# TÉCNICAS DE CLUSTERIZAÇÃO PARA PAÍSES POR INDICADORES DE CRESCIMENTO VERDE E CONTEXTO SOCIOECONÔMICO A PARTIR DE DADOS DE 2019 DA OCDE

Matheus Santos Dias (diassmatheus@outlook.com)

***Resumo:*** *Esse estudo se propõe a aplicar os algoritmos de k-means e agrupamento hierárquico aglomerativo, com o número de clusters definido a partir dos coeficientes de silhueta, para agrupar países de acordo com os indicadores de crescimento verde presentes na base de dados Green Growth da OCDE, com corte transversal para o ano de 2019, e por fim discutir os resultados levando em conta o contexto geopolítico dos países e os pontos de similaridades e de dissimilaridades entre cada cluster.*

***Palavras-chave:*** *Mineração de Dados; Machine Learning; Clusterização; Crescimento Verde.*

## 1. Introdução

Em um mundo cada vez mais informatizado, os dados se tornaram componentes essenciais a ponto de serem comumente anunciados como "o novo petróleo" por executivos e comunicadores, que fazem referência à frase original *"data is the new oil"* criada por Clive Humby, um matemático londrino especializado em ciência de dados.

No entanto, sem uma sistematização que extraia informações úteis, esses dados são de pouca valia. E é justamente isso que o processo de descoberta de conhecimento em bases de dados (em inglês, *Knowledge Discovery in Databases* - KDD) propõe. Os algoritmos de reconhecimento de padrões, que fazem parte do processo KDD, são capazes de realizar inúmeras tarefas visando diferentes objetivos (BHARATI; RAMAGERI, 2010). Entre tais tarefas, pode-se citar a generalização, classificação, agrupamento de dados, associação, visualização de dados, entre outras (JOTHI; RASHID; HUSAIN, 2015).

O agrupamento de dados, ou clusterização, é uma categoria de técnicas de aprendizado não supervisionado, pois utiliza dados não rotulados para o reconhecimento de padrões (ZENGIN et al., 2011), que nos permite descobrir estruturas ocultas em dados onde não sabemos a resposta certa antecipadamente. O objetivo destas técnicas é encontrar padrões e formar agrupamentos naturais de dados de forma que os itens no mesmo grupo sejam mais semelhantes entre si do que aqueles de grupos diferentes (RASCHKA, 2015).

Este artigo tem como objetivo aplicar algoritmos de agrupamento *k-means* e agrupamento hierárquico em uma base de dados de Indicadores de crescimento verde

da OCDE (Organização para a Cooperação e Desenvolvimento Econômico), com corte transversal para o ano de 2019, e analisar seus resultados.

## 2. Fundamentação Teórica

O algoritmo *k-means* é computacionalmente muito eficiente em comparação com outros algoritmos de agrupamento, além de extremamente fácil de implementar, o que pode explicar sua popularidade. Pertence à categoria de agrupamento baseado em protótipo. O agrupamento baseado em protótipo significa que cada grupo é representado por um protótipo, que pode ser o centróide (média) de pontos semelhantes com atributos contínuos ou o medóide (o ponto mais representativo ou de ocorrência mais frequente) no caso de atributos categóricos (RASCHKA, 2015).

Embora *k-means* seja acurado em identificar clusters de forma esférica, uma das desvantagens desse algoritmo de agrupamento é que temos que especificar o número de grupos k a priori. Uma escolha inadequada para k pode resultar em baixo desempenho (RASCHKA, 2015).

O procedimento iterativo do método *K-means* pode ser resumido pelas quatro etapas a seguir:

1. Escolhe aleatoriamente k centróides dos pontos de amostra como centros iniciais do grupo.
2. Atribuí cada amostra ao centróide mais próximo.
3. Move os centróides para o centro das amostras que foram atribuídas a ele.
4. Repete as etapas 2 e 3 até que a atribuição do cluster não mude ou uma tolerância definida pelo usuário ou um número máximo de iterações seja alcançado.

Outra tarefa de agrupamento que foi utilizada e que também é baseada em protótipo é o agrupamento hierárquico. As duas principais abordagens para agrupamento hierárquico são agrupamento hierárquico aglomerativo e divisivo. No agrupamento hierárquico divisivo, começamos com um cluster que abrange todas as amostras e dividimos iterativamente o cluster em clusters menores até que cada cluster contenha apenas uma amostra (RASCHKA, 2015). Neste trabalho será utilizado o agrupamento aglomerativo, que adota a abordagem oposta. Começamos com cada amostra como um cluster individual e mesclamos os pares mais próximos de clusters até que apenas um cluster permaneça (RASCHKA, 2015). Seu procedimento iterativo pode ser resumido pelas seguintes etapas:

1. Calcula a matriz de distância de todas as amostras.
2. Representa cada ponto de dados como um cluster único.
3. Mescla os dois clusters mais próximos com base na distância dos membros mais dissimilares (distantes).

