Faculdade de Ciências da Universidade do Porto

**Estatística e Análise de Dados**

**Projeto 2**

Hélder Vieira up201503395

Miguel Tavares up200902937

Porto, 13 de Junho de 2021

Conteúdo

[1. Introdução 3](#_Toc74475528)

[2. Análise de dados 3](#_Toc74475529)

[*3.* *Clustering* 5](#_Toc74475530)

[3.1 K-means clustering 5](#_Toc74475531)

[3.2 Hierarchical clustering 6](#_Toc74475532)

[3.2 Similaridade 9](#_Toc74475533)

[4. Classificação 10](#_Toc74475534)

[4.1 LDA 10](#_Toc74475535)

[4.2 MLR 11](#_Toc74475536)

[5. Conclusão 13](#_Toc74475537)

# Introdução

No mundo atual existem cada vez mais dados disponíveis sobre o dia a dia de um cidadão normal, quer seja através da utilização dos mais diversos dispositivos como através da utilização de serviços coletivos em grandes centros urbanos. De forma a tirar partido desta enorme quantidade de dados para chegar a conclusões viáveis em tempo útil, é necessário fazer uso dos métodos disponíveis não só para a redução de variáveis e explicação de variância, mas também para subdivisões e agregações para melhor interpretação dos dados. Com isto é possível aumentar o rendimento das análises dos dados.

Neste projeto foi proposto um estudo dos dados recorrendo aos diversos métodos de *clustering*. Na primeira parte deste relatório serão expostos os resultados obtidos através dos métodos, bem como a semelhança entre eles.

Na segunda parte serão construídos modelos de classificação através dos métodos da Análise Discriminante Linear (LDA) e da Regressão Logística Multinomial (MLR). Estes modelos serão construídos sobre a Análise de Componente Principais (PCA).

# Análise de dados

Para a realização deste trabalho foi escolhido um *dataset* que traduz o nível de felicidade, bem como outros indicadores, de diversos países, relativo a ano de 2019. Uma vez que cada entrada correspondia a um país, decidiu-se agrupar os dados pelo continente ao qual os países pertencem. Começando pelo *Score*, este é baseado nas respostas de um questionário sobre a avaliação da qualidade de vida da população. Na questão, conhecida como *Escada de Cantril* é pedido que se imagine uma escada com 10 degraus (0 em baixo e 10 no topo). O décimo degrau corresponde à melhor vida que o questionado poderia ter, e o primeiro, a pior.

Na Figura 1 podemos observar a média dos Scores dos diversos países pelo continente ao qual pertencem. Rapidamente se observa que continentes onde se encontram os países mais desenvolvidos são os que, em média, têm um *Score* mais elevado.

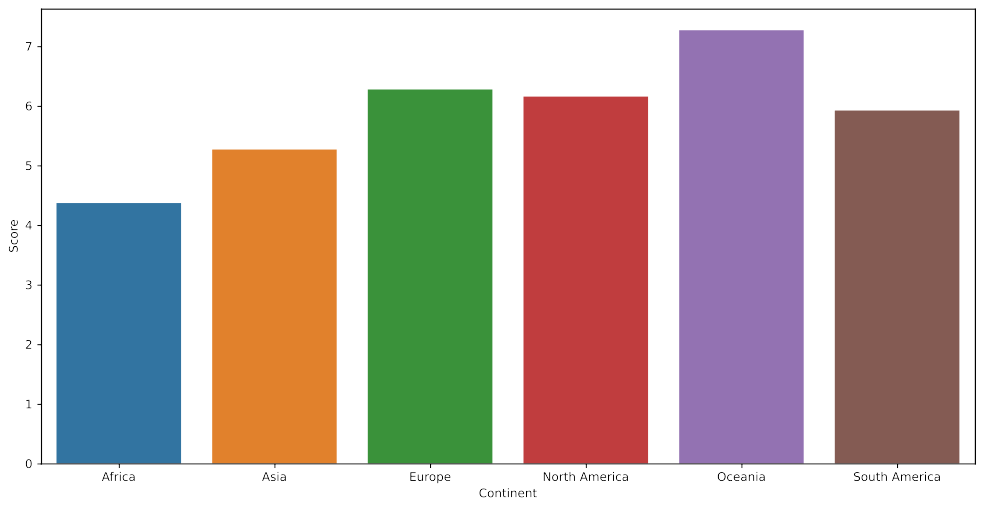


Figura 1- *Score de felicidade* médio por continente.

