics laboratory, focused on measuring the densities of plastic solids. The experiment involves analyzing various types of plastics by determining their densities through straightforward mass and volume measurements. The collected data are subsequently processed using clustering algorithms, including K-means, DBSCAN, and Gaussian Mixture Models, to classify the materials based on their intrinsic properties. This interdisciplinary approach bridges concepts from experimental physics and machine learning, offering students an integrative learning experience that emphasizes both data analysis and measurement methodologies. The findings highlight the effectiveness of practical applications in enhancing students' understanding of clustering algorithms while fostering critical thinking and analytical skills in data interpretation.
Keywords: Experimental physics, machine learning, clustering.

1. Introdução

A física experimental é um dos pilares fundamentais na formação de estudantes de física, oferecendo a oportunidade de consolidar conceitos teóricos por meio de experiências práticas e de desenvolver habilidades críticas, como planejamento experimental, coleta de dados, análise quantitativa e interpretação de resultados. Contudo, o avanço da ciência e da tecnologia tem transformado significativamente a maneira como os dados são gerados, analisados e interpretados, tornando essencial a integração de ferramentas modernas de análise ao ensino da física. Métodos de aprendizado de máquina e inteligência artificial, por exemplo, têm desempenhado um papel crescente em diversas áreas científicas e tecnológicas [1–5], permitindo a análise eficiente de grandes volumes de dados, a identificação de padrões complexos e a automação de processos decisórios.

Além disso, essas técnicas estão cada vez mais presentes no cotidiano, desde a recomendação de conteúdos em plataformas digitais até a otimização de sistemas de transporte e diagnóstico médico [6]. No campo científico, o aprendizado de máquina tem se tornado indispensável para resolver problemas complexos, como a análise de dados de grandes experimentos em física de partículas [7–9], a modelagem de sistemas físicos não lineares e a predição de propriedades de materiais [3, 10].

Diante desse cenário, a inserção de técnicas de aprendizado de máquina [11] tem sido tópico de inovação dentro do ensino de física [12, 13] visando a capacitação dos futuros profissionais a utilizar essas ferramentas de ciências de dados para enfrentar desafios científicos e tecnológicos do presente e do futuro. Este artigo apresenta uma proposta didática que integra conceitos tradicionais de física experimental, como a medição de densidades de sólidos plásticos, com o uso de algoritmos de aprendizado de máquina não supervisionado, como K-means [14], DBSCAN [15] e mistura gaussiana [16]. Essa abordagem interdisciplinar não apenas reforça os

*Endereço de correspondência: suaide@usp.br
Editor-Chefe: Marcello Ferreira

fundamentos experimentais e analíticos, mas também introduz os estudantes ao potencial transformador da inteligência artificial, preparando-os para lidar com problemas mais complexos e para explorar novas fronteiras da ciência e da tecnologia.

Este artigo está organizado em três principais etapas, que refletem o fluxo lógico da proposta didática. Inicialmente, descrevemos o processo de coleta de dados, onde são realizadas medidas de massa e volume de diferentes sólidos plásticos, todos com formato cilíndrico. O volume é calculado a partir das dimensões dos objetos, medindo seu diâmetro e altura com o auxílio de instrumentos básicos, como paquímetros e balanças de precisão. Esse procedimento fornece um conjunto de dados experimentais simples, ideal para introduzir conceitos de análise quantitativa.

Em seguida, realizamos uma análise tradicional baseada em estatística descritiva e inferencial, explorando como as propriedades físicas dos materiais podem ser comparadas e organizadas. Estatísticas como média e desvio padrão são calculados, permitindo aos alunos compreenderem a dispersão e a variabilidade nos dados coletados. Essa etapa também serve para destacar limitações de abordagens puramente estatísticas quando aplicadas a problemas que envolvem classificação de materiais.

Na terceira e última etapa, o foco recai sobre o problema de clusterização, no qual os dados experimentais são tratados com algoritmos de aprendizado de máquina não supervisionado. Abordamos diferentes métodos, incluindo K-means, DBSCAN e modelos de mistura gaussiana, explicando suas abordagens conceituais, vantagens e limitações. Por exemplo, discutimos como o K-means assume clusters de forma esférica e requer a definição do número de grupos a priori, enquanto o DBSCAN é mais robusto a outliers e identifica clusters de forma arbitrária, mas depende de parâmetros sensíveis, como a distância de vizinhança. Já os modelos de mistura gaussiana, por sua vez, são baseados em distribuições probabilísticas e oferecem maior flexibilidade, mas demandam maior capacidade computacional e compreensão matemática.

Ao longo do texto, são apresentados os resultados obtidos com cada abordagem, permitindo uma comparação crítica entre os métodos e evidenciando como a integração de aprendizado de máquina pode enriquecer o entendimento de conceitos físicos e analíticos. Dessa forma, o artigo propõe não apenas uma aplicação interdisciplinar, mas também um recurso pedagógico relevante para cursos de física que buscam incorporar ferramentas contemporâneas de ciência de dados.

2. Método Experimental

O experimento consiste na medida e análise de aproximadamente 200 cilindros plásticos confeccionados a partir de quatro materiais distintos: nylon, polietileno, acrílico e PVC (na Tabela 1 são mostradas as densidades típicas desses materiais, de acordo com a Ref. [17]). Os cilindros foram confeccionados em torno mecânico de forma a apresentam variações aleatórias em suas dimensões, de um cilindro a outro, com diâmetros variando entre 20 e 25 mm e alturas entre 10 e 25 mm. A escolha de cilindros, e apenas cilindros, facilita a construção das amostras, além de ter o mesmo efeito estatístico nas flutuações dos resultados por conta das precisões das medidas realizadas. Essa diversidade dimensional e material foi planejada para gerar um conjunto de dados suficientemente rico, com flutuações nas medidas adequadas à aplicação de métodos de análise estatística e aprendizado de máquina.

