Felipe Vaiano Calderan

Agrupamento com Supervisão para Seleção de Moléculas em Banco de Dados Químicos

São José dos Campos, SP

Felipe Vaiano Calderan

Agrupamento com Supervisão para Seleção de Moléculas em Banco de Dados Químicos

Trabalho de conclusão de curso apresentado ao Instituto de Ciência e Tecnologia – UNIFESP, como parte das atividades para obtenção do título de Bacharel em Ciência da Computação.

Universidade Federal de São Paulo – UNIFESP

Instituto de Ciência de Tecnologia

Bacharelado em Ciência da Computação

Orientador: Prof. Dr. Marcos Gonçalves Quiles

São José dos Campos, SP

Fevereiro de 2022

Felipe Vaiano Calderan

Agrupamento com Supervisão para Seleção de Moléculas em Banco de Dados Químicos

Trabalho de conclusão de curso apresentado ao Instituto de Ciência e Tecnologia – UNIFESP, como parte das atividades para obtenção do título de Bacharel em Ciência da Computação.

Trabalho aprovado em 11 de Fevereiro de 2022:

Prof. Dr. Marcos Gonçalves Quiles
Orientador

Prof. Dr. Fabio Augusto Faria
Convidado 1

Prof. Dr. Márcio Basgalupp
Convidado 2

São José dos Campos, SP
Fevereiro de 2022

Este trabalho é dedicado aos meus avós paternos Grimaldo e Valquiria e falecidos avós maternos José Carlos e Aparecida.

Agradecimentos

Agradeço, primeiramente, aos meus pais Adriano e Dinalva, que sempre foram de fundamental importância para meu crescimento pessoal e profissional, por todo o amor e atenção que me dão e por sempre tratarem meus estudos como prioridade máxima. Não chegaria onde estou hoje sem seus ensinamentos, conselhos, lições de vida e suporte. O mesmo posso dizer sobre meus avós paternos e falecidos avós maternos.

À minha irmã Natália, por estar ao meu lado não importa o que aconteça. Muitos dos momentos mais felizes da minha vida foram divididos diretamente com ela, sejam nas brincadeiras, jogos ou até mesmo quando não havia nada para fazer.

À Audrey, que tem estado ao meu lado como amiga desde 2015 e namorada desde 2016, por comemorar comigo nos meus momentos de sucesso e ajudar a me reerguer nos de fracasso. O mesmo para seus pais, irmão e parentes, que são minha segunda família.

Ao meu tio Alexandre, por ter me hospedado em sua residência como um filho e por ter me levado e trazido da faculdade durante os vários meses que não encontrei outra maneira de me transportar.

Agradeço ao meu orientador Prof. Dr. Marcos Quiles, por ter aceitado me orientar como aluno de Iniciação Científica e no Trabalho de Conclusão de Curso. Seus ensinamentos relacionados aos tópicos aqui abordados, vida acadêmica e burocracias serão levados para sempre comigo.

À todos meus professores: aqueles que me ensinaram o abecedário, aqueles que forneceram os conhecimentos necessários para que eu ingressasse numa faculdade pública e, claro, professores da Universidade Federal de São Paulo (UNIFESP) aos quais devo grande parte dos meus conhecimentos de Ciência da Computação.

À todos os funcionários da UNIFESP, que diariamente proporcionam um ambiente físico e virtual excelente para o desenvolvimento de todos os meus estudos.

Aos membros do grupo QTNano, em especial ao Prof. Dr. Juarez L. F. Da Silva (meu coorientador da Iniciação Científica), que são de grande importância não só para esse trabalho, mas também para minha Iniciação Científica, artigos que leio e escrevo e meus conhecimentos gerais sobre Ciência dos Materiais.

Agradeço também aos professores, funcionários e alunos de outras universidades que me receberam muito bem durante congressos e outros eventos acadêmicos que foram de grande importância para minha formação.

Aos meus colegas de turma, que dividiram mesas comigo na biblioteca, assim como

assentos no restaurante universitário inúmeras vezes. Nossas conversas acadêmicas e casuais são grandes fontes de inspiração para tudo o que faço e moldaram minha apreciação por certas áreas do conhecimento.

Finalmente, gostaria de agradecer à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) por fomentarem minhas pesquisas como aluno de Iniciação Científica.

"O futuro não pode ser predito, mas futuros podem ser inventados"
(Inventing the future, Dennis Gabor, 1963, p. 207, tradução nossa)

Resumo

Métodos de aprendizado de máquina, desde algoritmos não supervisionados a supervisionados, têm sido aplicados para resolver várias tarefas no domínio da Ciência dos Materiais, como predição de propriedades, projeto de novos compostos, modelos substitutos em simulações de dinâmica molecular, entre outras. No entanto, apesar de já haver avanços notáveis, o uso de modelos de Aprendizado de Máquina nesse domínio ainda está em seu estágio inicial. Com o objetivo de contribuir ainda mais para a área, e visando reduzir o custo computacional de triagem de materiais, este trabalho desenvolve um sistema que inclui algoritmos de agrupamento (K-Means) e otimização (força bruta e Basinhopping) para gerar um método de agrupamento supervisionado que pondera o conjunto de dados de acordo com a qualidade dos agrupamentos formados, seleciona amostras a serem testadas e, em seguida, fornece informações textuais e gráficas para facilitar as análises físico-químicas. Os resultados preliminares mostram que é um método viável de introduzir restrições na forma como os dados são agrupados e pode ser muito poderoso, uma vez que herda a ótima eficiência computacional dos métodos de agrupamento, enquanto também permite a construção de agrupamentos com propriedades específicas destacadas, o que muitas vezes é algo desejável.

Palavras-chaves: Agrupamento de dados, visualização de dados, agrupamento supervisionado, método de otimização

Abstract

Machine Learning methods, from unsupervised to supervised algorithms, have been applied to solve several tasks in the Materials Science domain, such as property prediction, design of new compounds, surrogate models in molecular dynamics simulations, among others. However, despite already having noticeable advances in the field, the use of Machine Learning models in the domain is still in its early stages. With the goal of further contributing to the area, and aiming to reduce the computational cost of material screening, this work develops a system that includes clustering (K-Means) and optimization (brute force and Basinhopping) algorithms in order to generate a supervised clustering method to weight the data set according to the quality of the clusters formed, select samples to be further tested and then provide textual and graphical information to facilitate physicochemical analyses. The preliminary results show that it is a viable method of introducing constraints in the way the data is clustered and can be very powerful, since it inherits the great computational efficiency of clustering methods but also allows clusters with specific properties highlighted to be built, which is often something desirable.

Key-words: Data clustering, Data visualization, Supervised clustering, Optimization method

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ACO	*Ant Colony Optimization* (otimização por colônia de formigas)
ARI	*Adjusted Rand Index* (Índice de Rand ajustado)
CBR	*Case-Base Reasoning* (Raciocínio baseado em casos)
CLI	*Command-line interface* (Interface de linha de comando)
CLINK	*Complete Linkage* (Conexão completa)
CSV	*Comma-Seperated Values* (Valores Separados por Vírgulas)
DBSCAN	*Density-based spatial clustering of applications with noise* (Agrupamento espacial baseado em densidade de para aplicações com ruído)
DFT	*Density Functional Theory* (Teoria do Funcional da Densidade)
GUI	*Graphical User Interface* (Interface Gráfica de Usuário)
ILS	*Iterative Label Spreading* (Espalhamento de rótulo iterativo)
ISODATA	*Iterative Self-Organizing Data Analysis Technique* (Técnica Iterativa de Análise de Dados Auto-Organizáveis)
KRR	*Kernel Ridge Regression* (Regressão da crista do kernel)
MADM	*Multiple Attribute Decision Making* (Tomada de decisão com múltiplos atributos)
MCDM	*Multi-criteria Decision Making* (Tomada de decisão com múltiplos critérios)
MODM	*Multiple Objective Decision Making* (Tomada de decisão com múltiplos objetivos)
MSE	*Mean Squared Error* (Erro quadrático médio)
NMI	*Normalized Mutual Information* (informação mútua normalizada)
OPTICS	*Ordering Points To Identify the Clustering Structure* (ordenando pontos para identificar a estrutura de agrupamento)
PCA	*Principal Component Analysis* (Análise das Componentes Principais)
PSO	*Particle Swarm Optimization* (otimização por enxame de partículas)

RNG	*Random Number Generator* (Gerador de números aleatórios)
SIMLES	*Simplified molecular-input line-entry system* (Sistema simplificado de entrada de linha de registro molecular)
SOM	*Self-Organizing Map* (Mapa auto-organizável)
t-SNE	*t-distributed Stochastic Neighbor Embedding* (Incorporação de Vizinho Estocástico com distribuição t)
WCSS	*Within-Cluster Square Sum* (Soma de quadrados intra-grupos)

Sumário

1 Introdução

1.1 Contextualização e Motivação

A grande disponibilidade de dados aliada ao alto poder computacional têm alavancado o uso de aprendizado de máquina em pesquisas de materiais (MONTAVON et al., 2013; BUTLER et al., 2018). Como um exemplo pode se citar a análise e descoberta de novos materiais que era tradicionalmente realizada por vias experimentais ou computacionais (HARMON, 2003; MCFARLAND; WEINBERG, 1999; LUDWIG, 2019; JAHAN et al., 2010) têm sido abordada sob a perspectiva de aprendizado de máquina (MATSUBARA et al., 2020; TABOR et al., 2018; LO et al., 2018).

Embora a análise computacional de materiais, por exemplo via dinâmica molecular ou cálculos DFT (Density Functional Theory), permita uma avaliação mais rápida e menos custosa em comparação a abordagem experimental, ela ainda se mostra proibitiva em função da explosão combinatória de materiais que podem ser gerados para um dado problema (JAHAN et al., 2010).

Uma forma de calcular a propriedade de diversos materiais reduzindo o custo computacional está na seleção de exemplos a serem simulados (ZIBORDI-BESSE et al., 2018). Essa seleção tem como objetivo encontrar exemplos (materiais) representativos dentre os diversos exemplos contidos num dado repositório. Materiais com estruturas/características semelhantes tendem a apresentar propriedades semelhantes. Nesse contexto, ao invés de realizar o cálculo de todas as combinações possíveis para uma data classe de materiais, o cientista de materiais pode conduzir a análise de apenas algumas exemplos e generalizar o resultado para outras estruturas. Para isso, é necessária alguma forma de selecionar quais amostras são consideradas como representativas da investigação.

Uma forma de realizar tal seleção é a partir do uso de algoritmos de similaridade baseados em distância Euclidiana (ZIBORDI-BESSE et al., 2018). Para isso, cada estrutura molecular é mapeada num vetor representando a distância de cada átomo ao centro de massa do sistema. Após, a similaridade entre estruturas é definida pelo inverso da distância Euclidiana entre os vetores que as representam.

Neste mesmo cenário, técnicas de aprendizado de máquina também podem ser empregadas para encontrar essas estruturas mais representativas em um universo de moléculas. Por exemplo, um método de agrupamento de dados pode ser empregado para identificar os grupos semelhantes existentes em um conjunto de dados e, após geração dos grupos, exemplos representativos podem ser selecionados para caracterizar cada grupo. Essa abordagem foi adotada num estudo recente desenvolvido pelo grupo de pesquisa QTNano (BATISTA et al., 2020).

Nesse trabalho, empregou-se a técnica de agrupamento *K-Means* (JAIN, 2010) para seleção de estruturas representativas.

1.2 Definição do Problema

A análise computacional de materiais é um procedimento demorado, podendo levar de minutos até meses (dependendo do tamanho das moléculas) (GILMER et al., 2017), pois os métodos encontrados na literatura, como DFT, são computacionalmente intensivos devido à complexidade dos cálculos que são realizados. Dessa forma, métodos para acelerar tais procedimentos têm sido estudados para que uma quantidade maior de resultados possa ser obtida num menor espaço de tempo, acelerando as pesquisas na área de ciência dos materiais (DAWSON et al., 2020; TRUSHIN; THIERBACH; GÖRLING, 2021; HERNANDEZ, 2015).

No contexto de aprendizado de máquina, embora um algoritmo não supervisionado de agrupamento de dados possa ser utilizado com esse propósito, o resultado do algoritmo depende diretamente de como os dados (compostos) estão representados (BUTLER et al., 2018). Normalmente, as moléculas/compostos são codificadas em vetores nos quais as dimensões representam características ou *fingerprints*. Uma vez codificados, os exemplos podem ser fornecidos como entradas para o modelo de aprendizado de máquina. Na sequência, os exemplos do conjunto são agrupados conforme sua semelhança definida pela função de similaridade sobre os vetores de características. Entretanto, se as características forem inadequadas ou de baixa qualidade, o resultado do processo de agrupamento pode ser precário (DASH; LIU, 2000).

Um outro empecilho é que algoritmos de agrupamento, como são tipicamente implementados, não levam em conta aspectos físico-químicos, como propriedades e interações conhecidas pelos especialistas, dos elementos presentes no conjunto de dados além do que é possível extrair diretamente a partir do processo estatístico. Isso não é o ideal, uma vez que existem nuances no repositório que não são capturados dessa forma, o que gera conjuntos de dados com informações insuficientes e, consequentemente, resultados inferiores.

Assim, visando melhorar o processo de seleção de exemplos representativos em problemas de ciências de materiais, este Trabalho de Conclusão de Curso (TCC) propõe uma técnica para aprimorar os resultados obtidos pelo processo de agrupamento de dados, via utilização de um algoritmo de agrupamento com supervisão. Para isso, uma técnica de otimização é incorporada ao processo de agrupamento de dados, permitindo que grupos mais significativos e relacionados ao problema em estudo sejam gerados. Uma vez que o agrupamento possa ser enviesado por uma propriedade físico-química, existem padrões subjacentes nos agrupamentos que evidenciam certos aspectos desejáveis pelo especialista.

1.3 Justificativas

O tema de aprendizado não supervisionado, ao longo da história, não recebeu a mesma atenção da comunidade científica quanto o aprendizado supervisionado, ainda mais considerando o *boom* das redes neurais artificiais e aprendizado profundo. No entanto, nos últimos tempos tem sido um tema bastante explorado, pois tem demonstrado resultados substanciais no ganho de informações sobre a estrutura e comportamento de sistemas descritos em conjuntos de dados. Sendo assim, é um tema em alta que merece ter suas aplicações analisadas e aplicadas em interfaces com outras áreas (JAIN; MURTY; FLYNN, 1999; JAIN, 2010).

Algoritmos de agrupamento pertencem a essa classe de aprendizado e têm bastante potencial de uso na área de ciências de materiais (mas não limitados a ela), pois podem selecionar elementos representativos que podem ser analisados em vez de todo o conjunto de dados. Ainda assim, as técnicas de agrupamento documentadas de forma mais recorrente na literatura não levam em conta as propriedades físico-químicas dos elementos nos conjuntos de dados desse domínio. Portanto, tem-se um cenário bastante fértil para desenvolver novos modelos capazes de lidar com tais limitações.

Vale mencionar, também, que existem bastante conjuntos de dados nesse domínio, muitos com propriedades bem descritas, mas que não são utilizadas por algoritmos não supervisionados devido a natureza desses métodos (lidam somente com as características e são completamente cegos com relação às propriedades). Isso é um grande desperdício de dados que poderiam ser utilizados para a obtenção de melhores resultados.

Sendo assim, unindo esses dois problemas: não consideração de aspectos físico-químicos e uma vasta diversidade de dados não explorados por algoritmos de agrupamento, tem-se boas razões para o desenvolvimento de um algoritmo de agrupamento com supervisão. Esses algoritmos utiliza todo o potencial de algoritmos não supervisionados, mas incorpora aspectos do aprendizado supervisionado, como a utilização das propriedades para melhorar os grupos a serem gerados.

Não bastassem essas razões, no campo de ciência dos materiais esse tipo de algoritmo é altamente desejável, uma vez que pode revelar ou realçar propriedades de interesse para os especialistas. Com isso, tem-se que esse trabalho é benéfico para a área de ciência dos materiais, aprendizado de máquina e computação aplicada no geral.

1.4 Objetivos: Geral e Específicos

O objetivo geral é investigar, via análise de agrupamento de dados e técnicas de otimização, a seleção de exemplos representativos em bases de dados de compostos químicos. Essa seleção tem como objetivo selecionar moléculas/compostos representativos em um dado repositório. Ao simular essa molécula, denominada representativa ou protótipo, espera-se estimar as

propriedades de um dado subconjunto de moléculas o qual possui características semelhantes à molécula selecionada. Assim, reduz-se a necessidade de realizar uma análise exaustiva de todos os compostos disponíveis no estudo.

Já os objetivos específicos, consistem em:

- Implementar e validar uma ferramenta (*package*) de seleção de moléculas com interface em Python.
- Realizar estudos de caso com conjuntos de dados públicos e com dados disponibilizados pelo grupo QTNano/USP

1.5 Sumário dos Resultados

Como principais resultados deste desenvolvimento, destacam-se a proposta de um método para seleção de melhores exemplos representativos de conjuntos de dados, ou seja, elementos que cubram da melhor forma possível o espaço químico, e a geração de uma ferramenta que realiza o processo automaticamente, dadas as configurações iniciais do usuário, além de sua respectiva implementação.

As análises realizadas a partir dos experimentos executados com cinco conjuntos de dados químicos mostram que a proposta é promissora, pois confirmam que exemplos mais significativos são selecionados ao comparar com a técnica convencional de seleção de representativos por algoritmos de agrupamento.

1.6 Organização do Documento

Este Trabalho de Conclusão de Curso está organizado da seguinte forma:

- O Capítulo 2 apresenta uma revisão bibliográfica de todos os tópicos relevantes para o desenvolvimento da ferramenta proposta. Nele, é feita uma breve introdução do estado da literatura pertinente e em seguida são mostrados cada um dos elementos que formam a ideia central desse trabalho: primeiramente a triagem de materiais, depois o aprendizado de máquina em ciência dos materiais, seguido pela caracterização dos dados, medidas de similaridade, algoritmos de agrupamento, medidas de qualidade de agrupamento e agrupamento supervisionado.
- O Capítulo 3 especifica exatamente como o produto final gerado por esse TCC é construído, quais são os tipos de entradas suportadas pelo programa, quais algoritmos são utilizados e como são implementados e, finalmente, como é a saída do programa.

- Já no Capítulo 4 são mostrados os resultados obtidos a partir dos conjuntos de dados especificados. Os resultados consistem de análises estatísticas e computacionais a partir dos dados textuais e gráficos gerados pelo programa e à parte.
- Finalmente, o Capítulo 5 apresenta as principais conclusões deste trabalho, sumariza as limitações e aponta possíveis trabalhos futuros.

2 Revisão Bibliográfica

2.1 Triagem de Materiais

Com o grande crescimento de quantidade e tipos de repositórios e materiais conhecidos (ROTH; FIELD; CLARK, 1994), tornou-se necessário criar rotinas para varrer, selecionar e ranquear todos esses materiais de forma que os mais interessantes para as aplicações buscadas pelos especialistas sejam encontrados. Essas rotinas são conhecidas como triagens de materiais.

Existem vários estudos abordando rotinas viáveis para esse propósito. Em (FARAG, 2015) tem-se a organização mais simples: basta fazer a triagem inicial, comparar as alternativas disponíveis e selecionar a solução ótima. Uma maneira levemente mais completa é mostrada em (KESTEREN; KANDACHAR; STAPPERS, 2007), onde é acrescentado o passo inicial de formular critérios desejados para o material buscado. Seguindo esse mesmo conjunto de procedimentos, em (CHINER, 1988) é adicionado um passo final de testes de verificação para averiguar se a solução ótima condiz com a formulação inicial. Já em (ASHBY et al., 2004), há o fornecimento de uma solução parecida, mas com uma abordagem eliminatória: em vez de tentar selecionar do conjunto inteiro de dados aqueles que melhor cumprem os requisitos, primeiro são eliminados aqueles que não cumprem.

Para a triagem em si, métodos bem distintos podem ser aplicados para atingir o mesmo objetivo (cada um com seus pontos fortes e fracos). Jahan em (JAHAN et al., 2010) cita vários deles, como: escolha na tentativa e erro, método de custo por unidade de propriedade, método do gráfico (ASHBY, M. F.; CEBON, D., 1993), método do questionário (que foi aprimorado ao longo dos anos) (FARAG, 1979; EDWARDS, 2005; PEDGLEY, 2009), ferramenta para escolha de materiais em produtos (KESTEREN; STAPPERS; BRUIJN, 2007), sistemas de seleção de materiais com ajuda computacional (DARGIE; PARMESHWAR; WILSON, 1982), sistemas com ajuda computacional e utilizando inteligência artificial, além dos métodos estatísticos convencionais (YU; KRIZAN; ISHII, 1993), sistemas baseados em conhecimento (BULLINGER; WARSCHAT; FISCHER, 1991; SAPUAN; ABDALLA, 1998), CBR (*Case-Base Reasoning* ou Raciocínio baseado em casos), que tenta solucionar novos problemas com base naqueles já resolvidos no passado (AMEN; VOMACKA, 2001) e rede neural artificial, que é uma das técnicas de aprendizado de máquina (JIANMIN, 2004; AMOIRALIS; GEORGILAKIS; GIOULEKAS, 2006; BALAKRISHNA et al., 2007).

Como cada método possui suas vantagens e desvantagens, não existe o método perfeito. Logo é preciso que o especialista conheça e compreenda bem o sistema que está sendo estudado e aplique uma abordagem sistemática para escolher qual o melhor método de triagem de materiais. Uma revisão sobre métodos para ajudar na escolha, como MCDM (*multi-criteria decision*

making ou tomada de decisão com múltiplos critérios), MODM (*multiple objective decision making*, ou tomada de decisão com múltiplos objetivos) e MADM (*multile attribute decision making*, ou tomada de decisão com múltiplos atributos) são revisados em (JAHAN et al., 2010).

2.2 Aprendizado de Máquina em Ciência dos Materiais

As aplicações de aprendizado de máquina em ciência dos materiais têm se tornado mais comum a cada dia, pois encontram ótimas aplicações em situações onde 1) as propriedades buscadas são muito difíceis ou custosas de serem descobertas com métodos tradicionais; 2) os fenômenos complexos ou não determinísticos proíbem o uso de soluções encontradas diretamente por equações; 3) os fenômenos estudados ainda não têm suas teorias e/ou equações matemáticas conhecidas (RAMPRASAD et al., 2017).

A aceleração da descoberta de novos compostos através de dados é um tema recorrente na literatura de aprendizado de máquina aplicado em ciência dos materiais. Hoje tem-se exemplos de predições de estruturas de cristais bem sucedidas (CURTAROLO et al., 2003) utilizando técnicas como PCA, regressão linear e matrizes de correlação; aceleração da descoberta de óxidos ternários através de um modelo treinado em dados experimentais (HAUTIER et al., 2010); descoberta de materiais com propriedades desejadas através de algoritmos genéticos enviesados por redes neurais (previsões de uma rede neural artificial progressivamente construída são empregadas para influenciar a evolução de um algoritmo genético) (PATRA et al., 2017) e muitos outros exemplos nesse segmento.

Outro uso comum é na predição de propriedades eletrônicas, como é o caso da predição de *bandgaps* de materiais inorgânicos através de redes neurais (ZHAOCHUN; RUIWU; NIANYI, 1998); e predição da eficiência termoelétrica por árvores de decisão (CARRETE et al., 2014), assim como por otimização Bayesiana (JU et al., 2017; YAMAWAKI et al., 2018).

Além de propriedades eletrônicas, encontra-se artigos (embora ainda poucos) expondo bons resultados quanto a predição de propriedades magnéticas. Dois grandes trabalhos nessa área incluem (SANVITO et al., 2017; PHAM et al., 2017). O primeiro documenta a construção de um repositório de ligas de Heusle. Tendo o conjunto construído e tratado, utilizou-se regressão linear para estimar as temperaturas de Curie e métodos de classificação para encontrar imãs e eletroimãs. O segundo trabalho utiliza KRR (*Kernel Ridge Regression* ou regressão da crista do kernel) para predizer corretamente os momentos magnéticos para as ligas de metais de transição lantanídeos.

Uma área sendo explorada recentemente é a transição de fase quântica em modelos de isoladores topológicos. Nela, é possível utilizar redes neurais para aprender estados topológicos, como é mostrado em (NIEUWENBURG; LIU; HUBER, 2017) e também métodos não supervisionados para a predição de transições de fase (NIEUWENBURG; LIU; HUBER, 2017; ZHAO; FU, 2019).

Também é visto um esforço para aplicação de aprendizado de máquina em supercondutividade, como em (OWOLABI; AKANDE; OLATUNJI, 2015), onde utilizou-se *Support Vector Regression* (regressão por vetor de suporte) para desenvolver um método de regressão que seja capaz de estimar a temperatura crítica de diferentes supercondutores; outro exemplo de trabalho nesse segmento é a combinação de mineração de dados em conjunto com algoritmo de florestas aleatórias para investigar mais de 16 mil supercondutores (STANEV et al., 2018).

Finalmente, existem aplicações de algoritmos não supervisionados, especificamente algoritmos de agrupamento, para encontrar representantes para as várias moléculas presentes em conjunto de dados químicos, como pode ser visto em (BATISTA et al., 2021) e é o segmento de aprendizado de máquina mais abordado nesse Trabalho de Conclusão de Curso.

2.3 Representação dos Dados

Moléculas geralmente são apresentadas em sua forma química por meio de uma cadeia (*string*) de símbolos ou por um grafo com os átomos e suas respectivas ligações. Informações complementares também podem estar disponíveis, como por exemplo a posição espacial de cada átomo da estrutura. Contudo, essa representação não é adequada como formato de entrada para técnicas de aprendizado de máquina, sejam elas supervisionadas ou não supervisionadas (BUTLER et al., 2018).

