IMPLEMENTAÇÃO DE INTELIGÊNCIA ARTIFICIAL UTILIZANDO AGRUPAMENTO HIERÁRQUICO NO DATASET DAS BANDEIRAS

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

Este trabalho apresenta um estudo sobre o uso do Agrupamento Hierárquico aplicado ao dataset Flags, disponível no *UCI Machine Learning Repository* (Dua e Graff 2019), que contém 194 registros e 30 atributos categóricos, numéricos e binários, relacionados às características de bandeiras de diferentes nações. Esses atributos incluem cores, formas geométricas, presença de símbolos específicos, além de variáveis categóricas como idioma e religião. O objetivo principal foi identificar grupos de países com características semelhantes em suas bandeiras. A análise utilizou o método de Ward (1963) e a métrica de distância euclidiana, resultando na formação de quatro clusters significativos. As distinções foram evidenciadas por meio de dendrogramas, e variáveis como Saltires, Population e Area destacaram-se na formação dos agrupamentos, conforme confirmado pela análise estatística ANOVA (Fisher 1925) e visualizações em histogramas. Os resultados demonstram a eficácia do Agrupamento Hierárquico na exploração de dados complexos e multidimensionais, oferecendo *insights* valiosos sobre padrões estruturais, culturais e regionais representados nas bandeiras.

Palavras-chave: bandeiras; clusterização; machine learning; agrupamento hierárquico.

INTRODUÇÃO

A pesquisa de agrupamento (clustering) é uma técnica amplamente utilizada em aprendizado de máquina para identificar padrões e agrupar dados com base em suas similaridades.

Este estudo analisa o dataset Flags, disponível no UCI Machine Learning Repository (Dua e Graff, 2019), que explora as características das bandeiras de várias nações. O conjunto de dados reúne 194 registros e 30 atributos, englobando variáveis categóricas, como idiomas e religiões, e variáveis numéricas, como o número de cores e elementos geométricos nas bandeiras, além de variáveis binárias, todos sem valores ausentes.

Neste trabalho, utilizamos o Agrupamento Hierárquico, uma abordagem de aprendizado não supervisionado que estrutura os dados em uma hierarquia de grupos. Esta metodologia permite uma representação visual das relações de similaridade entre os dados de forma escalonada, comumente ilustrada através de dendrogramas (Jain e Dubes 1988).

METODOLOGIA

O estudo teve como objetivo principal aplicar técnicas de agrupamento hierárquico (Müller e Guido 2016) para identificar padrões relevantes entre as bandeiras, permitindo a formação de *clusters* que representem características culturais, regionais ou históricas comuns.

Neste processo envolveu várias etapas metodológicas, incluindo a aplicação do método de agrupamento, visualização dos resultados por meio de dendrograma, gráficos, e análise detalhada das variáveis mais influentes em cada *cluster*.

O método de *Ward* foi utilizado por ser eficiente para minimizar a variância dentro dos *clusters* formados (Ward, 1963).

O dendrograma foi gerado utilizando a função “dendrogram” da biblioteca SciPy (Developers 2024), como descrito na documentação oficial. A técnica de agrupamento hierárquico é uma das abordagens essenciais para análise de dados (VanderPlas 2016) .

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

RESULTADOS E DISCUSSÕES

A importância desta pesquisa está em sua capacidade de identificar padrões ocultos nos dados, proporcionando novos insights que podem não ser imediatamente evidentes. O método permitiu a descoberta de relações subjacentes que podem ser relevantes para esse dataset a partir do agrupamento hierarquico foi gerado o dendrograma Figura 1, que mostra os clusters gerados e uma explicação simplicada da features mais importantes para as divisões realizadas.