Tabela 1: Densidades nominais dos materiais plásticos utilizados no experimento, de acordo com Ref. [17].

Material	Densidade Nominal (g/cm^3)
Polietileno	0.926–0.941
Nylon	1.13–1.15
Acrílico	1.18–1.20
PVC	1.36–1.45

As medidas foram realizadas utilizando-se equipamentos comuns, presentes em laboratórios didáticos de física. As massas dos cilindros foram medidas utilizando uma balança digital com precisão de 0,1 g, enquanto as dimensões (diâmetro e altura) foram determinadas com um paquímetro digital de precisão 0,1 mm. O volume de cada cilindro foi calculado utilizando a Eq. 1:

$$V = \pi r^2 h \tag{1}$$

onde r é o raio do cilindro e h sua altura. A densidade (ρ) de cada amostra foi então calculada com 2:

$$\rho = \frac{m}{V}, \tag{2}$$

onde m é a massa medida e V o volume calculado. Em um ambiente didático, essa tarefa poderia ser realizada pelos estudantes, permitindo que eles se familiarizem com o uso de instrumentos como balança e paquímetro, bem como com o processo de tabulamento e organização de dados. Em uma turma de 20 estudantes, poderia-se distribuir, aleatoriamente, 10 cilindros para cada um realizar as medidas. Esse tipo de experimento é realizado tradicionalmente distribuindo aos alunos um conjunto de objetos feito do mesmo material, de modo que cada aluno (ou grupo) deva identificar qual é o material recebido, através da análise de média e desvio padrão, comparando seus resultados aos valores esperados na Tabela 1. Contudo, a distribuição de cilindros de forma aleatória permite que cada aluno não consiga identificar o material recebido (a menos de diferenças visuais, como coloração e transparência, características de cada material), sugerindo a necessidade de processos mais avançados de processamento dos dados tomados. Os dados de cada aluno seriam, então, unificados em um

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2024-0443

único banco de dados, consistindo de características como massa, volume e densidade de cada cilindro. Seria possível também atribuir um rótulo a cada cilindro, permitindo que cada aluno possa, à posteriori, identificar suas medidas. Fica a critério do professor que conduz o experimento.

Nesse trabalho o processo experimental foi realizado pelos autores, conduzido com cuidado para minimizar erros sistemáticos, garantindo a repetição das medições em casos de discrepâncias significativas. Eventualmente, a realização de medidas em sala de aula por alunos pode apresentar maiores flutuações e desvios sistemáticos por conta da inexperiência deles em utilizar equipamentos e realizar medições. Os dados coletados constituem um conjunto realista, refletindo variações esperadas tanto nas propriedades dos materiais quanto nos processos de medição, sendo ideal para análises de clusterização e introdução de técnicas de aprendizado de máquina não supervisionado. Os dados utilizados nesse trabalho, bem como os programas de análise, estão disponíveis em [18]. Na impossibilidade de realizar o experimento, esse conjunto de dados pode ser utilizado livremente como ferramenta para exercitar os métodos e análise descritos nesse trabalho.

3. Apresentação e Avaliação Estatística dos Dados

Na Figura 1 são mostrados os dados obtidos durante o processo de medida. Nessa figura, o gráfico da massa medida em função do volume calculado para cada cilindro evidencia pelo menos três faixas com coeficientes angulares diferentes visualmente. Uma análise estatística, através de um ajuste de retas implicaria em fazer uma divisão desse conjunto de dados em subconjuntos. Nesse caso, há uma certa dificuldade em fazer essa separação, principalmente na região de volumes menores, onde as flutuações estatísticas não permitem uma clara separação entre as faixas.

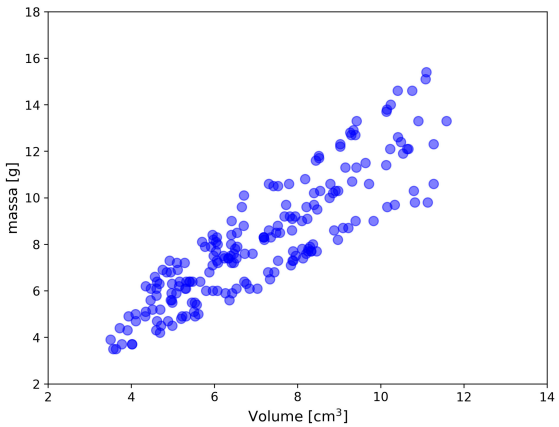

Figura 1: Distribuição das massas em função dos volumes para cada cilindro medido durante a realização do experimento.

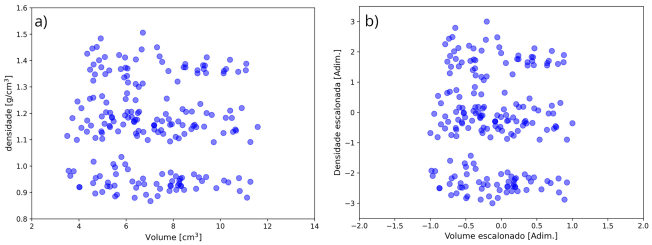

Figura 2: Distribuição das densidades dos cilindros em função do volume. a) distribuição de densidades em função do volume medidas experimentalmente. b) distribuição das densidades em função dos volumes após o escalonamento uniforme.

Uma segunda abordagem seria calcular a densidade para cada cilindro. Na Figura 2 é mostrado um gráfico da densidade em função do volume, onde fica mais evidente a presença de pelo menos 3 conjuntos com diferentes densidades. Mais uma vez, há uma certa dificuldade em separar esse conjunto em subgrupos para uma análise individual de cada um deles.