Para alimentar um modelo de aprendizado de máquina, primeiro é preciso transformar as informações a serem processadas em algo utilizável pelo algoritmo. Esses dados transformados, também conhecidos como descritores, são o conhecimento do computador sobre o universo dos dados passados, sendo assim, essa transformação afeta diretamente a performance (tanto em custo computacional, como em resultados) do modelo, logo é um processo que requer um bom esforço e dedicação de tempo por parte dos cientistas (KHATIB; JONG, 2020).

Em (SEKO; TOGO; TANAKA, 2018) é mostrado que descritores podem ser coisas simples, como número atômico, massa atômica, período, grupo na tabela periódica, primeira energia de ionização, segunda energia de ionização, afinidade eletrônica, Eletronegatividade de Pauling, eletronegatividade de Allen, raio de van der Waals, raio covalente, raio atômico, raio pseudopotencial para o orbital s, raio pseudopotencial para o orbital p, ponto de fusão, ponto de ebulição, densidade, volume molar, calor de fusão, calor de vaporização, condutividade térmica, calor específico ou até mesmo uma tabela binária que discrimine quais elementos químicos existem em quais moléculas. Além disso, descritores baseados em cálculos DFT, como volume, *band gap*, energia coesiva, constantes elásticas, constantes dielétricas, estruturas eletrônicas.

Como elementos químicos são todos compostos de átomos, é típico que descritores, conhecidos como atomísticos, utilizem matrizes para relacionar essas unidades. A Matriz de Coulomb (RUPP et al., 2012; HANSEN et al., 2015; BUTLER et al., 2018) usa os átomos de um sistema e suas respectivas distâncias para construir uma matriz de emparelhamento e pode

ser calculada seguindo a Equação 2.1, exibida logo abaixo.

$$C_{i,j} = \begin{cases} \frac{Z_i Z_j}{\|r_i - r_j\|} & \text{se } i \neq j \\ 0.5 Z_i^{2.4} & \text{se } i = j \end{cases} \tag{2.1}$$

no qual Z_i e Z_j são as cargas nucleares dos átomos i e j, respectivamente e r_i e r_j representam as coordenadas dos átomos i e j no espaço. A partir dessa matriz, algumas formas de representação vetorial podem ser ser obtidas: 1) *flattening* da matriz, no qual todos os elementos da matriz são considerados atributos; 2) representação a partir dos autovalores da matriz C; dentre outras formas, como a matriz de Coulomb ordenada (RUPP et al., 2012).

A Matriz de Soma de Ewald (FABER et al., 2015) é uma extensão da Matriz de Coulomb, mas para sistemas periódicos. Existe também a Matriz de Senos (FABER et al., 2015), que tem como proposta ser uma Matriz de Soma de Ewald, porém que selecione somente as características mais importantes para reduzir o custo computacional.

O MBTR (HUO; RUPP, 2018), que é o *Many-body Tensor Representation* ou representação por tensores de vários corpos é capaz de codificar tanto sistemas finitos, como periódicos. Isso é feito através da quebra desses sistemas em grupos, de diferentes tamanhos, de elementos. Já as ACSFs (BEHLER, 2011), que são as *Atom-centered Symmetry Functions* ou funções de simetria centradas no átomo condificam as configurações dos átomos ao redor de um átomo central através de funções de simetria. Existe também o SOAP (BARTÓK; KONDOR; CSÁNYI, 2013), *Smooth Overlap of Atomic Orbitals* ou Sobreposição suave de orbitais atômicos, que codifica o ambiente de uma estrutura atômica pela expansão de uma densidade atômica gaussiana baseada em harmônicos esféricos e funções de base radial. Para codificação de estruturas locais, também está presente na literatura o BOP (STEINHARDT; NELSON; RONCHETTI, 1983), *Bond-orientational Order Parameter* ou parâmetro de ordem de orientação de ligação.

A partir de representações como SMILES (*Simplified-Molecular-Input Line-Entry System*) (ANDERSON; VEITH; WEININGER, 1987), é possível extrair vários descritores moleculares. O pacote Mordred (MORIWAKI et al., 2018), por exemplo, disponibiliza rotinas para extração de mais de 1800 características dessa representação, desde matrizes de adjacência até calculadoras de carga topológica. Toda a capacidade do pacote está documentada em sua página do GitHub [1].

Alguns exemplos de pacotes de descritores muito utilizados onde aprendizado de máquina é aplicado no domínio de ciência dos materiais incluem o DScribe (HIMANEN et al., 2020), ML4Chem (KHATIB; JONG, 2020) e o previamente mencionado Mordred. Apesar de haver uma certa sobreposição entre os métodos implementados, eles trazem vantagens e desvantagens, além de interfaces diferentes o que pode ajudar no desenvolvimento do código.

[1] https://mordred-descriptor.github.io/documentation/master/

2.4 Medidas de Similaridade

Para comparar dois elementos de um repositório, é necessária uma medida de similaridade. Ao comparar dados quantitativos, a similaridade entre dois elementos é calculada a partir da distância entre eles: quanto menor a distância, maior a similaridade.

Em (CHA, 2007) é feita uma revisão sobre diversas medidas de similaridade/distância. Como são muitas, essa revisão aborda em detalhes uma de cada família de medidas. Começando pela família de Minkowski, tem-se a distância de Minkowski, que é dada pela equação

$$L_p(p,q) = \sqrt[n]{\sum_{i=1}^{d}(q_i - p_i)^n}, \tag{2.2}$$

que é uma generalização da distância Euclidiana (Equação 2.3), onde $n = 2$. Essa família inclui, também, as distâncias de Manhattan e Chebyshev.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{2.3}$$

Na família L_1, existem as distâncias de Søresen, Gower, Soergel, Kulczynski, Canberra e Lorentzian. A distância de Søresen, dada por

$$d_{sor}(p,q) = \frac{\sum_{i=1}^{d}|p_i - q_i|}{\sum_{i=1}^{d}(p_i + q_i)} \tag{2.4}$$

e encontra uso na área de ecologia (CHA, 2007).

A família de intersecções conta com a distância de intersecção, Wave Hedges, Czekanowski, Motyka, Kulczynski, Ruzicka e Tanimoto. Essa família calcula intersecções entre funções de densidade de probabilidade e é uma forma largamente utilizada de medidas de semelhança (DUDA; HART; STORK, 2000). A distância de intersecção é

$$d_{non-is}(p,q) = \frac{1}{2}\sum_{i=1}^{d}|p_i - q_i| \tag{2.5}$$

A próxima família de medidas é a do produto interno, que inclui a similaridade do produto interno, média harmônica, coeficiente dos cossenos, Kumar-Hassebrook, Jaccard e Dice. A similaridade do produto interno pode ser escrita como

$$S_{IP}(p,q) = p \bullet q = \sum_{i=1}^{d} p_i q_i, \tag{2.6}$$

e é interessante, pois o produto interno de dois vetores retorna um valor escalar, que pode ser utilizado como uma medida, já que é possível pensar nele como a justaposição dos vetores. As demais medidas nesse família trazem versões mais elaboradas desse conceito.

A soma das médias geométricas é referida como semelhança de fidelidade e origina a família de fidelidade. As equações integrantes são a semelhança de fidelidade em si, Bhattacharyya, Hellinger, Matusita e *Squared-chord*. A equação da semelhança de fidelidade é dada por

$$S_{Fid}(p,q) = \sum_{i=1}^{d} \sqrt{p_i q_i} \tag{2.7}$$

A próxima família é a família do χ^2 ou L^2. Os membros dela são a distância euclidiana quadrática, χ^2 de Pearson, χ^2 de Neyman, χ^2 quadrático, χ^2 probabilístico simétrico, divergência, Clark e χ^2 aditivo simétrico. A fórmula para o χ^2 de Pearson é

$$d_p(p,q) = \sum_{i=1}^{d} \frac{(p_i - q_i)^2}{q_i}, \tag{2.8}$$

que é a distância euclidiana quadrática, mas com todos os termos divididos por q_i. O χ^2 de Neyman divide por p_i e o χ^2 quadrático divide por $p_i + q_i$. Divergência é o χ^2 quadrático multiplicado por 2. Todos os membros são bem similares e partiram do χ^2 de Pearson, sendo que muitas das medidas foram criadas para lidar com a assimetria dele.

A última família é a da entropia de Shannon. Essa família tem como membros Kullback-Leibler, Jeffreys, divergência de K, Topsøe, Jensen-Shannon e diferença de Jensen. Todas elas vem do conceito de incerteza probabilística (entropia) de Shannon. A distância de Kullback-Leiber é definida como

$$d_{KL}(p,q) = \sum_{i=1}^{d} p_i ln \frac{p_i}{q_i} \tag{2.9}$$

Além de todas essas medidas, existem algumas que combinam mais de uma medida em uma única equação, como Taneja e Kumar-Johnson. Cada medida de similaridade tem seus pontos fortes e fracos e deve ser aplicada dependendo do contexto, logo, cabe ao cientista decidir qual a melhor para o caso estudado.

2.5 Algoritmos de Agrupamento e K-Means

Algoritmos de agrupamento são utilizados para fins exploratórios, tipicamente análise de padrões, agrupamento de dados, tomada de decisão, mineração de dados, recuperação de documentos e classificação de padrões (JAIN; MURTY; FLYNN, 1999).

O funcionamento básico desses algoritmos é, a partir de um conjunto de dados, identificar os grupos nele existentes. Elementos pertencentes a um grupo são mais similares entre si que com elementos de outros grupos. Vale mencionar, também, que isso é feito somente com as características dos elementos presentes no repositório, caso se trate de um algoritmo de agrupamento convencional. Isso é um traço de algoritmos de aprendizado de máquina não supervisionado.

Tome um conjunto de dados de átomos, por exemplo. Dois átomos que têm valores de características semelhantes, como eletronegatividade e massa, provavelmente apresentam propriedades mais semelhantes do que dois átomos que têm valores completamente diferentes.

O conjunto de características de um elemento é também conhecido como seu padrão (DUDA; HART et al., 1973). Matematicamente, um padrão é um vetor multidimensional de elementos que podem ser classificados como

- Características quantitativas, que podem ser contínuas (como comprimento, peso, volume), discretas (quantidade de pessoas numa sala) e de intervalo (duração de um processo)
- Características qualitativas/categóricas, que podem ser nominais/sem ordem (como cores) ou ordinais (cargo na hierarquia de uma empresa) (Chidananda Gowda; DIDAY, 1991)

Algoritmos de agrupamento utilizam medidas de similaridade/distância para decidir quais elementos são similares ou dissimilares. A medida mais amplamente utilizada é a distância Euclidiana da família de Minkowski, discutida na Seção 2.4. Em suma, tem-se que se

$$\sum_{v=1}^{n} d_v(A, B) < \sum_{v=1}^{n} d_v(A, C), \tag{2.10}$$

onde v representa o valor de cada uma das n características dos elementos no repositório, então A é mais semelhante a B do que C (em uma perspectiva puramente voltada a agrupamento de dados).

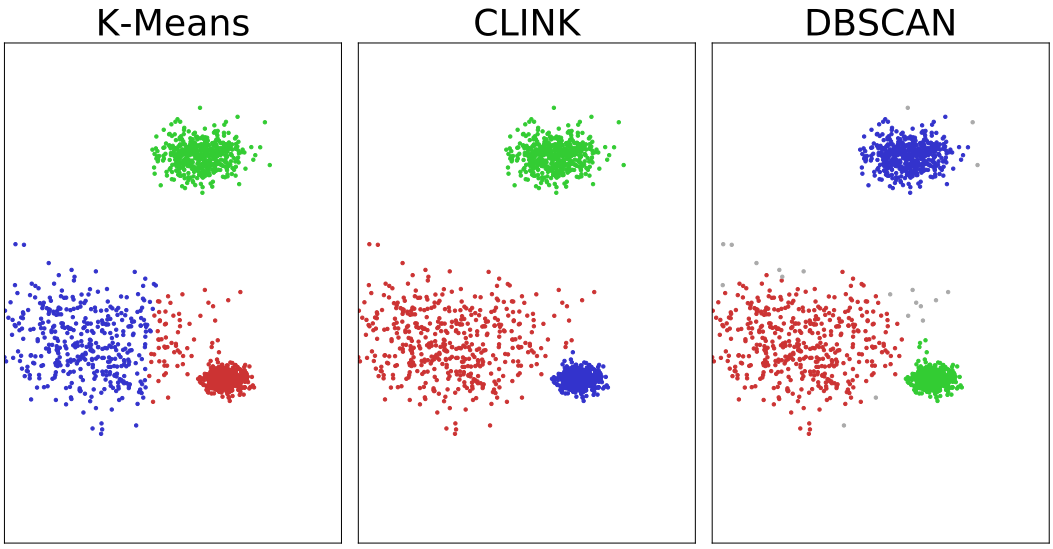
As medidas de similaridades lidam muito bem com valores numéricos, porém no caso dos elementos serem categóricos, é possível codificá-los em valores numéricos com técnicas como *One-hot encoding*, representação binária ou *hashing*. Levando em conta que essas representações alternativas podem impactar o uso de memória, custo computacional e eficácia das medidas de similaridade, é preciso avaliar, cuidadosamente, qual é a representação mais propícia para o problema (SEGER, 2018).

Com isso, é possível concluir que um dos fatores mais definidores de um algoritmo de agrupamento é como a medida de similaridade é aplicada: ela verifica a distância entre os elementos e os centros de massa? Compara a distância entre os próprios elementos para encontrar os pares mais próximos? Cada aplicação proposta gera um método único e qual é o melhor depende inteiramente da situação. A Figura 2.5 mostra como diferentes algoritmos de agrupamento agrupam, de maneira distinta, o mesmo conjunto de dados gerado aleatoriamente.

Uma representação matemática simples para um método de agrupamento genérico é dada pela seguinte equação:

$$K(\mathbb{X}) = \overrightarrow{\mathbb{L}} \tag{2.11}$$

Figura 1 – Resultados de diferentes métodos de agrupamento para um mesmo conjunto de dados



Fonte: o autor

onde K é o método de agrupamento, $\mathbb{X} \in (\mathbb{R}^n \times \mathbb{R}^m)$ é o conjunto de dados de entrada com n elementos e m características por elemento, e $\overrightarrow{\mathbb{L}} \in \mathbb{Z}_+^n$ é o vetor de rótulos, que contém o grupo atribuído a cada um dos n elementos.

Existem diversas técnicas de agrupamento na literatura. Começando com os algoritmos particionais, nessa categoria está presente o algoritmo de agrupamento mais utilizado devido a sua rapidez e simplicidade de implementação: o K-Means (MACQUEEN, 1967). Esse é o algoritmo utilizado neste TCC e é dado pelos seguintes passos:

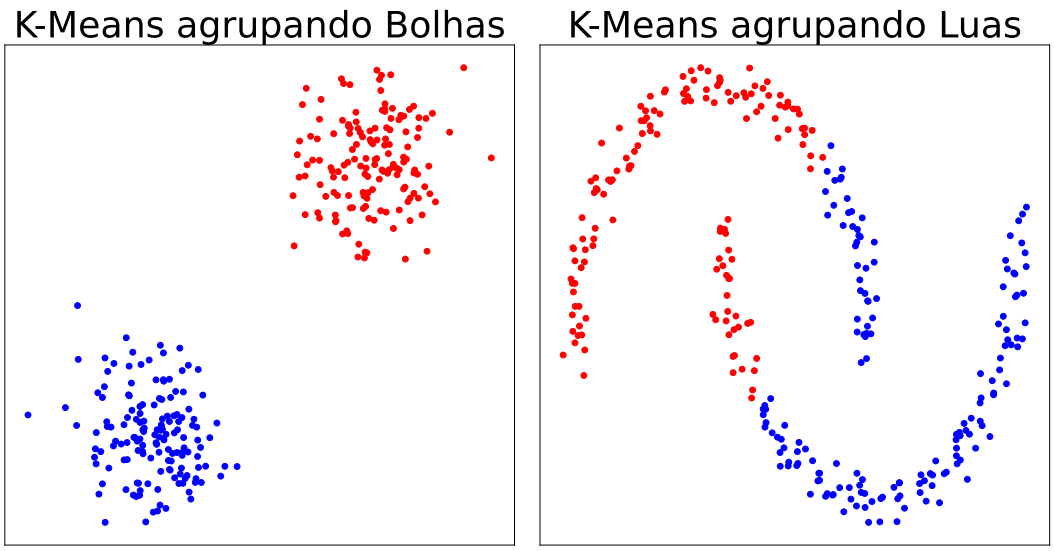
1. Escolha k centroides que coincidam com k elementos aleatórios do repositório
2. Atribua cada elemento ao centroide mais próximo (distância euclidiana)
3. Recalcule o centroide de cada grupo como o centro de massa de seus membros
4. Enquanto o critério de convergência não for atendido, repita a partir de 2 (MACQUEEN, 1967)

Vale ressaltar que, neste trabalho, o K-Means (implementação do scikit-learn) é utilizado como o método de agrupamento devido à sua ótima performance e aos bons resultados obtidos pelo grupo QTNano usando o algoritmo em pesquisas anteriores (BATISTA et al., 2021)

Embora o K-Means seja um bom algoritmo de agrupamento, ele tem suas limitações, assim como qualquer outra técnica. Duas das mais notáveis são:

- Precisa saber de antemão a quantidade de grupos a serem gerados
- Não é tão eficaz no agrupamento de dados não gaussianos, como mostrado na Figura 2

Figura 2 – K-Means agrupando conjuntos de dados de formatos diferentes



Fonte: o autor

Com o K-Means documentado, diversas variantes desse algoritmo começaram a emergir (ANDERBERG, 1973), onde algumas delas tentavam melhorar o posicionamento inicial das sementes, outras possibilitavam divisões e uniões de partições dos agrupamentos resultantes (por exemplo, ISODATA, *Iterative Self-Organizing Data Analysis Technique* ou técnica iterativa de análise de dados auto-organizáveis (BALL; HALL, 1965)). Outra possibilidade é mudar a função de critério, permitindo diferentes representações para um grupo, como é visto em (DIDAY, 1973; SYMONS, 1981).

Além dos algoritmos particionais, existem também os hierárquicos que fazem processos de aglomeração ou divisão de grupos a medida que as iterações avançam. A maioria dos algoritmos dessa família têm origem do algoritmo *Single-Linkage* (SNEATH; SOKAL et al., 1973), *Complete-Linkage* (KING, 1967) ou Variância Mínima (WARD, 1963; MURTAGH, 1983). Desses, os dois primeiros são bastante populares, sendo que diferem somente na maneira como caracterizam a similaridade entre pares de agrupamentos: o *Single-Linkage* considera que a distância entre dois agrupamentos é a distância entre os elementos mais próximos de cada um dos grupos, enquanto no *Complete-Linkage*, a distância considerada é entre os elementos mais distantes.

Uma outra família de algoritmos de agrupamento é a baseada em teoria dos grafos. A abordagem mais conhecida dessa classe é a baseada na construção da MST (*Minimal Spanning Tree* ou árvore geradora mínima) dos dados e posterior deleção das arestas mais longas para gerar os grupos (ZAHN, 1971). Proveniente da literatura de redes complexas, técnicas de detecção de comunidades também são vistas como ferramentas de agrupamento em grafos (PINHEIRO et al., 2020; FORTUNATO, 2010).

Também existem os *Mixture-Resolving and Mode-Seeking Algorithms* (algoritmos de resolução de mistura e busca de modo), cuja ideia básica é assumir que os padrões dos dados a

serem agrupados pertencem a alguma distribuição e, através dessa ideia, tentar identificar quais são os parâmetros dessa distribuição. Essa espécie de método de agrupamento pode ser vista em (DEMPSTER; LAIRD; RUBIN, 1977; MITCHELL et al., 1997).

Redes neurais artificiais também podem ser utilizadas para realizar agrupamento. Nessa classe tem-se a rede SOM *Self-Organizing Map* (mapa auto-organizável) (KOHONEN, 1989), modelo de teoria da ressonância adaptativa (CARPENTER; GROSSBERG, 1990), rede competitiva (ou *winner-take-all*) (JAIN; MAO; MOHIUDDIN, 1996), entre outras implementações.

Algoritmos de agrupamento que levam a densidade dos dados também estão presentes na literatura, como é o caso do DBSCAN (*Density-based spatial clustering of applications with noise* ou Agrupamento espacial baseado em densidade para aplicações com ruído) (ESTER et al., 1996), do OPTICS (*Ordering Points To Identify the Clustering Structure* ou ordenando pontos para identificar a estrutura de agrupamento) (ANKERST et al., 1999) e do ILS (*Iterative Label Spreading* ou Espalhamento de rótulo iterativo) (PARKER; BARNARD, 2019).

Finalmente, existem outras técnicas, por exemplo: técnicas que utilizam algoritmos genéticos, técnicas que utilizam inspirações interdisciplinares, como processos físicos ou técnicas baseadas em lógica nebulosa. Uma variedade de algoritmos de agrupamento é abordada em (JAIN; MURTY; FLYNN, 1999).

2.6 Medidas de Qualidade de Agrupamento

Apesar do resultado dos algoritmos de agrupamento poderem ser subjetivos, ou seja, dependem de um especialista no domínio para dizer se os grupos formados fazem sentido ou não, ainda assim existem técnicas estatísticas para extrair informações quantitativas sobre o quão eficaz foi um processo de agrupamento no ponto de vista puramente matemático.

O método mais clássico de medir qualidade de agrupamentos é conhecido por *Elbow Method* ou método do cotovelo (THORNDIKE, 1953), que essencialmente é um gráfico da WCSS (*within-cluster sum of the squares* ou soma dos quadrados intra-grupo) contra a quantidade de grupos. A WCSS é dada por:

$$\sum_{k=1}^{K} n_k \sum_{C_i=k} \sum_{j=1}^{d} (x_{ij} - \bar{x}_{kj})^2 \tag{2.12}$$

onde K é o número de grupos, n_k é o número de elementos no grupo k, C_i é o grupo ao qual o elemento i pertence, d é a quantidade de dimensões no padrão do elemento i e $\bar{x}_{kj}$ é a j-ésima característica no padrão do centroide do grupo k (BAIR, 2013). Caso o método do cotovelo seja eficaz para dado conjunto de dados e algoritmo de agrupamento, será possível observar um "cotovelo" no gráfico gerado. Esse "cotovelo" representa justamente a melhor quantidade de grupos, como é visível na Figura 3

Já o método mais comumente utilizado é a Silhueta (ou *Silhouette* (ROUSSEEUW, 1987). Esse método se baseia na comparação da compacidade e separação entre os grupos e é capaz de mostrar quais objetos estão estatisticamente dentro de seu grupo ideal, e quais estão ao acaso em um determinado local. A medida da silhueta é dada pela Equação 2.13 e varia entre -1 e 1. Quando é negativa indica que o elemento foi atribuído ao grupo incorreto, quando é próxima de 0 mostra que um dado elemento não está bem colocado e quando é positiva mostra que o elemento foi atribuído ao grupo correto. Quanto maior a magnitude da silhueta, mais acentuado são os efeitos e isso é possível ser observado na Figura 3.

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}, \quad 2 \leq K \leq |\mathbb{X}| - 1 \tag{2.13}$$

onde i é a amostra em análise e $|\mathbb{X}|$ é a quantidade de elementos no conjunto de dados.

Ainda sobre a silhueta, tem-se que a pontuação da silhueta (ou silhueta global) é a média de todos os coeficientes de silhueta, ou seja:

$$\mathbb{S} = \frac{\sum_{i=1}^{|\mathbb{X}|} S_i}{|\mathbb{X}|} \tag{2.14}$$

Uma outra medida para extrair a qualidade dos grupos formados é o critério da proporção da variância (CALIńSKI; HARABASZ, 1974), também conhecida como índice de Calińsk-Harabasz. Esse índice é definido como a proporção entre a dispersão intra-grupo e extra-grupo. O índice para um número K de grupos em um repositório $\begin{bmatrix} d_1 & d_2 & \cdots & d_N \end{bmatrix}^T$ com N elementos é matematicamente definido como:

$$CH = \frac{\frac{\sum_{k=1}^{K} n_k \|c_k - c\|^2}{K-1}}{\frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \|d_i - c_k\|^2}{N-K}} \tag{2.15}$$

onde n_k é o número de elementos no grupo k, c_k é o centroide do grupo k, c é o centroide de todos os elementos do conjunto de dados. A Figura 3 ilustra o índice mudando para várias quantidades de grupos. Quanto maior o valor do índice, melhor é a configuração.

Até agora, os métodos abordados não utilizaram nenhuma espécie de rotulação externa para definir a qualidade dos agrupamentos, porém caso essa informação esteja disponível, existem métodos que fornecem valores mais significativos utilizando-a. Um desses métodos é o ARI (*Adjusted Rand Index* ou índice de Rand ajustado) (HUBERT; ARABIE, 1985). O Índice de Rand é dado pela Equação 2.16.