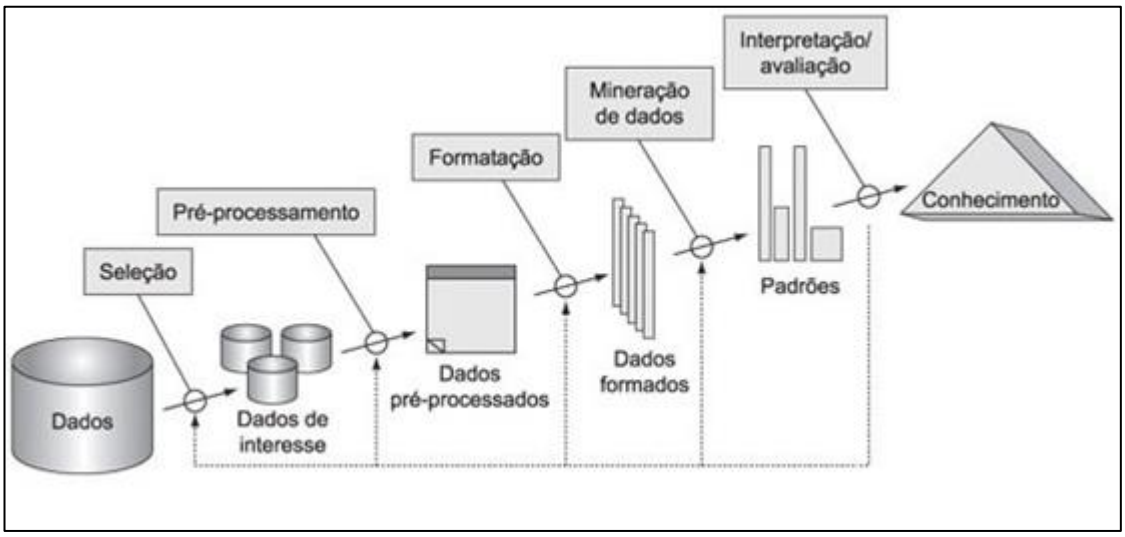
4. Atualiza a matriz de similaridade.
5. Repete as etapas 2 a 4 até que um único cluster permaneça.

## 3. Materiais e Métodos

Para a realização deste trabalho foi utilizada a linguagem de programação Python em sua versão 3.9, o ambiente de desenvolvimento integrado (IDE, Integrated Development Environment) Jupyter Notebook, o formato de troca de dados JavaScript Object Notation (JSON) e as bibliotecas de software Pandas, Matplotlib, Seaborn, NumPy, Scikit-learn, SciPy e Plotly.

O KDD é composto por cinco etapas. Inicialmente deve-se selecionar os dados de interesse. Em seguida é necessário o pré-processamento para limpar o conjunto de dados de ruídos ou valores faltantes. Na sequência é realizada a transformação dos dados já processados, com procedimentos como a normalização ou a padronização, por exemplo, que contribuem para um melhor desempenho do algoritmo de mineração. Aplica-se então o algoritmo de mineração de dados. Por fim, os resultados obtidos através do algoritmo são interpretados e avaliados. Um resumo deste processo pode ser observado na Figura 1.

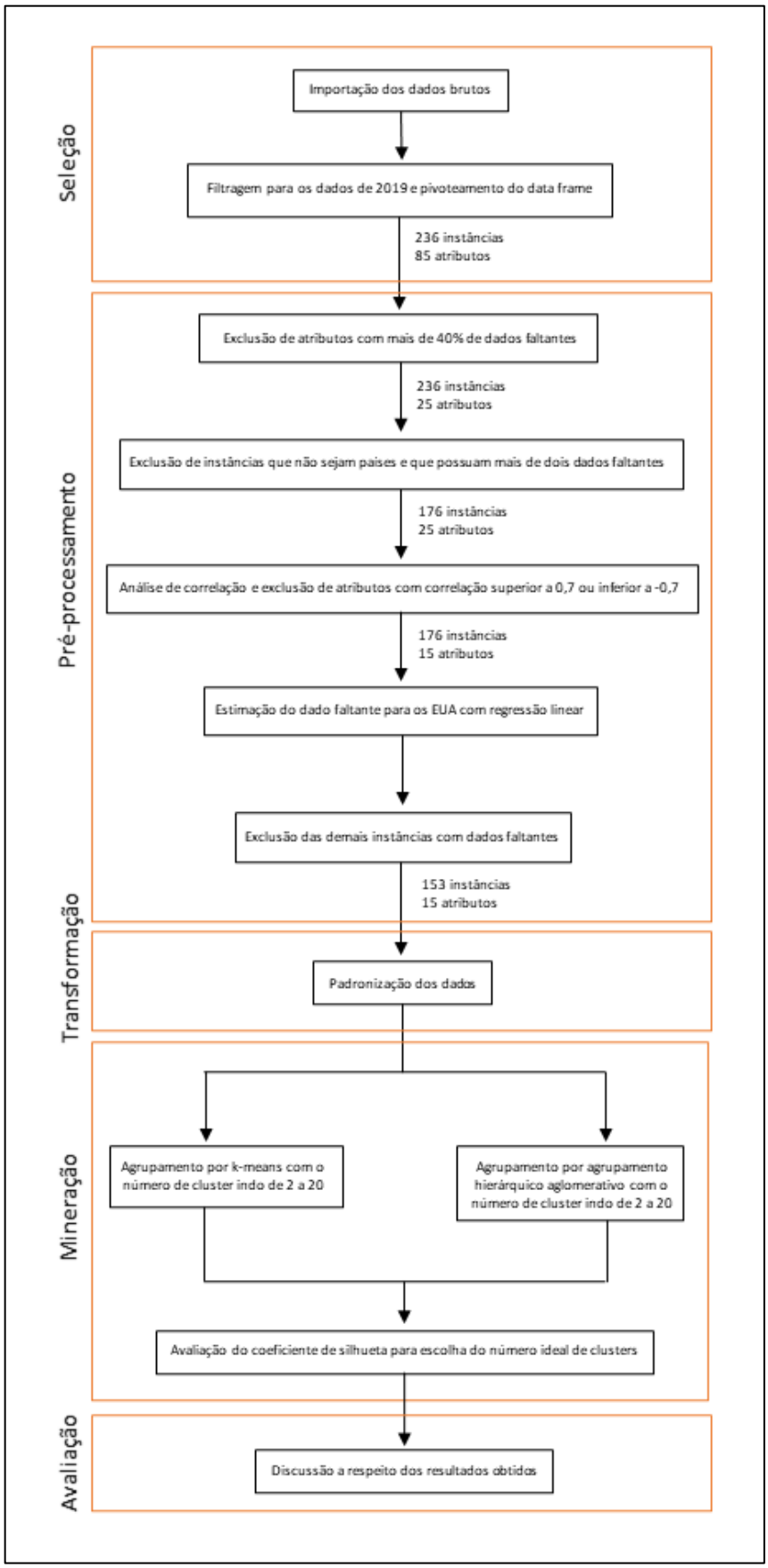
Figura 1 – O processo de descoberta de conhecimento em bases de dados




Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996)

As etapas realizadas neste trabalho podem ser vistas na Figura 2.

