IMPLEMENTAÇÃO DE INTELIGÊNCIA ARTIFICIAL UTILIZANDO K-MEANS NO AGRUPAMENTO DE VIDROS

Marcelo Pontes Rodrigues[1]; Marcelo de Oliveira Rego[2]
[1] Programa de Pós-Graduação em Tecnologias Aplicadas a Animais de Interesse Regional–UFPI;
[2]Programa de Doutorado em Cência da Computação Associação UFMA/UFPI/DCCMAPI/CCET
[1]marcelo.rodrigues@ufpi.edu.br; [2]oliveira@ufdpar.edu.br;

Resumo

A pesquisa apresenta a aplicação do algoritmo de agrupamento não supervisionado K-means no conjunto de dados "*Glass Identification*", disponível no UCI Machine Learning Repository (Dua e Graff 2019). O objetivo é explorar padrões químicos e físicos em amostras de vidro, identificando agrupamentos baseados em sua composição química e propriedades. A abordagem adotada inclui a normalização dos dados utilizando o método MinMaxScaler, seleção de clusters com base no índice de Davies-Bouldin (Davies e Bouldin 1979) e análise de atributos relevantes para a formação dos grupos. Os resultados indicam a eficácia do K-means (MacQueen 1967) em identificar padrões ocultos e diferenciar composições específicas.

Palavras-chave: vidro; clusterização; machine learning; aprendizado de máquina; k-means.

INTRODUÇÃO

O objetivo desta pesquisa é desenvolver um modelo de agrupamento não supervisionado de vidros com base em sua composição química. Utilizou-se o conjunto de dados "*Glass Identification*"do repositório UCI Machine Learning (Dua e Graff 2019), que contém informações detalhadas sobre a composição de diferentes tipos de vidros, como os usados em janelas e em ambientes especiais. Originalmente criado para fornecer uma base confiável e padronizada para investigações forenses e estudos em engenharia de materiais, este conjunto de dados permite a classificação automatizada de amostras, o que pode auxiliar na identificação rápida e precisa de materiais, economizando tempo e recursos. Neste estudo, exploramos o K-means (MacQueen 1967) para a análise e agrupamento dos tipos de vidro, com base nas propriedades químicas e físicas do material. O uso do K-means é especialmente adequado, pois este algoritmo é capaz de identificar grupos e padrões ocultos em dados multidimensionais, uma característica essencial para analisar a diversidade nas composições químicas deste dataset.

METODOLOGIA

Inicialmente, realizamos a normalização dos dados, plotamos e analisamos os gráficos (Figura1) bem como o índice de Davies Bulduin (Davies e Bouldin 1979) para diferentes quantidades de clusters (Tabela 1) sendo esta informação considerada para decidirmos por clusters igual a seis(k= 6) próximo do valor indicado pelo método do cotovelo (Thorndike 1953) e terceiro melhor valor indicado pelo gráfico da silhueta (Rousseeuw 1987).

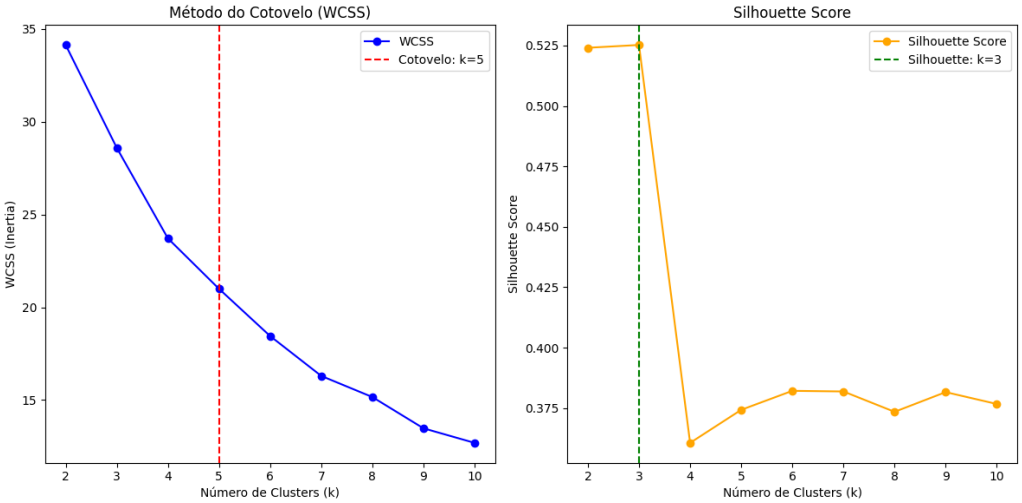

Figura 1: Método do Cotovelo e Silhueta Score

Quantidade	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**
DBI	1.2312	1.2060	1.0066	0.9679	1.0261

Tabela 1: Tabela do Índice de Davies-Buldin (DBI) para diferentes quantidades de clusters

Aplicamos o K-means que criou os seis clusters (Tabela 2).

Clusters	**0**	**1**	**2**	**3**	**4**	**5**
Quantidade	24	19	99	2	30	40

Tabela 2: Quantidade de elementos por clusters

O estudo foi conduzido com o objetivo de analisar os clusters formados pelo algoritmo K-means. Para isso, utilizamos gráficos de boxplot, que permitiram explorar a relação entre as variáveis e os clusters. Na (Figura 2), por exemplo, é apresentado o gráfico que mostra a distribuição do atributo RI em relação aos clusters.