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.16}$$

onde TP são os verdadeiros positivos, TN os verdadeiros negativos, FP os falsos positivos e FN são os falsos negativos. Já o ARI, é definido pela equação 2.17

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)} \tag{2.17}$$

onde RI é o índice de Rand, $max(RI)$ é o índice de Rand máximo e $E(RI)$ é o índice de Rand esperado. Novamente, é possível ver o índice em funcionamento na Figura 3

O outro método bastante utilizando quando existem rótulos externos para comparar com os obtidos pelos algortimos de agrupamento é o NMI (*Normalized Mutual Informatio* ou informação mútua normalizada) (KREER, 1957). Seja (X, Y) um par de variáveis aleatórias sobre o espaço $\mathbb{X} \times \mathbb{Y}$. Se sua distribuição conjunta for $P_{(X,Y)}$ e as distribuições marginais forem P_X e P_Y, a informação mútua é definida por

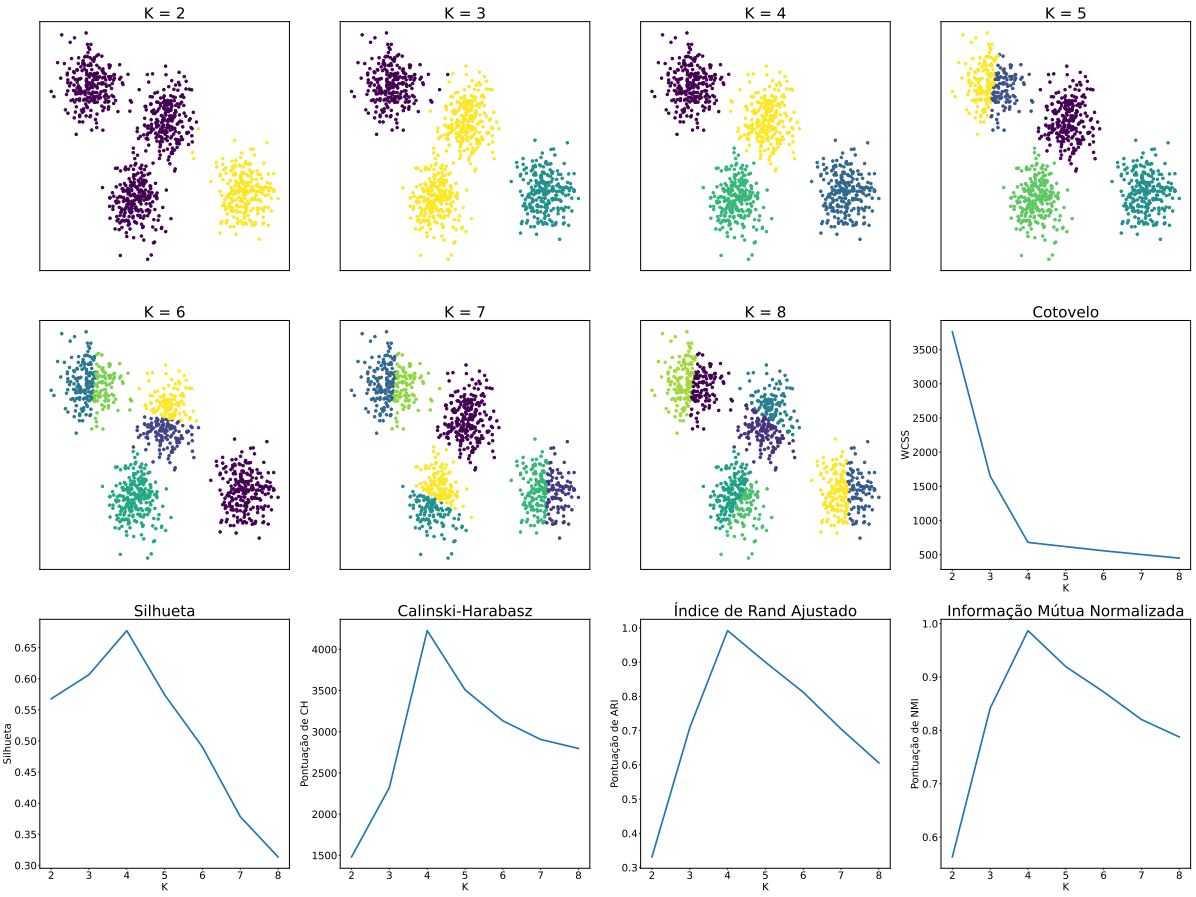
$$I(X;Y) = D_{KL}(P_{(X,Y)} \| P_X \otimes P_Y) \tag{2.18}$$

onde D_{KL} (Divergência de Kullback-Leibler) é dado por

$$D_{KL}(P\|Q) = \sum_{x \in \mathbb{X}} P(x) log\left(\frac{P(x)}{Q(x)}\right) \tag{2.19}$$

quando P e Q são definidas no mesmo espaço de probabilidade $\mathbb{X}$ (MACKAY, 2003).

Figura 3 – Diversos índices de qualidade ao variar o K do K-Means



Fonte: o autor

2.7 Algoritmos de Otimização e Basin-Hopping

Otimização é a área da matemática aplicada que lida com a otimização de um ou mais critérios (conhecidos como funções objetivo) com a finalidade de retornar a combinação ótima (que minimiza ou maximiza um resultado buscado) de valores para os critérios em questão (WEISE, 2009). Problemas com objetivos explícitos podem ser expressos como

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{min/max}\ f(x),\ x = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \in \mathbb{R}^n, \\ \text{sujeito a } \phi_j(x) = 0,\ j \in [1..M], \\ \psi_k(x) \geq 0,\ k \in [1..N] (YANG, 2008) \end{aligned} \tag{2.20}$$

onde $f(x)$ é a função de custo, x é o vetor de decisão (e $x_1, \cdots, x_n$ é um estado), $\phi_j(x)$ são restrições em termos de M igualdades e $\psi_k(x)$ são restrições em termos de N desigualdades.

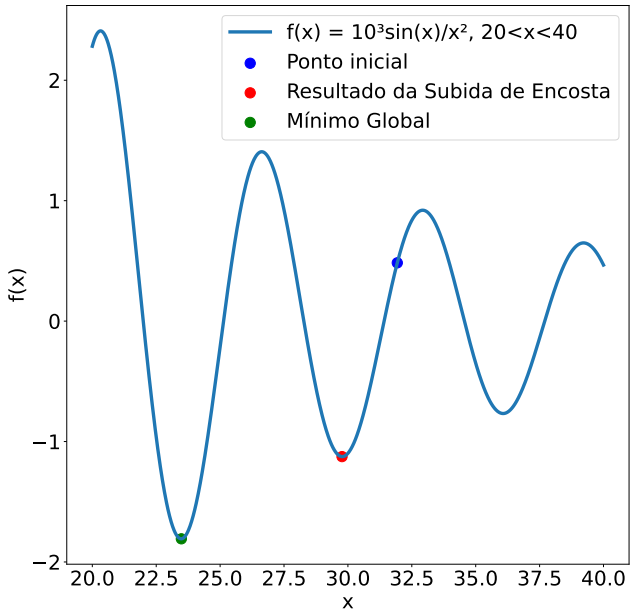
O algoritmo de subida de encosta (*hill-climbing*) é um exemplo bem simples de otimização. Dada uma iteração inicial i e um ponto inicial $x_i \in\in S \subset R^d$, onde d é a quantidade de características num padrão e S é um conjunto finito de elementos, o subida de encosta move-se para um ponto vizinho $x_{i+1} \in S$. Assim, é computada uma função custo $C_i = f(x_{i+1}) - f(x_i)$. No caso de uma minimização, se $C_i < 0$, o novo ponto x_{i+1} é aceito e na próxima iteração ele será utilizado como partida, caso contrário o ponto x_i continua como partida e será tentado um novo x_{i+1}. Para maximização o procedimento é o mesmo, porém o critério se torna $C_i > 0$. Sendo assim, é possível extrair que o algoritmo caminha no sentido que minimiza/maximiza a função custo dependendo do caso (LIM; RODRIGUES; ZHANG, 2006).

Embora *hill-climbing* seja muito fácil de implementar, ele deixa a desejar em problemas com várias soluções "boas", porque algumas delas podem ser apenas ótimos locais, ou seja, não são os melhores (mais otimizados) resultados para a função. Basicamente, algoritmos de busca local (como este) encontram, como o nome sugere, ótimos locais (como ilustrado na Figura 4).

Vale mencionar também que existem variantes do algoritmo de subida de encosta, como por exemplo a versão estocástica, que em vez de examinar todos os vizinhos (o que é impraticável em muitos casos, por exemplo: quando $|S|$ é infinito ou grande demais para ser computacionalmente inviável), seleciona vizinhos aleatoriamente e decide a aceitação dependendo da melhoria provinda de aceitar aquele vizinho (RUSSELL; NORVIG, 2009). Uma outra variante é a de aceitação tardia (BURKE; BYKOV, 2017).

Um outro algoritmo de otimização local é a busca tabu (GLOVER, 1986). Ela é levemente diferente da implementação da subida de encosta pois além de checar uma potencial solução para o problema (vizinho "bom"), checa também os vizinhos desse vizinho (que são similares com exceção de pequenos detalhes) na expectativa de encontrar uma solução melhorada. Outra diferença é que a busca tabu emprega dois critérios para melhorar a busca local: *movimentos piores*, que permitem com que o algoritmo caminhe para locais onde há uma perda de qualidade e *proibições*, que desencorajam o algoritmo de revisitar locais. Esses dois critérios

Figura 4 – Tentativa do *Hill-climbing* de otimizar a função f.



Fonte: o autor

ajudam a busca ficar menos presa em ótimos locais imediatos e a não perder tempo revisitando locais.

Já na categoria de otimização global é típico que algum elemento de aleatoriedade seja adicionado aos algoritmos de otimização para fazer com que a fuga de ótimos locais seja possível. Um algoritmo bem famoso de busca global é o *Simulated Annealing* (têmpera simulada) (KIRKPATRICK; GELATT; VECCHI, 1983; LAARHOVEN; AARTS, 1987), que se inspira no fenômeno físico de têmpera. Esse algoritmo introduz o conceito de *temperatura* de forma que seja uma medida de agitação do buscador. Quanto maior a temperatura, maior a chance do algoritmo saltar para um outro local no espaço de busca, potencialmente distante de onde estava anteriormente. À medida que as iterações avançam, a temperatura reduz até que os saltos ficam menos e menos frequentes fazendo com que, eventualmente, a têmpera simulada se torne um algoritmo de subida de encosta.

Um algoritmo de funcionamento similar à têmpera simulada é o *Basin-Hopping*. Esse algoritmo tem sido bem-sucedido em lidar com problemas multimodais multivariáveis (LEARY, 2000). O termo "Basin-Hopping" foi cunhado por Wales e usado para encontrar as estruturas de menor energia de aglomerados de Lennard-Jones contendo até 110 átomos (Wales; Doye, 1997).

A técnica usada pelo algoritmo para realizar buscas globais é, além de fazer os procedimentos de busca local padrão, perturbar a posição de busca atual de forma aleatória, mas controlada. Isso é feito para que haja uma maneira de escapar dos ótimos locais.

O critério de aceitação desempenha um grande papel no comportamento do Basin-Hopping. Em sua forma mais simples, o algoritmo sempre armazena o melhor valor encontrado até o ponto atual e rejeita qualquer outra opção que não seja tão boa (portanto, sendo um otimizador monotônico), mas o critério de aceitação geralmente é o critério de Metrópolis dos algoritmos de Monte Carlo.

O algoritmo 1 mostra um pseudo-código para a implementação monotônica do Basin-Hopping. Observe que para alterar o critério de aceitação, basta trocar a operação 7 pela desejada. Além disso, a condição de parada é geralmente um limite de contagem de iteração ou quando os ganhos fornecidos por manter o algoritmo em execução são muito pequenos.

Procedimento 1 Basin-Hopping Monotônico (OLSON et al., 2012)

1: $i \leftarrow 0$
2: $X_i \leftarrow$ ponto inicial aleatório no espaço de variáveis
3: $Y_i \leftarrow$ BUSCALOCAL(X_i)
4: **Enquanto** PARADA não satisfeita **faça**
5: $X_{i+1} \leftarrow$ PERTURBE(Y_i)
6: $Y_{i+1} \leftarrow$ BUSCALOCAL(X_{i+1})
7: **Se** $f(Y_{i+1}) < f(Y_i)$ **então**
8: $i \leftarrow i + 1$
9: **Fim Se**
10: **Fim Enquanto**

Existem, também, algoritmos de otimização global baseados em entidades ou processos da natureza. Um deles é o PSO (*Particle Swarm Optimization* ou otimização por enxame de partículas) (KENNEDY; EBERHART, 1995), onde cada partícula se movimenta pelo espaço de busca e é atraída pelo melhor valor encontrado por ela, porém também é guiada para o melhor valor encontrado pelo grupo. A tendência, a medida que as iterações passam, é que as partículas convirjam para o ótimo global da função. Esse algoritmo é uma abordagem metaheurística, já que não precisa de nenhum conhecimento prévio do universo e é capaz de buscar espaços bem largos.

Algoritmo de ACO (*Ant Colony Optimization* ou otimização por colônia de formigas) (DORIGO; BIRATTARI; STUTZLE, 2006) é um outro exemplo dessa classe de otimizadores globais. Feito para ser eficiente em realizar buscas de caminhos em grafos, o algoritmo utiliza o conceito de feromônio para fazer com que as formigas (agendes buscadores) convirjam para o ótimo global. Essencialmente, as formigas caminham aleatoriamente entre os nós e vão deixando feromônio com elas, que enviesa o percurso das demais formigas. Com o tempo, o feromônio começa a dissipar, então para que um caminho seja mantido, é preciso que várias formigas continuem passando por ele. Com a ajuda do feromônio e medidas heurísticas, eventualmente as formigas passam a caminhar pelo menor percurso no grafo.

Existem diversos outros algoritmos inspirados na natureza, como o algoritmo do morcego (YANG; GANDOMI, 2012) e algoritmo genético (MITCHELL, 1998), assim como exis-

tem diversos outros algoritmos mais voltados à matemática pura, como o método do gradiente descendente (LEMARÉCHAL, 2012) e o conjunto de métodos Quasi-Newton (BROYDEN, 1967). A área de otimização é gigantesca e contém diversos algoritmos diferentes, para propósitos distintos e com inspirações variadas.

Infelizmente, mesmo os algoritmos de busca global podem falhar em encontrar um ótimo global, mas, em geral, são capazes de retornar resultados melhores do que os algoritmos de busca local.

2.8 Agrupamento Supervisionado

Agrupamento supervisionado é uma classe especial de algoritmos de agrupamento que utilizam dados externos ao conjunto de características de forma a tentar obter agrupamentos mais significativos ou que atendem requisitos impostos pelos especialistas no domínio em questão. Quando apenas parte do conjunto de dados possui essa informação externa, tais algoritmos também são chamados de agrupamento semi-supervisionado (BAIR, 2013; VARGHESE; CAWLEY; HONG, 2018),

Uma generalização do K-Means descrita em (BASU; BANERJEE; MOONEY, 2002), chamada K-Means com restrição, utiliza de rótulos conhecidos de um subconjunto do conjunto completo de dados para definir as sementes iniciais (para esse algoritmo supõe-se que o número K de grupos é conhecido). Isso é interessante, pois remove a seleção inicial aleatória do K-Means, o que ajuda no custo computacional, pois normalmente o K-Means é rodado múltiplas vezes para evitar que um início aleatório ruim prejudique o resultado final, e também ajuda na convergência para uma configuração final melhor, já que por si só o K-Means tende a convergir para mínimos locais.

O K-Means com restrição possui uma variação conhecida como K-Means semeado (BASU; BANERJEE; MOONEY, 2002), que em vez de já atribuir todos os elementos com rótulos aos grupos respectivos, atribui só um de cada rótulo e os demais são atribuídos aos grupos rotulados seguindo a maneira convencional do K-Means. Essa variação é importante quando houver a possibilidade de rótulos incorretos/imprecisos. O K-Means com restrição não é capaz de lidar com rótulos incorretos e acaba gerando resultados ruins. Vale mencionar que Basu et al. não foram as únicas pessoas que propuseram uma transformação desse tipo, pois uma abordagem bem similar está presente em (GAYNOR; BAIR, 2012), embora seja mais voltada à aplicação e em agrupamentos esparsos.

Uma outra abordagem é utilizar relações entre elementos como restrição, em vez de rótulos externos, ou seja, nesse modelo existem informações sobre quais elementos precisam estar no mesmo grupo (relação *must-link*) e quais não podem estar no mesmo grupo (relação *cannot-link*); ou ainda, quais elementos devem pertencer a um grupo (relação *positive label*) ou não podem pertencer a um grupo (*negative label*) Uma implementação famosa é a descrita em

(WAGSTAFF et al., 2001), chamada de COP-KMEANS, cujo algoritmo é:

1. Atribua, aleatoriamente, cada elemento à um grupo
2. Para cada característica j e grupo k, calcule $\bar{x}_{kj}$ (média da característica j no grupo k).
3. Atribua cada elemento i a um novo grupo C_i, onde C_i se dá pela Equação 2.21
4. Repita 2. e 3. até que o algoritmo convirja.

$$C_i = arg\min_{k \in D_{ik}} \sum_{j=1}^{d} (x_{ij} - \bar{x}_{kj})^2 \tag{2.21}$$

onde D_{ik} é o conjunto de atribuições de elementos i a grupos k que não ferem nenhuma restrição *must-link* e *cannot-link* (se $D_{ik} = \emptyset$ em qualquer iteração, o COP-KMEANS falha).

Posteriormente, foi elaborada uma versão melhorada do COP-KMEANS chamada PCK-Means (BASU; BANERJEE; MOONEY, 2004). Essa versão permite que restrições fossem violadas em situações que existem fortes evidências que é uma restrição incorreta. O PCKMeans tenta minimizar uma versão modificada da versão objetivo para realizar o procedimento de "correção" de restrições potencialmente incorretas. Resultados similares podem ser obtidos ao alterar a medida de similaridade usada pelo K-Means de forma que elementos com relações *must-link* estejam espacialmente mais próximos e elementos com relações *cannot-link* estejam espacialmente mais distantes (KLEIN; KAMVAR; MANNING, 2002).

Apesar da grande maioria dos métodos com restrição serem variantes do K-Means ou outros algoritmos particionais, existem também variantes de métodos hierárquicos. Essas variantes devem considerar tipos de restrições diferentes/modificadas comparadas às mencionadas anteriormente pelo seguinte motivo: tome um algoritmo hierárquico aglomerativo, é inevitável que em algum ponto da hierarquia as restrições *must-link* sejam satisfeitas e as *cannot-link*, violadas (o inverso ocorre para abordagens divisivas).

Uma possibilidade é exigir que restrições *must-link* sejam atendidas na camada mais baixa da hierarquia e que restrições *cannot-link* não estejam na mesma hierarquia de agrupamentos (senão estariam sendo violadas). Para isso, é preciso gerar várias hierarquias diferentes de agrupamento, onde existe uma hierarquia para cada elemento que é parte de uma relação *cannot-link* (MIYAMOTO; TERAMI, 2010). Seguindo uma estratégia similar, é viável também exigir que certos elementos (os que possuírem relações *must-link*) sejam agrupados antes de quaisquer outros elementos (BADE; NURNBERGER, 2006) ou até mesmo requisitar que uma ordem de junção entre elementos seja cumprida (ZHAO; QI, 2010).

Na classe de agrupamento por densidade, tem-se a abordagem do *Iterative Label Spreading* (PARKER; BARNARD, 2019), discutido na seção 2.5, que pode utilizar rótulos para definir a quantidade de agrupamentos e qual será a semente inicial de cada um deles. Na versão

não-supervisionada desse método, somente uma semente inicial é definida e ela é o elemento mais próximo do centro de massa da totalidade do conjunto de dados.

3 Metodologia

3.1 Ferramentas para a Construção do Programa

O código-fonte do algoritmo proposto é inteiramente escrito em *Python 3* [1]. Essa decisão se deu ao fato de existirem muitas ferramentas matemáticas úteis e bibliotecas de Aprendizado de Máquina disponíveis para a linguagem, além de ser amplamente utilizada em diversos projetos do grupo. As bibliotecas utilizadas são as seguintes:

- **SciPy**: ecossistema baseado em Python de software de código aberto para matemática, ciências e engenharia. [2]
- **NumPy**: O pacote fundamental para computação científica com Python. Combina a linguagem Python rica e fácil de usar com a velocidade da linguagem C. [3]
- **scikit-learn**: a coleção mais conhecida de métodos de Aprendizado de Máquina para a linguagem, contendo algoritmos para realizar agrupamento, classificação, regressão e muitas outras operações. [4]
- **pandas**: ferramenta de código aberto para manipulação e análise de dados poderosa e fácil de usar. Suporta muitos formatos de conjuntos de dados locais e fornece estruturas de dados úteis. [5]
- **Matplotlib**: A principal biblioteca de gráficos para Python. Essencial para criar visualizações estáticas, animadas e interativas. [6]

3.2 Entrada de Dados e Formatação

A caixa de ferramentas espera um arquivo de configuração INI[7] como entrada. Este arquivo deve fornecer 3 conjuntos principais de informações:

1. Repositório de entrada, pasta de saída e semente aleatória (para tornar as execuções reproduzíveis, caso necessário).

[1] https://www.python.org/
[2] https://www.scipy.org/
[3] https://numpy.org/
[4] https://scikit-learn.org/
[5] https://pandas.pydata.org/
[6] https://matplotlib.org/
[7] https://docs.python.org/3/library/configparser.html

2. Parâmetros para o algoritmo de agrupamento *K-Means*.

3. Parâmetros para o algoritmo de otimização *Basin-Hopping*.

Após a entrada ter sido fornecida, **pandas** é usado para carregar o conjunto de dados informado, que deve estar no formato CSV (*Comma-Separated Values*, ou em português, Valores Separados por Vírgulas) (SHAFRANOVICH, 2005), onde cada linha representa um elemento e cada coluna, uma característica. Os dados carregados são convertidos em um DataFrame do Pandas, que é a estrutura de dados corretamente compreendida pelo algoritmo.

Se o agrupamento for supervisionado, a coluna de propriedade de viés é separada do conjunto de dados de forma que o K-Means utilize somente as características e o Basin-Hopping utilize somente a propriedade, caso contrário, todo o repositório é enviado para o K-Means, já que o segmento do Basin-Hopping não será executado.

3.3 Representação Vetorial de Moléculas

Como o programa desenvolvido tem como núcleo técnicas de aprendizado de máquina, é fundamental que os dados estejam formatados de maneira adequada para alimentar esse segmento de algoritmos, que seria um vetor de características provindo de um descritor. O descritor escolhido para maioria dos conjuntos de dados são os autovalores da Matriz de Coulomb. Nos casos onde esse descritor não é viável no ponto de vista dos especialistas na área, são utilizadas características recomendadas por eles.

3.4 Agrupamento Supervisionado

Na caixa de ferramentas desenvolvida, o agrupamento supervisionado acontece em duas etapas principais:

- É selecionado automaticamente o número de grupos K através de medidas de qualidade de agrupamento (como por exemplo, a *Silhouette Score* ou pontuação da silhueta) (ROUSSEEUW, 1987) e uma implementação de algoritmo de força bruta simples.

- É feita a otimização dos grupos em si, seguindo um enviesamento da formação dos grupos proporcionado pela deformação do espaço (alteração dos pesos das características) pelo algoritmo de otimização de busca global Basin-Hopping.

O funcionamento da primeira etapa é bastante simples: o usuário seleciona um valor máximo K de número de grupos que deseja encontrar e o algoritmo de força bruta executa o cálculo da medida de qualidade de agrupamento para as configurações finais de agrupamento

geradas a partir de cada um dos valores de K, começando de 2 e indo até o valor especificado. O valor de K que resulta na melhor qualidade (de acordo com o critério escolhido) é selecionado.

A segunda etapa é mais complexa. No programa desenvolvido, nenhuma restrição explícita (*must-link/cannot-link* ou *positive/negative labels*) é passada, em vez disso, deve ser fornecida qual propriedade será usada para gerar as restrições implicitamente. Como consequência, a propriedade usada para enviesar o procedimento de agrupamento é destacada nos agrupamentos formados.

Para o procedimento de agrupamento supervisionado, é utilizado o K-Means, que já foi apresentado na seção 2.6, como algoritmo de agrupamento e o Basin-Hopping, detalhado na seção 2.7, como algoritmo de otimização. Como já visto, o Basin-Hopping é capaz de realizar pesquisa global e isso é uma característica importante, pois para esse problema, em específico, é fácil cair em um mínimo local tentando ajustar a matriz de associação para que a função de variância seja otimizada. A lógica completa para o agrupamento supervisionado é mostrada a seguir.

Seja $\mathbb{X} \in (\mathbb{R}^n \times \mathbb{R}^m)$ o conjunto de dados de entrada e $\overrightarrow{\mathbb{B}} \in \mathbb{R}^n$ o vetor de propriedades (para gerar o viés) com n valores (um para cada elemento na matriz de entrada). Além disso, seja $\overrightarrow{\mathbb{W}} \in \mathbb{R}^m$ o vetor de pesos.

