Universidade Federal de Minas Gerais (UFMG)

Departamento de Ciência da Computação

Guilherme Mendes de Oliveira

guilhermemendes@ufmg.br

AGRUPAMENTO DE PAÍSES POR INDICADORES ECONÔMICOS

Belo Horizonte, MG – Brasil

2022

Guilherme Mendes de Oliveira

AGRUPAMENTO DE PAÍSES POR INDICADORES ECONÔMICOS

Trabalho prático de agrupamento da matéria de Mineração de Dados ministrada pelo professor Wagner Meira Jr.

Belo Horizonte, MG – Brasil

2022

1. Introdução

Este trabalho tem como objetivo a implementação de uma técnica de mineração de dados para agrupamento em uma base de dados. Para tal atividade foi utilizado o algoritmo ***KMeans*.**

Além disso a escolha para base de dados foi feita a partir World Data Bank um portal que disponibiliza múltiplos indicadores dos países relacionados ao seus respectivos desenvolvimento , a base foi escolhida a partir de uma dúvida: As classificações que aprendemos ao longo do currículo escolar nos anos de ensino fundamental entre países desenvolvidos, emergentes, subdesenvolvidos, além dos agrupamentos geográficos em regiões e continentes é algo assertivo, isto é há aspecto comum que de fato una esses grupos? .

O portal disponibiliza esses diversos indicadores em séries temporais, para tornar a aplicação da técnica e a análise dos resultados menos complexa foi selecionado apenas o ano de 2020 e os indicadores financeiros PIB e PIB Per Capita.

```
[ ] dfIndicadores.sample(10)
```

	Country Name	Country Code	2020 [YR2020] - GDP (current US$) [NY.GDP.MKTP.CD]	2020 [YR2020] - GDP growth (annual %) [NY.GDP.MKTP.KD.ZG]	2020 [YR2020] - GDP per capita (current US$) [NY.GDP.PCAP.CD]
31	Burundi	BDI	2780510624.64184	0.327156892638385	233.837510306669
13	Bahamas, The	BHS	9699500000	-23.8226075898758	24665.0968345675
12	Azerbaijan	AZE	42693000000	-4.30000010230737	4229.91064904503
77	Greenland	GRL	3075968328.69677	0.356531660011413	54570.3750190141
141	New Zealand	NZL	211734532308013	-1.25266452942354	41596.5055023403
39	Channel Islands	CHI	..	..	..
93	Ireland	IRL	425888950992003	5.86697550911293	85422.5428682266
34	Cameroon	CMR	40804449726.0184	0.491914742161129	1537.13021832773

Figura 1 – Amostra dos dados da base de indicadores.

Para auxiliar na avaliação e nas respostas a partir da classificação continental foi utilizada uma base que classifica os países em continentes e regiões, obtidas em repositório aberto no github.

```
dfContinentes.sample(5)
```

	country	code_2	code_3	country_code	iso_3166_2	continent	sub_region	region_code	sub_region_code
122	Latvia	LV	LVA	428	ISO 3166-2:LV	Europe	Northern Europe	150.0	154.0
113	Jordan	JO	JOR	400	ISO 3166-2:JO	Asia	Western Asia	142.0	145.0
234	United Kingdom of Great Britain and Northern I...	GB	GBR	826	ISO 3166-2:GB	Europe	Northern Europe	150.0	154.0
132	Madagascar	MG	MDG	450	ISO 3166-2:MG	Africa	Eastern Africa	2.0	14.0

Figura 2 – Amostra dos dados da base continental.

2. Motivação/ justificativa

A aplicação das técnicas de modelagem envolvidas no trabalho juntamente com a utilização de um algoritmo em uma base de dados que se aproxima de problemas reais encontrados no

cotidiano é além de uma forma de avaliação uma oportunidade de desenvolver todas as competências técnicas necessárias para aplicar futuramente esse conhecimento no “mundo real” ou seja uma forma de ganhar experiência analítica dos resultados e competências técnicas de programação para exploração, limpeza dos dados , aplicação do algoritmo e extração de métricas para avaliação dos resultados.

3. Objetivo

A aplicação das técnicas de agrupamento nessa base de dados dos países, busca simular uma melhor forma de entender a similaridade entre os países em dado ano com base em algum aspecto .

