Análise de Componentes Principais e Clusters

Estatística Multivariada

Sofia Begonha Morgado

Faculdade de Ciências e Tecnologia

Universidade Nova de Lisboa

17 de janeiro de 2022

Sumário

1 Introdução

1.1 Análise de Componentes Principais

A análise de componentes principais é uma técnica multivariada que consiste na realização de transformações lineares a uma série de variáveis correlacionadas, permitindo a obtenção de novas variáveis não correlacionadas [1].

A análise de componentes principais tem como objetivo a redução da dimensionalidade de um dataset, preservando a sua variabilidade. Esta técnica consiste na procura de combinações lineares, as quais se denominam componentes principais (CP), de um conjunto inicial de p variáveis. Estas combinações lineares são calculadas tal que a primeira CP corresponda a um vetor com a direção coincidente com a maior variância da amostra. De seguida, a segunda CP é aquela que, não estando relacionada com a primeira, melhor retém a variabilidade restante, e assim sucessivamente, até à obtenção de p CPs [2].

Para a implementação deste método, calculamos os valores e vetores próprios da matriz de covariâncias ou de correlações do dataset. Os vetores próprios dão-nos a CP e os valores próprios a proporção da variabilidade retida pela CP correspondente. As novas variáveis são obtidas através da combinação linear dos vetores próprios com as variáveis iniciais.

A seleção da quantidade de CPs a reter pode ser feita a partir de diversas metodologias. Entre elas, encontramos o Critério de Kaiser e o Scree Plot. O Método de Kaiser consiste na retenção das CPs cujo valor próprio é superior à média de todos os valores próprios, dado que estas correspondem às CPs que retêm mais variabilidade do que as variáveis originais [1].

O Scree Plot é um gráfico que ordena as CP por ordem decrescente da grandeza do valor próprio correspondente, o que equivale à ordem decrescente da quantidade de variabilidade retida. O método por análise do Scree Plot consiste em selecionar as CPs até àquela em que observamos uma alteração acentuada da inclinação da linha desenhada. No entanto, este método pode ser subjetivo e insuficiente para a seleção do melhor número de CP [3]. Pode, assim, complementar outros métodos de seleção.

1.2 Análise de Clusters

A análise de clusters é uma metodologia de análise multivariada que corresponde a agrupar a amostra em conjuntos de dados semelhantes. Há varias características que podem definir estes clusters. Os clusters podem ser hierárquicos, quando existem clusters de nível superior que incluem um ou mais clusters de nível inferior, ou não hierárquicos, quando todos os clusters pertencem ao mesmo nível.

Os clusters hierárquicos podem ainda ser divididos em clusters divisivos ou de partição ou aglomerativos ou de agregação. Neste trabalho desenvolvemos diferentes métodos de agregação. Os clusters hierárquicos de agregação são realizados de forma a que, inicialmente, cada exemplo

da amostra pertence a um cluster, e estes vão-se unindo a cada iteração. Existem várias formas de unir estes clusters, tendo em conta os diferentes métodos usados.

Dos métodos não hierárquicos será implementado o k-means. É um método de clustering baseado num protótipo, o centróide. O cluster é definido pela distância de cada ponto ao centróide, e cada ponto pertence a um, e só um, cluster.

Uma das desvantagens deste método é o facto de, no caso em que um ponto esteja numa posição intermédia entre dois clusters, ele ser forçosamente classificado como pertencente a um destes clusters. Esta classificação pode não ser tão clara quando os clusteres não estão bem delimitados.

Outra desvantagem é que este método não exclui outliers dos clusters obtidos, dado que todos os pontos são classificados. Mais ainda, estes outliers podem influenciar a posição do centroíde [4].

Este método assume ainda que os clusteres apresentam uma forma esférica, dado que procuramos a distância ao centróide, o que pode não corresponder ao dataset.

No entanto, é um método mais eficiente que os métodos hierárquicos e podem ser obter bons resultados.

1.3 Objetivos

O objetivo deste trabalho é encontrar clusters que separem os países tendo em conta o seu contexto socioeconómico, dado pelas variáveis do dataset.

Em primeiro lugar, o objetivo será, através da implementação da Análise de Componentes Principais, reduzir a dimensionalidade do nosso dataset. Por fim, a divisão dos dados no melhor número de clusters, aplicando vários métodos.