Figura 2 – Fluxograma das atividades realizadas neste trabalho



O banco de dados utilizado foi o Green Growth da OCDE que contém indicadores selecionados para monitorar o progresso em direção ao crescimento verde, para apoiar a formulação de políticas públicas e informar o público em geral. O banco sintetiza dados e indicadores em uma ampla gama de domínios. Seus indicadores foram selecionados de acordo com critérios bem especificados e incorporados em uma estrutura conceitual, que é distribuída em torno de quatro eixos para capturar as características principais de crescimento verde, sendo eles “produtividade ambiental e de recursos”, “base de ativos naturais”, “dimensão ambiental da qualidade de vida”, “oportunidades econômicas e respostas de políticas públicas”, além de indicadores de contexto socioeconômico dos países(OECD, 2020).

A base originalmente possui dados desde o ano de 1990, no entanto para a tarefa de agrupamento optou-se por um corte transversal para o ano de 2019. Após a importação dos dados e da filtragem apenas para o ano de 2019, um data frame com 236 instâncias (países, blocos econômicos, fóruns internacionais ou agrupamentos geográficos) e 85 atributos (indicadores) foi estruturado a partir da biblioteca Pandas.

O primeiro problema identificado a ser contornado na fase de pré-processamento foi a alta quantidade de dados faltantes, 12.189 no total, ou seja, mais de 60% dos dados. Foi adotado como critério a exclusão dos atributos com menos de 60% de dados válidos, dessa forma o data frame passou de 85 para 25 atributos. Dentre as instâncias foram excluídas todas que não eram referentes a países e também as que possuíam mais de dois dados faltantes, fazendo com que o número de instâncias caísse de 236 para 176.

Depois disso realizou-se a análise de correlações entre os atributos a partir do método de Pearson. Sempre que dois atributos apresentaram um coeficiente de correlação superior a 0,7 ou inferior a -0,7 um dos dois foi excluído. Ao final o número de indicadores caiu de 25 para 15.

Neste ponto constatou-se que os Estados Unidos da América ainda possuíam um dado faltante para o indicador “Emissões de CO2 de transporte aéreo por unidade do PIB”. Um dos critérios adotados nessa etapa de pré-processamento foi a não exclusão de países com grande importância econômica ou política no cenário global, por tanto optou-se pela estimação desse dado para os EUA. A série histórica de 2013 a 2018 dos EUA foi levantada para esse indicador e, a partir desses dados, com a biblioteca Scikit-learn, foi feita a regressão linear para estimar o dado faltante. Posteriormente todas as demais instâncias que ainda tinham dados faltantes foram excluídas, sendo que o data frame ficou com a dimensão final de 153 instâncias e 15 atributos. Os atributos remanescentes podem ser observados na Tabela 1. A descrição detalhada de cada

indicador e informações de como foram calculados, podem ser acessados no web site https://www.oecd-ilibrary.org/ no Dataset Green growth indicators.

**Tabela 1** – Indicadores utilizados nas tarefas de agrupamento

| **Indicador** | **Eixo conceitual** |
|---|---|
| Emissões de CO2 de transporte aéreo por unidade do PIB | Produtividade ambiental e de recursos |
| Mudança da temperatura anual da superfície, desde 1951-1980 | Base de ativos naturais |
| Mortalidade por exposição ao ozônio | Dimensão ambiental da qualidade de vida |
| Mortalidade por exposição ao chumbo | Dimensão ambiental da qualidade de vida |
| Mortalidade por exposição a PM2.5 | Dimensão ambiental da qualidade de vida |
| Exposição média da população a PM2.5 | Dimensão ambiental da qualidade de vida |
| Porcentagem da população exposta a mais de 10 µg / $m^3$ | Dimensão ambiental da qualidade de vida |
| Mortalidade por exposição ao radônio residencial | Dimensão ambiental da qualidade de vida |
| PIB real, índice 2000 = 100 | Contexto socioeconômico |
| PIB real per capita | Contexto socioeconômico |
| População | Contexto socioeconômico |
| Densidade populacional, habitantes por $km^2$ | Contexto socioeconômico |
| População por faixa etária 15 a 64 anos, % do total | Contexto socioeconômico |
| Rede de migração | Contexto socioeconômico |
| Mulheres, % da população total | Contexto socioeconômico |

Com a biblioteca Seaborn foram plotados boxplots para todos esses atributos, afim de uma melhor compreensão da distribuição dos dados. Verificou-se que o conjunto de dados possuía um alto número de outliers.