Esse entendimento por exemplo pode ser utilizado ainda que de forma genérica para a tomada de decisão de uma empresa que busca expandir suas instalações, um indivíduo que quer mudar de localidade, um fundo que planeja investir capital e quer uma alternativa a dado país podendo escolher outro similar de seu mesmo grupo ou até mesmo uma organização que busca promover equidade pode se basear naquele grupo de países que está em pior momento econômico a fim de organizar ações,, acordos e políticas para tratar esse problema.

4. Metodologia

A metodologia utilizada foi inspirada no CRISP-DM, inspirada porque uma das etapas, mais especificamente a do entendimento do negócio, é realizada de forma simulada uma vez que que proposição de um problema só é possível após uma verificação preliminar da base de dados, ou ela é proposta e partir disso que os dados são buscados. Além disso, ela está direcionada a ser um problema que seja possível explorar e propor soluções com base em uma técnica específica, nesse caso o agrupamento.

Além disso, a etapa de *deployment* também não é pertinente uma vez que é uma prática e nesse caso não há um problema real a ser tratado, para as demais etapas todas foram possíveis e foram aplicadas para a conclusão do trabalho prático.

O fluxo de entendimento, preparação, modelagem e avaliação podem ser vistos aqui.

5. Desenvolvimento
5.1. Business Understanding

Essa etapa compreende o entendimento do problema, como foi propositivo a partir da base de dados e já direcionado para a aplicação de agrupamento esta etapa não se aplica.

5.2. Data Understanding

A base utilizada possui um total de 217 linhas e 5 colunas e não possui linhas com valores nulos em nenhum de seus atributos,no entanto apresenta strings ('..') que sinalizam dados faltantes em alguns países, mas não em uma quantidade que impacte no prosseguimento do trabalho ao serem expurgados na próxima etapa a base ainda terá 201 registros válidos. Portanto, no entendimento da base de dados verificamos que a mesma está apta para a aplicação da técnica de mineração uma vez que possui uma quantidade considerável de registros .

5.3. Data Preparation

Na etapa de preparação inicialmente foi feito o expurgo dos registros que não tem os valores dos indicadores que serão utilizados como "medida de similaridade" para a aplicação da técnica de agrupamento.

Além disso foram renomeadas as colunas para deixar mais claro a que se refere o atributo e realizada a conversão dos valores numéricos para *float* com apenas duas casas decimais.

5.4 Modelagem

Na etapa de modelagem da base para a aplicação do algoritmo de agrupamento (*KMeans*), foi realizado o procedimento de Normalização nos campos numéricos a fim de tratar problemas relacionados a escala das ordens de grandeza dos valores que serão utilizados como "medida de similaridade" foram selecionadas os campos de Pib Per Capita e a Taxa de Crescimento do PIB em relação ao último ano.

```
dfNorm = dfIndicadores.copy()
dfNorm['PIB'] = (dfNorm['PIB'] - dfNorm['PIB'].mean())/ dfNorm['PIB'].std()
dfNorm['Taxa de Crescimento PIB'] = (dfNorm['Taxa de Crescimento PIB'] - dfNorm['Taxa de Crescimento PIB'].mean())/ dfNorm['Taxa de Crescimento PIB'].std()
dfNorm['PIB per Capita'] = (dfNorm['PIB per Capita'] - dfNorm['PIB per Capita'].mean())/ dfNorm['PIB per Capita'].std()
```

Figura 3 – Processo de normalização dos atributos numéricos.

Para a seleção do número de agrupamentos mais assertiva foi computada a distância quadrática intra grupo para uma quantidade de 1 a 10. Escolhido como parâmetro 6 *clusters*, pois foi o valor que apresentou um melhor ganho em relação ao número anterior e por mais que uma quantidade de clusters maior represente uma menor distância intra cluster que o valor escolhido não é um ganho maior que o ganho da escolha de 6 em relação a 5.

Foram utilizadas como métricas para calibrar o parâmetro do número de *clusters* a distância intra cluster e o Índice de Calinski-Harabraz, ambos avaliando a comparação do valor k com o valor da métrica de k-1

Figura 4 – Gráfico de distâncias intracluster x número de clusters testados.