Existem outros factores que estão fortemente relacionados, mas sem impacto, com o Score, tais como o GDP per capita (PIB per capita), *Social support* (apoios sociais), *Healthy life expectancy* (esperança média de vida*), Freedom to make life choices* (liberdade), *Generosity* (generosidade) *e Perception of corruption* (Corrupção). É importante realçar que todos estes factores são também resultados de questionários, expecto o PIB per capita e a esperança média de vida. Uma nota para a variável Corrupção. Esta não representa o quão corrupto é um país, mas sim a capacidade da população em detectar/identificar corrupção.

Como observável na Figura 3, todos os indicadores seguem o *Score*. Relativamente à generosidade e corrupção, a Oceânia apresenta valores de média muito mais elevados que os restantes continentes devido ao facto de a sua amostra ser constituída por apenas dois países de elevado índice de desenvolvimento.

Na figura 2 (abaixo ilustrada) podemos observar a relação entre as diversas variáveis.

Como observável na Figura 2, todos os indicadores seguem o Score. Para concluir, através de uma análise rápida dos dados agrupados por continente, fica retida a ideia de que os países considerados desenvolvidos vão obter resultados mais elevados, e que esses mesmo países se encontram, na sua maioria, no hemisfério norte.

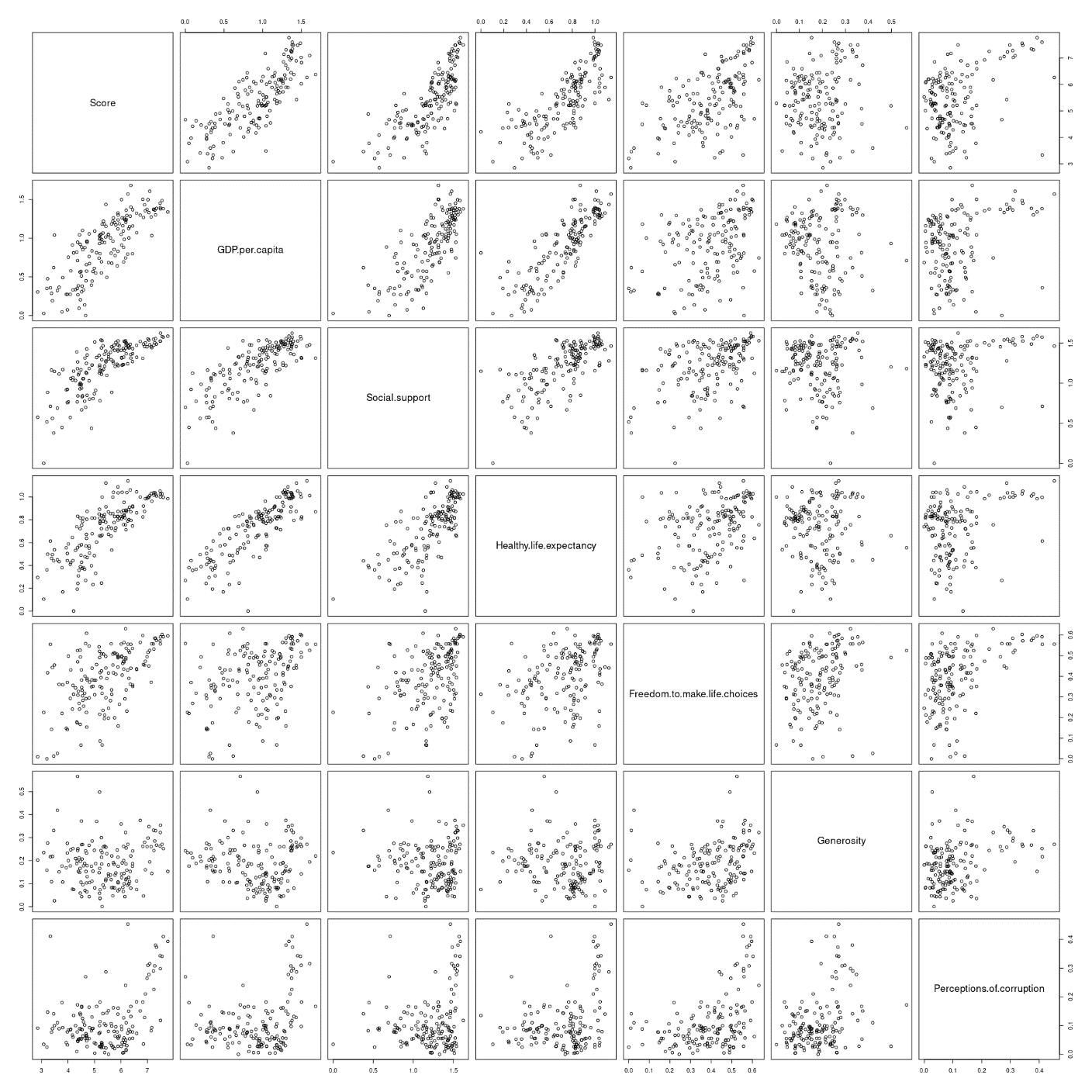


Figura 2- Relação entre as diversas varíaveis do *dataset*, note-se que uma grande parte aparenta possuir uma relação de proporcionalidade direta com o *Score*.

# *Clustering*

## 3.1 K-means clustering

Um dos métodos de *clustering* utilizados foi o *k-means clustering*. Este método é bastante utilizado, pois computacionalmente é pouco exigente, e é bastante simples de se implementar. Este baseia-se em agrupar os dados consoante a sua similaridade.

Dado que um dos hiperparâmetros é o número de *clusters* a utilizar, o primeiro passo foi obter o número ótimo dos mesmos, recorrendo ao *Elbow Method*.

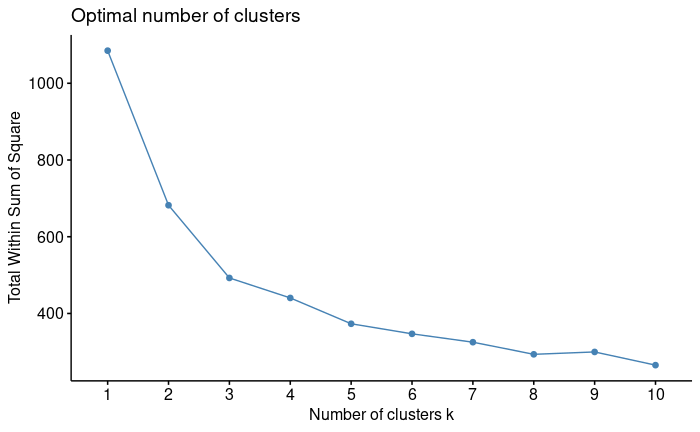


Figura 3- Gráfico utilizado para a determinação do número ideal de *clusters,* número de *clusters*  vs. *total within sum of squares.*

Observando o gráfico da figura 3, observa-se que o ‘*elbow’* *point* é em *k=3*. Uma vez, obtido o número ótimo de *clusters,* procedeu-se ao *clustering* dos dados.

Analisando rapidamente o resultado obtido na figura 4, vemos que num dos *clusters* se encontram apenas países menos desenvolvidos e noutro, países desenvolvidos. Contudo, no segundo *cluster,* a divisão não foi tão boa, conseguindo-se observar países dos dois lados do espectro no mesmo *cluster*.