Nesse caso, uma abordagem natural seria avaliar esse conjunto de dados estatisticamente, ou seja, avaliar a distribuição das medidas e compará-la a alguma função densidade de probabilidade. Na Figura 3 é mostrado o histograma das densidades obtidas experimentalmente, onde é perceptível visualizar uma distribuição tri-modal. Esse histograma apresenta o comportamento de uma distribuição normal, reforçado pelo teste de Kolmogorov-Smirnov (KS) a partir do valor-p calculado em cada intervalo dos dados (conforme mostrado na Tabela 2), onde podemos ajustar uma curva gaussiana para então extrair o valor médio juntamente e o desvio padrão do conjunto de dados, sendo a incerteza desses valores provenientes do ajuste das curvas gaussianas pelo algoritmo *curve_fit* da biblioteca scipy [19]. Esses valores, que estão apresentados na Tabela 2, são então comparados aos valores da Tabela 1 de modo a identificar os componentes plásticos dos cilindros estudados.

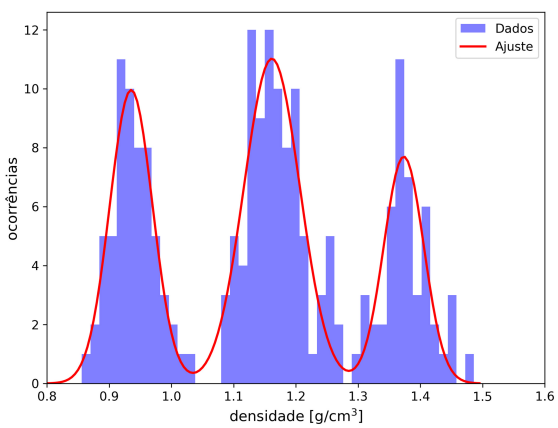

Figura 3: Histograma com as densidades medidas experimentalmente e um ajuste com três funções gaussianas.

Tabela 2: Resultado do ajuste de três gaussianas aos dados de densidades mostrados na Figura 3.

Valor médio	Desvio padrão	Valor-p teste KS	Material identificado
0.935 ± 0.003	0.034 ± 0.004	0.907	Polietileno
1.161 ± 0.004	0.045 ± 0.004	0.594	Nylon ou Acrícilo
1.373 ± 0.005	0.032 ± 0.005	0.576	PVC

4. Métodos de Aprendizado Não Supervisionado e Clusterização

Apesar do procedimento descrito anteriormente ser suficiente para que possamos avaliar globalmente o conjunto de dados, ainda resta a pergunta de como poderíamos proceder para subdividir esse conjunto em amostras separadas, cada uma delas representando o mesmo tipo de material. Após a apresentação e análise estatística inicial dos dados, foram aplicados métodos de aprendizado de máquina não supervisionado com o objetivo de identificar automaticamente padrões e separar o conjunto de dados em subconjuntos, sem a necessidade de rótulos prévios. Esse processo de clusterização é particularmente útil para explorar agrupamentos naturais no conjunto de dados, como os associados aos diferentes tipos de plástico com base em suas densidades e volumes.

Nesse trabalho, usamos três algoritmos de clusterização, com características distintas. o k-means, DBSCAN e Mistura Gaussiana. Cada um desses métodos possui vantagens e desvantagens, a depender das características do conjunto de dados. Contudo, uma etapa importante anterior ao uso de qualquer um desses algoritmos é a preparação (pré-processamento) dos dados.

4.1. Pré-processamento dos dados: necessidade e técnicas

O pré-processamento é uma etapa fundamental no uso de métodos de aprendizado de máquina, especialmente em problemas de clusterização, onde a qualidade dos dados influencia diretamente a eficácia dos algoritmos. Neste experimento, a necessidade de pré-processar os dados surgiu devido às diferentes escalas das variáveis medidas: massa (gramas), volume (cm^3) e densidade (g/cm^3). Variáveis com magnitudes muito distintas podem dominar as métricas de distância (como a distância euclidiana) utilizadas por muitos algoritmos de clusterização, resultando em agrupamentos enviesados. Portanto, o escalonamento dos dados é fundamental para equalizar as variáveis e garantir que cada dimensão tenha uma contribuição equilibrada no processo de clusterização.

4.1.1. Escalonamento uniforme: conceito e aplicação

O escalonamento uniforme é uma técnica de pré-processamento que transforma todas as variáveis para uma escala comum, mantendo suas distribuições relativas. Isso garante que nenhuma variável exerça influência desproporcional sobre os algoritmos que utilizam distâncias para determinar clusters. Uma abordagem comum é normalizar ou padronizar os dados:

- **Normalização (Min-Max Scaling):** Transforma os dados para que fiquem dentro de um intervalo fixo, geralmente $[0, 1]$, usando a fórmula 3:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

  Aqui, $x_{\min}$ e $x_{\max}$ são os valores mínimo e máximo da variável. Essa técnica é útil quando se deseja preservar a proporcionalidade entre os valores, como no caso de medidas de massa e volume.
- **Padronização (Z-Score Scaling):** Converte os dados para uma distribuição com média zero e desvio padrão unitário, usando a fórmula 4:

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

  Onde μ é a média e σ é o desvio padrão da variável. A padronização é mais apropriada quando as variáveis têm distribuições gaussianas ou há outliers que poderiam distorcer os resultados de uma normalização simples.

4.1.2. Importância do escalonamento no contexto do experimento

No contexto do experimento, o escalonamento foi necessário para garantir que as variáveis massa, volume e densidade fossem tratadas de forma equitativa pelos algoritmos de clusterização. Por exemplo:

- No **K-means**, a métrica de distância euclidiana poderia ser dominada por variáveis com maiores magnitudes (como massa), resultando em clusters que refletem mais essas variáveis do que a densidade, que tem uma magnitude menor.
- No **DBSCAN**, o parâmetro de vizinhança (ε) depende diretamente das distâncias entre os pontos. Sem escalonamento, a escolha de ε seria influenciada pela variável com maior escala, dificultando a identificação de clusters densos.
- No **GMM**, a probabilidade de cada ponto pertencer a um cluster é influenciada pela covariância das variáveis. Variáveis não escalonadas podem levar a matrizes de covariância distorcidas, afetando o ajuste dos modelos.