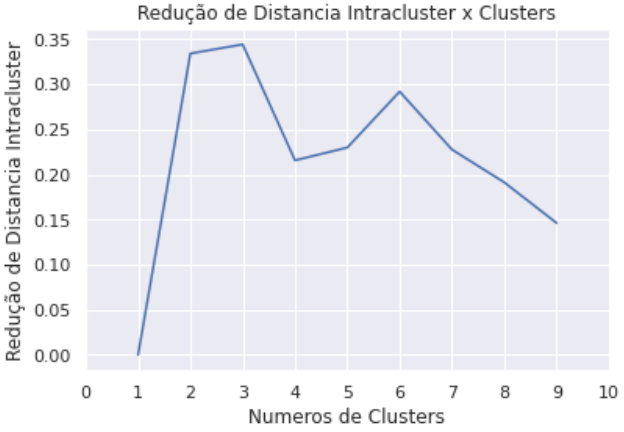

Figura 5 – Gráfico de redução de distâncias intracluster (em percentual) x número de clusters testados.

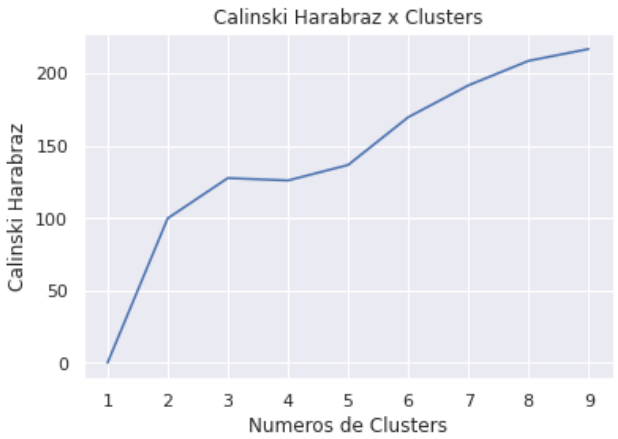

Figura 6 – Gráfico do índice Calinski-Harabraz x número de clusters testados.

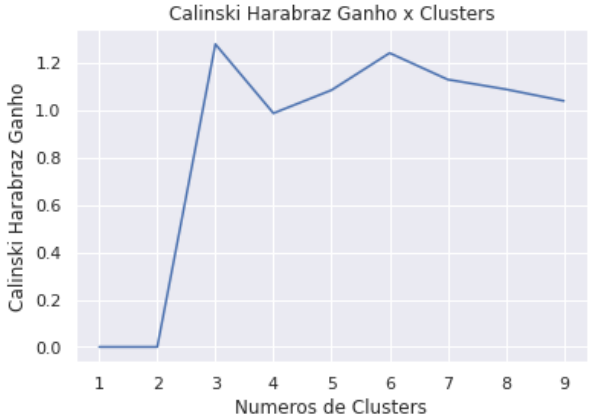

Figura 7 – Gráfico do índice ganho do índice Calinski-Harabraz (de k em relação a k-1) x número de clusters testados.

6. Resultados experimentais e análise

6.1. Evaluation

Os agrupamentos numa primeira impressão foram satisfatórios ao observarmos o gráfico de dispersão dos pontos com a classificação obtida pelo uma vez que vemos uma divisão bem clara e a princípio com pouco ruído.

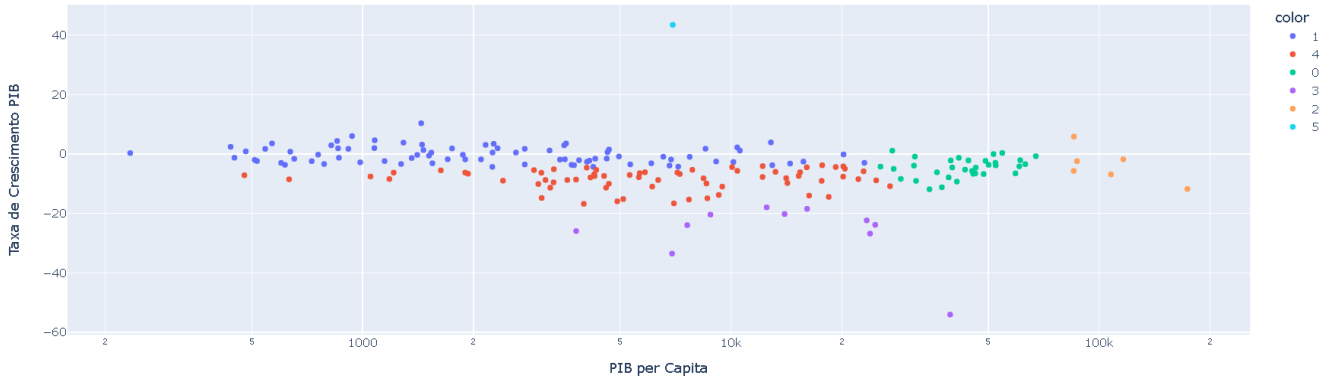

Figura 8 – Gráfico de dispersão dos pontos com o agrupamento.