## 3.2 Hierarchical clustering

Outro método de *clustering* é o *Hierarchical clustering*. A grande vantagem deste método é a sua não necessidade de se especificar o número de *clusters.* Contudo, recorrendo novamente ao *Elbow Method*, determinou-se o número óptimo de *clusters* para este método de *clustering.*

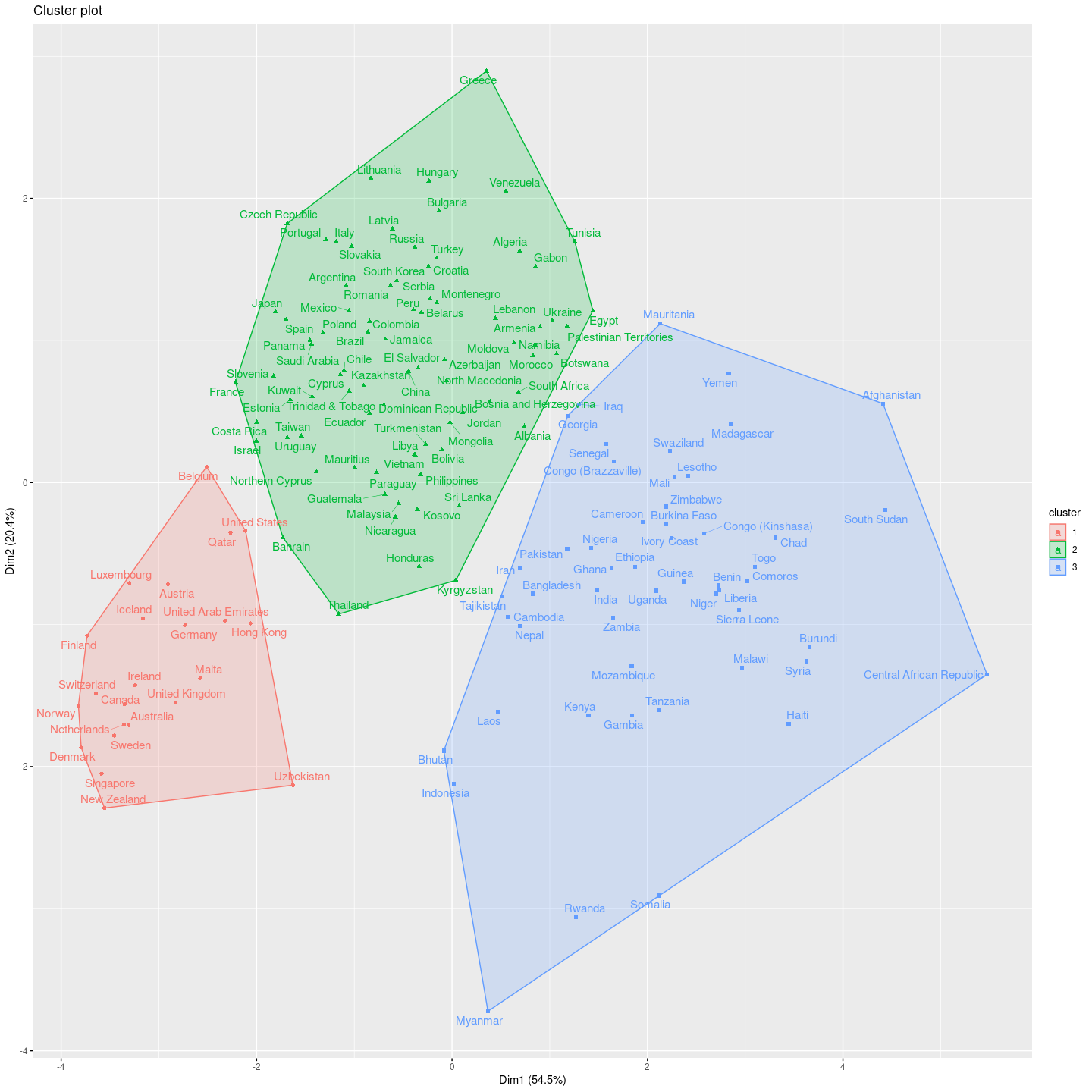


Figura 4- *Clusters* obtidos através de *k-means clustering*.

Pela análise da figura 5, verifica-se que novamente, *k=3* é o número ideal.

Uma das vantagens deste método de *clustering* é a possibilidade de obtermos dendrogramas da divisão obtida. É importante referir que o método de *linkage* utilizado foi o de ‘*Ward’*. Embora a primeira abordagem tenha sido o de *Average,* observou-se que dois dos três *clusters* iriam conter apenas países muito afastados dos restantes, ou seja, *outliers*.

