Pré-processamento global de dados: Uma análise do desenvolvimento humano

Felipe Arruda Brito
Dep. de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
felipeabrito@alu.ufc.br

Gustavo Vieira de Andrade
Dep. de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
avgustavo@alu.ufc.br

Lucas de Albuquerque P. Oliveira
Dep. de Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
lucasde@alu.ufc.br

***Resumo*—Diversos indicadores econômicos e sociais são utilizados para tentar entender como os países se encaixam a nível global. Nesse intuito, o presente estudo tem por objetivo performar um pré-processamento de dados em alguns dos principais indicadores de desenvolvimento humano, a fim de entender como diferentes países se comparam e se comportam em diferentes índices. Como fonte dos dados utilizados, optou-se pelo Gapminder, analisando 11 preditores para cada país, a saber: Expectativa de vida ao nascer, Produto Interno Bruto *per capita*, Número de Habitantes (população), Coeficiente de Gini, Acesso ao Saneamento Básico, Mortalidade Infantil até 5 anos, Força de Trabalho entre 15 e 64 anos, Porcentagem de trabalhadores na indústria, Índice de Desenvolvimento Humano (IDH), Total de Assassinatos por Ano e Geração de Energia Elétrica por Pessoa. Além disso, os países foram agrupados em 6 classes, cada uma correspondendo a um continente: América do Norte, América do Sul, Europa, África, Ásia e Oceania. Os resultados observados demonstram uma descrição dos dados para o ano de 2019. Além disso, para cada preditor, foi realizada uma análise monovariada como média, desvio-padrão e assimetria. Também foi performada uma análise monovariada orientada à classe, além de análises bivariada incondicional e multivariada incondicional. O código-fonte da análise é disponibilizado em Python, linguagem escolhida para o pré-processamento.**

***Index Terms*—Análise de Dados, Pré-processamento, Gapminder, Desenvolvimento**

I. Introdução

A análise do desenvolvimento humano pode depender de vários fatores, como expectativa de vida, qualidade da saúde no país, acesso à educação básica, empregabilidade, segurança, dentre outros fatores.

Nesse sentido, vê-se a importância de entender como esses fatores estão interligados entre si, em uma análise abrangente para diversos países, a fim de compreender não só como esses índices estão presentes satisfatoriamente em um estado, como também entender como os continentes em si comportam-se como classes, para uma análise holística.

O presente trabalho busca responder ao questionamento do que os países desenvolvidos tem em comum uns com os outros, a partir do entendimento de como cada classe, ou seja, continente, se comporta. Para tanto, foram escolhidos 11 preditores para as 208 amostras de países, separadas em 6 classes de continentes.

A presente análise baseia-se em um aprendizado não supervisionado. Isto é, não existe uma variável de resposta para supervisionar a análise, pois cada vetor de medida x_i não possui uma resposta associada y_i [4].

A partir dos dados aqui analisados, futuros estudos para desenvolvimento de modelos poderão ser realizados. Após o entendimento dos dados e dos objetivos da análise, será possível a construção de modelos [3].

II. Métodos

Em primeiro lugar, os dados serão adquiridos e descritos mediante o número *N* de observações e escolha dos *D* preditores da análise. Em seguida, as amostras serão agrupadas em *L* classes e contabilizadas para cada uma, entendendo a distribuição.

Adiante, será performada uma análise incondicional monovariada de cada um dos *D* preditores. Para tanto, serão plotados histogramas e *box-plots* para cálculo da média μ_d, desvio padrão σ_d e assimetria γ_d, para $d = 1, \ldots, D$. Cada uma dessas estatísticas será calculada por meio do código-fonte.

Em seguida, a análise será condicional às *L* classes dos continentes. Cada um dos descritores será analisado com os mesmos padrões de média, desvio-padrão e assimetria, mas, desta vez, levando em consideração o conjunto das amostras. A metodologia terá por norte a utilização de histogramas e *box-plots*, além do cálculo estatístico.

Posteriormente, os preditores serão analisados conjuntamente, dois a dois, verificando a relação entre estas variáveis. Como metodologia, serão utilizados gráficos de dispersão para cada combinação de variáveis, a fim de identificar, visualmente, a possibilidade de relação entre elas. Também será montada uma tabela de correlação com cores.