Verificando a distribuição dos valores de cada atributo usado no agrupamento em relação ao seu grupo observamos que de modo geral há uma boa separação dos valores.

Os países que apresentam um valor mais alto de Pib Per Capita estão concentrados no *cluster* 2, as taxas de crescimento do Pib para os *clusters* 2 e 0 estão com uma distribuição bem similar, no entanto vemos a distinção em relação ao seu Pib Per Capita.

Os *clusters* 1,3,4 tem certa interseção considerando o Pib Per Capita, porém os países encontram em momentos de desenvolvimento distintos, os países do *cluster* 1 estão com uma taxa de crescimento maiores que 0,os do *cluster* 3 de modo geral estão em uma recessão alta e os países do *cluster* 4 em baixa recessão.

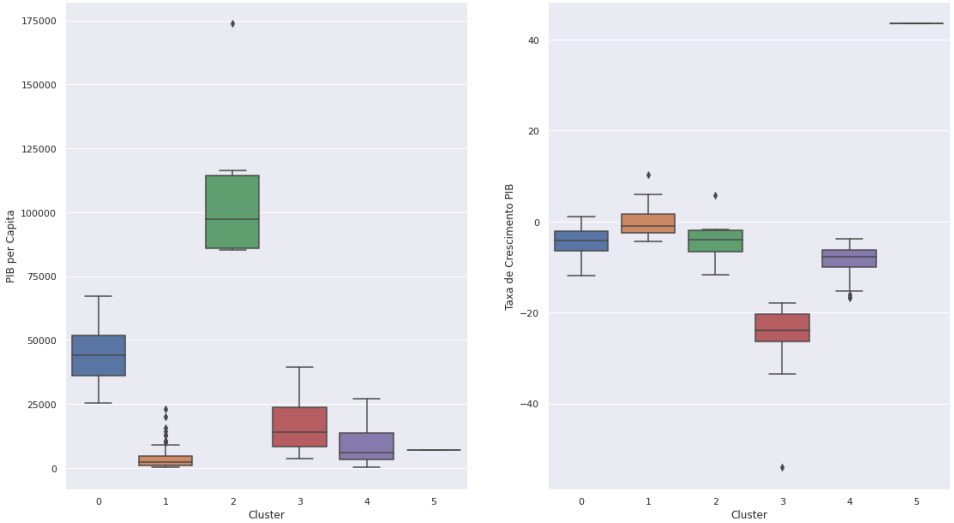

Figura 9 – Distribuição dos valores de Pib Per Capita e Taxa de Crescimento do PIB de cada *cluster*.

Uma das questões que surgiram após a validação da proposta e numa análise preliminar da base foi se essa classificação econômica segue algum padrão de classificação continental ou de regiões geográficas, intuitivamente era de se esperar que os países da Europa fossem os mais ricos e estivessem o melhor nível de crescimento, os fazendo ser alocados no mesmo cluster.

No entanto, ao vermos a distribuição dos países de cada *cluster* em relação ao continente a que pertencem vemos que não temos uma concentração de um continente ou região em um único cluster.

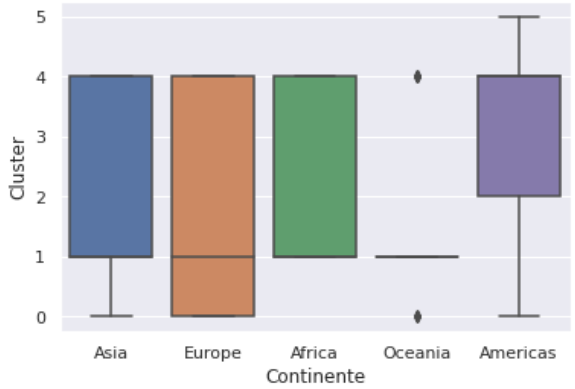

Figura 10 – Distribuição dos *clusters* de cada país em relação aos continentes.