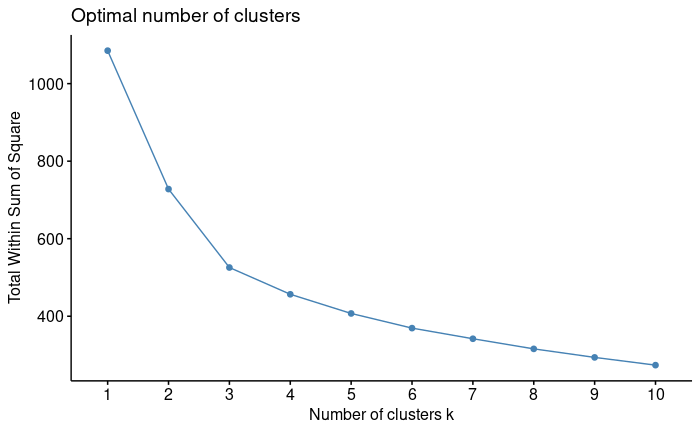
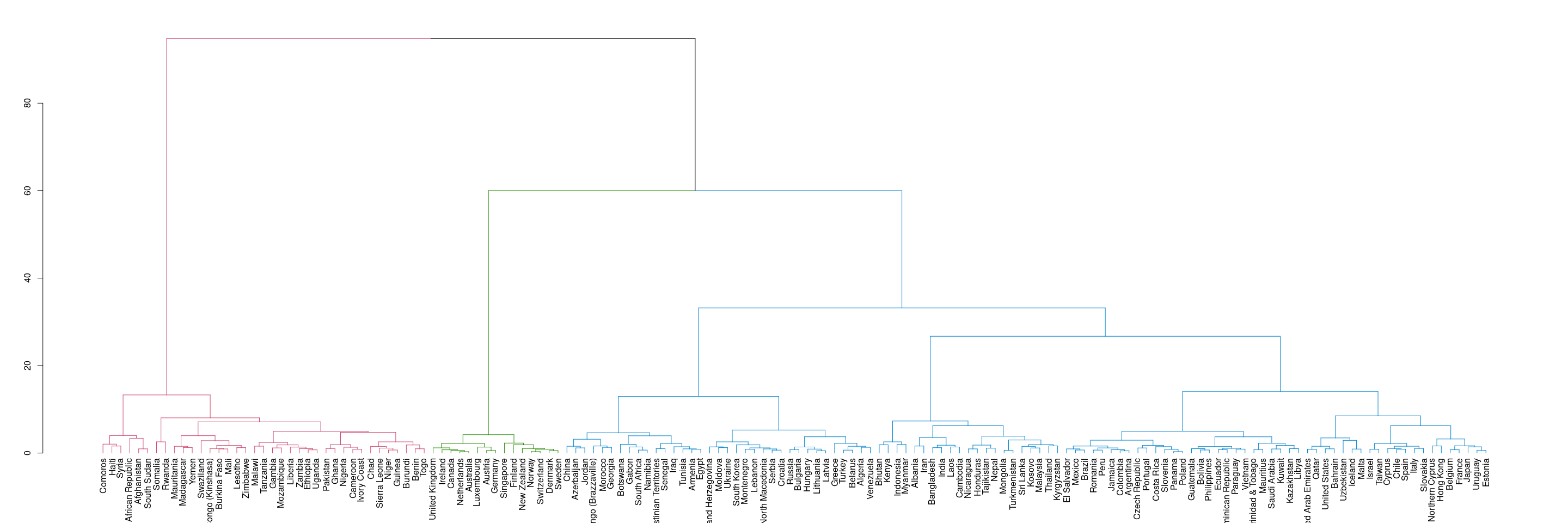


Figura 5- Gráfico utilizado para a determinação do número ideal de *clusters,* número de *clusters*  vs. *total within sum of squares.*

Na figura 6 é possível observar o dendrograma resultante da aplicação de *hierarchical clustering*.