2 Dataset

O dataset utilizado para este trabalho denomina-se Unsupervised Learning on Country Data, do site Kaggle [5]. Este dataset inclui dados socioeconómicos e de saúde de 167 países.

Para cada país, encontramos 9 variáveis: mortalidade infantil, exportações, gastos em saúde per capita, importações, salário médio, inflação, esperança média de vida, índice de fertilidade e produto interno bruto per capita.

3 Metodologia

Em primeiro lugar, realizarei uma análise de componentes principais (ACP) de forma a reduzir a dimensionalidade e selecionar as componentes que melhor representam as variáveis originais (que melhor preservam a sua variabilidade).

Em primeiro lugar, irei avaliar a adequabilidade da realização de ACP através do estudo das correlações e das correlações parciais e também por implementação do teste de esfericidade.

A normalização dos dados para ACP deve ser cautelosa pois queremos reter a maior variabilidade possível, e desta forma perdemos o maior peso das variáveis com maior variância. Para decidir se as variáveis devem ser normalizadas, começarei por comparar a unidade em que são expressas e a sua variância (diagonal da matriz de covariâncias). Caso os valores de variância sejam muito dispares entre si, pelas variáveis apresentarem diferentes ordens de grandeza, estas devem ser normalizadas.

Em seguida, determinarei os valores e vetores próprios da matriz de covariancias dos dados, de forma a obter as combinações lineares que definem os CP.

Irei avaliar qual a percentagem de variabilidade retida e selecionar o número de componentes principais mais adequada. Numa primeira fase, usando o critério de Kaiser, seguido da a observação de um scree plot.

Irei avaliar a mvárias formas de separar estes dados em clusters, usando clusters hierárquicos com agregação, avaliando o método simples, completo e de Ward; decidi utilizar este método porque permite avaliar apenas à posteriori quantos clusters utilizar e agrupar os países em subgrupos (por exemplo, dentro dos países "desenvolvidos", tentar encontrar se existe alguma forma de os subagrupar).

Por outro lado, irei implementar o método k-means; para selecionar o melhor número de clusters, irei aplicar o Elbow method e também avaliar o silhouette score. Os resultados obtidos por estes métodos serão dispostos num gráfico para apreciação.

4 Resultados

4.1 Análise de componentes principais

4.1.1 Análise da Adequabilidade da ACP

O primeiro passo para a realização da Análise de Componentes Principais (ACP) passa pelo estudo da adequabilidade desse método.

Para isso, foram utilizados 3 processos: o teste de esfericidade, a análise da correlação das variáveis e, por fim, a análise da correlação parcial das variáveis.

Teste de esfericidade O teste de esfericidade é realizado com o intuito de esclarecer se existe esfericidade das variáveis. Isto é, se a matriz de correlações dos dados é igual à matriz identidade, caso extremo em que as CP serão iguais às variáveis originais. Assim, esta metodologia permite esclarecer se há ou não evidências de que não há correlação entre as nossas variáveis, necessária a uma boa ACP [1].

A nossa hipotese nula é de que existe esta igualdade, e realizamos este teste para procurar se existem ou não evidências que comprovem a esta nossa H_0.

O teste foi realizado, e os valores obtidos foram o valor de U^* (estatística de teste) igual a 1174.28 e o valor de p igual a 0. Assim, podemos rejeitar a hipotese nula de que existe esfericidade das variáveis, e, portanto, concluir a adequabilidade da ACP.

Correlação das variáveis A análise da correlação das variáveis originais do nosso dataset é imprescindível para concluirmos se poderemos realizar uma boa ACP, pois é necessário que existam correlações entre as variáveis originais. Em primeiro lugar, foi calculada a percentagem de correlações entre variáveis superiores a 0.5. Foi obtido um valor de 27.78%.

Correlação parcial das variáveis As correlações parciais representam as interações diretas entre duas variáveis condicionadas por todas as restantes variáveis [6]. Para a análise da correlação parcial das variáveis a matriz de covariâncias inicial foi convertida usando a função 'cor2pcor' do R. De seguida, foi calculada a percentagem de correlações parciais inferiores a 0.5, e o resultado obtido foi 88.89%.

4.1.2 Obtenção das componentes principais

Tendo em conta as diferentes unidades em que as variáveis são expressas e a análise da sua variância, que encontramos na tabela 1, foi decidido proceder à normalização dos dados.