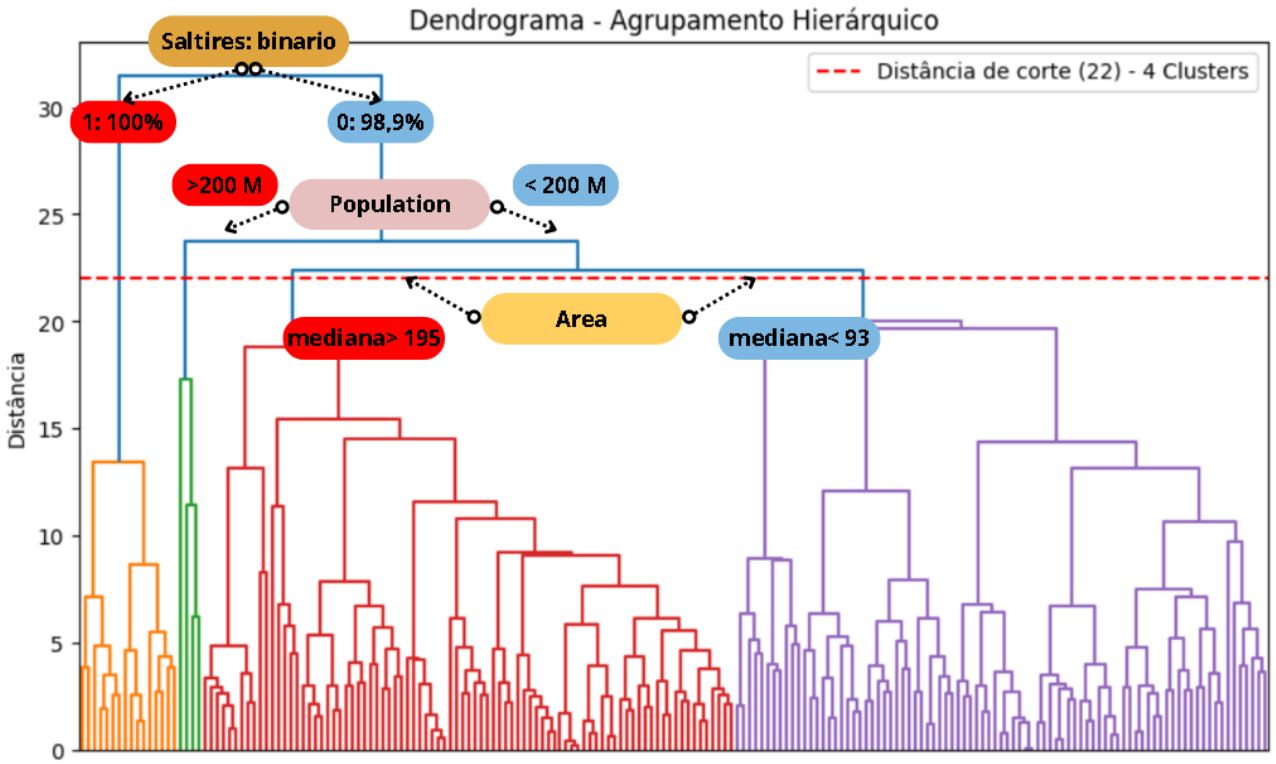

Figura 1: Dendrograma do Dataset Flags - Este dendrograma exibe a estrutura de agrupamento do dataset Flags, identificando quatro clusters principais no ponto de corte com distância de 22. Os clusters são visualmente diferenciados por cores: Cluster 1 em laranja, Cluster 2 em verde, Cluster 3 em vermelho e Cluster 4 em lilás. Observa-se que a variável "Saltires"teve um papel crucial na primeira bifurcação dos clusters. Subsequentemente, "Population"promoveu uma divisão adicional, seguida por "Area", que definiu a separação final dos grupos.

Para identificar as features mais influentes na distinção dos clusters formados pelo agrupamento hierárquico, empregamos a análise de variância (ANOVA). Esta técnica estatística, amplamente utilizada em estudos de clustering, avalia se existem diferenças significativas entre as médias dos grupos. Assim, a ANOVA permite determinar quais variáveis têm maior impacto na formação dos grupos, proporcionando insights valiosos sobre os padrões observados nos dados (Fisher 1925).

Realizamos um teste ANOVA que classificou as variáveis em ordem decrescente de importância, com base nos seguintes valores de F-estatística: Saltires (459.79), Crosses (89.09), Language (32.66), Topleft (13.49), Quarters (22.65), Religion (15.94), e Area (62.14). A análise visual dessas características, por meio de histogramas, confirma sua significativa influência na segregação dos clusters como pode ser demonstrado no histograma Figura 2.