Figura 6- Dendrograma resultante de hierarchical clustering



Como podemos observar, em conjunto com a representação gráfica dos *clusters* da figura 7*,* vemos que a divisão obtida é semelhante ao obtido no *k-means*. Um *cluster* mais pequeno em que apenas países desenvolvidos se encontram presentes e um *cluster* ligeiramente maior com apenas países menos desenvolvidos. Para terminar, podemos ver que o grande *cluster* continua a ter países dos dois lados do espectro, tal como obtido no *k-means.*

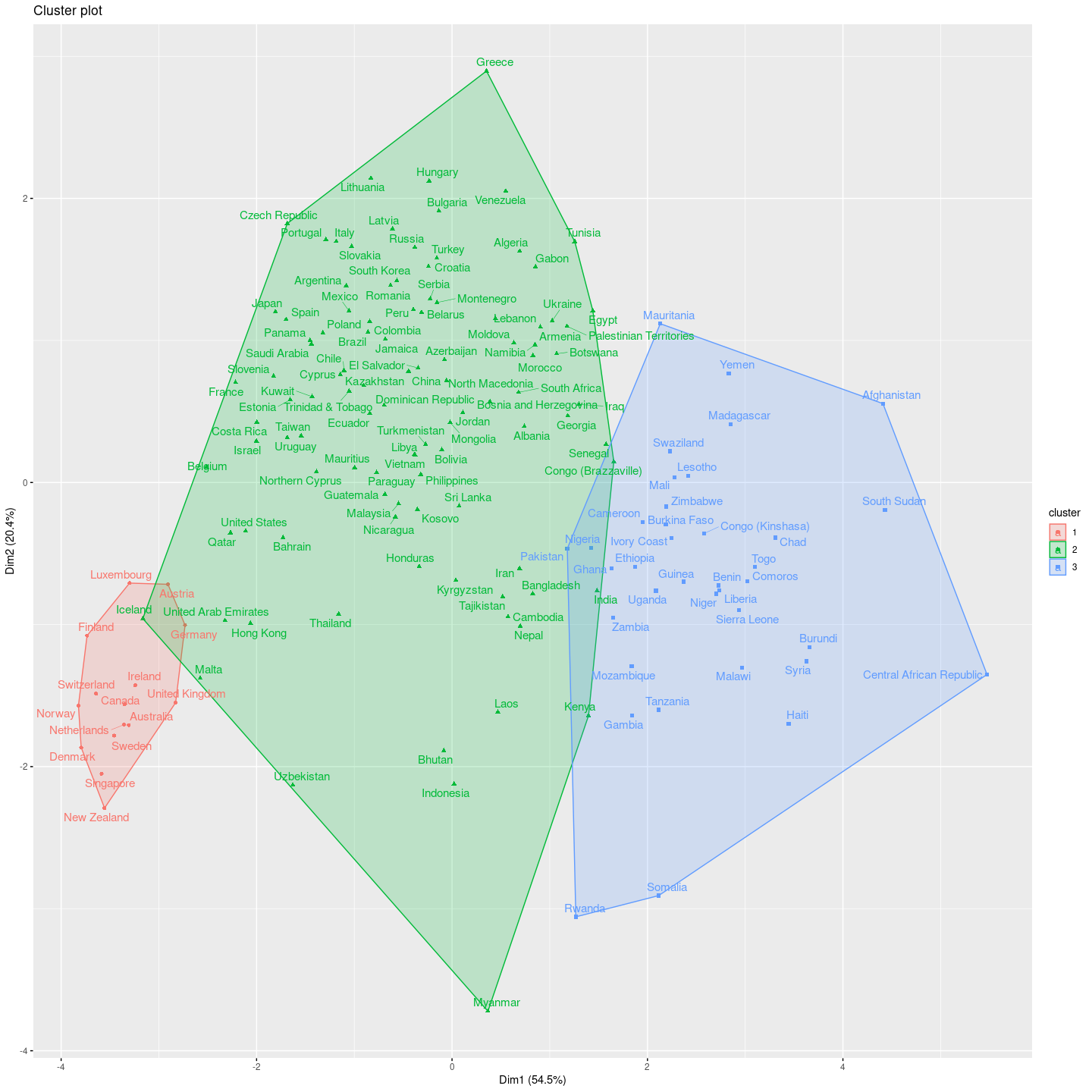


Figura 7- *Clusters* obtidos por *hierarchical clustering*.