De seguida, foi calculada a matriz de covariâncias e selecionados os valores e vetores próprios. A primeira componente principal, correspondente à combinação linear obitada pelo vetor próprio correspondente ao maior valor próprio, está definida na equação 1.

Tabela 1: Variância das variáveis do dataset

Variável	Unidades	Variância
Child_mort	Permilagem	$1.62 * 10^3$
Exports	Percentagem	$7.51 * 10^2$
Health	Percentagem	$7.55 * 10^0$
Imports	Percentagem	$5.86 * 10^2$
Income	Dolares	$3.72 * 10^8$
Inflation	Percentagem	$1.12 * 10^2$
Life_exp	Anos	$7.91 * 10^1$
Total_fer	Nascimentos/Mulher	$2.29 * 10^0$
GDPP	GDP/População	$3.36 * 10^8$

$$y = 0.42x_1 - 0.28x_2 - 0.15x_3 - 0.16x_4 - 0.39x_5 + 0.19x_6 - 0.42x_7 + 0.40x_8 - 0.39x_9 \quad (1)$$

4.1.3 Seleção das Componentes Principais

A tabela 2 foi criada de forma a resumir os dados acerca da variabilidade retida por cada componente principal. Para cada CP, encontramos o valor próprio correspondente bem como a variância retida, obtida pela divisão do valor próprio pela soma de todos os valores próprios, bem como a variância cumulativa e o desvio padrão.

Tabela 2: Informação da análise de componentes principais. Legenda: Val_prop, valor próprio; σ, desvio padrão; σ^2, variância; σ^2 cum, variância cumulativa

CP	Val_prop	σ	σ^2	σ^2cum
1	4.14	2.03	0.46	0.46
2	1.55	1.24	0.17	0.63
3	1.17	1.08	0.13	0.76
4	0.99	0.99	0.11	0.87
5	0.66	0.81	0.07	0.95
6	0.22	0.47	0.02	0.97
7	0.11	0.34	0.01	0.98
8	0.09	0.30	0.01	0.99
9	0.07	0.26	0.01	1.00

Pelo Critério de Kaiser selecionam-se as CP cujo valor próprio é superior à média dos valores próprios da matriz de covariâncias. Tendo em conta que as variáveis foram normalizadas previ-

amente à analise de componentes principais, aplicamos o Critério de Kaiser selecionando as CP cujo valor próprio é superior a 1 [1]. Assim, nesta primeira fase selecionamos as 3 primeiras CPs.

De seguida, podemos analisar o Scree Plot 1. A análise do Scree Plot permite a seleção das CP anteriores ao ponto em que a inclinação da linha desenhada no gráfico diminui acentuadamente, tornando-se aproximadamente horizontal. Este ponto demonstra que a variabilidade retida pela CP à direita desse ponto é muito reduzida em relação às anteriores.

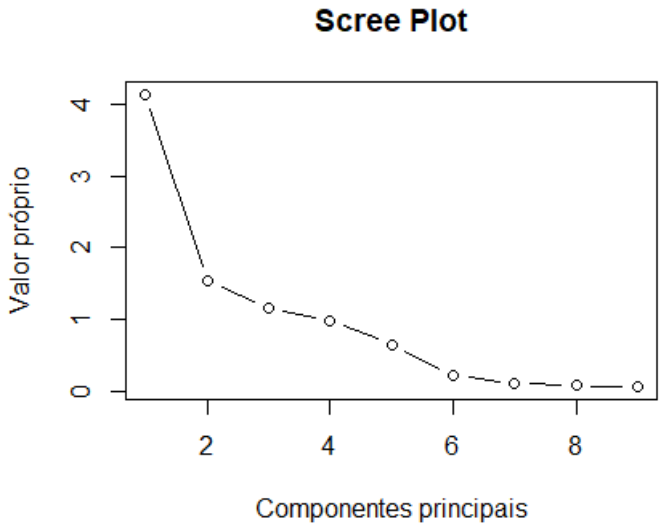

Figura 1: Scree plot

Apesar de uma primeira mudança de direção na CP 2, a linha não se torna tendencialmente horizontal, e voltamos a observar uma nova defleção na CP 6, a partir da qual a linha se torna quase horizontal. Por este motivo, decidi reter as três primeiras CPs já selecionadas pelo Critério de Kaiser.