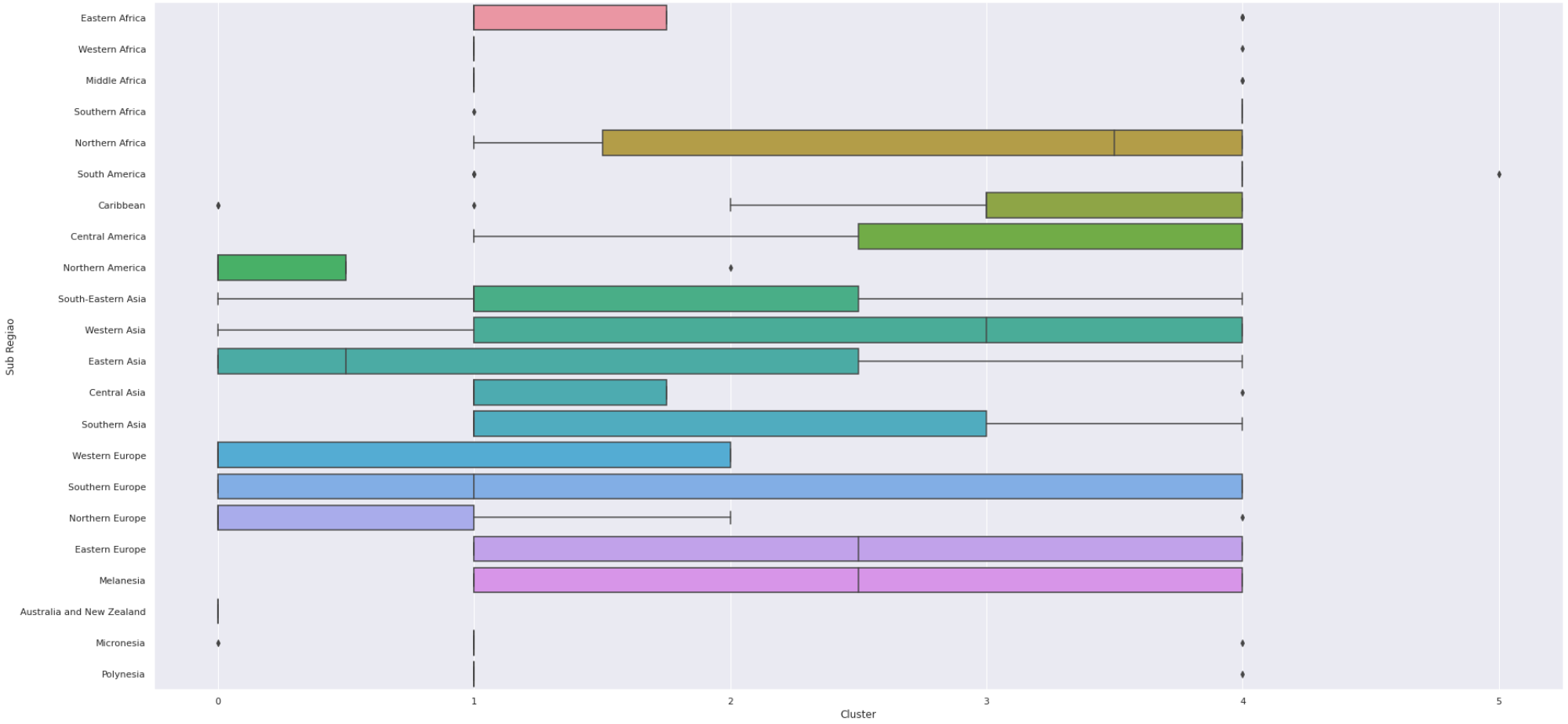

Figura 11 – Distribuição dos *clusters* de cada país em relação às sub-regiões.

Para finalizar ao observarmos o *Silhouette Coeficient* que mede a diferença entre a distância média dos pontos de um cluster aos pontos do cluster mais próximo em relação a distância média intracluster obtemos um valor de 0.453, em relação a escala de valores [-1,0,1] é satisfatório.

7. Conclusões e perspectivas

O resultado do trabalho prático foi bem satisfatório, foi possível explorar algumas das técnicas de calibração do parâmetro k que diz respeito ao número de *clusters* e práticas além do cálculo dessas métricas a forma de analisá-las. Além disso, foi possível extrapolar analiticamente o resultado do agrupamento, formulando e investigando uma hipótese levantada na análise exploratória preliminar dos dados.

8.Referências

Zaki, M and Meira, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Capítulo 13: Representative-based Clustering;

Zaki, M and Meira, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Capítulo 17: Clustering Validation;

<https://scikit-learn.org/stable/modules/clustering.html#> . Acesso em: 28 de outubro de 2022.

<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics> . Acesso em: 30 de outubro de 2022.