O primeiro passo é aplicar os pesos a $\mathbb{X}$. Isso é feito multiplicando os valores de cada coluna j do conjunto de dados pelo valor do índice j no vetor de pesos. Para escrever este procedimento matematicamente, define-se um operador $\bullet$:

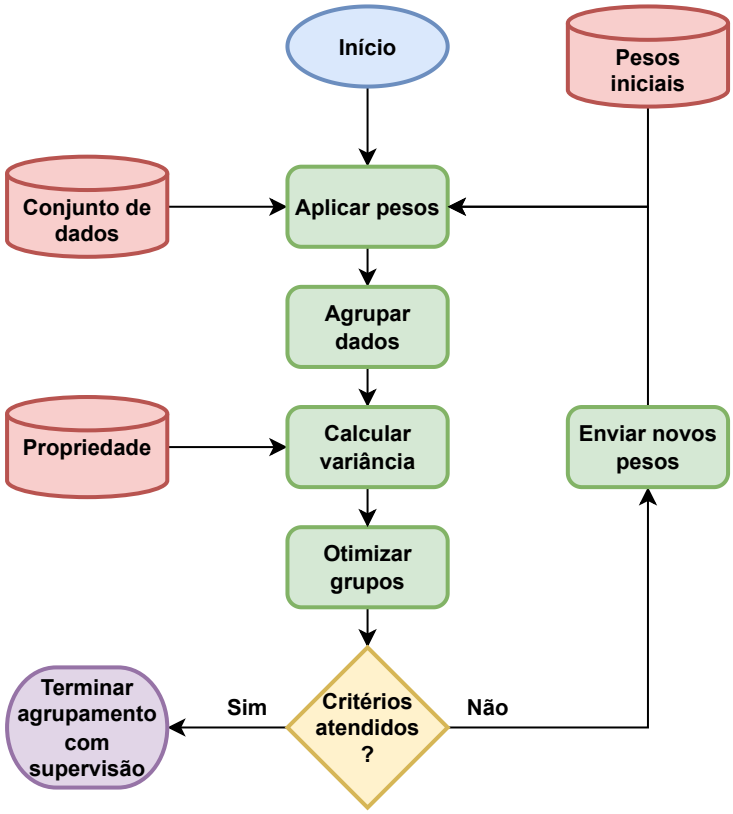
$$\begin{aligned} &A \bullet \overrightarrow{B} = A \cdot diag(\overrightarrow{B}), \\ &diag(\overrightarrow{B}) = \sum_{i=1}^{m} (e_i e_i^T)(e_i^T \overrightarrow{B}) \in (\mathbb{R}^m \times \mathbb{R}^m), \\ &A \in (\mathbb{R}^n \times \mathbb{R}^m),\ \overrightarrow{B} \in \mathbb{R}^m,\ e_i \text{ é o } i\text{-ésimo vetor da base} \end{aligned} \tag{3.1}$$

Dessa forma:

$$\mathbb{X} \bullet \overrightarrow{\mathbb{W}} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \bullet \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} = \begin{bmatrix} x_{11}w_1 & \cdots & x_{1m}w_m \\ \vdots & \ddots & \vdots \\ x_{n1}w_1 & \cdots & x_{nm}w_m \end{bmatrix} \tag{3.2}$$

O próximo passo é alimentar o método de agrupamento com a matriz de características ponderada para obter o vetor de rótulos. Assim, de acordo com a equação 2.11: $K(\mathbb{X} \bullet \overrightarrow{\mathbb{W}}) = \overrightarrow{\mathbb{L}} \in \mathbb{Z}_+^n$. Com isso, o vetor de rótulos $\overrightarrow{\mathbb{L}}$ e o vetor da propriedade escolhida $\overrightarrow{\mathbb{B}}$ são conhecidos,

Figura 5 – Fluxograma do agrupamento supervisionado



Fonte: o autor

então é possível obter a soma da variância intra-grupo de $\overrightarrow{\mathbb{B}}$:

$$V(\overrightarrow{\mathbb{L}}, \overrightarrow{\mathbb{B}}) = \sum_{j=1}^{k} \sigma^2 \left(\begin{bmatrix} c_1 & \cdots & c_n \end{bmatrix}^T \right),$$
$$c_i = \begin{cases} \overrightarrow{\mathbb{B}}_i, \text{ if } \overrightarrow{\mathbb{L}}_i = j \\ 0, \text{ caso contrário} \end{cases}, \ i \in [1..n], \tag{3.3}$$

k é o número de grupos

Então, todas as partes estão bem definidas, permitindo com que o algoritmo de otimização seja executado. Seguindo o conjunto de equações 2.20, é obtido que

$$\underset{\overrightarrow{\mathbb{W}} \in \mathbb{R}^m}{min/max}\ f(\overrightarrow{\mathbb{W}}) = V(\overrightarrow{\mathbb{L}}, \overrightarrow{\mathbb{B}}) \in \mathbb{R}_+, \tag{3.4}$$

sujeito aos parâmetros do algoritmo de otimização

Finalmente, o algoritmo de otimização é executado até que o número máximo de iterações seja alcançado ou os critérios de aceitação sejam atendidos. A Figura 5 mostra um diagrama que resume todo o processo.

Uma vez que o agrupamento supervisionado tenha encerrado, são escolhidos como amostras representativas os elementos do repositório mais próximos dos últimos protótipos (sementes) fornecidos pelo K-Means.

3.5 Projeções PCA e t-SNE

Ao fim do programa é gerada uma série de gráficos, que são detalhadamente descritos no Capítulo 4, para que os especialistas em ciência dos materiais possam interpretar os resultados.

Uma informação gráfica importante para apresentar é a distribuição dos elementos no espaço de recursos e a qual grupo eles são atribuídos. Infelizmente, é impraticável traçar um gráfico de dispersão m-dimensional (onde m é o número de recursos) facilmente compreensível por seres humanos, portanto, os dados devem ser projetados em duas ou três dimensões.

Para fazer isso, o programa oferece duas soluções de projeção: a PCA (*Principal Component Analysis* ou Análise das Componentes Principais), que é um método confiável e bastante testado na literatura e o t-SNE (*t-distributed Stochastic Neighbor Embedding* ou Incorporação de Vizinho Estocástico com distribuição t), que é o estado da arte em projeções. Ambas as implementações são da biblioteca scikit-learn.

O PCA recebe a tabela multidimensional como entrada e, em seguida, encontra um novo conjunto de variáveis ortogonais (as *componentes principais*) que explicam a maior parte da variância nos dados, removendo colunas inter-correlacionadas além da realização de outros procedimentos matemáticos. Uma vez que é possível obter um número arbitrário (até a quantidade de dimensões que existe originalmente) de componentes principais, selecionar as duas mais importantes permite a construção de gráficos 2D significativos. (ABDI; WILLIAMS, 2010)

O outro método, t-SNE (MAATEN; HINTON, 2008) se difere da PCA em quatro pontos principais: é um método não-linear, probabilístico, considera apenas a vizinhança local (visto que tenta ao máximo manter elementos semelhantes próximos) e, como limitação, é um método transdutivo, visto que não permite a projeção de dados não existentes no conjunto de treinamento.

Como é possível observar na Figura 6, a não linearidade permite que o t-SNE entenda padrões mais complexos subjacentes aos dados (nesse caso, o conjunto Iris[8]) apresentados, gerando projeções que tendem a separar melhor elementos distintos.

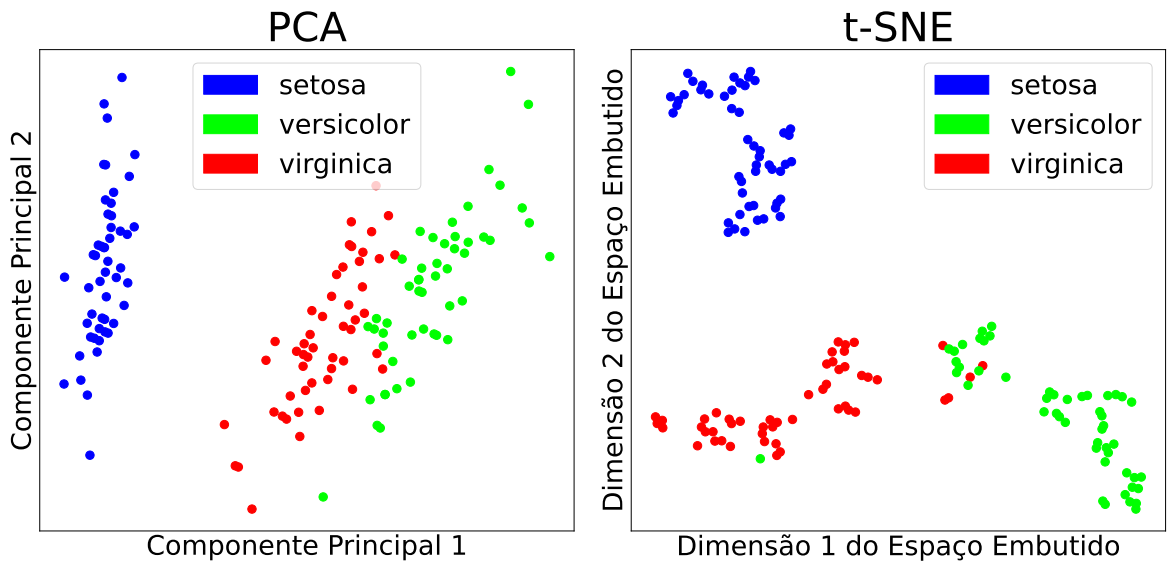
3.6 Saída do Programa

Após o programa ter executado todo o procedimento de agrupamento supervisionado, ele mostra as informações relevantes que podem ser posteriormente utilizadas para análises quantitativas ou qualitativas. A saída textual da caixa de ferramentas permite ao usuário saber:

- O melhor valor K de grupos encontrado para o K-Means

[8] https://archive.ics.uci.edu/ml/datasets/Iris

Figura 6 – Dataset Iris projetado em 2D através de PCA e t-SNE



Fonte: o autor

- As amostras representativas selecionadas do repositório
- A variância intra-grupo da coluna de propriedade
- Os melhores pesos encontrados pelo algoritmo de otimização
- Qual elemento é atribuído a qual grupo
- Explicação da razão de variância da PCA
- O tempo de execução

Além disso, se o programa for iniciado com a saída verbosa ativada, as informações sobre a variância intra-grupo atual e a do *Silhouette Score* são exibidas em um emulador de terminal enquanto o programa está sendo executado, fornecendo mais informações sobre o que está acontecendo o tempo todo.

Os gráficos gerados ao final do programa (que têm suas aplicações práticas discutidas em mais detalhes no Capítulo 4) são os seguintes:

1. Métricas de 4 medidas de qualidade de agrupamento
2. Gráfico de coordenadas paralelas
3. Gráfico de radar mostrando os pesos finais escolhidos pelo algoritmo de otimização
4. Gráfico de dispersão 2D do conjunto de dados não ponderados projetado em 2D através de PCA

5. Gráfico de dispersão 2D do conjunto de dados não ponderados projetado em 2D através de t-SNE
6. Gráfico de dispersão 2D do conjunto de dados ponderados projetado em 2D através de PCA
7. Gráfico de dispersão 2D do conjunto de dados ponderados projetado em 2D através de t-SNE
8. *Boxplots* das variâncias intra-cluster do conjunto de dados não ponderados
9. *Boxplots* das variâncias intra-cluster do conjunto de dados ponderados
10. Comparação dos *boxplots* das variâncias intra-cluster dos conjuntos de dados ponderados e não ponderados

O gráfico 1 contém 4 sub-gráficos cujos eixos X são os números de grupos e eixo Y os valores das medidas de qualidade para determinada configuração final com a quantidade de grupos dada pelo eixo X. No caso, os sub-gráficos mostram os valores para as medidas: *Elbow*, *Silhouette*, *Calinski-Harabasz* e *Davies-Bouldin*. A interpretação de cada medida é diferente: na *Elbow*, o cotovelo (se bem visível) mostra a melhor quantidade de grupos, na Silhouette e *Calinski-Harabasz*, quanto maior o valor, melhor a qualidade, enquanto na *Davies-Bouldin*, quanto menor o valor, melhor.

O gráfico 2 é outra forma de representar dados agrupados. A forma como os dados são representados permite verificar a propagação, ou não, dos grupos através das dimensões. Embora o gráfico de Coordenadas Paralelas possam ser considerado "sem perdas", pode acabar não sendo uma opção viável para repositórios de alta dimensão e com informações muito divergentes, devido a uma "superlotação" da figura, principalmente se houver muitos grupos. O gráfico em si é simples: as características estão no eixo x e seus valores correspondentes estão no eixo y. Linhas são traçadas ao longo dos meridianos das características para cada elemento e as cores mostram qual grupo está atribuído a um determinado elemento.

O gráfico 3 consiste em uma circunferência onde todos os recursos estão igualmente espaçados ao longo do perímetro, então, dentro dele existem várias circunferências concêntricas com números de ponto flutuante anexados. Esses números formam a escala de pesos finais. Assim, cada coluna do conjunto de dados ponderados pode ser representada como um par $c = (\theta, r)$, onde o ângulo θ está relacionado a qual recurso está sendo representado e a distância do centro (ou raio) r representa o seu peso final correspondente.

Os gráficos 4, 5, 6 e 7 são autoexplicativos considerando a seção anterior, mas é importante ressaltar que uma projeção é apenas uma representação dos dados, ou seja, pode não capturar todas os detalhes do conjunto de dados original, uma vez que os dados (caso sejam de dimensão superior) são espremidos em duas dimensões.

Os gráficos 8 e 9 mostram nos eixos X os grupos e nos eixos Y o *boxplot* para as propriedades. São ordenados de acordo com a mediana de cada grupo para esse critério.

Finalmente, o gráfico 10 mostra a mesma coisa que os gráficos 8 e 9, mas com os grupos intercalados e comparados (colocados ao lado um do outro) de acordo com a média da propriedade.

3.7 Repositórios

Para o desenvolvimento deste TCC, são utilizados dados produzidos pelo grupo QT-Nano e outros grupos de pesquisa da área. Os seguintes repositórios de dados estão sendo considerados:

- $CeZrO_4$ e nanoligas de PtTM: Este repositório foi usado anteriormente para investigar as propriedades energéticas, estruturais e eletrônicas de nanoclusters mistos de Cério e Zircônio, que são muito importantes na nanocatálise (FELíCIO-SOUSA et al., 2019). Consiste em um conjunto tratado de óxidos mistos $Ce_{14}ZrO_{30}$ a $CeZr_{14}O_{30}$ com 1646 estruturas geradas a partir da variação na quantidade de átomos Ce e Zr, explorando diferentes padrões de substituição, como ter mais Ce no centro da partícula e Zr na superfície ou regiões mais ricas em Ce em um hemisfério e Zr no outro.

- Nanoclusters de Cu_n: Este conjunto foi construído para ampliar o conhecimento sobre a morfologia de nanoclusters metálicos. As geometrias obtidas por meio da ferramenta interna Revised Basin-Hopping Monte Carlo (rBHMC) (RONDINA; DA SILVA, 2013) para Cu_{55} nanoclusters, onde as interações são descritas via ReaxFF (van Duin et al., 2001) com parâmetros abordados em (NIELSON et al., 2005). Em dez execuções de rBHMC, um total de 1048 geometrias Cu_{55} foram geradas, todas correspondendo a mínimos locais na superfície de energia potencial desses sistemas.

- Nanoligas Core-Shell baseadas em Pt de 55 átomos: Consiste em nanooligas de metal de transição baseadas em 330 Pt usando descritores físico-químicos derivados da adsorção e ativação de CO_2 em nanoclusters de 55 átomos, a saber, Pt_nTM_{55-n}, onde $n = 0, 13, 42, 55$ e o TM = Fe, Co, Ni, Cu, Ru, Rh, Pd, Ag, Os, Ir e Au. O projeto que utilizou este repositório foi uma triagem *ab initio* baseada em cálculos da Teoria do Funcional da Densidade das estruturas nele contidas. É uma tarefa significativa entender melhor a ativação de CO_2, que é relevante para a conversão de CO_2 em metanol, ácido fórmico, metano e outros compostos (MENDES et al., 2021).

- Desidrogenação de CH_4 em clusters de TM_{13}: Todas as 770 estruturas neste conjunto de dados são formadas combinando as possibilidades para *n* em CH_n + (4-*n*) H e as possibilidades para aglomerados de metais de transição de 13 átomos de Fe, Co, Ni e Cu,

explorando diferentes configurações espaciais para os átomos. O Repositório foi usado anteriormente para investigar a desidrogenação de CH_4 nesses clusters por meio de cálculos da Teoria do Funcional da Densidade e outros métodos, como um esforço para identificar as principais características físico-químicas para modular CH_4 desidrogenação particularmente em regime de nanoescala, onde novo efeitos podem promover a quebra da ligação C-H (ANDRIANI; MUCELINI; Da Silva, 2020).

- QM9: Repositório com 134 mil pequenas moléculas orgânicas estáveis compostas de Carbono, Hidrogênio, Oxigênio, Nitrogênio e Flúor do universo químico GDB-17 com 166 bilhões de moléculas orgânicas. Contém geometrias mínimas em energia, frequências harmônicas correspondente, momentos dipolo, polarizabilidade, com as energias, entalpias e energias livres de atomização. Todas as propriedades foram calculadas no nível B3LYP/6-31G(2d,p) de química quântica. Com isso, tem-se que este conjunto de dados fornece propriedades químicas quânticas para um espaço químico relevante, consistende e compreensivo de moléculas orgânicas pequenas. (RAMAKRISHNAN et al., 2014; RUDDIGKEIT et al., 2012)

3.8 Análise dos Resultados

Considerando os respositórios acima, a ferramenta desenvolvida nesse TCC possui dois objetivos principais: a) realizar a tarefa de agrupamento de dados enviesada pela informação externa (target) fornecida pelo especialista; e b) a partir dos agrupamentos formados, devolver um conjunto de exemplos representativos que possam caracterizar a região (agrupamento).

Para análise, os diversos gráficos gerados, assim como medidas de qualidade de agrupamento e soma da variância intra-cluster são considerados. Além dessas informações, também é levado em conta um regressor linear, que é detalhado no Capítulo 4, criado para medir a qualidade das amostras representativas selecionadas.

4 Resultados

Este capítulo apresenta os principais resultados obtidos com os cinco conjuntos introduzidos na Seção 3.7. Além dos gráficos gerados pelo programa, mais dois gráficos por conjunto de dados foram produzidos separadamente. Esses gráficos mostram o resultado da regressão linear feita utilizando os exemplos representativos como conjunto de treino e os pontos reais como conjunto de teste. Especificamente, o eixo X representa a primeira componente principal (PCA-1) e o eixo Y o indica o valor da propriedade aprendida pelo preditor.

É um fato conhecido na ciência de dados que quanto maior a qualidade do conjunto de treino, maiores são as chances de gerar um bom preditor. Levando esse conhecimento para a análise proposta por esses gráficos, pode-se extrair que quanto melhor a qualidade dos exemplos representativos (conjunto de treino), melhor será o desempenho do regressor linear, em outras palavras, erros quadráticos médios (MSE) menores indicam representantes melhores.

Vale ressaltar que somente no conjunto $CeZrO_4$ e Nanoligas de PtTM são mostrados e comentados todos os gráficos. Para os demais conjuntos, os gráficos estão nos Apêndices, exceto para o caso onde o gráfico é tão relevante que merece ocupar o espaço desta seção, e os comentários não são feitos para cada um dos gráficos, mas sim para os mais relevantes.

4.1 Protocolo Experimental

A Tabela 1 mostra um sumário de cada um dos repositórios utilizados. **CeZrO4_exc** representa o conjunto de dados $CeZrO_4$ e nanoligas de PtTM, **Cu_n** é o repositório Nanoclusters de Cu_n, **PtTM_exc** representa Nanoligas Core-Shell baseadas em Pt de 55 átomos, **CHnTM_H4** é o conjunto Desidrogenação de CH_4 em clusters de TM_{13} com somente as moléculas de 4 hidrogênios e **QM9** é o conjunto QM9 com somente 5000 das moléculas de 18 átomos. Nas propriedades, "E" significa Energia, sendo que **Energia de Excesso** é a energia liberada quando há a ligação de átomos para formar novas ligações mais estáveis, **Energia Total** é a soma das energias potencial e cinética no sistema, **Energia de Adsorção** é a energia decrescente enquanto dois materiais são combinados sob o processo de adsorção e **Energia Livre à 298K** é a quantidade máxima de trabalho de não expansão que pode ser extraída de um sistema termodinamicamente fechado com temperatura constante de 298K. No repositório CHnTM_H4 os atributos escolhidos foram as distâncias dos átomos de carbono e hidrogênio até os respectivos metais de transição mais próximos, abreviados por "Dist dos C e H até TM mais próximo".

A Tabela 2 mostra as configurações utilizadas para a realização de cada um dos experimentos. **Semente do RNG** (gerador de números aleatórios) é o que define a sequência de

Tabela 1 – Sumário de cada um dos conjuntos de dados utilizados

Conjunto	# Exemplos	Atributos	Propriedade
CeZrO4_exc	1646	Autovalores da Matriz de Coulomb	E Excesso
Cu_n	1048	Autovalores da Matriz de Coulomb	E Total
PtTM_exc	330	Autovalores da Matriz de Coulomb	E Excesso
CHnTM_H4	398	Dist dos C e H até TM mais próximo	E Adsorção
QM9_18	5000	Autovalores da Matriz de Coulomb	E Livre 298K

Fonte: o autor

valores pseudo-aleatórios que são utilizados pelos algoritmos do programa, como por exemplo o K-Means e o Basin-Hopping. **Seleção de K** é o modo com o qual o valor de K está sendo escolhido (automático implica na otimização para sugerir o melhor K, enquanto manual é totalmente controlado pelo usuário). **Medida de qualidade** define qual a medida de qualidade a ser utilizada para a otimização do valor de K. **Exato/máximo K** mostra a quantidade máxima de K a ser encontrada pelo otimizador (caso esteja ativado) ou o valor exato de K, no caso do otimizador não estar em uso. **Propriedade** mostra qual a propriedade que está sendo utilizada para fazer a supervisão do sistema. **Máximo de iterações** indica, como o nome diz, o máximo de iterações que serão rodadas pelo Basin-Hopping. **Maior passo** aponta qual a maior distância possível que o Basin-Hopping pode se deslocar, numa iteração, na vizinhança de pontos. **Temperatura inicial** é o valor de temperatura com o qual Basin-Hopping inicia e, finalmente, **Paciência** indica o máximo de iterações que o Basin-Hopping pode rodar sem melhorar o valor a ser otimizado, se esse valor for ultrapassado, a otimização termina. Assim como explicado na descrição da Tabela 1, as diferentes energias estão abreviadas.

Tabela 2 – Configurações utilizadas para a realização de cada um dos experimentos

Parâmetro	CeZrO4_exc	Cu_n	PtTM_exc	CHnTM_H4	QM9_18
Semente do RNG	321	321	321	321	321
Seleção de K	Auto	Auto	Auto	Auto	Manual
Medida de qualidade	Silhueta	Silhueta	Silhueta	Silhueta	Silhueta
Exato/máximo K	20	50	20	20	6
Propriedade	E Excesso	E Total	E Excesso	E Adsorção	E Livre 298K
Máximo de iterações	300	500	500	300	500
Maior passo	1	1	1	1	1
Temperatura inicial	1000	1000	1000	1000	1000
Paciência	100	250	250	100	250

Fonte: o autor

4.2 $CeZrO_4$ e Nanoligas de PtTM

A Tabela 3 apresenta os principais resultados para o conjunto $CeZrO_4$ e Nanoligas de PtTM. Especificamente, apresenta-se o número de agrupamentos, soma das variâncias intra-

grupo antes e depois da otimização e, por fim, os MSEs dos regressores treinados com os exemplos selecionados pelas abordagens não supervisionada e supervisionada.

Tabela 3 – Sumário dos resultados para o repositório $CeZrO_4$ e Nanoligas de PtTM

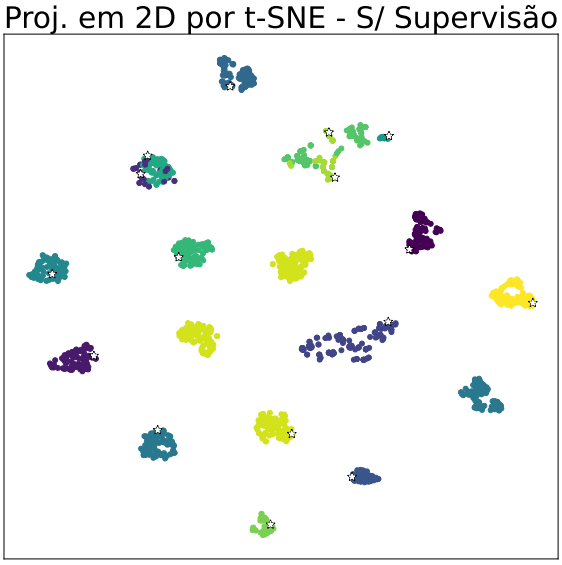
Descrição	Valor
Melhor número K de grupos/representantes encontrado	16
Soma das variâncias intra-grupo antes da otimização	17.028679
Soma das Variâncias intra-grupo depois da otimização	16.835181
MSE da Regressão Linear antes da otimização	24.032095
MSE da Regressão Linear depois da otimização	14.270303

Fonte: o autor

Figura 7 – Projeção 2D (PCA) do conjunto CeZrO4 não supervisionado. Os eixos X e Y representam as duas componentes principais do conjunto de dados

Fonte: o autor

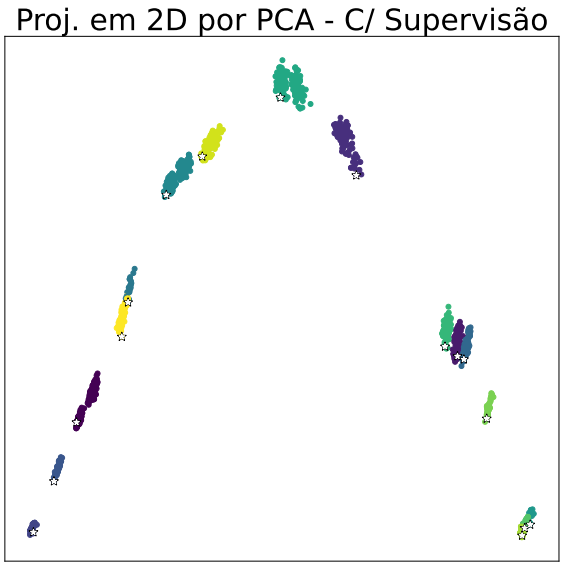
Figura 8 – Projeção 2D (t-SNE) do conjunto CeZrO4 não supervisionado. Os eixos X e Y representam as dimensões do espaço embutido



Fonte: o autor

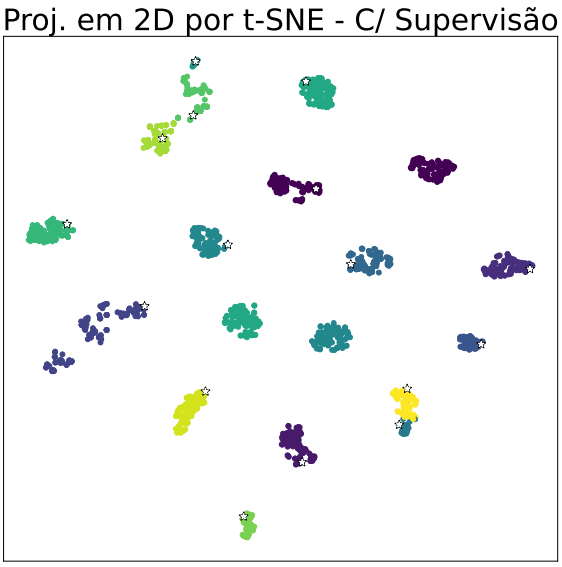
Observando as Figuras 7 e 9, é possível notar uma reestruturação no espaço (após otimização) de forma que alguns grupos que antes estavam separados na projeção, acabam ficando mais unidos. Isso é particularmente notado no caso do grupo verde claro no gráfico pré-otimização, que se encontra quebrado em 3 partes distintas, enquanto no gráfico pós-otimização não existe nenhum grupo quebrado em 3 partes na projeção. Também é possível observar a distorção nas componentes principais, uma vez que no primeiro caso o conjunto lembra uma reta descendente com um degrau, enquanto no segundo caso parece muito uma parábola. Infelizmente não é fácil extrair qualquer conclusão significativo das Figuras 8 e 10.