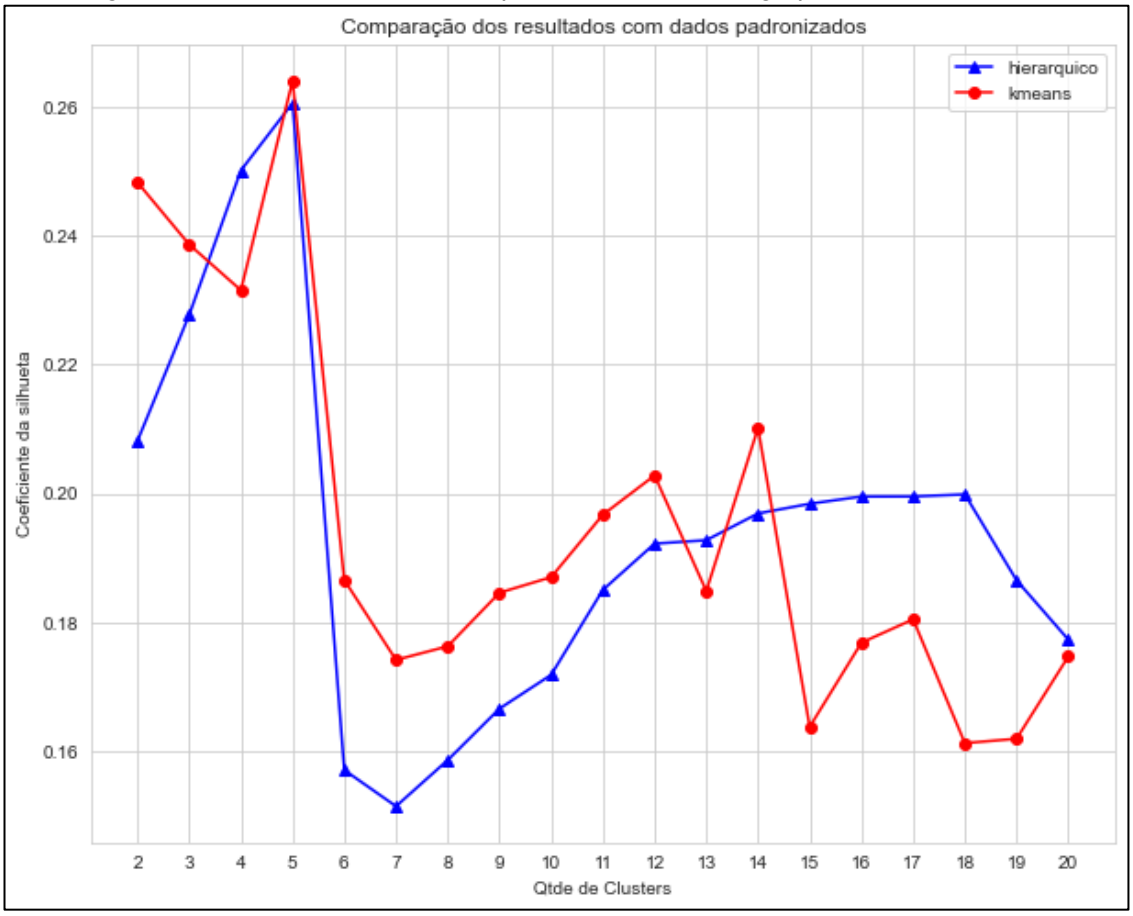
Essa análise foi importante para a fase seguinte do KDD, a transformação dos dados. Considerando que os indicadores possuem ordens de grandeza diferentes, a normalização ou padronização dos dados é uma pratica para evitar que o algoritmo fique enviesado para as variáveis que possuam maior ordem de grandeza. Foi levado em consideração que devido ao alto número de outliers, a normalização poderia achatar a real variabilidade dos dados. Optou-se, portanto, pela padronização dos dados através da função StandardScaler da biblioteca Scikit-learn. Outro insight que a análise de distribuição dos dados proporcionou foi que o método DBScan, que encontra clusters

baseado na densidade de observações em determinada região, muito provavelmente não seria uma boa opção para agrupar esse conjunto de dados.

## 4. Resultados e Discussão

Com os dados padronizados, foi realizada a tarefa de clusterização pelos métodos k-means e agrupamento hierárquico aglomerativo. Como citado na seção de Fundamentação Teórica, uma das limitações do método k-means é ter que definir a priori o número de clusters. Foi gerado o coeficiente de silhueta para ambos os métodos com o número de clusters variando de 2 a 20, vide Figura 3. Constatou-se que o número ideal de clusters para esse conjunto de dados em ambos os métodos é igual a cinco, e é aos agrupamentos com esse número de clusters que as posteriores considerações farão referência.

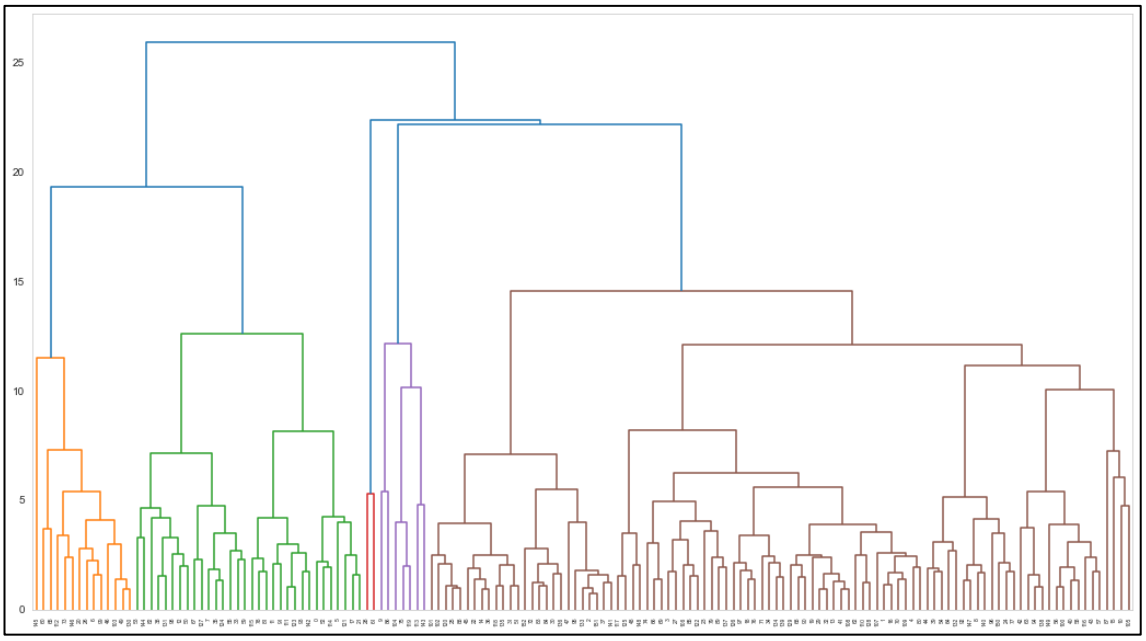
Figura 3 – Coeficientes de silhueta para os métodos de agrupamento




A partir da biblioteca scikit-learn foi criado o dendrograma da clusterização através do agrupamento hierárquico. Dendrograma é um diagrama de árvore que exibe os grupos formados por agrupamento de observações em cada passo e em seus níveis de

similaridade. O nível de similaridade é medido ao longo do eixo vertical e as diferentes observações são listadas ao longo do eixo horizontal.