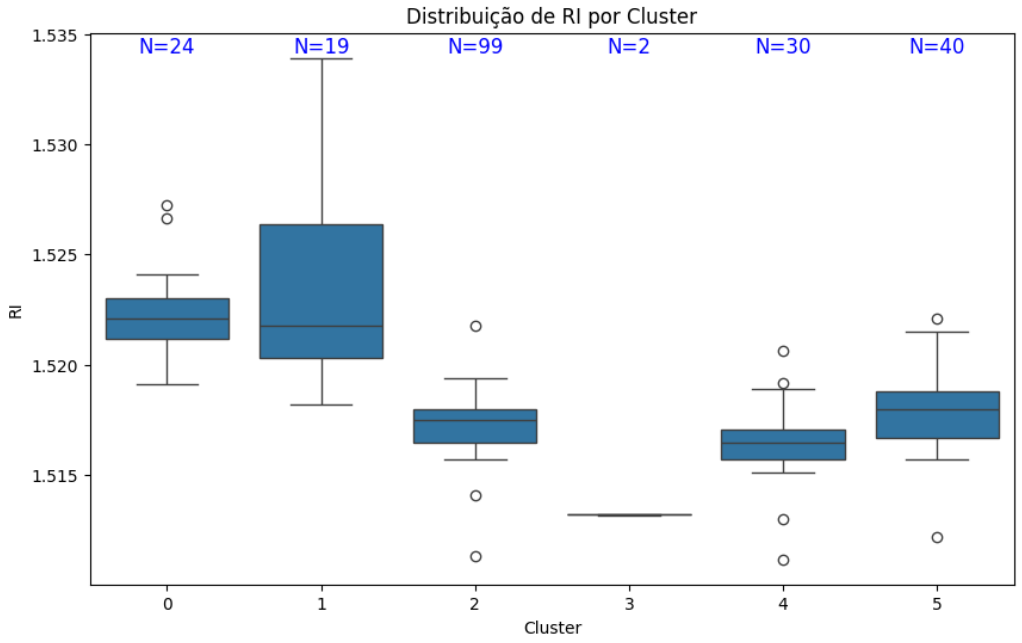

Figura 2: Variabilidade do atributo RI por clusters

Observamos na figura acima que a variabilidade do atributo RI entre os elementos alocados no cluster 2 é a menor em comparação com os demais clusters e a que deixa menos autliers proporcionalmente. Esse comportamento sugere que o atributo RI desempenhou um papel significativo na alocação dos elementos nesse cluster, então associamos de forma excludente o atributo ao cluster utilizando por critério essa "menor variabilidade". Fizemos o mesmo estudo para todos os atributos e produzimos a (Tabela 3) abaixo que mostra quais atributos nesta nossa perspectiva foram de maior importância para cada cluster.

Clusters	**0**	**1**	**2**	**3**	**4**	**5**
Atributos importantes	RI>1520	Ba	RI<1520, Al, Si, Ca	(todas)- Si	Mg, K, Fe	Na

Tabela 3: Atributos julgados mais relevantes por cluster.

O estudo foi implementado na linguagem python e está disponível no repositório do GitHub (Pontes e Rego 2024)

RESULTADOS E DISCUSSÕES

A aplicação do K-means resultou na formação de seis clusters distintos, permitindo a identificação de padrões específicos relacionados aos atributos do conjunto de dados. Conforme descrito na (Tabela 2), foi possível inferir os atributos mais relevantes para a composição de cada cluster:

O **cluster 0** destacou-se por reunir elementos com altos valores de RI (>1520), sugerindo que esse atributo foi crucial na segmentação desses dados. O **cluster 1** apresentou uma associação significativa com o atributo Ba, indicando que a presença desse elemento é predominante nesse grupo. O **cluster 2** foi definido por uma combinação de atributos, com Al, Si e Ca desempenhando papéis importantes, o que pode refletir interações químicas específicas entre esses componentes. O **cluster 3** apresentou uma dependência mais abrangente, com relevância de quase todos os atributos, exceto Si, apontando para uma composição mais variada. O **cluster 4** agrupou elementos caracterizados por baixos valores de RI (<1520) e a relevância de Mg, K e Fe, indicando que essas propriedades tiveram impacto na formação do grupo. Por fim, o **cluster 5** destacou-se pela predominância do atributo Na, indicando sua importância única na segregação desse cluster. Essas inferências destacam a capacidade do K-means em identificar padrões complexos e diferenciar grupos com base em combinações específicas de atributos. Esse tipo de análise é fundamental para explorar a diversidade química do material estudado, permitindo insights detalhados sobre como as propriedades individuais contribuem para a segmentação do conjunto de dados.

CONCLUSÃO

Os resultados destacam a utilidade do K-means para explorar dados complexos e multidimensionais, proporcionando insights valiosos sobre padrões químicos e físicos do material estudado.

Referências

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, 1979.

DUA, D.; GRAFF, C. *UCI Machine Learning Repository*. 2019. Irvine, CA: University of California, School of Information and Computer Science. Disponível em: <https://archive.ics.uci.edu/ml/index.php>.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, v. 1, n. 14, p. 281–297, 1967.

PONTES, M.; REGO, M. *Repositório de Inteligência Artificial utilizando Árvore de Decisão no Diagnóstico Câncer de Mama*. 2024. Acessado em: 21 de novembro de 2024. Disponível em: <https://github.com/infopontes/inteligencia_artificial/>.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, n. 1, p. 53–65, 1987.

THORNDIKE, R. L. Who belongs in the family? *Psychometrika*, v. 18, n. 4, p. 267–276, 1953.