4.1.3. Escolha do escalonamento uniforme

Para este estudo, foi adotada a **normalização uniforme (Min-Max Scaling)**, dado que as variáveis massa, volume e densidade apresentavam escalas distintas que precisavam ser ajustadas para um intervalo

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2024-0443

comum. A normalização permitiu que todas as variáveis contribuíssem igualmente para as métricas de distância utilizadas pelos algoritmos de clusterização, garantindo que nenhuma variável dominasse as demais. Além disso, a preservação da proporcionalidade entre os valores foi crucial para manter a integridade das relações entre massa, volume e densidade durante a análise. A Figura 2-b mostra o gráfico da densidade em função do volume após aplicarmos o escalonamento uniforme. Nesse caso, os volumes foram escalonados no intervalo $[-1, 1]$ enquanto as densidades, calculadas pela divisão da massa pelo volume (eq. 2) antes do escalonamento, foram escalonadas no intervalo $[-3, 3]$, de modo que cada cluster tivesse aproximadamente a mesma largura nessas duas variáveis. Alguns algoritmos, como o k-means, possuem maior robustez se os clusters identificados tiverem forma mais próxima de circunferências ou esferas. Por conta disso fizemos a escolha desses intervalos de escalonamento.

A aplicação do Min-Max Scaling facilitou a identificação precisa dos clusters, uma vez que todas as variáveis estavam na mesma escala, melhorando a performance dos algoritmos de clusterização utilizados. Essa escolha de técnica de escalonamento demonstrou ser eficaz para equilibrar a influência das diferentes variáveis e maximizar a capacidade dos métodos de aprendizado de máquina em detectar agrupamentos naturais nos dados experimentais.

A adoção de técnicas de escalonamento demonstra como o pré-processamento é crucial para garantir a robustez dos métodos de aprendizado de máquina e para explorar ao máximo o potencial dos dados experimentais.

4.2. Algoritmos de clusterização utilizados

Neste trabalho, foram aplicados três algoritmos de clusterização com abordagens distintas para identificar padrões nos dados experimentais: **K-means**, **DBSCAN** e **Modelos de Mistura Gaussiana (GMM)**. Como variável de entrada para os modelos utilizamos x_1 = volume escalonado, e x_2 = densidade escalonada. Cada método apresenta características próprias que atendem a diferentes necessidades e cenários de análise. A seguir, descrevemos os princípios e características principais de cada algoritmo.

4.2.1. K-means

O **K-means** é um algoritmo de clusterização baseado na minimização da soma das distâncias quadradas entre os pontos de dados e os centros dos clusters. O procedimento pode ser resumido nas seguintes etapas:

1. Escolha inicial dos k centros dos clusters, seja aleatoriamente ou com base em um critério pré-definido.
2. Atribuição de cada ponto ao cluster cujo centro esteja mais próximo (com base na distância euclidiana).
3. Atualização dos centros dos clusters como a média dos pontos pertencentes a cada cluster.
4. Repetição das etapas 2 e 3 até que os centros dos clusters não se alterem significativamente ou um critério de convergência seja atingido.

Apesar de sua simplicidade e eficiência computacional, o K-means exige que o número de clusters (k) seja especificado previamente, o que pode ser uma limitação em cenários exploratórios. Além disso, o algoritmo é sensível a outliers, que podem distorcer os centros dos clusters.

4.2.2. DBSCAN (e OPTICS)

O **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) é um método baseado na densidade dos dados, que identifica clusters como regiões densas separadas por áreas de baixa densidade. Este algoritmo é particularmente robusto contra outliers e não exige a especificação do número de clusters. Os principais conceitos utilizados pelo DBSCAN são:

- ε **(raio de vizinhança):** Define a distância máxima para que um ponto seja considerado vizinho de outro.
- ***minPts*:** Determina o número mínimo de pontos necessários para formar um cluster.

O DBSCAN classifica os pontos em três categorias:

1. *Core points*: Pontos com pelo menos *minPts* vizinhos em um raio ε.
2. *Border points*: Pontos que estão dentro do raio ε de um *core point*, mas têm menos de *minPts* vizinhos.
3. *Noise points*: Pontos que não atendem a nenhum dos critérios acima.

Este algoritmo é eficiente para dados com clusters de forma arbitrária, mas sua performance depende de uma escolha cuidadosa dos parâmetros ε e *minPts*. O algoritmo OPTICS (Ordering Points To Identify the Clustering Structure), também utilizado nesse trabalho, é uma extensão do DBSCAN, projetado para superar algumas limitações deste último. No DBSCAN, é necessário definir ε antes de executar o algoritmo, e clusters serão formados apenas com base nesse único valor. No OPTICS, os clusters são determinados de forma hierárquica, explorando uma faixa de valores de ε, o que evita a necessidade de escolher um valor fixo e permite identificar clusters em diferentes escalas de densidade. Assim, o OPTICS é projetado para lidar com dados que contêm clusters de densidade variável, graças à sua abordagem hierárquica.

4.2.3. Modelos de mistura gaussiana (GMM)

O **GMM** (Gaussian Mixture Model) é um modelo probabilístico que assume que os dados são gerados por uma combinação de distribuições gaussianas. Cada cluster é representado por uma gaussiana, caracterizada por sua média (μ) e matriz de covariância (Σ). O GMM utiliza o algoritmo **EM** (Expectation-Maximization) para ajustar os parâmetros do modelo:

1. Na etapa de *Expectation*, calcula-se a probabilidade de cada ponto pertencer a cada cluster, com base nos parâmetros atuais (μ e Σ).
2. Na etapa de *Maximization*, atualizam-se os parâmetros das distribuições gaussianas para maximizar a probabilidade total dos dados.
3. Repete-se o processo até que a convergência seja atingida.