4.2 Análise de clusters

Após a criação de uma matriz composta pelas combinações lineares para as 3 primeiras CPs, foi realizada uma análise de clusters. Em primeiro lugar, foi realizada uma análise de clusters hierárquica e, por fim, foi utilizado o Método K-means.

4.2.1 Clustering hierárquico

Foi utilizado um método hierárquico de aglomeração. Para isso, foram testados vários métodos de agregação. Em primeiro lugar, foi testado o método de ligação simples, que consiste na união dos dois clusters que apresentem uma distância mínima entre os dois pontos mais próximos (Figura 2).

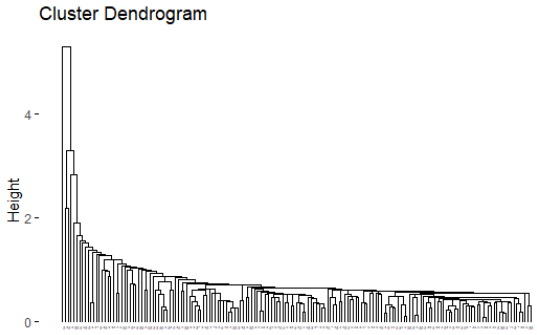

Figura 2: Cluster hierárquico utilizando o método simples

Como podemos observar pelo dendrograma obtido, este método de clusters não foi adequado. Não só os clusters são formados, na sua maioria, pela adição de mais um país ao cluster de maior dimensões, como, quando dividimos os dados, por exemplo, em 2 clusters, obtemos um cluster com um único país. Por este motivo, outros métodos foram testados (Figura 3).

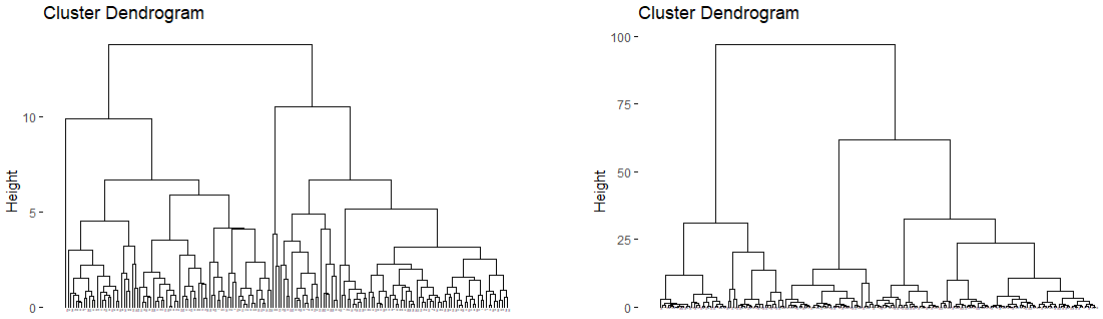

Figura 3: Dendrograma da análise de clusters hierárquico de agregação pelo método completo (esquerda) e Ward (direita)

Assim, foi aplicado o método de ligação completa, que consiste em unir os dois clusters que apresentem os pontos mais distantes mais próximo um do outro. Podemos observar como já conseguimos distinguir um nível com 2 clusters e um nível com 4 clusters em que a distância necessária para o agrupamento de mais clusters é ainda elevada.

Por fim, o método de Ward, que é o que melhor majora esta distância necessária ao próximo agrupamento, como observamos nos níveis correspondentes a 2 e 3 clusters. Este método agrupa os clusters com o objetivo de diminuir a heterogeneidade entre grupos, minorando a variância entre clusters [7].

4.2.2 Método K-means

Seleção do número de clusters O modelo k-means é um método de clustering não-hierárquico em que o número de clusters deve ser definido pelo utilizador. Por este motivo, a procura do melhor valor de clusters é um problema importante na sua implementação, dado que não existe uma solução universal. Existem vários métodos para realizar este estudo. Em primeiro lugar, a observação do dendograma dos resultados de clustering hierarchico pode ser útil, no entanto, muito moroso. Outros métodos, como o Elbow Method e o Silhouette score podem ser usados [8]. Este dois métodos avaliam aspetos diferentes e podem ser usados em conjunto, como é realizado neste trabalho (Figura 4).