## 3.2 Similaridade

De maneira a comparar os dois métodos de *clustering,* recorreu-se à análise de similariedade de *Jaccard* e *Rand*. Na tabela 1 podemos observar os resultados.

Tabela 1- Similaridade dos *clusters* através das métricas de Jaccard e Rand.

|  |  |
| --- | --- |
| Jaccard | Rand |
| 0.609 | 0.780 |

# Classificação

Os dados utilizados no processo de classificação são provenientes de análise via Análise de Componentes Principais (PCA) da qual resultaram duas componentes (que explicam aproximadamente 75% da variância). Foi realizada a divisão estratificada do conjunto de dados em dois subconjuntos: treino (60%) e teste (40%) sendo a variável alvo o continente.

Os métodos utilizados para classificação foram Regressão Logística Multinominal (MLR) e Análise Discriminante Linear (LDA) que são descritos nos subcapítulos seguintes.

## 4.1 LDA

O método de LDA foi aplicado através da função *lda()* do *package* ‘MASS’. O modelo foi treinado tendo como base o subconjunto de treino.

Os coeficientes obtidos para o discriminante linear 1 foram de -1.40 para a 1ª componente da PCA e -0.39 para a 2ª componente da PCA. Relativamente ao discriminante linear 2 os valores obtidos foram de 0.26 (1ª componente) e -0.95 (2ª componente).

Foi então, realizado a classificação sobre o subconjunto de treino tendo-se obtido a matrix de confusão representada na tabela 2. A *accuracy* obtida foi de 0.5579.

Tabela 2- Matrix de confusão para o dataset de treino através de LDA.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Real |  |  |  |
|  |  | Africa | Asia | Europe | North America | Oceania | South America |
|  | Africa | 20 | 6 | 0 | 1 | 0 | 0 |
|  | Asia | 7 | 13 | 4 | 2 | 0 | 1 |
| Previsto | Europe | 0 | 9 | 19 | 5 | 1 | 5 |
|  | North America | 0 | 0 | 0 | 0 | 0 | 0 |
|  | Oceania | 0 | 0 | 0 | 0 | 1 | 0 |
|  | South America | 0 | 0 | 1 | 0 | 0 | 0 |

Na tabela 3, é possível observar valores estatísticos relevantes específicos das classes.

Tabela 3- Medidas estatísticas da classificação específicas de classe no dataset de treino através de LDA.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Africa | Asia | Europe | North America | Oceania | South America |
| Sensibilidade | 0.7407 | 0.4643 | 0.7917 | 0 | 0.50000 | 0 |
| Especificidade | 0.8971 | 0.7910 | 0.7183 | 1 | 1 | 0.99 |
| *Accuracy* Balanceada | 0.820 | 0.628 | 0.755 | 0.50 | 0.75 | 0.494 |

Foi também realizada classificação no subconjunto de teste tendo-se obtido a matriz de confusão que se pode observar na tabela 3, a *accuracy* obtida foi de 0.6721.

Tabela 4- Matrix de confusão para o dataset de teste através de LDA.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Real |  |  |  |
|  |  | Africa | Asia | Europe | North America | Oceania | South America |
|  | Africa | 15 | 0 | 0 | 0 | 0 | 0 |
|  | Asia | 2 | 12 | 2 | 2 | 0 | 2 |
| Previsto | Europe | 1 | 6 | 14 | 3 | 0 | 2 |
|  | North America | 0 | 0 | 0 | 0 | 0 | 0 |
|  | Oceania | 0 | 0 | 0 | 0 | 0 | 0 |
|  | South America | 0 | 0 | 0 | 0 | 0 | 0 |

Verifica-se que o modelo de LDA apresenta uma forte tendência para classificar corretamente países Europeus e Africanos. Quanto a países da América do Norte e do Sul, verifica-se que a classificação destes tem tendência a ser incorreta, sendo que estes são geralmente classificados pelo modelo como pertencentes à Ásia ou Europa, relativamente à Oceânia a classificação também mostra-se imprecisa, devido, provavelmente, à sua baixa representatividade no conjunto de dados relativamente aos outros continentes.