Diferentemente do K-means, o GMM considera a forma e a orientação dos clusters, sendo mais flexível ao ajustar agrupamentos elípticos. Além disso, ele fornece uma análise probabilística, permitindo estimar a probabilidade de um ponto pertencer a um determinado cluster. Entretanto, o GMM também exige a especificação prévia do número de clusters e é mais sensível a inicializações inadequadas.

5. Resultados e Discussão

Na Figura 4 são mostrados os resultados de cada algoritmo de clusterização utilizado nesse trabalho. Cada algoritmo atribui um rótulo para cada medida ao seu respectivo cluster (0, 1 ou 2, já que temos três clusters nítidos). A cor vermelha foi atribuída ao cluster 0, verde ao 1 e azul ao cluster 2. Como o rótulo depende de parâmetros de inicialização dos algoritmos, geralmente aleatórios, não há correspondência de cores entre um algoritmo e outro. No caso do algoritmo de misturas gaussianas (GMM) também são mostradas as gaussianas que representam a distribuição de probabilidade de cada cluster. Os resultados apresentados são aqueles após otimização dos parâmetros de configuração de cada algoritmo (normalmente chamados de hiperparâmetros no jargão de aprendizado de máquina). Os algoritmos K-means e GMM sempre atribuem um evento a um determinado cluster. Por outro lado, DBSCAN e OPTICS, que são baseados em densidades, podem não atribuir um evento a um determinado cluster. Nesse caso, os cilindros que não foram associados a um certo cluster, são considerados outliers e aparecem, nas figuras correspondentes, como pontos cinzas.

Os algoritmos K-means e GMM foram inicializados com 3 clusters, por ser o número mais razoável

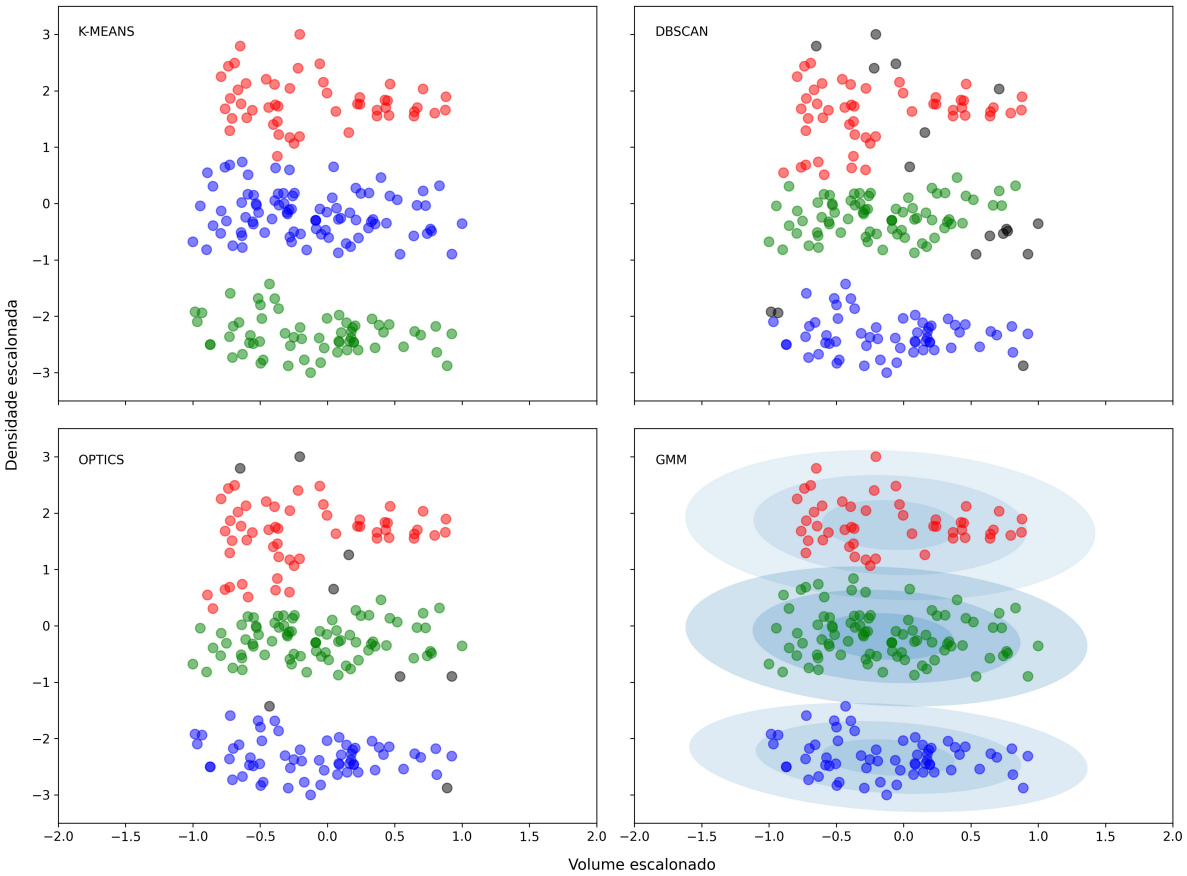

Figura 4: Resultado dos algoritmos de clusterização utilizados nesse trabalho. As cores representam a atribuição de cada medida a um respectivo cluster. Nos algoritmos DBSCAN e OPTICS há a possibilidade de outliers, representados pela cor cinza.