Figura 4 – Dendrograma da clusterização por agrupamento hierárquico




E para melhor visualização dos resultados obtidos através desses algoritmos de clusterização, foi plotado no mapa-mundi a distribuição dos 5 clusters obtidos pelos dois métodos.

Figura 5 – Clusters obtidos por k-means plotados no mapa-mundi

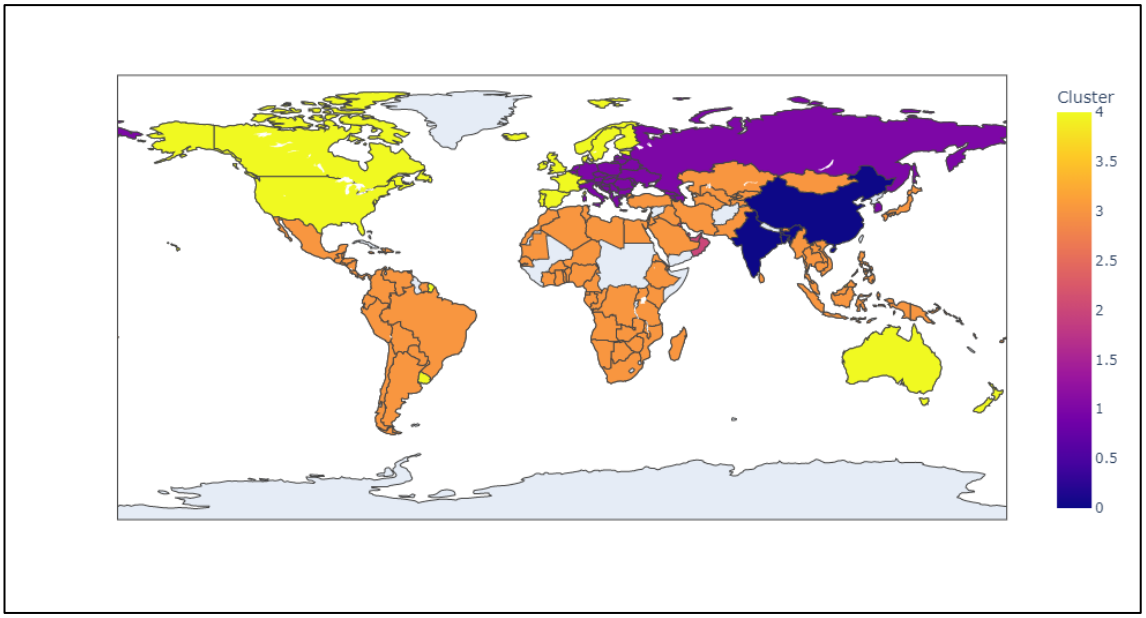
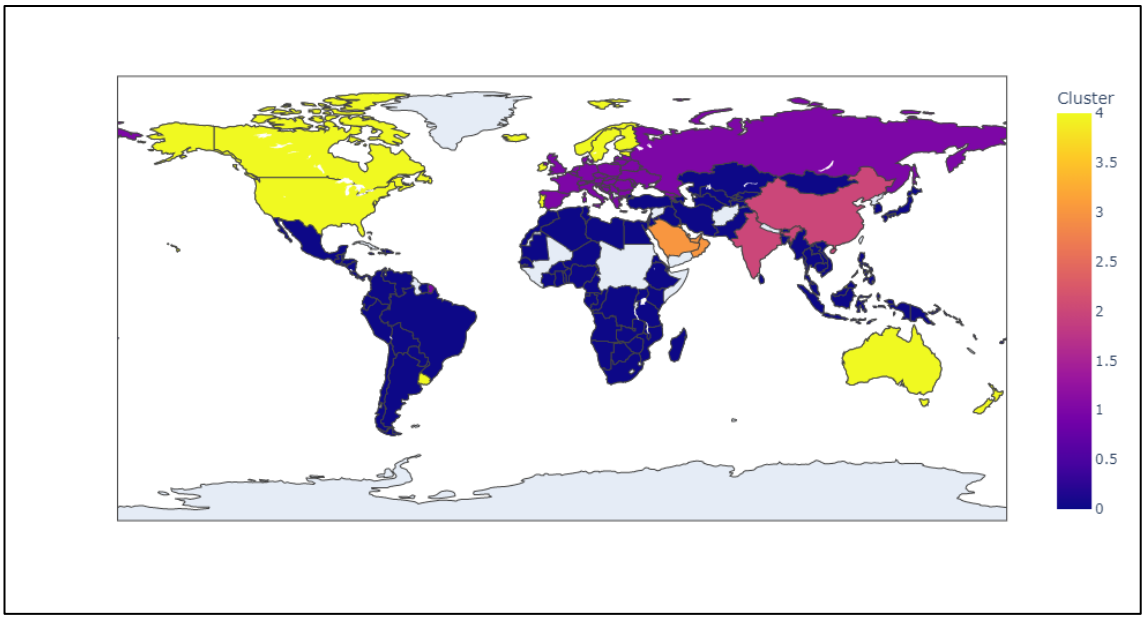