Figura 9 – Projeção 2D (PCA) do conjunto CeZrO4 supervisionado. Os eixos X e Y representam as dimensões do espaço embutido



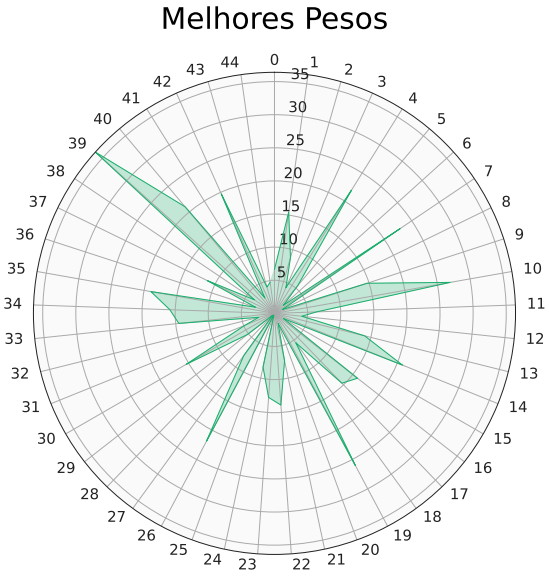
Fonte: o autor

Figura 10 – Projeção 2D (t-SNE) do conjunto CeZrO4 supervisionado. Os eixos X e Y representam as dimensões do espaço embutido



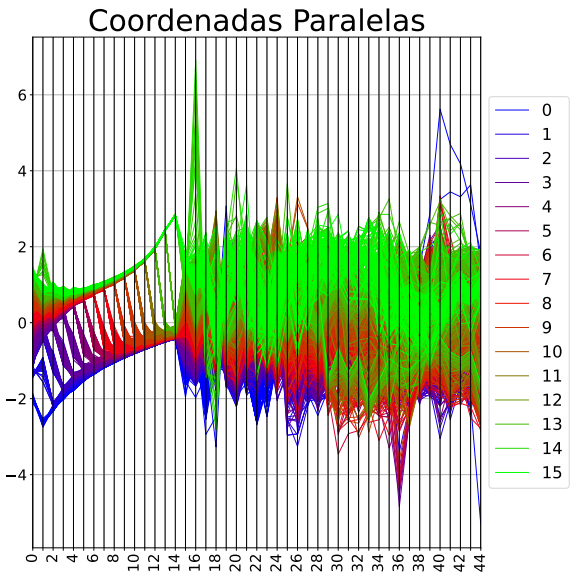
Fonte: o autor

Figura 11 – Gráfico de radar do conjunto CeZrO4 supervisionado. Os valores externos representam cada um dos autovalores e os internos os pesos atribuídos a eles



Fonte: o autor

Figura 12 – Coordenadas paralelas do conjunto CeZrO4. O eixo X representa cada um dos autovalores, o eixo Y seus valores e as cores das linhas são os grupos aos quais os elementos pertencem.



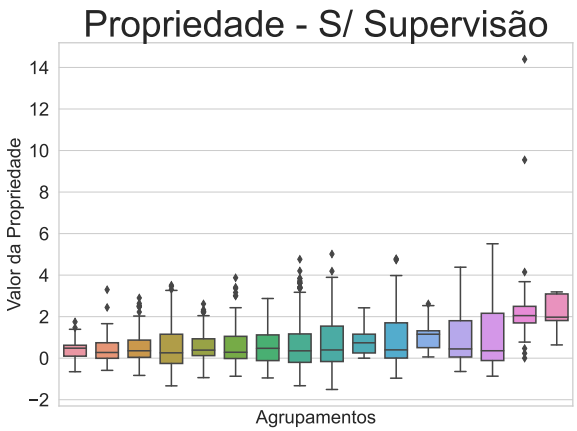
Fonte: o autor

É perceptível o efeito da otimização no gráfico de radar, Figura 11, pois os pesos de cada característica estão bem definidos e claramente alguns autovalores da Matriz de Coulomb têm mais importância que outros, por exemplo: o autovalor 39 é o mais relevante para o viés de

acordo com a propriedade selecionada, já o autovalor 38 é o caso completamente exposto.

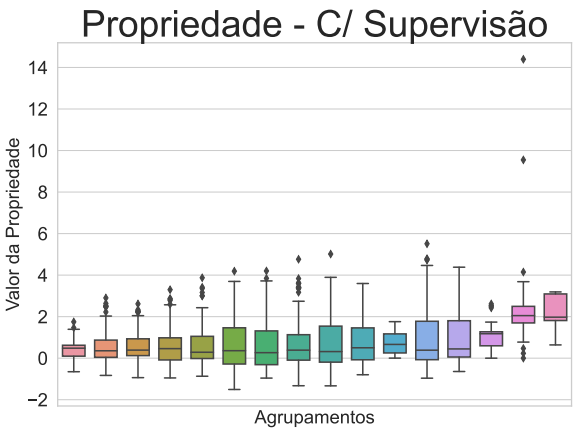
No gráfico de coordenadas paralelas, Figura 12, é possível observar uma interessante estruturação dos valores dos elementos de cada grupo, especialmente considerando os primeiros autovalores (começo do eixo X). É observável que a maneira como os grupos foram separados casa muito bem com os autovalores, implicando que esses tem um impacto substancial na formação dos grupos.

Figura 13 – Boxplots do conjunto CeZrO4 não supervisionado. O eixo X representa cada um dos grupos formados e o eixo Y mostra os valores para os boxplots



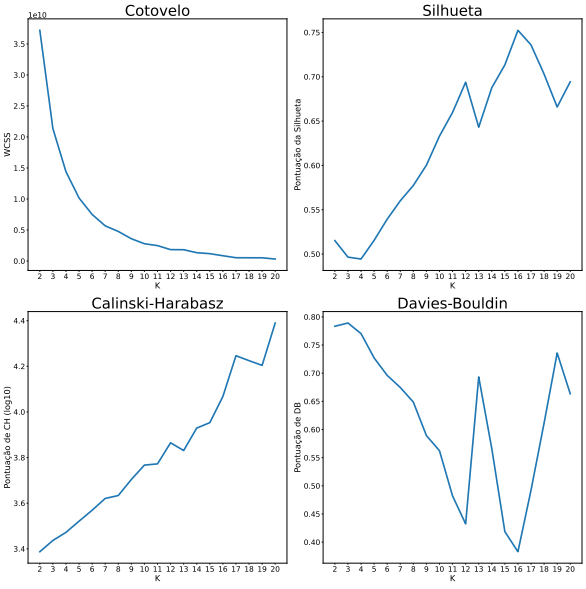
Fonte: o autor

Figura 14 – Boxplots do conjunto CeZrO4 supervisionado. O eixo X representa cada um dos grupos formados e o eixo Y mostra os valores para os boxplots



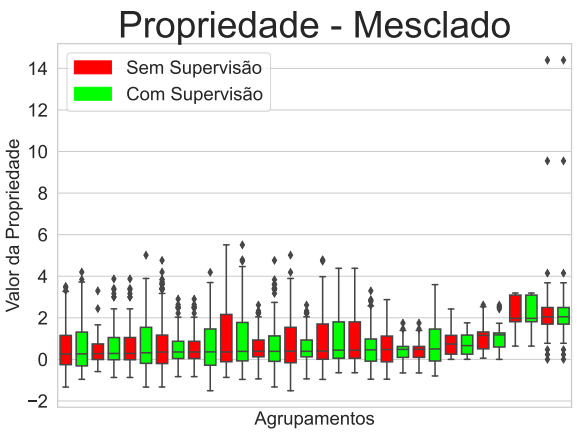
Fonte: o autor

Figura 15 – Métricas de qualidade de agrupamento do conjunto CeZrO4. Os eixos X são os valores de K e os eixos Y os valores das métricas de qualidade de agrupamento para os determinados valores de K



Fonte: o autor

Figura 16 – Boxplots do conjunto CeZrO4 não supervisionado e supervisionado intercalados. O eixo X representa cada um dos grupos formados e o eixo Y mostra os valores para os boxplots

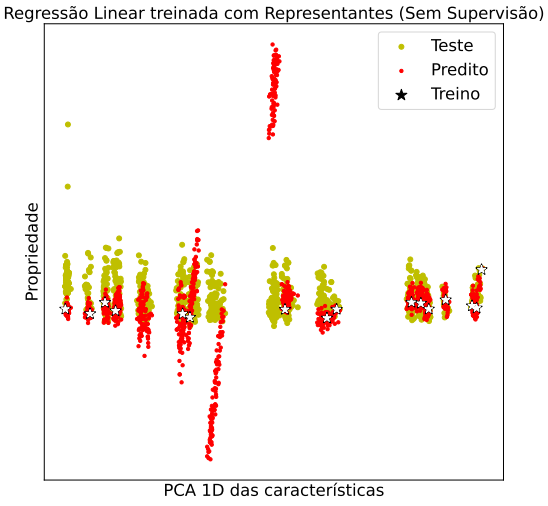


Fonte: o autor

Já nas Figuras 13, 14 e 16, existe, no geral, a redução das variâncias intra-grupo após a otimização ter sido rodada, ou seja, obtendo a altura média dos boxplots antes e depois da otimização, o segundo caso resulta num valor mais baixo. Esse fato é mais facilmente observado através dos dados textuais de variância intra-grupo onde é descrita uma redução de 1.14%.

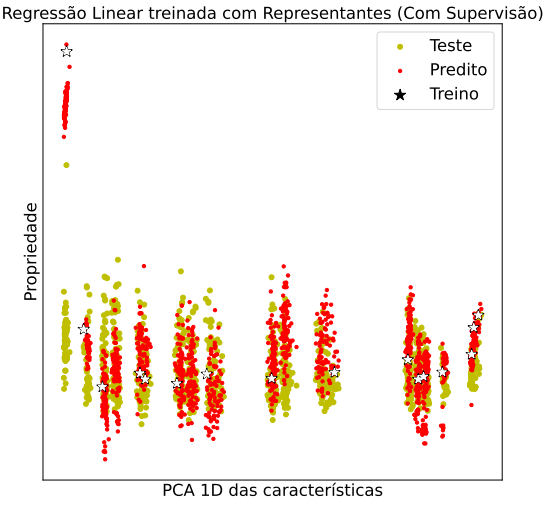
Olhando para o gráfico das métricas de qualidade de agrupamento (Figura 15), é notável que ambas as medidas *Silhouette* e *Davies-Bouldin* elegem 16 como a melhor quantidade de grupos, de acordo com a configuração final de agrupamentos para cada quantidade de grupos a serem gerados.

Figura 17 – Regressão Linear do conjunto CeZrO4 não supervisionado. O eixo X mostra os valores da projeção do espaço de características em 1 dimensão, enquanto o eixo Y revela os valores para cada uma das amostras. Os pontos amarelos mostram os valores reais das propriedades, os vermelhos mostram os preditos e as estrelas correspondem ao conjunto de treino, que são os exemplos representativos selecionados



Fonte: o autor

Figura 18 – Regressão Linear do conjunto CeZrO4 supervisionado. O eixo X mostra os valores da projeção do espaço de características em 1 dimensão, enquanto o eixo Y revela os valores para cada uma das amostras. Os pontos amarelos mostram os valores reais das propriedades, os vermelhos mostram os preditos e as estrelas correspondem ao conjunto de treino, que são os exemplos representativos selecionados



Fonte: o autor

Finalmente, vale notar o quão melhor a versão ponderada (ou supervisionada) foi na escolha de representantes, de acordo com os dados textuais: uma redução de mais de 40% no MSE. Isso fica bastante visível nas figuras 17 e 18, pois a versão supervisionada corresponde muito melhor aos valores reais.

4.3 Nanoclusters de Cu_n

Abaixo, a Tabela 4 revela os resultados para o conjunto Nanoclusters de Cu_n. Apresenta-se o número de agrupamentos, soma das variâncias intra-grupo antes e depois da otimização e

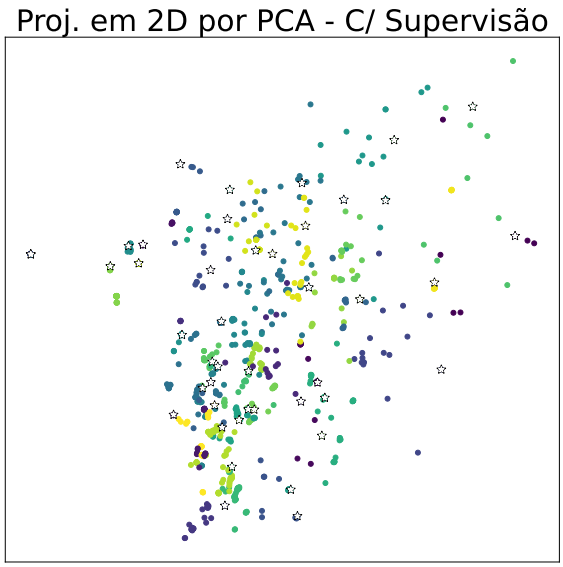
os MSEs do regressores treinados com os exemplos selecionados pelas abordagens não supervisionada e supervisionada.

Tabela 4 – Sumário dos resultados para o conjunto Nanoclusters de Cu_n

Descrição	Valor
Melhor número K de grupos/representantes encontrado	44
Soma das variâncias intra-grupo antes da otimização	923.994675
Soma das Variâncias intra-grupo depois da otimização	888.016251
MSE da Regressão Linear antes da otimização	53.023410
MSE da Regressão Linear depois da otimização	98.202393

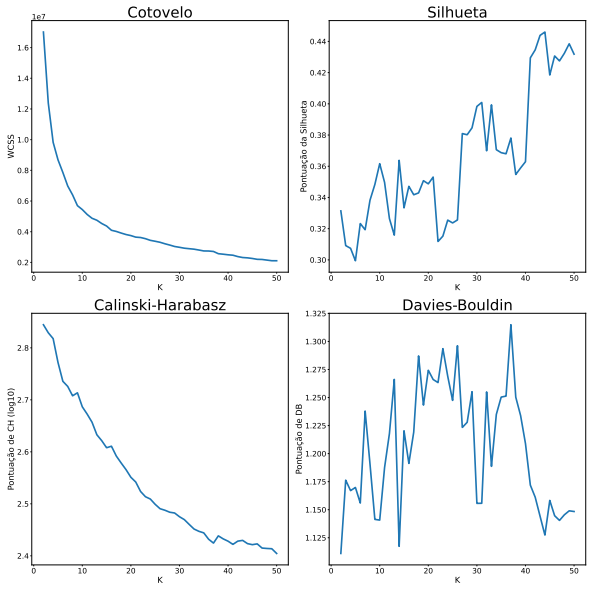
Fonte: o autor

Figura 19 – Projeção 2D (PCA) do conjunto Cu_n supervisionado



Fonte: o autor

Figura 20 – Métricas de qualidade de agrupamento do conjunto Cu_n



Fonte: o autor

Esse conjunto de dados é um "caso ruim" para o programa desenvolvido, pois os dados são extremamente difíceis de agrupar de maneira adequada, devido a não existência de regiões de baixa densidade entre os grupos. O resultado dessa dificuldade se torna aparente pelo grande embaralhamento de grupos nas projeções (como por exemplo na Figura 19). Isso faz com que as métricas de qualidade tendam sempre a preferir quantidades cada vez maiores de grupos, novamente visível pelas projeções e também na Figura 20.

Apesar de ter sido encontrado o valor $K = 44$ por ambos *Silhouette* e *Davies-Bouldin*, esse é um valor não significativo, pois é praticamente o maior valor possível, dada a limitação de $K = 50$, e não há indicativos de um pico claro em nenhuma das duas medidas de qualidade concordantes, ou seja, o valor tende a mudar com K máximo maior.

O resultado disso é que, apesar da soma da variância intra-grupos ter sido reduzida em 3.89% após a otimização, essa redução de variância não se traduz em exemplos representativos melhores, como é notável pelo aumento em 85.20% do MSE da regressão linear.

4.4 Nanoligas Core-Shell baseadas em Pt de 55 átomos

Agora, tem-se os resultados referentes ao repositório Nanoligas Core-Shell baseadas em Pt de 55 átomos resumidos na Tabela 5. Nela são mostrados o número de agrupamentos, soma das variâncias intra-grupo antes e depois da otimização, além dos MSEs dos regressores lineares treinados com os exemplos selecionados pelas abordagens não supervisionada e supervisionada.

Tabela 5 – Sumário dos resultados para o conjunto Nanoligas Core-Shell baseadas em Pt de 55 átomos

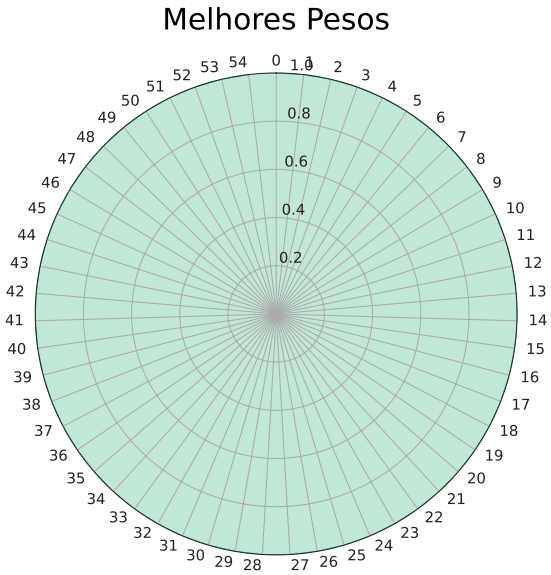
Descrição	Valor
Melhor número K de grupos/representantes encontrado	3
Soma das variâncias intra-grupo antes da otimização	0.027594
Soma das Variâncias intra-grupo depois da otimização	0.027594
MSE da Regressão Linear antes da otimização	0.009169
MSE da Regressão Linear depois da otimização	0.009169

Fonte: o autor

Figura 21 – Projeção 2D (PCA) do conjunto PtTM supervisionado

Fonte: o autor

Figura 22 – Gráfico de radar do conjunto PtTM supervisionado. Todos os pesos são igual a 1, ou seja, as versões não supervisionada e supervisionada do agrupamento são idênticas



Fonte: o autor

Enquanto o conjunto da Seção 4.3 é um “caso ruim” para o programa, este conjunto é bom. A otimização de achar o melhor valor de K tem duas boas opções: uma apontada pela

silhueta e *Davies-Bouldin*, que é $K = 3$ e outra apontada por *Calinski-Harabasz* e método do cotovelo, que é $K = 5$ (de acordo com a Figura 41 presente no Apêndice). Tal observação é perceptível na projeção bidimensional por PCA (vide Figura 21) e t-SNE, embora essa segunda projeção penda completamente para 5 grupos.

Uma implicação de um conjunto simples de ter seus elementos agrupados é que a supervisão torna-se desnecessária ou redundante. A Figura 22 revela que após a otimização, todas as características tem exatamente a mesma relevância, ou seja, o algoritmo de otimização não encontrou nada melhor do que os valores iniciais. Tal fenômeno também pode ser observado através das projeções: tanto PCA como t-SNE não sofrem alterações quando é feita a comparação entre pré-otimização e pós-otimização.

Portanto, neste repositório, a única utilidade do programa foi encontrar o melhor valor para K, de acordo com a medida de qualidade de agrupamento escolhida (aqui, a silhueta.)

4.5 Desidrogenação de CH_4 em clusters TM_{13}

A Tabela 6 expõe os resultados referentes ao conjunto Desidrogenação de CH_4 em clusters TM_{13}. Nela são mostrados o número de agrupamentos, soma das variâncias intra-grupo antes e depois da otimização e, finalmente, os MSEs dos regressores lineares treinado com os exemplos selecionados pelas abordagens não supervisionada e supervisionada.

Tabela 6 – Sumário dos resultados para o conjunto Desidrogenação de CH_4 em clusters TM_{13}

Descrição	Valor
Melhor número K de grupos/representantes encontrado	2
Soma das variâncias intra-grupo antes da otimização	28.520969
Soma das Variâncias intra-grupo depois da otimização	27.126402
MSE da Regressão Linear antes da otimização	109.911412
MSE da Regressão Linear depois da otimização	18.337151

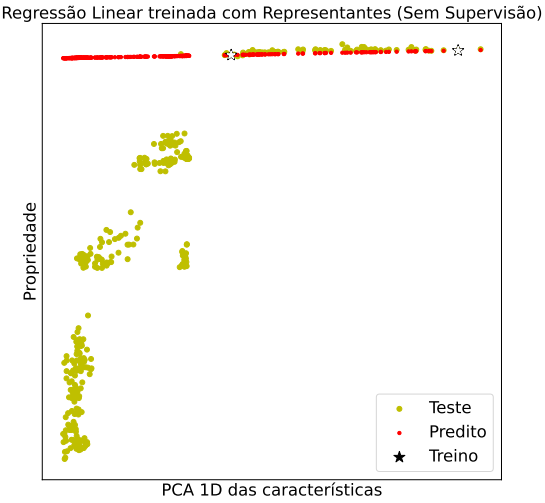
Fonte: o autor

Antes de comentar sobre este repositório propriamente dito, vale explicar que originalmente ele é composto por moléculas com 0, 1, 2, 3 e 4 Hidrogênios a serem adsorvidas nos metais de transição. Para fins de evitar estender demais a seção, somente as de 4 Hidrogênios estão sendo consideradas, pois são as que geram os resultados mais interessantes. De acordo com os especialistas, não faz sentido tratar todas as variações de CH como uma coisa só, então foi necessário rodar o programa para cada uma das variações individualmente.

Uma outra observação relevante quanto a este conjunto é que não estão sendo utilizados os autovalores da Matriz de Coulomb como características e sim a sugestão do especialista, que seria a distância dos átomos de Carbono e Hidrogênio até o metal de transição mais próximo.

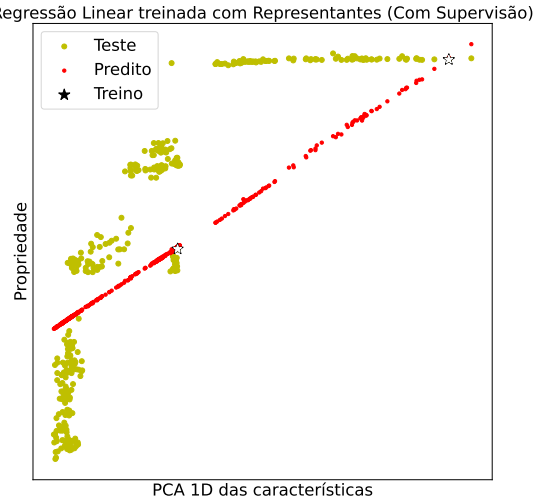
Nesse conjunto de dados acontece uma boa transformação no espaço, especialmente notável nas projeções t-SNE (Figuras 46 e 48), porém a alteração na variância não é tão cho-

Figura 23 – Regressão Linear do conjunto CH_nTM não supervisionado



Fonte: o autor

Figura 24 – Regressão Linear do conjunto CH_nTM supervisionado



Fonte: o autor

cante, tendo reduzido modestos 4.89%. Além disso, novamente existe uma disputa entre dois grupos de medidas de qualidade de agrupamento: *Silhouette* e *Davies-Bouldin* votam $K = 2$, enquanto *Calinski-Harabasz* e *Elbow* votam $K = 8$ e $K \approx 8$, respectivamente, conforme é possível observar na Figura 53.

O mais interessante aqui é a escolha de exemplos representativos pré-otimização e pós-otimização. Na Figura 23 os valores preditos pelo regressor formam praticamente uma reta na parte superior projeção, o que não é interessante, pois existe uma grande massa de pontos que se encontram na parte inferior, causando um erro quadrático médio alto. Isso é melhorado drasticamente após a otimização, onde um representante se mantém exatamente o mesmo, mas o outro acaba ficando muito mais próximo a massa de pontos na projeção, reduzindo o MSE em 83.32%.

4.6 QM9

Como último repositório analisado, tem-se o QM9. A Tabela 7 expõe os principais resultados relacionados ao conjunto, contendo o número de agrupamentos, soma das variâncias intra-grupo antes e depois da otimização e, por último, os MSEs dos regressores lineares treinado com os exemplos selecionados pelas abordagens não supervisionada e supervisionada.

No conjunto QM9, a quantidade de grupos/representantes K foi definida como 6 manualmente, a partir de inspeção visual das projeções PCA (Figuras 55 e 57) e t-SNE (Figuras 56 e 58). A definição manual é suportada pela ferramenta e pode ser utilizada pelo especialista caso julgue que a automática não está fazendo um bom trabalho em identificar a quantidade ideal de

Tabela 7 – Sumário dos resultados para o conjunto Desidrogenação de CH_4 em clusters TM_{13}

Descrição	Valor
Número K de grupos/representantes	6
Soma das variâncias intra-grupo antes da otimização	27.995522
Soma das Variâncias intra-grupo depois da otimização	22.388382
MSE da Regressão Linear antes da otimização	94.712168
MSE da Regressão Linear depois da otimização	26.794895

Fonte: o autor

grupos, independentemente da medida de qualidade de agrupamento.

É importante mencionar que foram utilizados somente 5000 elementos (todos possuindo 18 átomos) do total de 134 mil elementos disponíveis no repositório, com a finalidade de reduzir o custo computacional do experimento. Dessa maneira, é preciso levar em consideração que não é uma amostragem representativa do conjunto completo. Ainda assim, não foi uma escolha enviesada: os 5000 elementos foram selecionados aleatoriamente e o motivo de utilizar moléculas com 18 átomos é a grande frequência com que ocorrem no conjunto de dados.