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2024-0443

visualmente. No caso dos algoritmos DBSCAN e OPTICS, que têm como principal característica o agrupamento baseado em densidades, eles foram inicializados de modo a minimizarem o número de outliers e, como resultado, também fornecem 3 clusters como resultado. Além disso, é possível ver como parece haver alguns pontos no cluster superior que parecem mais razoáveis de pertencerem ao cluster central. Isso acontece pois o critério de atribuição é fortemente baseado em densidades e há alguma, embora pouca, intersecção entre esses dois clusters. No caso do K-means e do GMM, parece haver uma seleção mais compatível com aquilo que podemos observar visualmente. A escolha de um modelo específico de clusterização não é automática e é sempre importante olhar vários modelos para desenvolver alguma intuição a respeito deles. No final, o julgamento e análise do cientista é indispensável nesse processo. Os algoritmos apenas refletem as escolhas que o cientista toma.

Além da análise de vários algoritmos, a escolha dos hiperparâmetros de inicialização de cada um deles também é um fator crucial no processo de análise. Nesses casos, as escolhas de número de clusters e densidades afetam os resultados de cada algoritmo. Há várias técnicas estatísticas disponíveis para ajudar nesse processo. Vamos discutir algumas dessas técnicas para os algoritmos K-means e GMM por parecerem mais interessantes ao conjunto de dados disponível.

A escolha do número de clusters (k) é uma etapa crucial ao utilizar algoritmos como K-means e Modelos de Mistura Gaussiana (GMM), pois influencia diretamente a qualidade da clusterização e a interpretação dos resultados. No entanto, determinar k de forma adequada requer a utilização de métricas quantitativas e técnicas de validação. A seguir, descrevemos as abordagens utilizadas neste trabalho.

5.1. Critério da informação (AIC e BIC)

Para os Modelos de Mistura Gaussiana (GMM), critérios baseados em informações estatísticas, como o Critério de Informação de Akaike (AIC) [20] e o Critério de Informação Bayesiano (BIC) [21], são amplamente utilizados. Ambos avaliam o ajuste do modelo aos dados, penalizando modelos mais complexos. Suas definições são dadas por 5 e 6:

$$\mathrm{AIC} = -2\ln(\mathcal{L}) + 2p \tag{5}$$

$$\mathrm{BIC} = -2\ln(\mathcal{L}) + p\ln(n) \tag{6}$$

onde $\mathcal{L}$ é a verossimilhança do modelo, p é o número de parâmetros do modelo, e n é o número de amostras. O número ideal de clusters é aquele que minimiza o valor de AIC ou BIC, equilibrando o ajuste e a simplicidade do modelo.

A Figura 5 mostra o critério de informação aplicado ao algoritmo GMM, onde é possível notar que a escolha de 3 clusters parece ser aquela mais adequada ao conjunto de dados disponível. Mostramos também o resultado esperado desse algoritmo quando executado para 2 ou 4 clusters. É possível notar que o resultado difere fortemente do que imaginamos visualmente que o conjunto de dados represente.

5.2. Método do cotovelo (Elbow Method)

O método do cotovelo [22] é uma técnica visual frequentemente utilizada para determinar o número de clusters em algoritmos como o K-means. Ele baseia-se no cálculo da soma das distâncias quadradas intra-cluster (WSS, do inglês *Within-Cluster Sum of Squares*), definida pela Eq. 7:

$$WSS = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{7}$$

onde C_i é o conjunto de pontos pertencentes ao cluster i e μ_i é o centroide do cluster i. À medida que o número de clusters aumenta, WSS tende a diminuir, pois os pontos ficam mais próximos de seus respectivos centróides. No

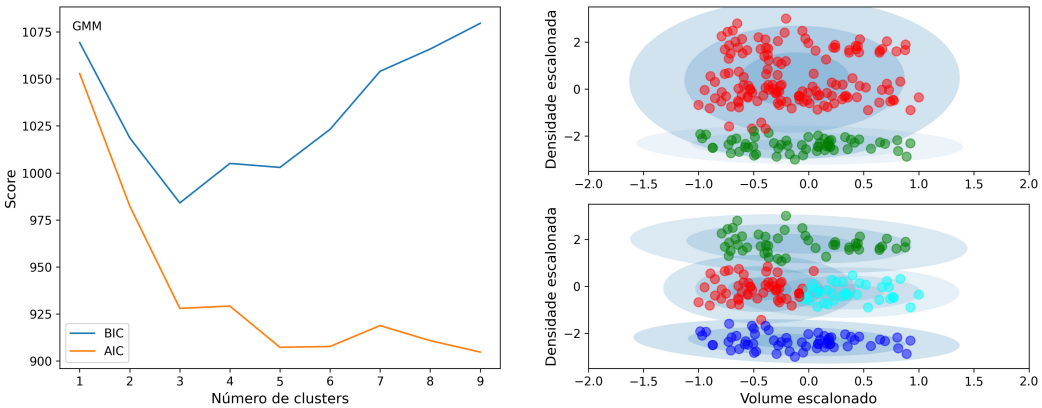

Figura 5: Índices AIC e BIC aplicados ao método de mistura gaussianas (GMM). Notem que o número de clusters igual a 3 parece ser aquele no qual há a melhor clusterização dos dados. Do lado direito mostramos o método GMM aplicado em configurações de 2 ou 4 clusters. Ver texto para discussão.

entanto, há um ponto em que a redução no WSS torna-se marginal, indicando o número ideal de clusters. Esse ponto é identificado visualmente como o "cotovelo" em um gráfico de WSS contra k.

5.3. Score de silhueta (Silhouette Score)

O índice de silhueta [23] mede a qualidade da clusterização com base na separação entre os clusters e na coesão interna de cada cluster. Para cada ponto i calcula-se, usando a Eq. 8:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{8}$$

onde $a(i)$ é a distância média de i para os pontos dentro do mesmo cluster e $b(i)$ é a distância média de i para os pontos no cluster mais próximo.

O valor de $s(i)$ varia entre -1 e 1, onde valores próximos de 1 indicam uma boa atribuição ao cluster. A média de $s(i)$ para todos os pontos fornece o score de silhueta (*Silhouette Score*), que pode ser usado para comparar diferentes valores de k, escolhendo aquele com maior valor.