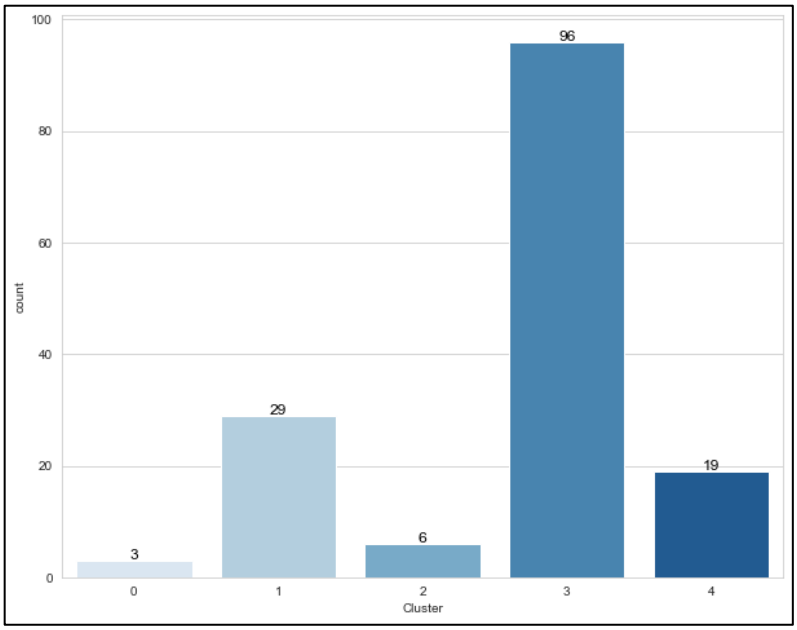
Figura 6 – Clusters obtidos por agrupamento hirárquico plotados no mapa-mundi




Nota-se que apesar de algumas mudanças para países da Europa e do Oriente Média, a maioria dos países foi agrupado da mesma forma em ambos os métodos. Todas as demais considerações de discussão dos resultados serão a respeito do agrupamento pelo método k-means.

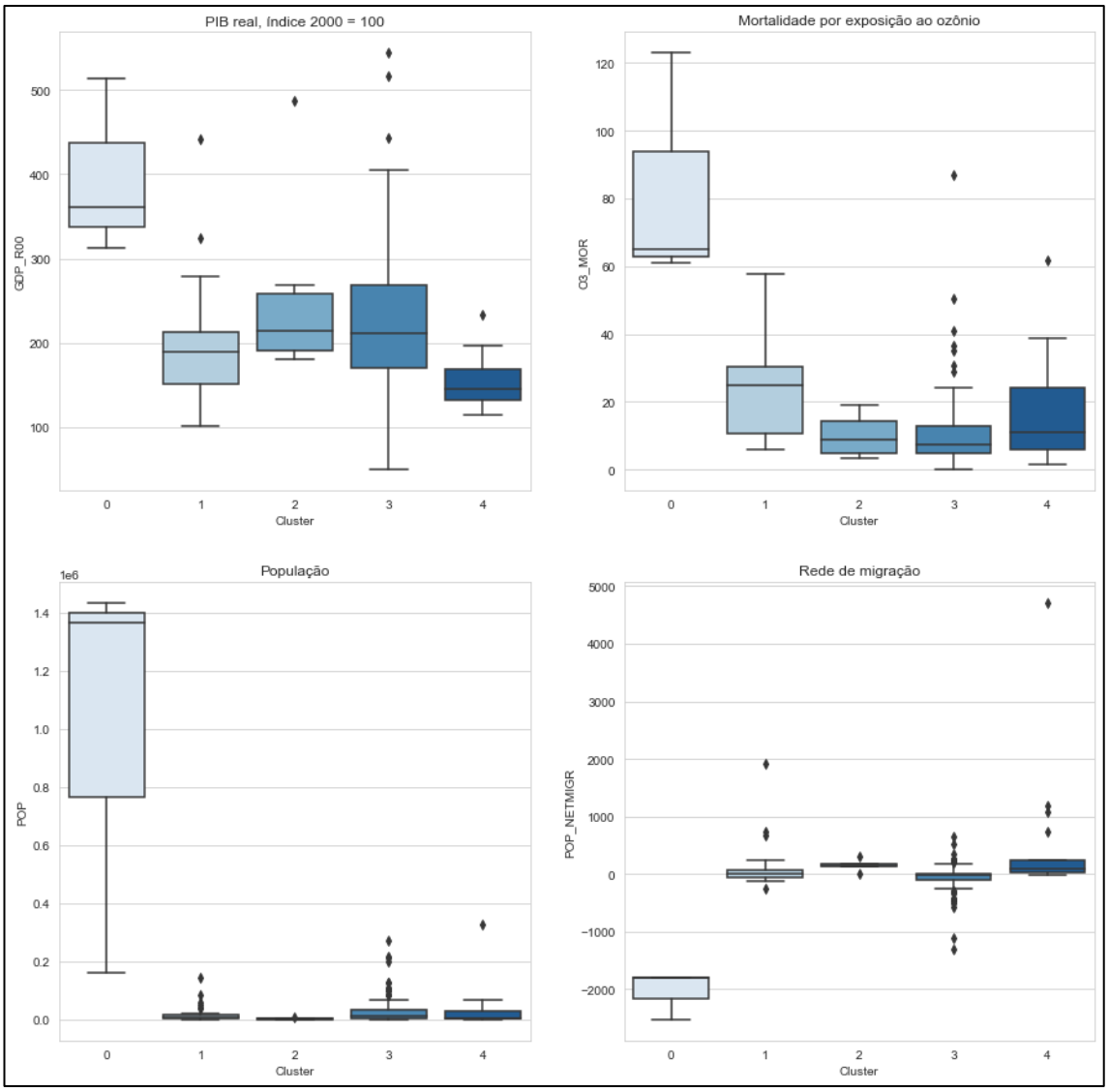
Uma aferição possível a partir da distribuição dos clusters no mapa-mundi e que se confirma na realidade é o desbalanceamento da quantidade de países por cluster. A distribuição dessas quantidades pode ser observada na Figura 7.