## 4.2 MLR

O método de MLR foi aplicado através da função *multinom* do *package* ‘nnet’.

O modelo foi treinado e realizou-se a classificação do conjunto de treino, cuja matrix de confusão obtida encontra-se representada na tabela 5. A *accuracy* obtida foi de 0.6105.

Tabela 5- Matrix de confusão para o dataset de treino através de MLR.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Real |  |  |  |
|  |  | Africa | Asia | Europe | North America | Oceania | South America |
|  | Africa | 23 | 6 | 1 | 1 | 0 | 0 |
|  | Asia | 4 | 14 | 3 | 3 | 0 | 2 |
| Previsto | Europe | 0 | 7 | 20 | 4 | 1 | 4 |
|  | North America | 0 | 0 | 0 | 0 | 0 | 0 |
|  | Oceania | 0 | 1 | 0 | 0 | 1 | 0 |
|  | South America | 0 | 0 | 0 | 0 | 0 | 0 |

Na tabela 6 é possível observar valores estatísticos relevantes específicos das classes.

Tabela 6- - Medidas estatísticas da classificação específicas de classe obtidas no dataset de treino através de MLR.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Africa | Asia | Europe | North America | Oceania | South America |
| Sensibilidade | 0.85 | 0.5 | 0.83 | 0 | 0.5 | 0 |
| Especificidade | 0.88 | 0.82 | 0.77 | 1 | 0.99 | 1 |
| *Accuracy* Balanceada | 0.87 | 0.66 | 0.80 | 0.5 | 0.74 | 0.5 |

Foi também realizada classificação dos países no subconjunto de teste, cuja matriz de confusão encontra-se exibida na tabela 6. A *accuracy* obtida foi de 0.623.

Tabela 7- Matrix de confusão para o dataset de teste através de MLR.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Real |  |  |  |
|  |  | Africa | Asia | Europe | North America | Oceania | South America |
|  | Africa | 16 | 3 | 0 | 0 | 0 | 0 |
|  | Asia | 2 | 11 | 4 | 2 | 0 | 1 |
| Previsto | Europe | 0 | 4 | 11 | 3 | 0 | 3 |
|  | North America | 0 | 0 | 0 | 0 | 0 | 0 |
|  | Oceania | 0 | 0 | 1 | 0 | 0 | 0 |
|  | South America | 0 | 0 | 0 | 0 | 0 | 0 |

Na tabela 8, é possível observar valores estatísticos relevantes específicos das classes na classificação do subconjunto de teste.

Tabela 8- Medidas estatísticas da classificação específicas de classe obtidas no dataset de teste através de MLR.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Africa | Asia | Europe | North America | Oceania | South America |
| Sensibilidade | 0.89 | 0.61 | 0.69 | 0 | 0 | 0 |
| Especificidade | 0.93 | 0.82 | 0.78 | 1 | 0 | 1 |
| *Accuracy* Balanceada | 0.91 | 0.70 | 0.73 | 0.5 | 0 | 0.5 |

De forma similar ao que se verificou na LDA, a MLR também apresenta baixa tendência a atribuir a classificação correta a países pertencentes a continentes com pouca representatividade, nomeadamente a Oceânia, América do Norte e América do Sul. Relativamente aos restantes continentes a classificação dos mesmos é satisfatória.

# Conclusão

A realização deste trabalho permitiu pôr em prática conhecimentos de diversas temáticas, nomeadamente *clustering* e classificação.

A execução de *clustering* foi realizada por dois métodos distintos, sendo que em ambos o número ideal de clusters foi o mesmo (3). Os *clusters* obtidos por ambos os métodos mostraram ter um grau de similaridade elevado.

Relativamente à classificação foram utilizados dois modelos distintos (LDA e MLR), sendo que as *accuracies* obtidas rondaram os 60% para ambos. Os resultados obtidos para estes foram bastante similares, sendo que se verificou uma baixa tendência a atribuir a classificação correta a continentes com baixa representatividade.