A Figura 6 Mostra o método do cotovelo e o score de silhueta aplicados aos dados com o algoritmo K-means. Para esses dois testes, a escolha de 3 clusters parece ser realmente a mais razoável. Nessa figura, à direita, são mostrados os resultados do K-means para 2 e 4 clusters e, visualmente, percebemos que essas escolhas não parecem descrever bem os dados.

Embora as métricas descritas acima sejam aplicáveis a ambos os métodos, o GMM oferece maior flexibilidade devido à sua capacidade de modelar clusters elípticos e considerar a probabilidade de pertencimento de cada ponto a múltiplos clusters. Em contrapartida, o K-means assume clusters esféricos e atribuições rígidas de pontos, o que pode simplificar a escolha de k, mas reduzir a acurácia em dados complexos.

Muitas vezes o conjunto de dados é multiparamétrico, podendo haver um número grande de parâmetros envolvidos e a visualização de clusters se torna muito difícil, senão impossível. Nesse caso, o uso de métricas desse tipo para avaliar a qualidade dos hiperparâmetros do algoritmo utilizado é imprescindível.

A Figura 7 mostra o histograma das densidades experimentais medidas após serem rotuladas pelo algoritmo de clusterização. Nesse caso, utilizamos o resultado do K-means como forma de separar os diversos grupos. Nota-se uma boa separação. É interessante ressaltar que, apesar de termos 4 materiais distintos na amostra, somos capazes de separar os dados em apenas três grupos claros. Isso por conta de que a densidade do Nylon

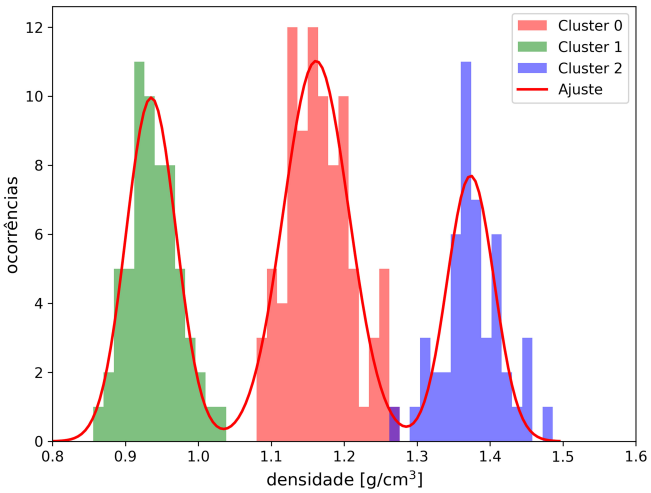

Figura 7: Histograma com as densidades medidas experimentalmente, separadas através do algoritmo K-means, e um ajuste com três funções gaussianas.

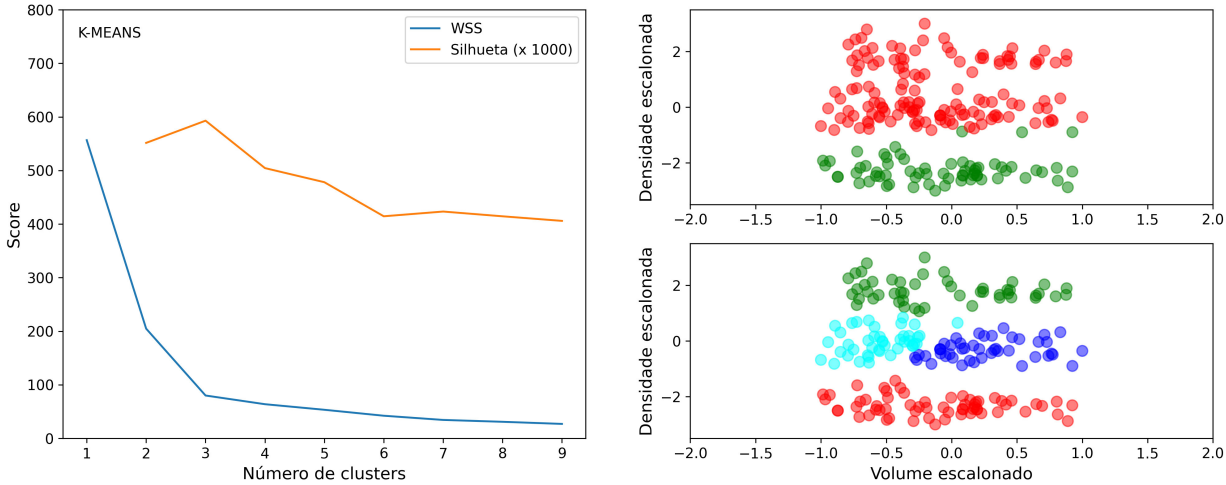

Figura 6: Índices WSS e Score de Silhueta aplicados no algoritmo K-means. O número de cluster igual a 3 parece ser aquele onde esses scores fornecem valores mais adequados, compatível com aquilo observado visualmente nos dados. As figuras da direita exemplificam o uso do K-means para 2 ou 4 clusters.

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2024-0443

e PVC serem muito próximas e os instrumentos de medida utilizados nesse trabalho não possuírem precisão suficiente para separar esses dois materiais. Além disso, a qualidade dos dados pode sofrer uma ligeira redução quando obtidas pelos alunos devido à inexperiência com instrumentos de medida. Esse é um tópico interessante de ser discutido. Pretendemos aplicar essa metodologia em breve com uma turma de alunos e verificar as possíveis divergências. Em resumo, embora métodos de aprendizado de máquina sejam cada vez mais importantes no processo de análise de dados, a qualidade dos dados – nesse caso relacionada à precisão dos instrumentos – continua sendo um fator fundamental no processo investigativo. O pré-processamento e a análise criteriosa envolvendo mais de um algoritmo também é essencial durante o processo para que as conclusões realizadas não sejam enviesadas por conta de uma limitação imposta por um determinado método utilizado.