Figura 7 – Quantidade de países por cluster



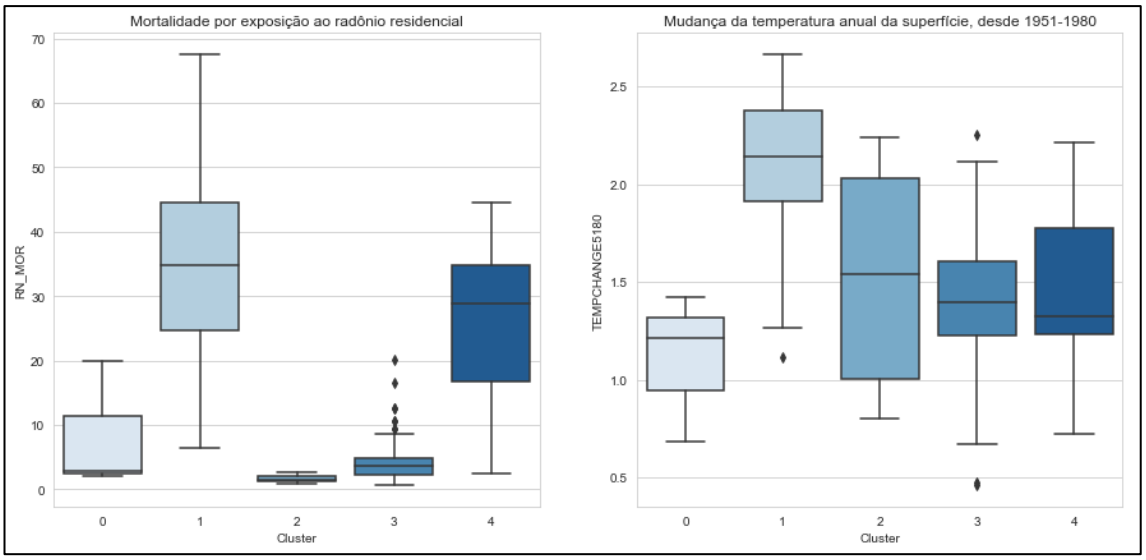
O cluster 0, que possui a menor quantidade de países, é composto por China, Índia e Bangladesh. Esses países asiáticos além apresentarem similaridades demográficas, possuem todos uma alta taxa de crescimento econômico. Bangladesh, que já foi considerado o país menos desenvolvido do mundo em termos econômicos per capita, graças a um boom econômico, associado a avanços em educação e saúde pública e um menor índice de vulnerabilidade, deve se livrar, até 2024, do selo de Países Menos Desenvolvidos (PMD) da Organização das Nações Unidas (ONU). Ademais, os países do cluster 0 apresentam uma alta mortalidade por exposição ao ozônio no nível do solo quando comparados com os outros clusters. A taxa de rede de migração, que representa o número de imigrantes menos o número de emigrantes, nesse cluster é menor que a dos demais, o que indica uma maior emigração nesses países. Algumas comparações podem ser observadas na Figura 8.

Figura 8 – Boxplots de indicadores de interesse para o cluster 0



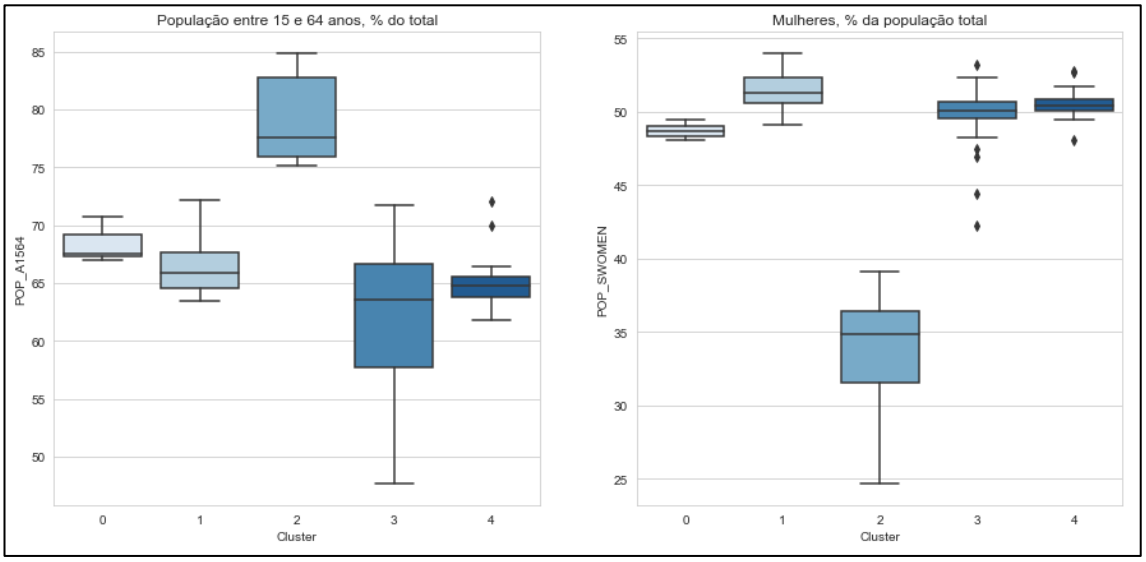
O cluster 1 é composto por 29 países, sendo a maioria deles ex-membros da extinta URSS (União das Repúblicas Socialistas Soviéticas), com algumas exceções como Grécia e Coréia, por exemplo. Muitos destes países enfrentam uma realidade ambiental muito deteriorada desde a queda da URSS.

Figura 9 – Boxplots de indicadores de interesse para o cluster 1




O cluster 2 é composto por seis países, sendo eles Catar, Emirados Árabes Unidos, Bahrein, Kuwait, Omã e Maldivas. Com exceção das Maldivas, todos os países desse cluster são do Oriente Médio. De acordo com os indicadores analisados, os maiores pontos de similaridade entre esses países são de cunho demográfico.