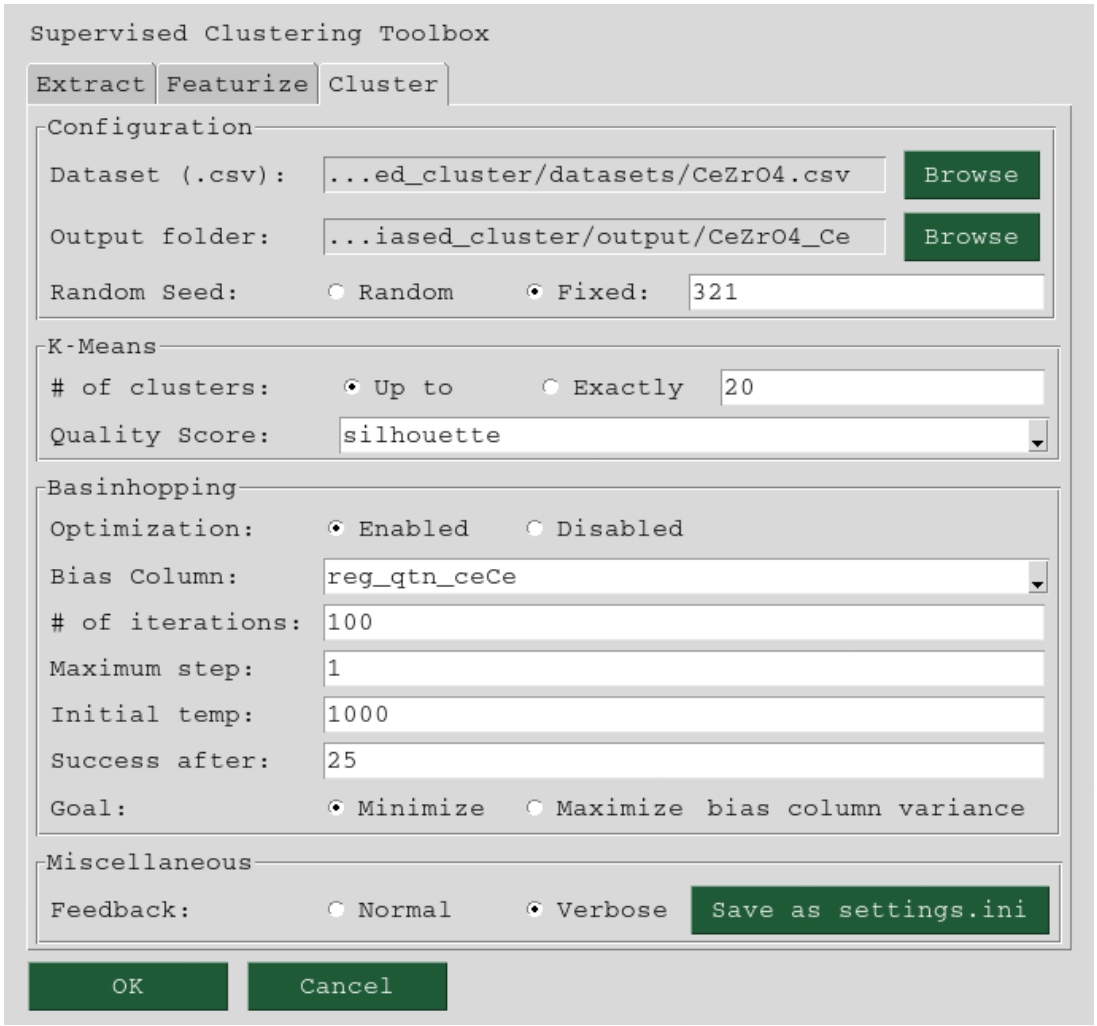
Os resultados para esse conjunto são bastante interessantes: há uma redução de 20.03% na soma das variâncias intra-grupo e ótimos 71.71% no erro quadrático médio da regressão linear.

4.7 Interface gráfica (GUI) para o programa

Uma interface gráfica (mostrada na Figura 25 foi desenvolvida para a ferramenta, de forma a tornar a utilização mais simples e acessível. Através dela é possível executar localmente o programa ou gerar arquivos de configuração que podem ser posteriormente utilizados para execução em um computador de alto desempenho.

Até o momento, somente a aba “Cluster”, que é a que lida com todos os algoritmos descritos até agora, está implementada, pois as demais estão fora do escopo desse TCC e podem ser implementadas em projetos futuros.

Figura 25 – Interface gráfica (GUI) da ferramenta



Fonte: o autor

5 Conclusão

Nesse trabalho foi desenvolvida uma ferramenta que implementa um algoritmo de agrupamento com supervisão que fornece a base para a seleção de moléculas representativas de conjuntos de dados químicos, com o objetivo de reduzir a quantidade de cálculos custosos, como DFT, a serem realizados. A parte fundamental do programa é composta por dois algoritmos principais: o algoritmo de agrupamento K-Means e o de otimização para busca global Basin-Hopping. O K-Means, por si só, já é capaz de oferecer os representantes, porém sob supervisão do Basin-Hopping é capaz de fornecer melhores opções.

Os resultados obtidos revelam que o método é eficaz, pois, através de análises a partir de gráficos e valores numéricos (especificamente as somas das variâncias intra-grupo e erros quadráticos médios das regressões lineares) observou-se substanciais melhoras na escolha de moléculas representativas. Isso é especialmente interessante levando em conta que os conjuntos de dados possuem estruturas bem distintas, revelando que o algoritmo é robusto e versátil, em outras palavras, consegue se sair bem em ambientes diversos.

A caixa de ferramentas, fruto da implementação do algoritmo de agrupamento supervisionado, é bastante acessível, que é uma qualidade bastante relevante considerando que o público-alvo não são pessoas da computação, mas sim pessoas da área de física, química e ciência dos materiais. Possui interface gráfica para utilização em computadores com ambientes gráficos e também interface por linha de comando, para utilização em computadores remotos de alto desempenho. A configuração do algoritmo pode ser feita através de um arquivo de configuração ou da interface gráfica. Vale mencionar que o programa faz uso de múltiplos núcleos de processamento e é construído em cima de uma fundação sólida, composta de bibliotecas constantemente atualizadas pelos seus mantenedores.

Como projetos ou avanços futuros relacionados a este trabalho de conclusão de curso tem-se a análise dos resultados de maneira qualitativa, por especialistas no domínio, além de quantitativa. Pode-se mencionar também a finalização da caixa de ferramentas, que abrange a implementação da aba “Extract”, que lida com extração de características a partir do *output* de diferentes programas utilizados recorrentemente em ciência dos materiais, como FHI-Aims, LAMMPS e VASP, e a aba “Featurize”, que fornece opções para adequar os dados obtidos através da extração de características aos modelos de aprendizado de máquina utilizados no programa. Por fim, pode-se estender o sistema para o cenário semi-supervisionado, no qual apenas uma fração do conjunto de dados terá a informação externa associada para enviesamento dos resultados.

Referências

ABDI, H.; WILLIAMS, L. J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, p. 433–459, 2010. Citado na página 51.

AMEN, R.; VOMACKA, P. Case-based reasoning as a tool for materials selection. *Materials Design*, v. 22, p. 353–358, 08 2001. Citado na página 29.

AMOIRALIS, E.; GEORGILAKIS, P.; GIOULEKAS, A. An artificial neural network for the selection of winding material in power transformers. In: . [S.l.: s.n.], 2006. p. 465–468. ISBN 978-3-540-34117-8. Citado na página 29.

ANDERBERG, M. R. The broad view of cluster analysis. *Cluster analysis for applications*, Elsevier, p. 1–9, 1973. Citado na página 37.

ANDERSON, E.; VEITH, G.; WEININGER, D. *Smiles: A line notation and computerized interpreter for chemical structures*. [S.l.], 1987. Citado na página 32.

ANDRIANI, K. F.; MUCELINI, J.; Da Silva, J. L. F. Methane dehydrogenation on 3d 13-atom transition-metal clusters: A density functional theory investigation combined with spearman rank correlation analysis. *Fuel*, v. 275, p. 117790, 2020. ISSN 0016-2361. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016236120307857>. Citado na página 55.

ANKERST, M. et al. Optics: Ordering points to identify the clustering structure. In: . [S.l.]: ACM Press, 1999. p. 49–60. Citado na página 38.

ASHBY, M. et al. Selection strategies for materials and processes. *Materials Design*, v. 25, n. 1, p. 51–67, 2004. ISSN 0261-3069. Disponível em: <https://www.sciencedirect.com-/science/article/pii/S0261306903001596>. Citado na página 29.

ASHBY, M. F.; CEBON, D. Materials selection in mechanical design. *J. Phys. IV France*, v. 03, p. C7–1–C7–9, 1993. Disponível em: <https://doi.org/10.1051/jp4:1993701>. Citado na página 29.

BADE, K.; NURNBERGER, A. Personalized hierarchical clustering. In: *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*. [S.l.: s.n.], 2006. p. 181–187. Citado na página 45.

BAIR, E. Semi-supervised clustering methods. *WIREs Computational Statistics*, v. 5, n. 5, p. 349–361, 2013. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics-.1270>. Citado 2 vezes nas páginas 38 e 44.

BALAKRISHNA, A. et al. Computer aided material selection processes in concurrent engineering using neural networks. v. 88, p. 20–23, 10 2007. Citado na página 29.

BALL, G. H.; HALL, D. J. *ISODATA, a novel method of data analysis and pattern classification*. [S.l.], 1965. Citado na página 37.

BARTÓK, A. P.; KONDOR, R.; CSÁNYI, G. On representing chemical environments. *Phys. Rev. B*, American Physical Society, v. 87, p. 184115, May 2013. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevB.87.184115>. Citado na página 32.

BASU, S.; BANERJEE, A.; MOONEY, R. Semi-supervised clustering by seeding. In: *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*. [S.l.: s.n.], 2002. Citado na página 44.

BASU, S.; BANERJEE, A.; MOONEY, R. J. Active semi-supervision for pairwise constrained clustering. In: SIAM. *Proceedings of the 2004 SIAM international conference on data mining*. [S.l.], 2004. p. 333–344. Citado na página 45.

BATISTA, K. E. A. et al. Ab initio investigation of co2 adsorption on 13-atom 4d clusters. *Journal of Chemical Information and Modeling*, v. 60, n. 2, p. 537–545, 2020. PMID: 31917570. Disponível em: <https://doi.org/10.1021/acs.jcim.9b00792>. Citado na página 23.

BATISTA, K. E. A. et al. Energy decomposition to access the stability changes induced by co adsorption on transition-metal 13-atom clusters. *Journal of Chemical Information and Modeling*, v. 61, n. 5, p. 2294–2301, 2021. PMID: 33939914. Disponível em: <https://doi.org/10.1021/acs.jcim.1c00097>. Citado 2 vezes nas páginas 31 e 36.

BEHLER, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, v. 134, n. 7, p. 074106, 2011. Disponível em: <https://doi.org/10.1063/1.3553717>. Citado na página 32.

BROYDEN, C. G. Quasi-newton methods and their application to function minimisation. *Mathematics of Computation*, JSTOR, v. 21, n. 99, p. 368–381, 1967. Citado na página 44.

BULLINGER, H.-J.; WARSCHAT, J.; FISCHER, D. Knowledge-based system for material selection for design with new materials. *Knowledge-Based Systems*, v. 4, n. 2, p. 95–102, 1991. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii-/095070519190013R>. Citado na página 29.

BURKE, E. K.; BYKOV, Y. The late acceptance hill-climbing heuristic. *European Journal of Operational Research*, v. 258, n. 1, p. 70–78, 2017. ISSN 0377-2217. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0377221716305495>. Citado na página 41.

BUTLER, K. T. et al. Machine learning for molecular and materials science. *Nature*, v. 559, n. 7715, p. 547–555, 2018. Disponível em: <https://doi.org/10.1038/s41586-018-0337-2>. Citado 3 vezes nas páginas 23, 24 e 31.

CALIńSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics*, Taylor Francis, v. 3, n. 1, p. 1–27, 1974. Disponível em: <https://www.tandfonline-.com/doi/abs/10.1080/03610927408827101>. Citado na página 39.

CARPENTER, G. A.; GROSSBERG, S. Art 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks*, v. 3, n. 2, p. 129–152, 1990. ISSN 0893-6080. Disponível em: <https://www.sciencedirect.com/science/article/pii-/089360809090085Y>. Citado na página 38.

CARRETE, J. et al. Nanograined half-heusler semiconductors as advanced thermoelectrics: An ab initio high-throughput statistical study. *Advanced Functional Materials*, v. 24, n. 47, p. 7427–7432, 2014. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm-.201401201>. Citado na página 30.

CHA, S.-H. Comprehensive survey on distance/similarity measures between probability density functions. *Int. J. Math. Model. Meth. Appl. Sci.*, v. 1, 01 2007. Citado na página 33.

Chidananda Gowda, K.; DIDAY, E. Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, v. 24, n. 6, p. 567–578, 1991. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/003132039190022W>. Citado na página 35.

CHINER, M. Planning of expert systems for materials selection. *Materials Design*, v. 9, n. 4, p. 195–203, 1988. ISSN 0261-3069. Disponível em: <https://www.sciencedirect.com/science-/article/pii/0261306988900313>. Citado na página 29.

CURTAROLO, S. et al. Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.*, American Physical Society, v. 91, p. 135503, Sep 2003. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevLett.91.135503>. Citado na página 30.

DARGIE, P. P.; PARMESHWAR, K.; WILSON, W. R. D. Maps-1: Computer-aided design system for preliminary material and manufacturing process selection. *Journal of Mechanical Design*, v. 104, p. 126–136, 1982. Disponível em: <https://doi.org/10.1115/1.3256302>. Citado na página 29.

DASH, M.; LIU, H. Feature selection for clustering. In: SPRINGER. *Pacific-Asia Conference on knowledge discovery and data mining*. [S.l.], 2000. p. 110–121. Citado na página 24.

DAWSON, W. et al. Complexity reduction in density functional theory calculations of large systems: System partitioning and fragment embedding. *Journal of Chemical Theory and Computation*, v. 16, n. 5, p. 2952–2964, 2020. PMID: 32216343. Disponível em: <https://doi.org/10.1021/acs.jctc.9b01152>. Citado na página 24.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 39, n. 1, p. 1–22, 1977. Disponível em: <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j-.2517-6161.1977.tb01600.x>. Citado na página 38.

DIDAY, E. The dynamic clusters method in nonhierarchical clustering. *International Journal of Computer & Information Sciences*, Springer, v. 2, n. 1, p. 61–88, 1973. Citado na página 37.

DORIGO, M.; BIRATTARI, M.; STUTZLE, T. Ant colony optimization. *IEEE Computational Intelligence Magazine*, v. 1, n. 4, p. 28–39, 2006. Citado na página 43.

DUDA, R. O.; HART, P. E. et al. *Pattern classification and scene analysis*. [S.l.]: Wiley New York, 1973. Citado na página 35.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification (2nd Edition)*. 2. ed. [S.l.]: Wiley-Interscience, 2000. Hardcover. ISBN 0471056693. Citado na página 33.

EDWARDS, K. Selecting materials for optimum use in engineering components. *Materials Design*, v. 26, n. 5, p. 469–473, 2005. ISSN 0261-3069. Disponível em: <https://www-.sciencedirect.com/science/article/pii/S0261306904001682>. Citado na página 29.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *kdd*. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231. Citado na página 38.

FABER, F. et al. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, v. 115, n. 16, p. 1094–1101, 2015. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.24917>. Citado na página 32.

FARAG, M. M. *Materials and process selection in engineering*. [S.l.: s.n.], 1979. Citado na página 29.

FARAG, M. M. Quantitative methods of materials selection. In: ____. *Mechanical Engineers' Handbook*. American Cancer Society, 2015. cap. 15, p. 1–22. ISBN 9781118985960. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118985960.meh115>. Citado na página 29.

FELíCIO-SOUSA, P. et al. Ab initio insights into the structural, energetic, electronic, and stability properties of mixed cenzr15no30 nanoclusters. *Phys. Chem. Chem. Phys.*, The Royal Society of Chemistry, v. 21, p. 26637–26646, 2019. Disponível em: <http://dx.doi.org/10-.1039/C9CP04762J>. Citado na página 54.

FORTUNATO, S. Community detection in graphs. *Physics Reports*, v. 486, n. 3, p. 75–174, 2010. ISSN 0370-1573. Disponível em: <https://www.sciencedirect.com/science/article/pii-/S0370157309002841>. Citado na página 37.

GAYNOR, S.; BAIR, E. Identification of biologically relevant subtypes via preweighted sparse clustering. bepress, 2012. Citado na página 44.

GILMER, J. et al. Neural message passing for quantum chemistry. In: PMLR. *International conference on machine learning*. [S.l.], 2017. p. 1263–1272. Citado na página 24.

GLOVER, F. Future paths for integer programming and links to artificial intelligence. *Computers Operations Research*, v. 13, n. 5, p. 533–549, 1986. ISSN 0305-0548. Applications of Integer Programming. Disponível em: <https://www.sciencedirect.com/science/article/pii-/0305054886900481>. Citado na página 41.

HANSEN, K. et al. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The Journal of Physical Chemistry Letters*, v. 6, n. 12, p. 2326–2331, 2015. PMID: 26113956. Disponível em: <https://doi.org/10.1021/acs-.jpclett.5b00831>. Citado na página 31.

HARMON, L. Experiment planning for combinatorial materials discovery. *Journal of Materials Science*, v. 38, n. 22, p. 4479–4485, 2003. Disponível em: <https://doi.org/10.1023-/A:1027325400459>. Citado na página 23.

HAUTIER, G. et al. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chemistry of Materials*, v. 22, n. 12, p. 3762–3767, 2010. Disponível em: <https://doi.org/10.1021/cm100795d>. Citado na página 30.

HERNANDEZ, S. *Development of Methods for Reducing the Cost of Density Functional Theory and Time-Dependent Density Functional Theory*. Tese (Doutorado) — UCLA, 2015. Citado na página 24.

HIMANEN, L. et al. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, v. 247, p. 106949, 2020. ISSN 0010-4655. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010465519303042>. Citado na página 32.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, n. 1, p. 193–218, 1985. Citado na página 39.

HUO, H.; RUPP, M. *Unified Representation of Molecules and Crystals for Machine Learning*. 2018. Citado na página 32.

JAHAN, A. et al. Material screening and choosing methods – a review. *Materials & Design*, v. 31, n. 2, p. 696 – 705, 2010. ISSN 0261-3069. Disponível em: <http://www-.sciencedirect.com/science/article/pii/S0261306909004361>. Citado 3 vezes nas páginas 23, 29 e 30.

JAIN, A.; MAO, J.; MOHIUDDIN, K. Artificial neural networks: a tutorial. *Computer*, v. 29, n. 3, p. 31–44, 1996. Citado na página 38.

JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, v. 31, n. 8, p. 651 – 666, 2010. ISSN 0167-8655. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR). Disponível em: <http://www-.sciencedirect.com/science/article/pii/S0167865509002323>. Citado 2 vezes nas páginas 24 e 25.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Computing Surveys*, v. 31, n. 3, p. 264–323, 1999. Citado 3 vezes nas páginas 25, 34 e 38.

JIANMIN, L. X. C. G. Z. W. W. Y. Z. K. W. Intelligent expert system used in gear material selection and its heat treatment. *Acta Metall Sin*, Acta Metall Sin, v. 40, n. 10, p. 1051, 2004. Disponível em: <https://www.ams.org.cn/EN/abstract/article_4061.shtml>. Citado na página 29.

JU, S. et al. Designing nanostructures for phonon transport via bayesian optimization. *Phys. Rev. X*, American Physical Society, v. 7, p. 021024, May 2017. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevX.7.021024>. Citado na página 30.

KENNEDY, J.; EBERHART, R. Particle swarm optimization. In: *Proceedings of ICNN'95 - International Conference on Neural Networks*. [S.l.: s.n.], 1995. v. 4, p. 1942–1948 vol.4. Citado na página 43.

KESTEREN, I. E. H. van; STAPPERS, P. J.; BRUIJN, J. C. M. de. Materials in product selection: Tools for including user-interaction aspects in materials selection. international journal of design. *International Journal of Design*, v. 1, n. 3, p. 41–55, 2007. Disponível em: <http://www.ijdesign.org/index.php/IJDesign/article/view/129/78>. Citado na página 29.

KESTEREN, I. V.; KANDACHAR, P.; STAPPERS, P. J. Activities in selecting materials from the perspective of product designers. *International Journal of Design Engineering*, v. 25, n. 1, p. 83–103, 2007. Disponível em: <http://www.inderscience.com/offer.php?id=15337>. Citado na página 29.

KHATIB, M. E.; JONG, W. A. de. *ML4Chem: A Machine Learning Package for Chemistry and Materials Science*. 2020. Citado 2 vezes nas páginas 31 e 32.

KING, B. Step-wise clustering procedures. *Journal of the American Statistical Association*, Taylor Francis, v. 62, n. 317, p. 86–101, 1967. Disponível em: <https://www.tandfonline-.com/doi/abs/10.1080/01621459.1967.10482890>. Citado na página 37.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. *Science*, American Association for the Advancement of Science, v. 220, n. 4598, p. 671–680, 1983. ISSN 00368075. Disponível em: <http://www.jstor.org/stable/1690046>. Citado na página 42.

KLEIN, D.; KAMVAR, S. D.; MANNING, C. D. *From Instance-level Constraints to Space-Level Constraints: Making the Most of Prior Knowledge in Data Clustering*. [S.l.], 2002. Disponível em: <http://ilpubs.stanford.edu:8090/528/>. Citado na página 45.

KOHONEN, T. *Self-Organization and Associative Memory: 3rd Edition*. Berlin, Heidelberg: Springer-Verlag, 1989. ISBN 0387513876. Citado na página 38.

KREER, J. A question of terminology. *IRE Transactions on Information Theory*, v. 3, n. 3, p. 208–208, 1957. Citado na página 40.

LAARHOVEN, P. van; AARTS, E. Simulated annealing. In: *Simulated Annealing: Theory and Applications*. [S.l.]: Springer, Dordrecht, 1987. p. 7–15. Citado na página 42.

LEARY, R. H. Global optimization on funneling landscapes. *J. of Global Optimization*, Kluwer Academic Publishers, USA, v. 18, n. 4, p. 367–383, dez. 2000. ISSN 0925-5001. Disponível em: <https://doi.org/10.1023/A:1026500301312>. Citado na página 42.

LEMARÉCHAL, C. Cauchy and the gradient method. *Doc Math Extra*, v. 251, n. 254, p. 10, 2012. Citado na página 44.

LIM, A.; RODRIGUES, B.; ZHANG, X. A simulated annealing and hill-climbing algorithm for the traveling tournament problem. *European Journal of Operational Research*, v. 174, n. 3, p. 1459–1478, 2006. ISSN 0377-2217. Disponível em: <https://www.sciencedirect.com-/science/article/pii/S0377221705003206>. Citado na página 41.

LO, Y.-C. et al. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, v. 23, n. 8, p. 1538 – 1546, 2018. ISSN 1359-6446. Disponível em: <http:/-/www.sciencedirect.com/science/article/pii/S1359644617304695>. Citado na página 23.

LUDWIG, A. Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. *npj Computational Materials*, v. 5, n. 1, p. 70, 2019. Disponível em: <https://doi.org/10.1038-/s41524-019-0205-0>. Citado na página 23.

MAATEN, L. van der; HINTON, G. Visualizing high-dimensional data using t-sne. 2008. Citado na página 51.

MACKAY, D. J. *Information theory, inference and learning algorithms*. [S.l.]: Cambridge university press, 2003. Citado na página 40.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: CAM, L. M. L.; NEYMAN, J. (Ed.). *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.]: University of California Press, 1967. v. 1, p. 281–297. Citado na página 36.

MATSUBARA, M. et al. Identifying superionic conductors by materials informatics and high-throughput synthesis. *Communications Materials*, v. 1, n. 1, p. 5, 2020. Disponível em: <https://doi.org/10.1038/s43246-019-0004-7>. Citado na página 23.

MCFARLAND, E. W.; WEINBERG, W. Combinatorial approaches to materials discovery. *Trends in Biotechnology*, v. 17, n. 3, p. 107 – 115, 1999. ISSN 0167-7799. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016777999801275X>. Citado na página 23.

MENDES, P. C. D. et al. Ab initio screening of pt-based transition-metal nanoalloys using descriptors derived from the adsorption and activation of co2. *Phys. Chem. Chem. Phys.*, The Royal Society of Chemistry, v. 23, p. 6029–6041, 2021. Disponível em: <http://dx.doi.org/10.1039/D1CP00570G>. Citado na página 54.

MITCHELL, M. *An introduction to genetic algorithms*. [S.l.]: MIT press, 1998. Citado na página 43.

MITCHELL, T. M. et al. Machine learning. McGraw-hill New York, 1997. Citado na página 38.

MIYAMOTO, S.; TERAMI, A. Semi-supervised agglomerative hierarchical clustering algorithms with pairwise constraints. In: *International Conference on Fuzzy Systems*. [S.l.: s.n.], 2010. p. 1–6. Citado na página 45.

MONTAVON, G. et al. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, IOP Publishing, v. 15, n. 9, p. 095003, sep 2013. Citado na página 23.

MORIWAKI, H. et al. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics*, v. 10, n. 1, p. 4, 2018. Disponível em: <https://doi.org/10.1186/s13321-018-0258-y>. Citado na página 32.

MURTAGH, F. A Survey of Recent Advances in Hierarchical Clustering Algorithms. *The Computer Journal*, v. 26, n. 4, p. 354–359, 11 1983. ISSN 0010-4620. Disponível em: <https://doi.org/10.1093/comjnl/26.4.354>. Citado na página 37.