6. Conclusões

Este trabalho propôs um experimento prático de medição de densidades de sólidos plásticos como uma ferramenta didática para introduzir alunos de física aos conceitos fundamentais de aprendizado de máquina não supervisionado. Ao combinar aspectos de um experimento simples de física, possível de ser realizado sem muitos recursos, com técnicas modernas de análise de dados, o experimento oferece uma abordagem acessível e introdutória para explorar a aplicação de métodos de clusterização em dados reais.

A análise dos dados foi realizada por meio do uso de diferentes algoritmos de clusterização, como K-means, DBSCAN (e OPTICS) e Modelos de Mistura Gaussiana (GMM), permitindo que os alunos comparassem diretamente as abordagens e compreendessem as vantagens, limitações e suposições de cada método. Além disso, o experimento destacou a importância de um pré-processamento cuidadoso, como o escalonamento dos dados, que é essencial para o bom desempenho desses algoritmos.

Um ponto central do trabalho foi o processo de escolha de hiperparâmetros, como o número de clusters no K-means e no GMM, bem como os parâmetros de densidade no DBSCAN. O uso de métricas quantitativas, como o índice de silhueta, a soma das distâncias intra-cluster, e critérios de informação como AIC e BIC, foi fundamental para avaliar a qualidade das configurações escolhidas. Essa abordagem sistemática não apenas melhorou os resultados da clusterização, mas também proporcionou aos alunos uma experiência prática na validação de modelos, um componente crítico no aprendizado de máquina.

Por meio de gráficos intuitivos e análises estatísticas, os alunos seriam capazes de visualizar como os algoritmos segmentam os dados e identificar padrões inerentes às propriedades físicas dos materiais analisados, neste caso, a densidade de plásticos comuns. Essa integração entre experimentação e modelagem computacional oferece uma introdução concreta e instigante a conceitos de aprendizado de máquina, reduzindo a barreira inicial de entrada para este campo.

Em suma, o experimento descrito neste trabalho representa um caminho eficiente e prático para introduzir alunos de física ao aprendizado de máquina, com ênfase em métodos não supervisionados. Além de expandir a compreensão de conceitos básicos de física e análise de dados, essa abordagem prepara os alunos para enfrentar desafios científicos mais complexos, onde a interseção entre física e ciência de dados desempenha um papel cada vez mais significativo.

Referências

[1] S.S. Pattanaik, B.K. Das, R. Tripathy, B.K. Prusty, M.K. Parida, S.R. Tripathy, A.K. Panda, B. Ravindran e R. Mukherjee, Scientific Reports **14**, 28765 (2024).

[2] L. Wang, Physical Review B **94**, 195105 (2016).

[3] Y. Wang, B. Seo, B. Wang, N. Zamel, K. Jiao e X.C.Adroher, Energy and AI **1**, 100014 (2020).

[4] N.Y. Khanday e S.A. Sofi, Biomedical Signal Processing and Control **69**, 102814 (2021).

[5] S.K. Singh, A.K. Tiwari e H.K. Paliwal, Engineering Analysis with Boundary Elements **155**, 62 (2023).

[6] M.M. Ahsan, S.A. Luna e Z. Siddique, Healthcare **10**, 541 (2022).

[7] M.D. Schwartz, Harvard Data Science Review **3**, 1 (2021).

[8] J. Kvita, P. Baroň, M. Machalová, R. Přívara, R.Vodák e J. Tomeček, arXiv:2410.13904 (2024).

[9] E. Richter-Was, T. Yerniyazov e Z. Was, arXiv:2411.06216 (2024).

[10] U. Syed, F. Cunico, U. Khan, E. Radicchi, F. Setti, A.Speghini, P. Marone, F. Semenzin e M. Cristani, arXiv:2411.13953 (2024).

[11] K. Faceli, A.C. Lorena, J. Gama e A.C.P.L.F Carvalho, *Inteligência Artificial – Uma Abordagem de Aprendizado de Máquina* (LTC, Rio de Janeiro, 2021), 2 ed.

[12] H. Ferreira, E.F. Almeida, W. Espinosa-García, E. Novais, J.N.B. Rodrigues e G.M. Dalpian, Revista Brasileira de Ensino de Física **44**, e20220214 (2022).

[13] J.H.S. Prates, C.S.L. Costa, E.F.G. Pessoa, G.P. Mattedi, H.B. Monteiro, J.V.B. Jesus, L.F.Y. Namioka, M.S. Huoya e Y.A. Matui, Revista Brasileira de Ensino de Física **45**, e20230207 (2023).

[14] S. Lloyd, IEEE Transactions on Information Theory **28**, 129 (1982).

[15] R.F. Ling, The Computer Journal **15**, 326 (1972).

[16] C.M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, Berlin, Heidelberg, 2006).

[17] R.C. Weast (ed.), *CRC Handbook of Chemistry and Physics: A Ready-reference Book of Chemical and Physical Data* (CRC Press, Boca Raton, 1975), 56 ed.

[18] A.A.P. Suaide, disponível em: https://github.com/suaide/RBEFdensidade.

[19] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright et al., Nature Methods **17**, 261 (2020).

[20] H. Akaike, em: *Selected Papers of Hirotugu Akaike*, editado por E. Parzen, K. Tanabe e G. Kitagawa (Springer, New York, 1998).

[21] E. Wit, E. van den Heuvel e J.W. Romeijn, Statistica Neerlandica **66**, 217 (2012).

[22] D.J. Ketchen e C.L. Shook, Strategic Management Journal **17**, 441 (1996).

[23] P.J. Rousseeuw, Journal of Computational and Applied Mathematics **20**, 53 (1987).