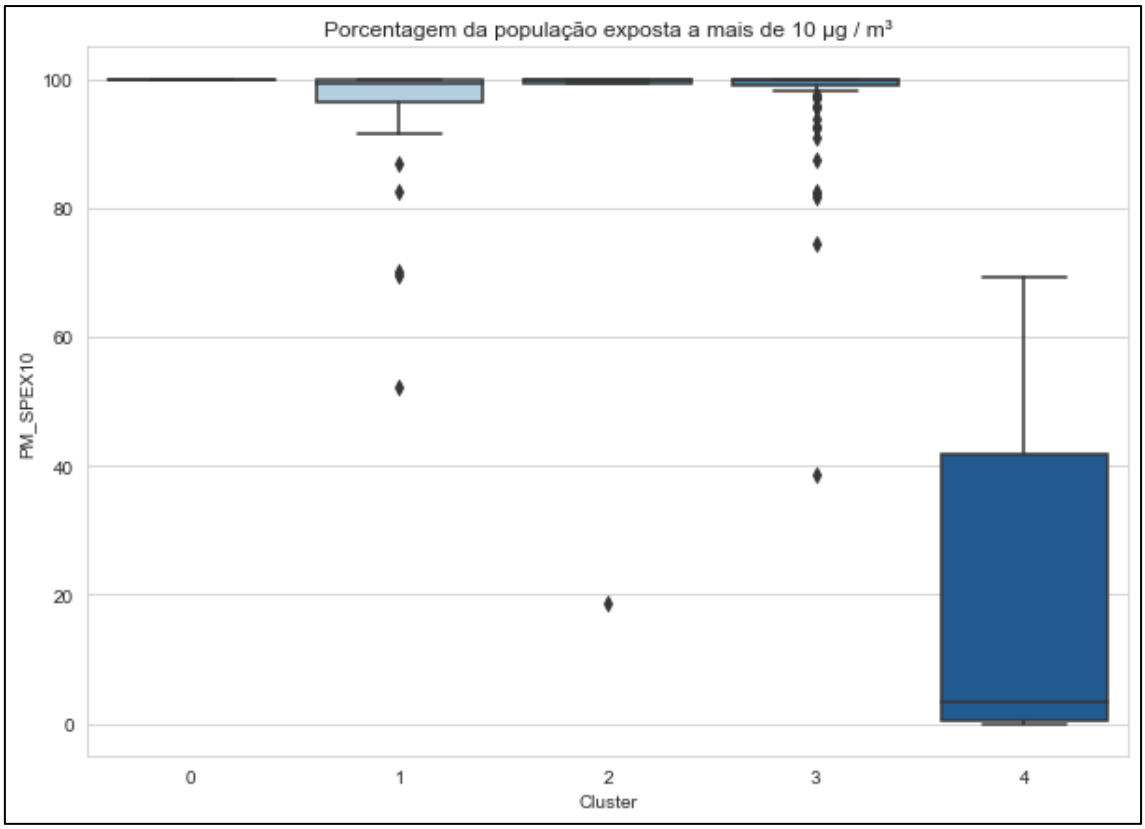
Figura 10 – Boxplots de indicadores de interesse para o cluster 2



O cluster 3 foi o que agrupou a maior quantidade de países, totalizando 96 países. Majoritariamente esses países são do que estudos pós-coloniais e transnacionais chamam de Sul Global. São, no geral, economias subdesenvolvidas ou em desenvolvimento, agroexportadoras, com baixa emissão de carbono e que têm uma estrutura social e econômica com grandes desigualdades em padrões de vida, expectativa de vida ou acesso a recursos. Os dados desse cluster apresentaram maior variabilidade se comparados aos demais, por possuir uma quantidade significativamente maior de instâncias.

Por fim, o cluster 4 agrupou 19 países. Entre eles EUA, Canadá, Austrália e países da Europa Ocidental. Países ricos, com acesso a tecnologias avançadas, sistemas políticos estáveis e alta expectativa de vida. O principal insight proveniente da análise dos dados desse cluster foi a baixa taxa de exposição a partículas com concentrações anuais que excedem 10 microgramas por metro cúbico.

Figura 10 – Boxplots do indicador de interesse para o cluster 4



## 5. Conclusões

Esse estudo aplicou os algoritmos de clusterização k-means e agrupamento hierárquico aglomerativo para identificar padrões entre países a partir dos dados de indicadores de crescimento verde da OCDE, definindo o número ideal de clusters através dos coeficientes de silhueta.

Por conta principalmente do alto número de dados faltantes, a maioria dos indicadores levados em consideração na etapa de mineração dos dados foram do eixo conceitual dimensão ambiental da qualidade de vida e do contexto socioeconômico.

Considerando que uma das finalidades das tarefas de agrupamento é descobrir estruturas ocultas nos dados, pode-se afirmar que os resultados desse trabalho foram satisfatórios, desde o cálculo da melhor quantidade de grupos até o agrupamento em si. Os agrupamentos possibilitaram alguns insights, como a baixa taxa de exposição a partículas com concentrações anuais que excedem 10 microgramas por metro cúbico para o cluster 4, que é composto por países desenvolvidos como EUA, Canadá, Austrália e países da Europa Ocidental.

Como limitação do trabalho, pode-se citar o baixo número de atributos dos eixos conceituais “produtividade ambiental e de recursos” e “base de ativos naturais”, e a ausência de atributos do eixo conceitual “oportunidades econômicas e respostas de políticas públicas”. A estimação de alguns indicadores desses eixos conceituais é uma possibilidade para estudos futuros.

## Referências Bibliográficas

BHARATI, M.; RAMAGERI. DATA MINING TECHNIQUES AND APPLICATIONS. **Researchgate**, 2010. v. 1, p. 301–305, 2010. ISSN 0976-5166.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Advances in knowledge discovery and data mining. In: **From Data Mining to Knowledge Discovery: An Overview**. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996.

JOTHI, N.; RASHID, N. A.; HUSAIN, W. Data Mining in Healthcare. **Procedia Computer Science**, 2015.v. 72, p. 306-313, 2015. ISSN 1877-0509.

OECD. OECD **Green Growth Indicators** The OECD Green Growth database contains selected indicators for monitoring progress. n. March, p. 68, 2020.

RASCHKA, SEBASTIAN. **Python Machine Learning**. Birmingham – Mumbai, 2015. ISBN 978-1-78355-513-0.

ZENGIN, K. et al. A sample study on applying data mining research techniques in educational science: Developing a more meaning of data. **Procedia Social and Behavioral Sciences**, 2011. v. 15, p. 4028–4032, 2011. ISSN 1877-0428.