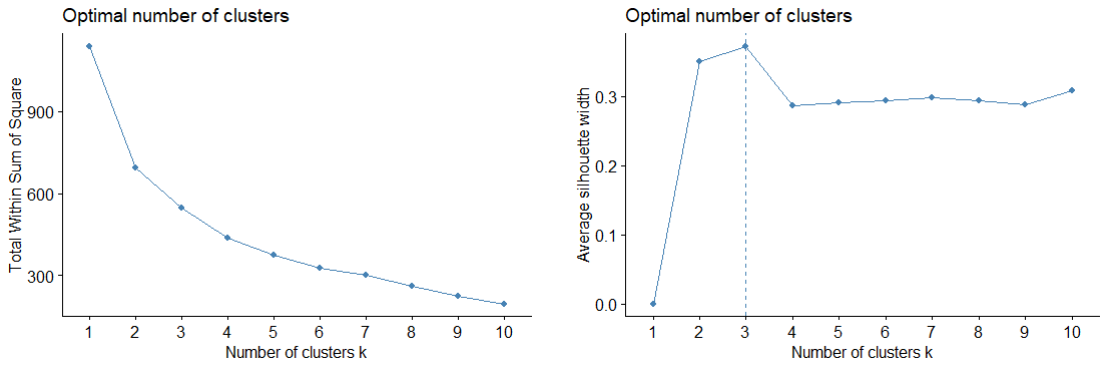

Figura 4: Gráfico para realização do Elbow Method (esquerda) e para procurar o melhor Silhouette score (direita)

O Elbow method pretende avaliar a variabilidade intragrupo para cada número de clusters k, de forma gráfica. Para isso, é realizada a soma dos valores da distância de cada ponto do cluster ao centróide. O método passa por encontrar um 'cotovelo' no gráfico, que representa um valor de k para o qual a variabilidade intragrupo diminui bruscamente, tornando-se, a partir desse ponto, desprezíveis as diferenças na variabilidade. Analisando na figura 4 o gráfico da esquerda, observamos que é difícil o discernir um 'cotovelo', pois a diferença de variabilidade é constante. É possível ainda colocar a hipotese de que, se esta diferença de variabilidade é dada de forma tão constante, então os clusters poderão estar mal definidos.

Por outro lado, podemos ainda realizar a análise do Silhouette Score. O Silhouette Score avalia a qualidade do clustering, tendo em conta outros aspetos do clusters, como variabilidade, espessura e variação de tamanho [8] [9].

O gráfico da direita na figura 4 mostra a variação do Sillhoutte com a variação do valor de k. Podemos observar que $k = 3$ permite o melhor valor de Silhouette Score, e, consequentemente, o melhor clustering. Por fim, foram obtidos os valores de silhouette score para cada exemplo da amostra. Os resultados encontram-se em anexo, na figura 6.

Obtenção dos clusters Após a escolha do melhor número de clusters a realizar, foi implementado o método k-means com $k = 3$.

Foi criada uma dataframe composta pelo nome dos países, as suas coordenadas no sistema composto pelas 3 primeiras componentes principais e o cluster atribuido ao respetivo país.

Foi criado um gráfico em 3 dimensões que representa os 3 clusters (Figura 5). Cada ponto corresponde a um país e os clusters podem ser distinguidos pela cor de cada ponto: a vermelho, os países classificados como pertencentes ao cluster 1, com 40 elementos; a azul, o cluster 2, com 80 elementos; e a amarelo, o cluster 3, com 47 elementos, para um total de 167 países.