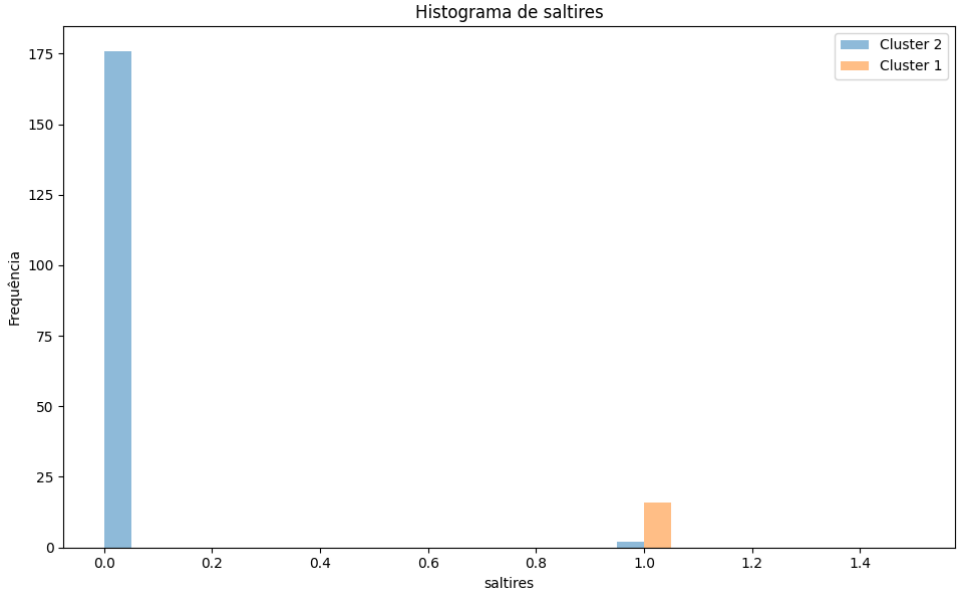

Figura 2: Histograma de Saltires - Ilustra a relevância desta variável na determinação inicial dos clusters.

CONCLUSÃO

Os resultados evidenciam a eficácia do Agrupamento Hierárquico na exploração de dados complexos e multidimensionais, permitindo identificar padrões estruturais no conjunto de dados e oferecendo insights valiosos sobre similaridades culturais e regionais representadas nas bandeiras. Concluímos que a análise do dendrograma, combinada com ferramentas estatísticas e representações visuais, como histogramas, é essencial para avaliar a qualidade dos agrupamentos e validar a consistência dos clusters identificados.

Neste estudo, foram gerados quatro clusters significativos com uma distância superior a 20. Dois desses clusters apresentaram ”braços”longos no dendrograma, destacando sua clara distinção em relação aos demais, mesmo com o cluster 2 sendo mais próximo (ou ”irmão”) dos clusters 3 e 4. Por outro lado, os clusters 3 e 4 exibem maior proximidade, evidenciada pelo menor tamanho de seus braços no dendrograma, como ilustrado na Figura 1.

Referências

DEVELOPERS, S. *Hierarchical clustering*. 2024. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html>. Accessed: 2024-12-08.

DUA, D.; GRAFF, C. *UCI Machine Learning Repository - Flags Data Set*. 2019. Acessado em: 08/12/2024. Disponível em: <https://archive.ics.uci.edu/ml/datasets/Flags>.

FISHER, R. A. Statistical methods for research workers. *Oliver and Boyd*, p. 43–57, 1925.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. [S.l.]: Prentice-Hall, Inc., 1988.

MüLLER, A. C.; GUIDO, S. *Introduction to Machine Learning with Python*. Sebastopol, CA: O'Reilly Media, 2016. ISBN 978-1-4493-9786-4.

PONTES, M.; REGO, M. *Repositório de Inteligência Artificial utilizando Árvore de Decisão no Diagnóstico Câncer de Mama*. 2024. Acessado em: 10 de dezembro de 2024. Disponível em: <https://github.com/infopontes/inteligencia_artificial/>.

VANDERPLAS, J. *Python Data Science Handbook*. Sebastopol, CA: O'Reilly Media, 2016. ISBN 978-1-4493-8289-0.