NIELSON, K. D. et al. Development of the reaxff reactive force field for describing transition metal catalyzed reactions, with application to the initial stages of the catalytic formation of carbon nanotubes. *J. Phys. Chem. A*, v. 109, p. 493–499, 2005. Citado na página 54.

NIEUWENBURG, E. P. L. van; LIU, Y.-H.; HUBER, S. D. Learning phase transitions by confusion. *Nature Physics*, v. 13, p. 435–439, 2017. Disponível em: <https://doi.org/10.1038-/nphys4037>. Citado na página 30.

OLSON, B. et al. Basin hopping as a general and versatile optimization framework for the characterization of biological macromolecules. *Advances in Artificial Intelligence*, v. 2012, 12 2012. Citado na página 43.

OWOLABI, T. O.; AKANDE, K. O.; OLATUNJI, S. O. Estimation of superconducting transition temperature tc for superconductors of the doped mgb2 system from the crystal lattice parameters using support vector regression. *Journal of Superconductivity and Novel Magnetism*, v. 28, p. 75–81, 2015. Disponível em: <https://doi.org/10.1007/s10948-014-2891-7>. Citado na página 31.

PARKER, A. J.; BARNARD, A. S. Selecting appropriate clustering methods for materials science applications of machine learning. *Advanced Theory and Simulations*, v. 2, n. 12, p. 1900145, 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/adts-.201900145>. Citado 2 vezes nas páginas 38 e 45.

PATRA, T. K. et al. Neural-network-biased genetic algorithms for materials design: Evolutionary algorithms that learn. *ACS Combinatorial Science*, v. 19, n. 2, p. 96–107, 2017. PMID: 27997791. Disponível em: <https://doi.org/10.1021/acscombsci.6b00136>. Citado na página 30.

PEDGLEY, O. Influence of stakeholders on industrial design materials and manufacturing selection. *International Journal of Design*, v. 3, p. 1–15, 04 2009. Citado na página 29.

PHAM, T. L. et al. Machine learning reveals orbital interaction in materials. *Science and Technology of Advanced Materials*, Taylor Francis, v. 18, n. 1, p. 756–765, 2017. PMID: 29152012. Disponível em: <https://doi.org/10.1080/14686996.2017.1378060>. Citado na página 30.

PINHEIRO, G. A. et al. A graph-based clustering analysis of the qm9 dataset via smiles descriptors. In: GERVASI, O. et al. (Ed.). *Computational Science and Its Applications – ICCSA 2020*. Cham: Springer International Publishing, 2020. p. 421–433. ISBN 978-3-030-58799-4. Citado na página 37.

RAMAKRISHNAN, R. et al. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, Nature Publishing Group, v. 1, p. 140022, 2014. Citado na página 55.

RAMPRASAD, R. et al. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, v. 3, n. 54, 2017. Disponível em: <https://doi.org/10-.1038/s41524-017-0056-5>. Citado na página 30.

RONDINA, G. G.; DA SILVA, J. L. F. Revised basin-hopping Monte Carlo algorithm for structure optimization of clusters and nanoparticles. *J. Chem. Inf. Model.*, American Chemical Society, v. 53, n. 9, p. 2282–2298, Sept. 2013. Disponível em: <http://dx.doi.org/10.1021-/ci400224z>. Citado na página 54.

ROTH, R.; FIELD, F.; CLARK, J. P. Materials selection and multi-attribute utility analysis. *Journal of Computer-Aided Materials Design*, v. 1, p. 325–342, 1994. Citado na página 29.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii-/0377042787901257>. Citado 2 vezes nas páginas 39 e 48.

RUDDIGKEIT, L. et al. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, ACS Publications, v. 52, n. 11, p. 2864–2875, 2012. Citado na página 55.

RUPP, M. et al. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, American Physical Society, v. 108, p. 058301, Jan 2012. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevLett.108.058301>. Citado 2 vezes nas páginas 31 e 32.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3rd. ed. USA: Prentice Hall Press, 2009. ISBN 0136042597. Citado na página 41.

SANVITO, S. et al. Accelerated discovery of new magnets in the heusler alloy family. *Science Advances*, American Association for the Advancement of Science, v. 3, n. 4, 2017. Disponível em: <https://advances.sciencemag.org/content/3/4/e1602241>. Citado na página 30.

SAPUAN, S.; ABDALLA, H. A prototype knowledge-based system for the material selection of polymeric-based composites for automotive components. *Composites Part A: Applied Science and Manufacturing*, v. 29, n. 7, p. 731–742, 1998. ISSN 1359-835X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1359835X98000499>. Citado na página 29.

SEGER, C. *An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing*. 2018. 34 p. (TRITA-EECS-EX, 2018:596). Citado na página 35.

SEKO, A.; TOGO, A.; TANAKA, I. Descriptors for machine learning of materials data. In: *Nanoinformatics*. [S.l.]: Springer, Singapore, 2018. p. 3–23. Citado na página 31.

SHAFRANOVICH, Y. Common format and mime type for comma-separated values (csv) files. 2005. Disponível em: <https://www.rfc-editor.org/info/rfc4180>. Citado na página 48.

SNEATH, P. H.; SOKAL, R. R. et al. *Numerical taxonomy. The principles and practice of numerical classification.* [S.l.: s.n.], 1973. Citado na página 37.

STANEV, V. et al. Machine learning modeling of superconducting critical temperature. *npj Computational Materials*, v. 4, p. 29, 2018. Disponível em: <https://doi.org/10.1038/s41524-018-0085-8>. Citado na página 31.

STEINHARDT, P. J.; NELSON, D. R.; RONCHETTI, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B*, American Physical Society, v. 28, p. 784–805, Jul 1983. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevB.28.784>. Citado na página 32.

SYMONS, M. J. Clustering criteria and multivariate normal mixtures. *Biometrics*, JSTOR, p. 35–43, 1981. Citado na página 37.

TABOR, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature Reviews Materials*, v. 3, n. 5, p. 5–20, 2018. Disponível em: <https://doi.org/10.1038/s41578-018-0005-z>. Citado na página 23.

THORNDIKE, R. L. Who belongs in the family. *Psychometrika*, p. 267–276, 1953. Citado na página 38.

TRUSHIN, E.; THIERBACH, A.; GÖRLING, A. Toward chemical accuracy at low computational cost: Density-functional theory with σ-functionals for the correlation energy. *The Journal of Chemical Physics*, AIP Publishing LLC, v. 154, n. 1, p. 014104, 2021. Citado na página 24.

van Duin, A. C. T. et al. Reaxff: A reactive force field for hydrocarbons. *J. Phys. Chem. A*, v. 105, p. 9396–9409, 2001. Citado na página 54.

VARGHESE, A.; CAWLEY, M.; HONG, T. Supervised clustering for automated document classification and prioritization: a case study using toxicological abstracts. *Environment Systems and Decisions*, Springer, v. 38, n. 3, p. 398–414, 2018. Citado na página 44.

WAGSTAFF, K. et al. Constrained k-means clustering with background knowledge. In: . [S.l.: s.n.], 2001. p. 577–584. Citado na página 45.

Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *Journal of Physical Chemistry A*, v. 101, n. 28, p. 5111–5116, jul. 1997. Citado na página 42.

WARD, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, Taylor Francis, v. 58, n. 301, p. 236–244, 1963. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845>. Citado na página 37.

WEISE, T. Global optimization algorithms-theory and application. *Self-Published Thomas Weise*, 2009. Citado na página 41.

YAMAWAKI, M. et al. Multifunctional structural design of graphene thermoelectrics by bayesian optimization. *Science Advances*, American Association for the Advancement of Science, v. 4, n. 6, 2018. Disponível em: <https://advances.sciencemag.org/content/4/6-/eaar4192>. Citado na página 30.

YANG, X.-S. *Introduction to Mathematical Optimization: From Linear Programming to Metaheuristics*. [S.l.]: Cambridge International2 Science Publishing, 2008. Citado na página 41.

YANG, X.-S.; GANDOMI, A. H. Bat algorithm: a novel approach for global engineering optimization. *Engineering computations*, Emerald Group Publishing Limited, 2012. Citado na página 43.

YU, J.-C.; KRIZAN, S.; ISHII, K. Computer-aided design for manufacturing process selection. *Journal of Intelligent Manufacturing*, v. 4, p. 199–208, 01 1993. Citado na página 29.

ZAHN, C. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, C-20, n. 1, p. 68–86, 1971. Citado na página 37.

ZHAO, H.; QI, Z. Hierarchical agglomerative clustering with ordering constraints. In: *2010 Third International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2010. p. 195–199. Citado na página 45.

ZHAO, X.; FU, L. Machine learning phase transition: An iterative proposal. *Annals of Physics*, Elsevier BV, v. 410, p. 167938, Nov 2019. ISSN 0003-4916. Disponível em: <http://dx.doi.org/10.1016/j.aop.2019.167938>. Citado na página 30.

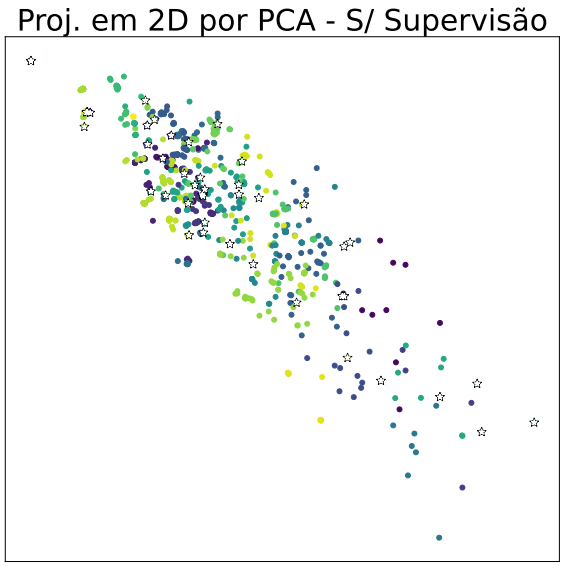
ZHAOCHUN, Z.; RUIWU, P.; NIANYI, C. Artificial neural network prediction of the band gap and melting point of binary and ternary compound semiconductors. *Materials Science and Engineering: B*, v. 54, n. 3, p. 149–152, 1998. ISSN 0921-5107. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0921510798001573>. Citado na página 30.

ZIBORDI-BESSE, L. et al. Physical and chemical properties of unsupported (mo2)n clusters for m = ti, zr, or ce and n = 1–15: A density functional theory study combined with the tree-growth scheme and euclidean similarity distance algorithm. *The Journal of Physical Chemistry C*, v. 122, n. 48, p. 27702–27712, 2018. Disponível em: <https://doi.org/10.1021-/acs.jpcc.8b08299>. Citado na página 23.

Apêndices

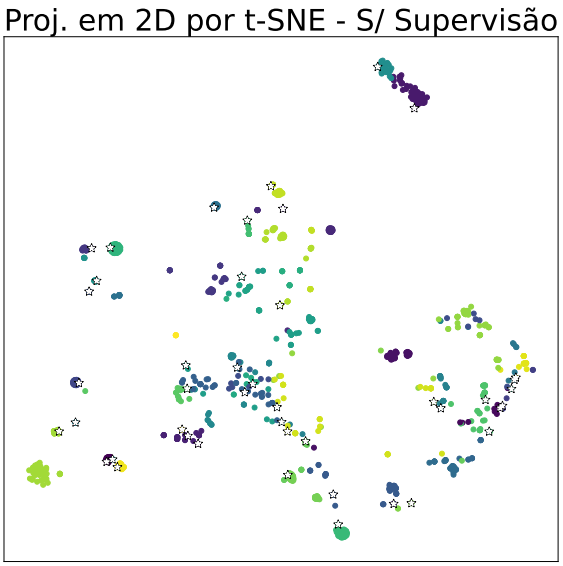
APÊNDICE A – Figuras Nanoclusters de Cu_n

Figura 26 – Projeção 2D (PCA) do conjunto Cu_n não ponderado



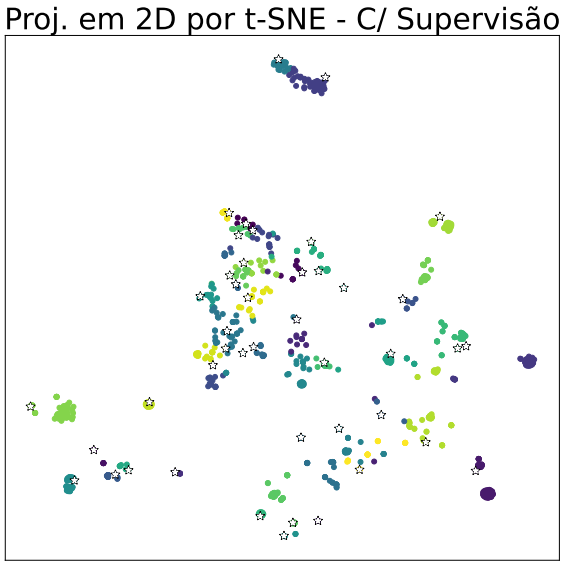
Fonte: o autor

Figura 27 – Projeção 2D (t-SNE) do conjunto Cu_n não ponderado



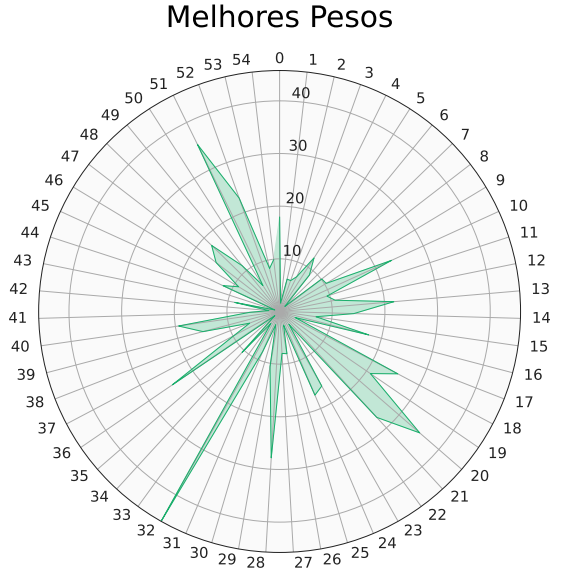
Fonte: o autor

Figura 28 – Projeção 2D (t-SNE) do conjunto Cu_n ponderado



Fonte: o autor

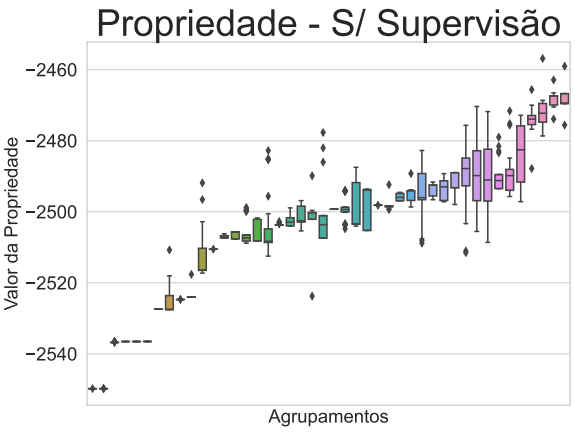
Figura 29 – Gráfico de radar do conjunto Cu_n ponderado



Fonte: o autor

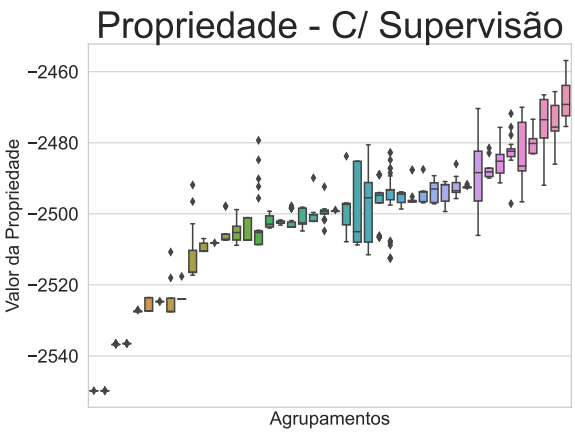
Não é possível gerar figura de coordenadas paralelas de boa qualidade para esse conjunto, devido à superlotação de linhas no plot decorrente da quantidade alta de grupos formados.

Figura 30 – Boxplots do conjunto Cu_n não ponderado



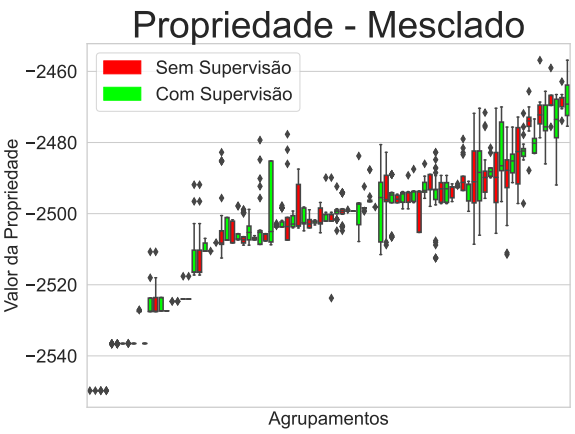
Fonte: o autor

Figura 31 – Boxplots do conjunto Cu_n ponderado



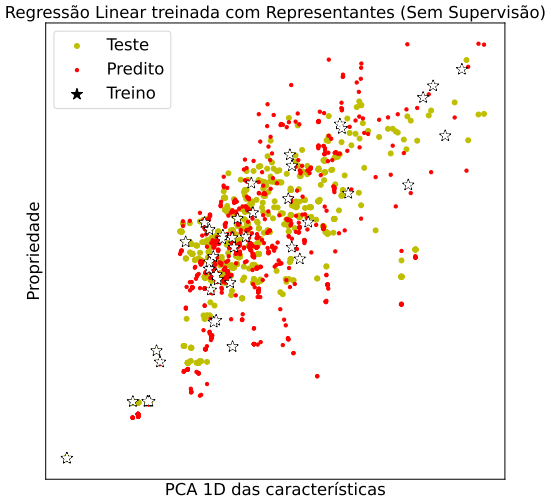
Fonte: o autor

Figura 32 – Boxplots do conjunto Cu_n não ponderado e ponderado



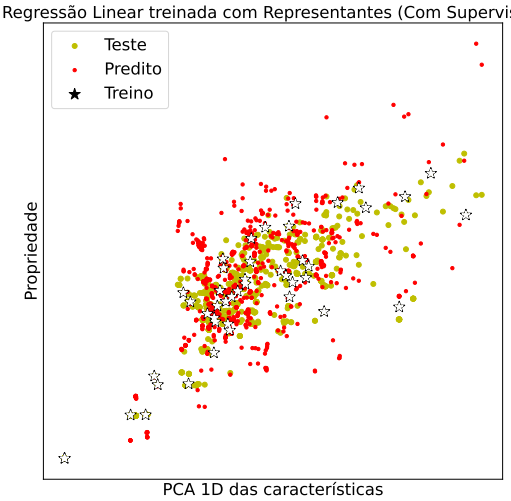
Fonte: o autor

Figura 33 – Regressão Linear do conjunto Cu_n não ponderado



Fonte: o autor

Figura 34 – Regressão Linear do conjunto Cu_n ponderado



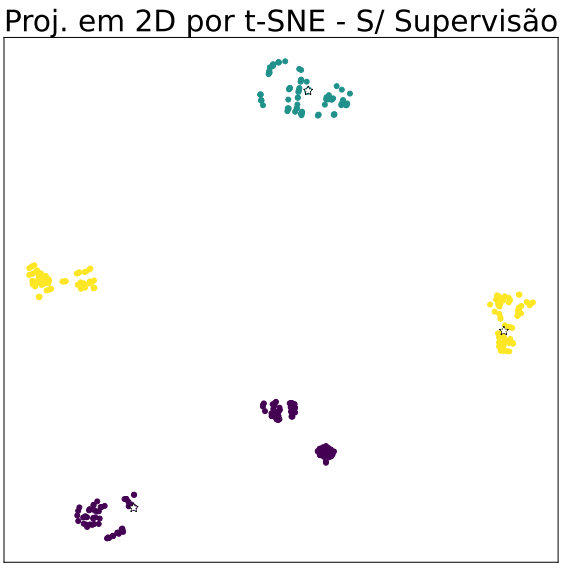
Fonte: o autor

APÊNDICE B – Figuras Nanoligas Core-Shell baseadas em Pt de 55 átomos

Figura 35 – Projeção 2D (PCA) do conjunto PtTM não ponderado

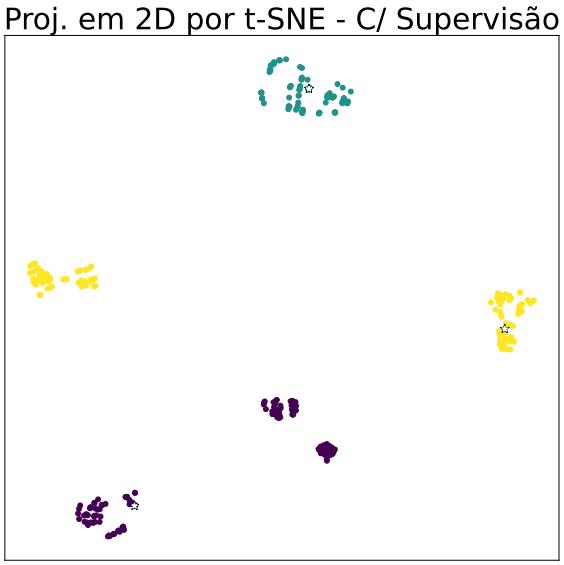
Fonte: o autor

Figura 36 – Projeção 2D (t-SNE) do conjunto PtTM não ponderado



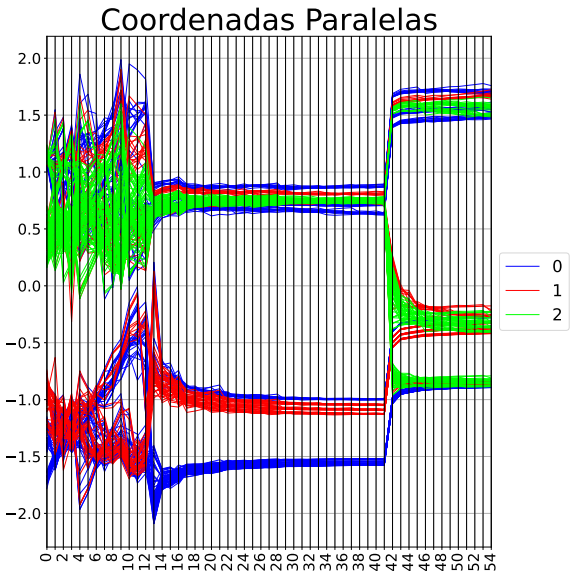
Fonte: o autor

Figura 37 – Projeção 2D (t-SNE) do conjunto PtTM ponderado



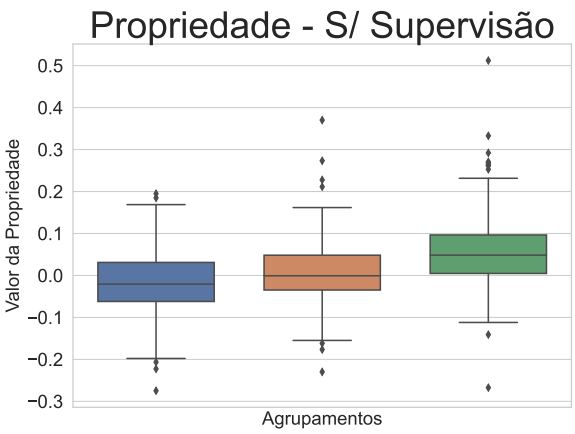
Fonte: o autor

Figura 38 – Gráfico de coordenadas paralelas do conjunto PtTM ponderado



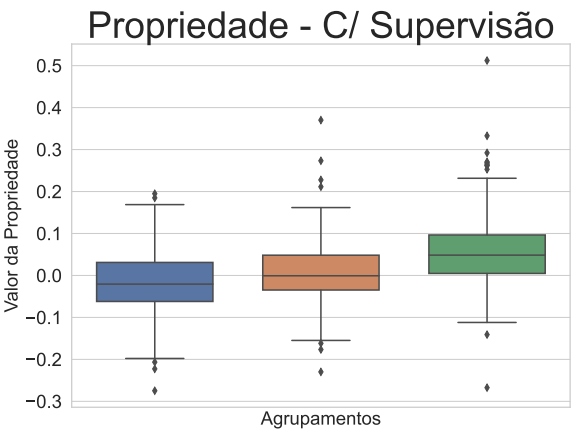
Fonte: o autor

Figura 39 – Boxplots do conjunto PtTM não ponderado



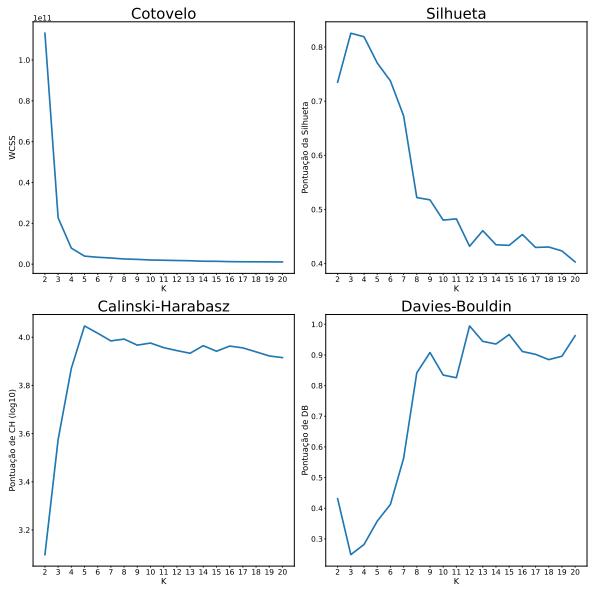
Fonte: o autor

Figura 40 – Boxplots do conjunto PtTM ponderado



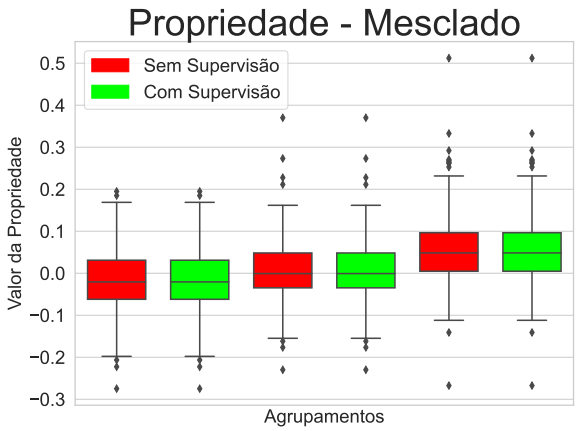
Fonte: o autor

Figura 41 – Métricas de qualidade de agrupamento do conjunto PtTM



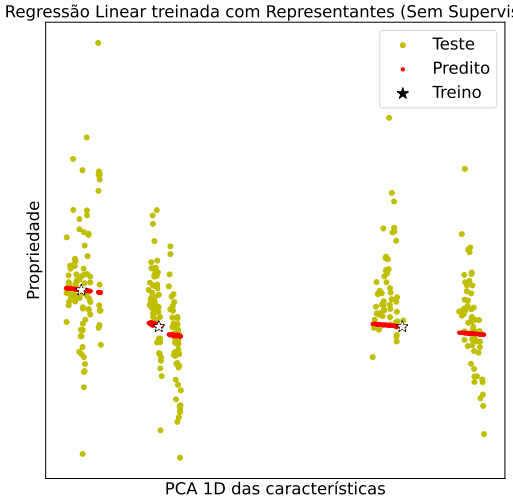
Fonte: o autor

Figura 42 – Boxplots do conjunto PtTM não ponderado e ponderado



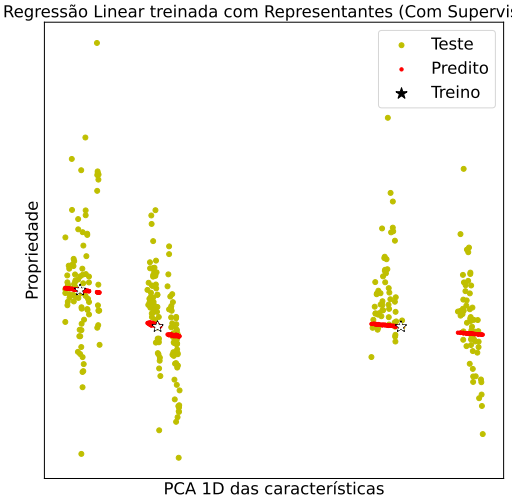
Fonte: o autor

Figura 43 – Regressão Linear do conjunto PtTM não ponderado



Fonte: o autor

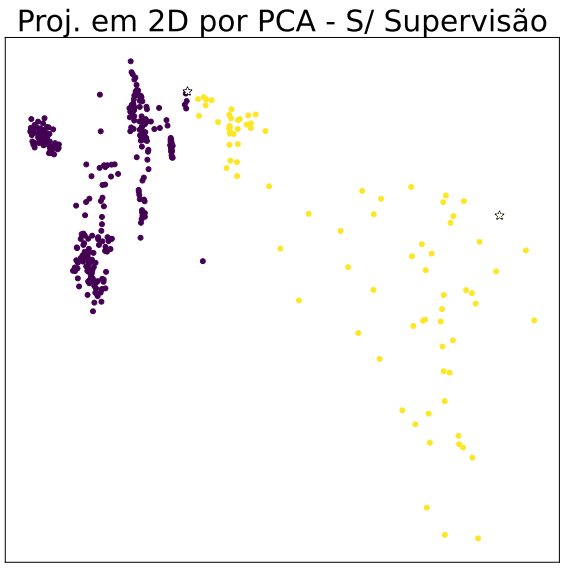
Figura 44 – Regressão Linear do conjunto PtTM ponderado