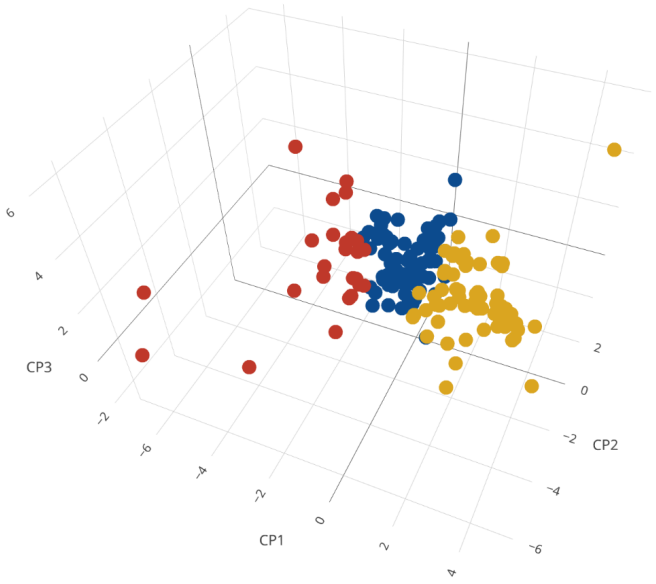

Figura 5: Representação 3D dos clusters obtidos através do método de K-means. Vermelho - Cluster 1; Azul - Cluster 2; Amarelo - Cluster 3

Como podemos observar na figura 5, os vários pontos representativos de cada país apresentam continuidade, e os clusters não estão bem separados entre si. No entanto, é possível definir 3 clusters que não se sobrepõe.

Relembrando, o método k-means é um método em que todos os pontos são incluidos num dos clusters, não há pontos considerados 'noise'. Podemos colocar a questão: será que são valores outliers que não deviam ser colocados neste cluster e que influenciam excessivamente a posição do seu centróide?

As figuras 7, 8 e 9, presentes no anexo II, demonstram a representação gráfica dos clusters a partir da visualização do gráfico em relação a cada um dos eixos.

Podemos reparar pela observação destas figuras que os clusters estão bem delimitados quando observamos na perspetiva de duas dimensões que correspondem à CP1 e CP2 (Figura 7). Por outro lado, observamos pontos correspondentes ao cluster 1 (vermelho), distantes do restante cluster.

Quando observamos o gráfico da perspetiva que corresponde à CP1 e CP2, podemos observar

como ainda é fácil distinguir o cluster 3 (amarelo) dos restantes, no entanto, encontramos já sobreposição entre o cluster 1 (vermelho) e 2 (azul).

No caso oposto, quando observamos os clusters num plano composto pelos eixos CP2 e CP3, torna-se dificil de encontrar a separação dos clusters. Estas figuras evidenciam a importância relativa e decrescente das CP obtidas.

A tabela 7 (Anexo III) apresenta um conjunto de países selecionados aleatóriamente do nosso dataset final. Foi incluído o seu cluster obtido através do método k-means e o seu SES (socioeconomic status), obtido de uma outra base de dados do site Kaggle [10] para o ano de 2010.

O valor do SES representa a percentagem de pessoas no mundo que vive em condições socioeconómicas inferiores à média desse país.

Esta tabela pretende ser apenas ilustrativa dos clusters obtidos, dado que a comparação entre os dois datasets apresenta algumas limitações. Em primeiro lugar, alguns países do nosso dataset não se encontram classificados com o SES (todos eles pertencentes ao cluster 3 para a nossa amostra aleatória). Por outro lado, o nosso dataset não refere o ano de obtenção dos dados.

No entanto, podemos observar que todos os países classificados no cluster 3 apresentam um score SES inferior a todos os países do cluster 1 ou 2, para esta amostra, mostrando que o cluster 3 apresenta países com um nível socioeconómico tendencialmente inferior aos restantes países.

Seria interessante realizar uma avaliação da correlação entre os nossos clusters e este ou outros scores que meçam o nível socioeconómico dos países.

5 Conclusões

O objetivo deste projeto foi separar os países do nosso dataset em clusters consoante o seu status socioeconómico, baseado em 9 variáveis originais.

Em primeiro lugar, foi procurado diminuir a dimensionalidade do nosso dataset através no método da análise de componentes principais (ACP). Foi realizado um estudo da adequabilidade da ACP e a variância retida pelas componentes principais (CP) foi estudada de forma a oselecionarmos o melhor número. Foi possível reter 76% da variabilidade do dataset selecionando as 3 primeiras CP. Foi criado um novo dataset com a projeção da amostra nos novos eixos (as CP).

Por outro lado, foi realizada a análise de cluster através de 3 métodos hierárquicos de aglomeração, obtendo 3 dendogramas. O método de ligação 'Ward' demonstrou ser aquele que melhor separa os nossos clusters, sendo possível obter 2 ou 3 clusters com uma boa distância de aglomeração. Foi também aplicado o Método K-means. O melhor número de clusters foi obtido por combinação dos métodos do Scree Plot e do Silhouette Score. Obtiveram-se 3 clusters que foram projetados num gráfico em 3 dimensões para avaliação. Por fim, sugere-se a validação destes dados por comparação a outros datasets que avaliem status socio-económico dos países.