Fonte: o autor

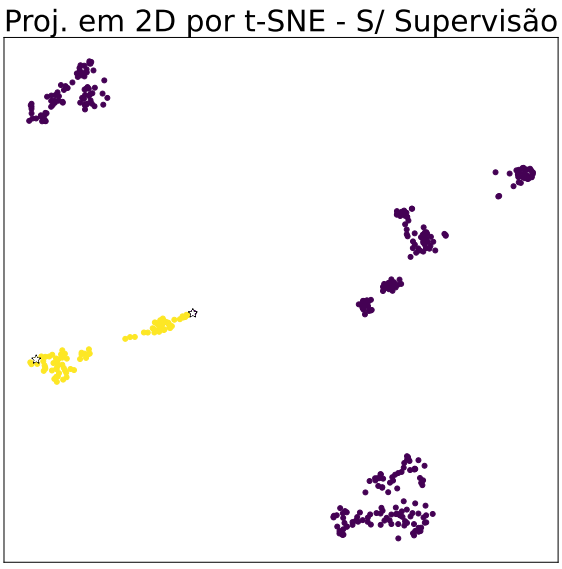
APÊNDICE C – Desidrogenação de CH_4 em clusters TM_{13}

Figura 45 – Projeção 2D (PCA) do conjunto CH_nTM não ponderado



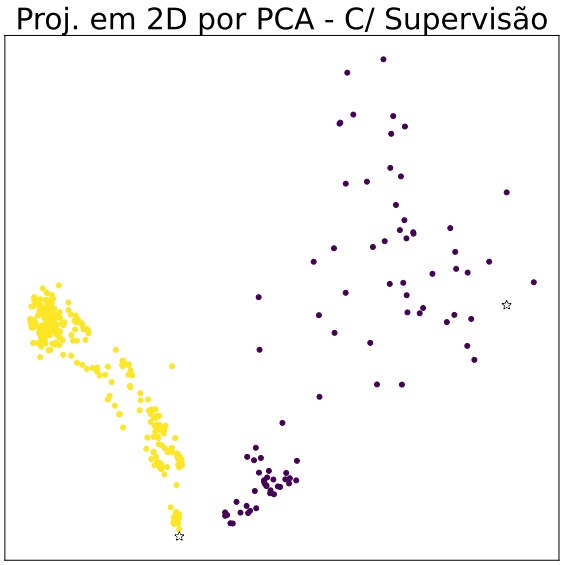
Fonte: o autor

Figura 46 – Projeção 2D (t-SNE) do conjunto CH_nTM não ponderado



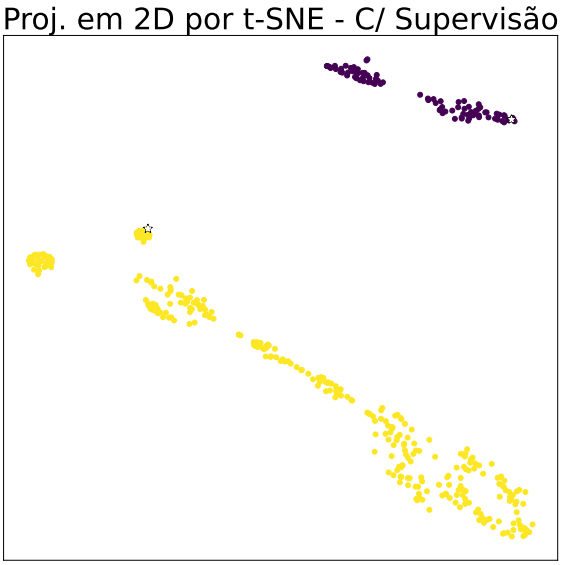
Fonte: o autor

Figura 47 – Projeção 2D (PCA) do conjunto CH_nTM ponderado



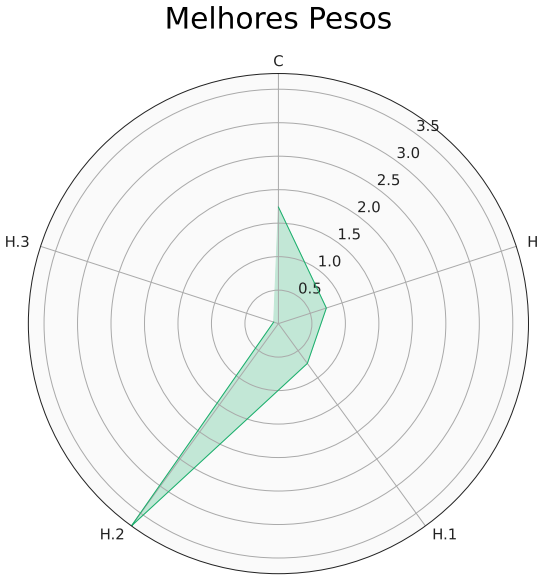
Fonte: o autor

Figura 48 – Projeção 2D (t-SNE) do conjunto CH_nTM ponderado



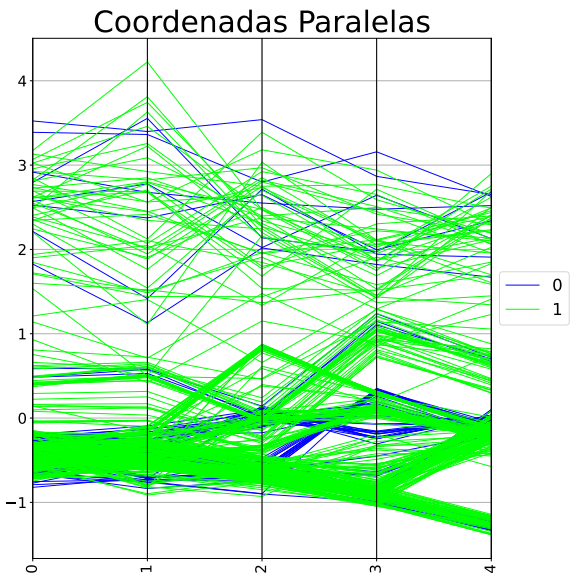
Fonte: o autor

Figura 49 – Gráfico de radar do conjunto CH_nTM ponderado



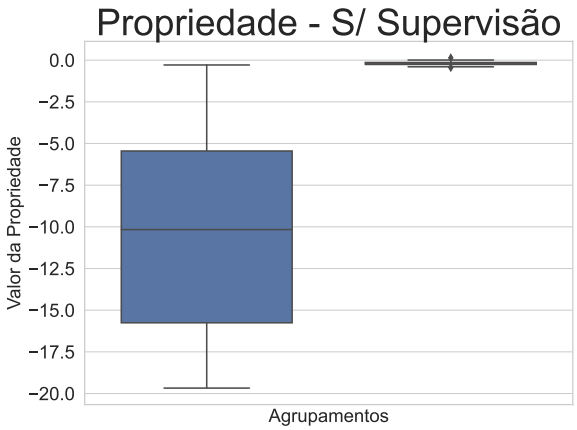
Fonte: o autor

Figura 50 – Gráfico de coordenadas paralelas do conjunto CH_nTM ponderado



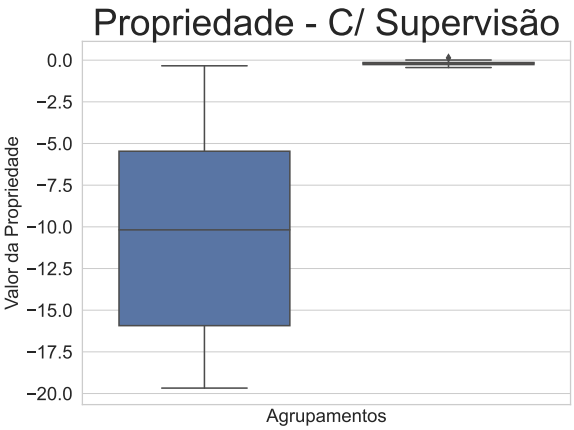
Fonte: o autor

Figura 51 – Boxplots do conjunto CH_nTM não ponderado



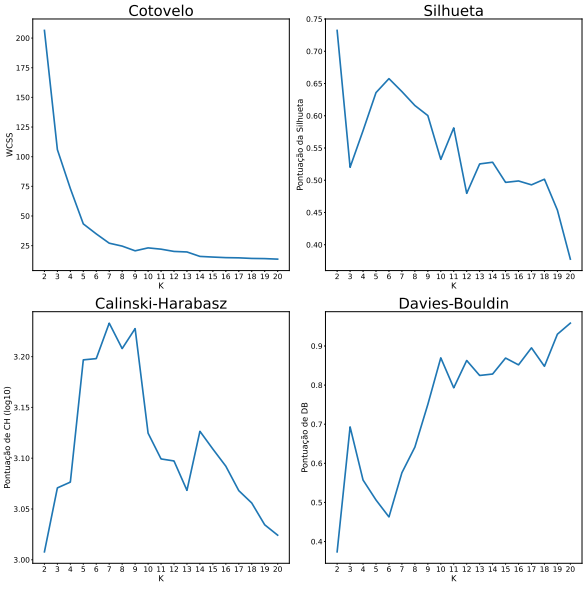
Fonte: o autor

Figura 52 – Boxplots do conjunto CH_nTM ponderado



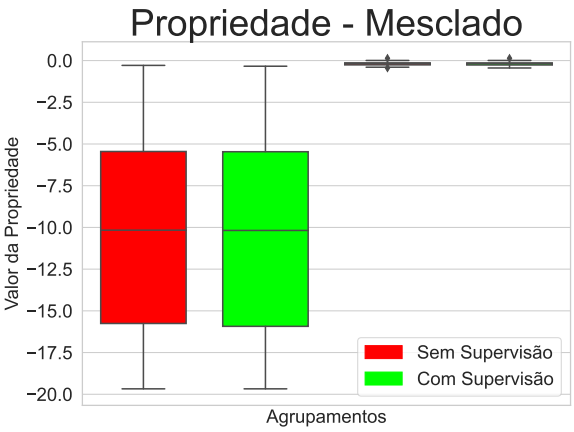
Fonte: o autor

Figura 53 – Métricas de qualidade de agrupamento do conjunto CH_nTM



Fonte: o autor

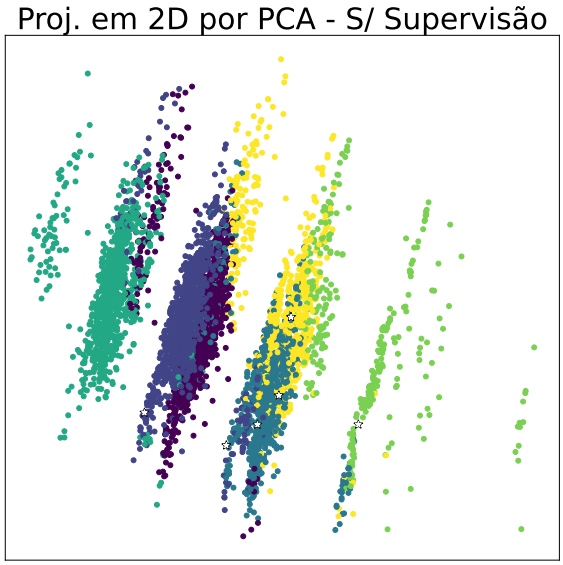
Figura 54 – Boxplots do conjunto CH_nTM não ponderado e ponderado



Fonte: o autor

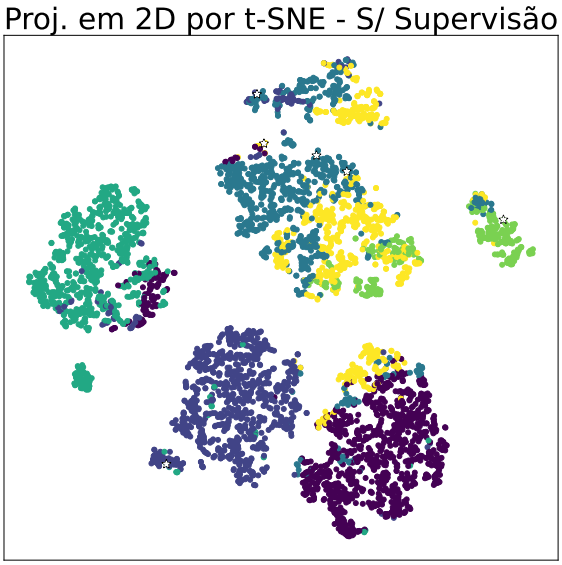
APÊNDICE D – QM9

Figura 55 – Projeção 2D (PCA) do conjunto QM9 não ponderado



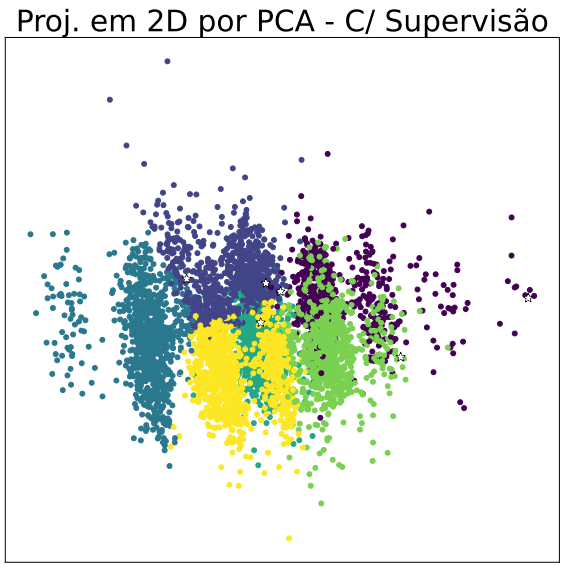
Fonte: o autor

Figura 56 – Projeção 2D (t-SNE) do conjunto QM9 não ponderado



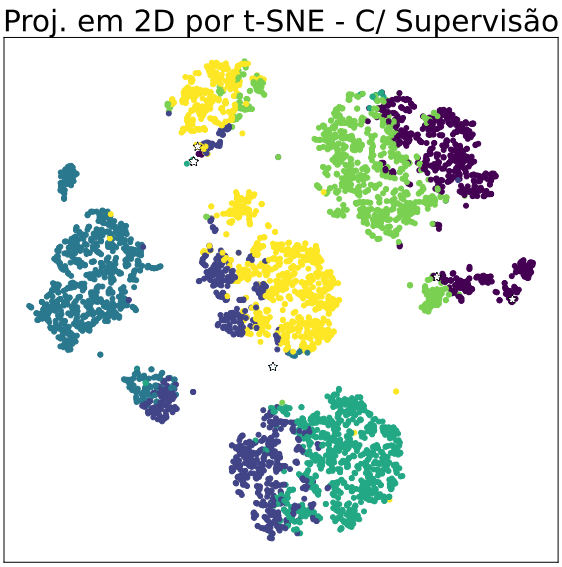
Fonte: o autor

Figura 57 – Projeção 2D (PCA) do conjunto QM9 ponderado



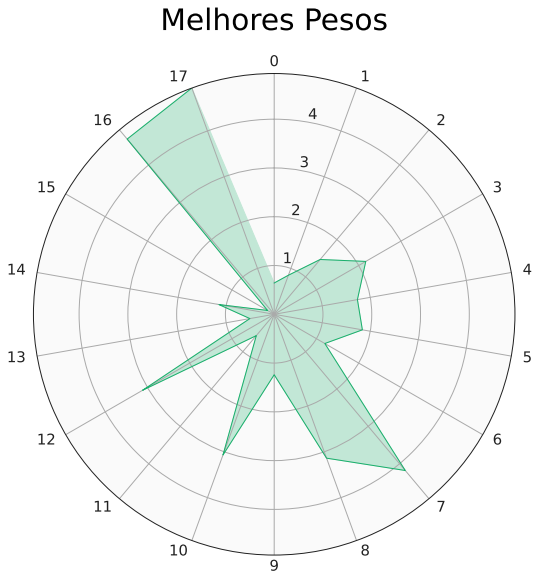
Fonte: o autor

Figura 58 – Projeção 2D (t-SNE) do conjunto QM9 ponderado



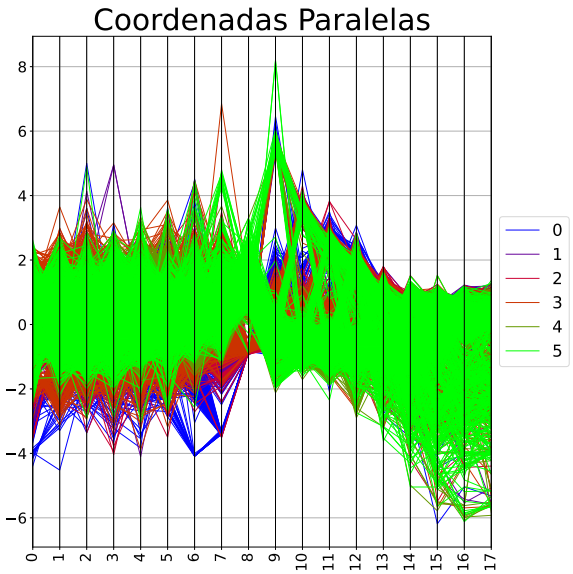
Fonte: o autor

Figura 59 – Gráfico de radar do conjunto QM9 ponderado



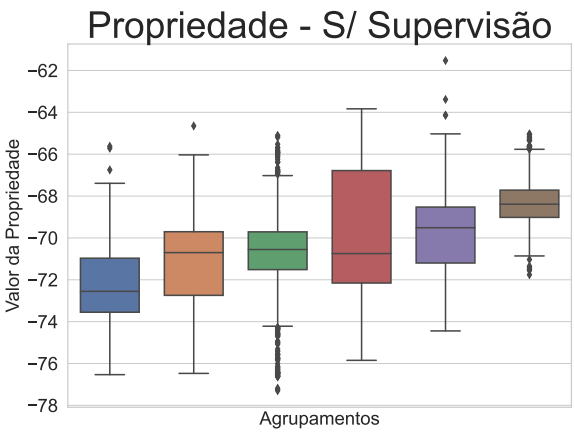
Fonte: o autor

Figura 60 – Gráfico de coordenadas paralelas do conjunto QM9 ponderado



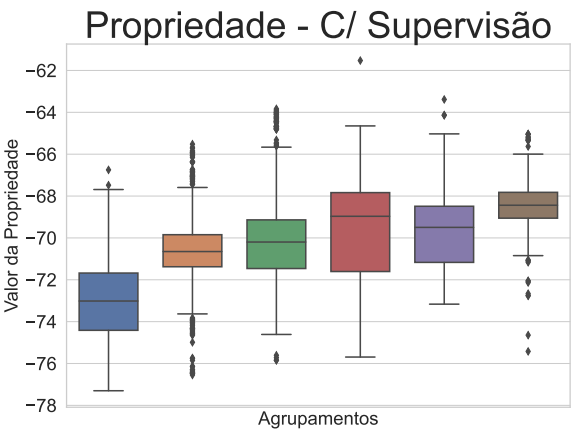
Fonte: o autor

Figura 61 – Boxplots do conjunto QM9 não ponderado



Fonte: o autor

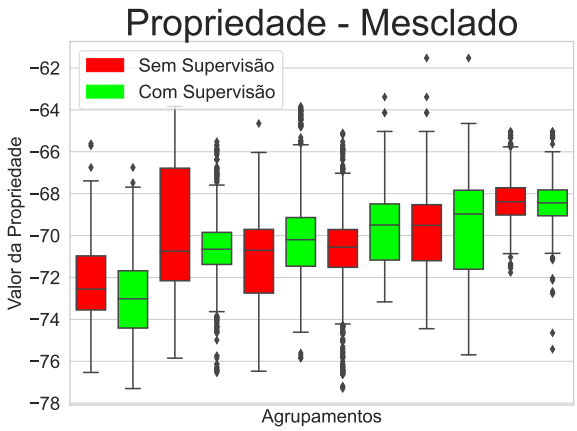
Figura 62 – Boxplots do conjunto QM9 ponderado



Fonte: o autor

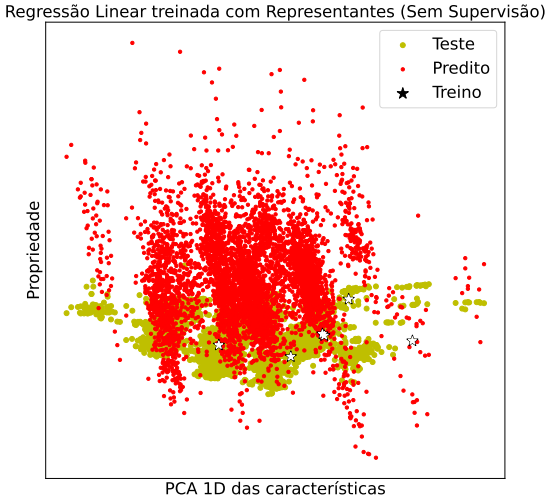
O valor de 6 grupos foi inserido manualmente de acordo com inspeção visual das projeções PCA e t-SNE dos dados, logo o gráfico que mostra os valores das métricas de qualidade de agrupamento para cada valor de K não foi gerado.

Figura 63 – Boxplots do conjunto QM9 não ponderado e ponderado



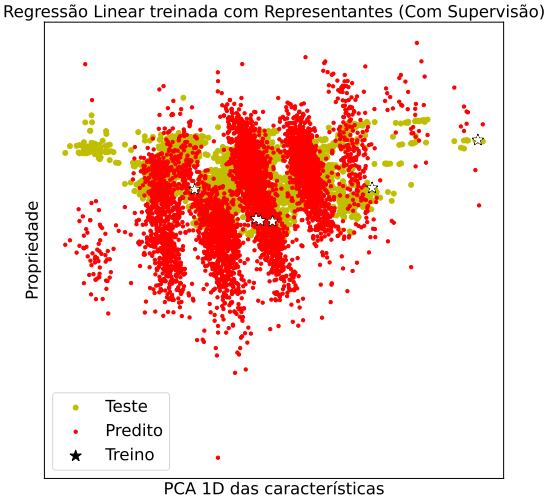
Fonte: o autor

Figura 64 – Regressão Linear do conjunto QM9 não ponderado



Fonte: o autor

Figura 65 – Regressão Linear do conjunto QM9 ponderado



Fonte: o autor