6 Referências

Referências

[1] Hongyu, K., Sandanielo, V., Junior, G.: Análise de Componentes Principais: resumo teórico, aplicação e interpretação. Engineering and Science (2015)

[2] Ringnér, M.: What is principal component analysis? Nature (2008)

[3] Ledesma, R., Valero-Mora, P., Macbeth, G.: The Scree Test and the Number of Factors: A Dynamic Graphics Approach. The Spanish Journal of Psychology (2015)

[4] Marsland, S.: Machine Learning: An Algorithmic Perspective. First edition. Chapman Hall/CRC (2009)

[5] https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data

Last accessed 16.01.2022

[6] https://www.rdocumentation.org/packages/corpcor/versions/1.6.9/topics/cor2pcor

Last accessed 16.01.2022

[7] Majerova, I., Nevima, J.: The measurement of human development using the Ward method of cluster analysis. Journal of International Studies (2017)

[8] Shahapure, k., Nicholas, C.: Cluster Quality Analysis Using Silhouette Score. e University of Maryland, Baltimore County (2020)

[9] Kumar, A.: Elbow Method vs Silhouette Score – Which is Better? Data Analytics (2021)

[10] https://www.kaggle.com/sdorius/globses

Last accessed 17.01.2022

7 Anexo I - Silhouette Score por elemento da amostra

Na figura abaixo encontramos os valores de Silhuoette Score para cada um dos elementos do nosso dataset quando inseridos num cluster obtido pelo Método K-means com k=3. Foram obtidos 3 clusters: cluster 1, com 40 elementos e Silhouette Score médio de 0.12; cluster 2, com 80 elementos e Silhouette Score médio de 0.36; cluster 3, com 47 elementos e Silhouette Score médio de 0.32.

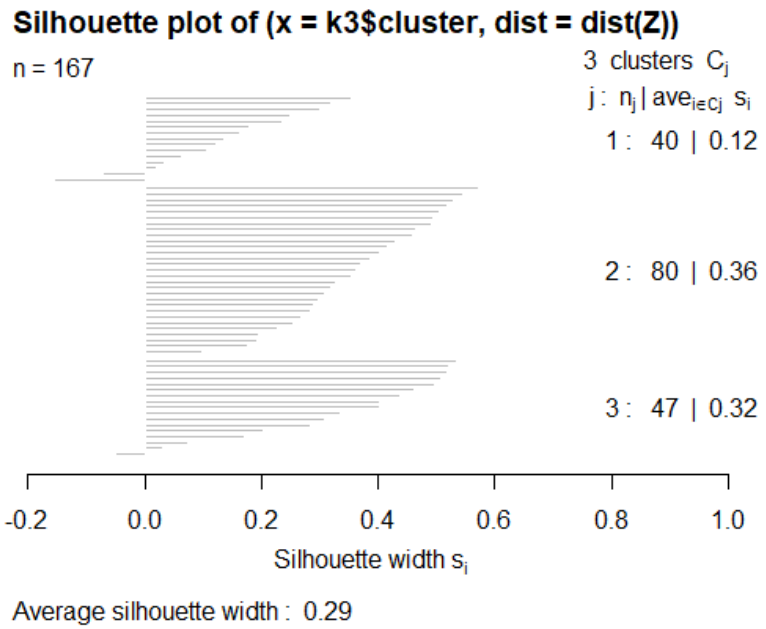

Figura 6: Valores de Silhouette Score para cada um dos elementos dos clusters. Foram obtidos 3 clusters, com 40, 80 e 47 elementos, respetivamente. O Silhouette Score dos exemplos do primeiro clusters são inferiores aos do segundo e terceiro.

8 Anexo II - Representações dos clusters obtidos pelo método K-means

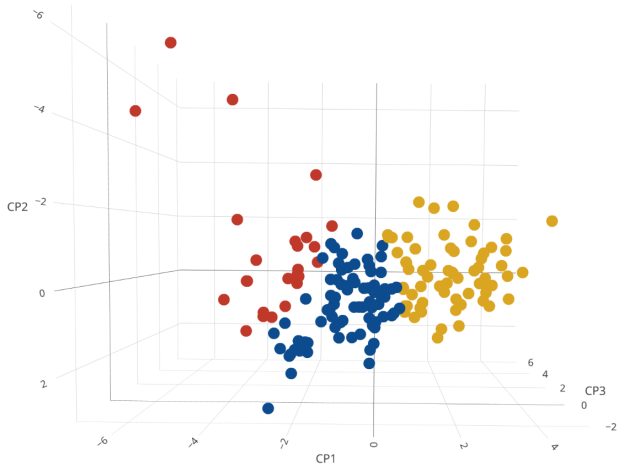

Figura 7: Representação dos clusters obtidos através do método de K-means, permitindo a visualização da CP1 e CP2. Vermelho - Cluster 1; Azul - Cluster 2; Amarelo - Cluster 3.

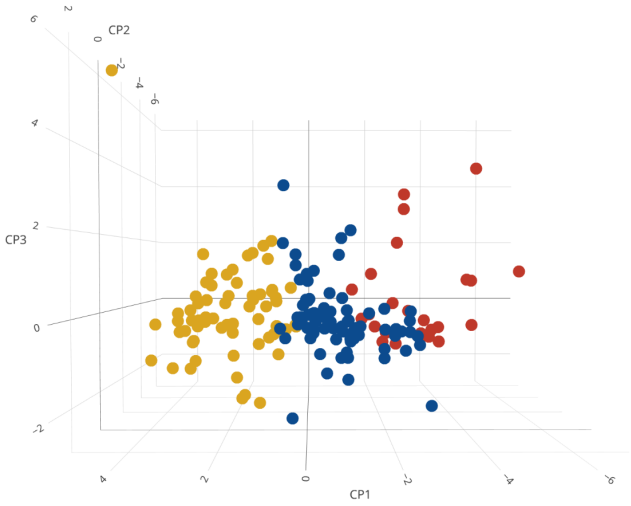

Figura 8: Representação dos clusters obtidos através do método de K-means, permitindo a visualização da CP1 e CP3. Vermelho - Cluster 1; Azul - Cluster 2; Amarelo - Cluster 3.

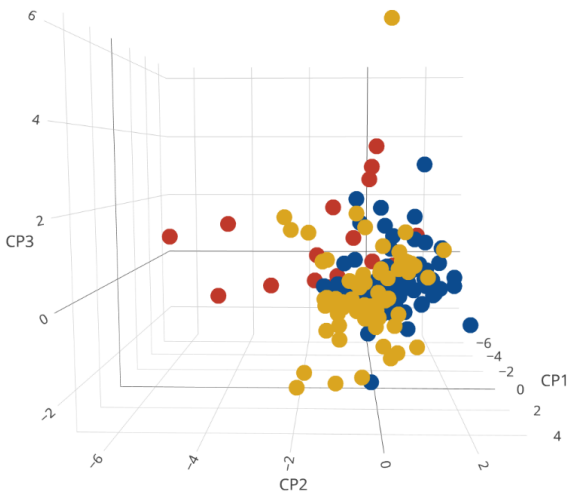

Figura 9: Representação dos clusters obtidos através do método de K-means, permitindo a visualização da CP2 e CP3. Vermelho - Cluster 1; Azul - Cluster 2; Amarelo - Cluster 3.

9 Anexo III - Tabela com os clusters obtidos por K-means e SES

Tabela 3: Comparação do cluster obtido para uma amostra aleatória dos países e o Socioeconomic Status

País	**Cluster**	**SES**
United Kingdom	2	85.16819
Barbados	2	74.515846
Equatorial Guinea	3	NA
Panama	2	78.675011
Cyprus	1	84.794022
Yemen	3	19.492294
Venezuela	2	60.248875
Tunisia	2	56.942608
Libya	2	68.144432
Lithuania	2	83.806778
United Arab Emirates	1	89.092285
Luxemboug	1	91.404564
Solomon Islands	3	NA
Malta	1	84.051285
Iran	2	76.485092
Burkina Faso	3	2.2252226
Armenia	2	62.782436
Turkmenistan	3	NA
Japan	2	89.416603
Kenya	3	27.877026
Vanuatu	3	NA
Moldova	2	46.91198
Pakistan	3	23.425098
Italy	2	85.016357
Sudan	3	7.4338622