Por fim, em uma análise multivariada incondicional, será aplicada a técnica PCA (Análise de Componentes Principais), que ajuda a reduzir a complexidade de conjuntos de dados, identificando e destacando os padrões mais importantes. Os 2 componentes principais com maiores autovalores serão utilizados para confeccionar um gráfico de dispersão, que servirá de análise das características da base de dados.

III. Resultados

Para cada um dos objetivos metodológicos propostos, os resultados serão analisados seguindo os exemplos encontrados na literatura referenciada, bem como as práticas padrões para a linguagem Python, que serão detalhadas conforme cada caso.

A. *Descrição dos dados*

Optou-se por utilizar a base de do Gapminder [2], criado com o intuito de evitar visões estereotipadas e antiquadas sobre o mundo. A plataforma justifica uma melhor aquisição de seus dados em virtude das mudanças advindas do mundo em desenvolvimento, já que existem diversas colaborações internacionais entre países e instituições para troca de informações.

Cada amostra é um dos 208 países da base de dados, para o ano de 2019. Este ano foi escolhido por ser o mais completo em número de dados pois, para vários preditores, não há informação posterior a esta data.

Foram escolhidos 11 preditores, basilares para esta análise:

- Expectativa de vida ao nascer - Quantos anos, em média, uma pessoa de um determinado país tende a viver, considerada a sua data de nascimento,
- Produto Interno Bruto *per capita* - O Produto Interno Bruto de cada países dividido pelo seu número de habitantes,
- Número de Habitantes (população) - A quantidade total de habitantes de um determinado país,
- Coeficiente de Gini - Índice que mede o grau de concentração de renda em determinado grupo, medindo a desigualdade entre países. Quanto menor o índice, menor a desigualdade,
- Acesso a Saneamento Básico - Qual a porcentagem da população de um determinado país tem acesso à saneamento básico,
- Mortalidade Infantil até 5 anos - Quantas crianças de até 5 anos morrem a cada 1000 nascimentos,
- Força de Trabalho entre 15 e 64 anos - Porcentagem de trabalhadores nessa faixa etária,
- Porcentagem de trabalhadores na indústria - Porcentagem de trabalhadores inseridos no mercado de trabalho do segundo setor produtivo,
- Índice de Desenvolvimento Humano (IDH) - Índice estatístico que classifica os países de acordo com o desenvolvimento humano. Inclui, em sua análise, a indicação de renda *per capita*, expectativa de vida, escolaridade, dentre outros
- Total de Assassinatos por Ano - Número bruto de assassinatos em cada país em cada ano,
- Geração de Energia Elétrica por Pessoa - Geração Elétrica em Kilowatt-horas por pessoa em cada uma das amostras.

Além disso, 6 classes foram agrupadas:

- América do Norte e América Central (North America) - Região situada no hemisfério norte, ligada ao sul pela América Central. Composta por 34 países, estende-se do Ártico ao Panamá,
- América do Sul - 12 países, delimitados ao Norte pelo Panamá.
- Europa - 42 países, delimitados a oeste pelo Oceano Atlântico e a leste pelos Montes Urais
- África - 53 países, delimitados pelo Mar Mediterrâneo ao Norte, pelo Oceano Índico ao sudeste, e pelo Oceano Atlântico ao oeste
- Ásia - Maior continente em área e população, composto por 48 países. Localizado primariamente nos hemisférios oriental e norte, é delimitado pelos Montes Urais a oeste e pelo Oceano Pacífico a leste.
- Oceania - 19 países, localizados no hemisfério sul, entre o Oceano Índico e o Oceano Pacífico. Inclui a Austrália, a Nova Zelândia, e diversas ilhas.

Essa é, portanto, a descrição dos dados utilizados.

B. *Análise monovariada incondicional*

Iniciando a análise monovariada incondicional, há o plot dos histogramas para cada um dos preditores. A partir da análise do código [5], é possível ter acesso a todos os histogramas. Inicialmente, veja-se os 6 primeiros preditores:

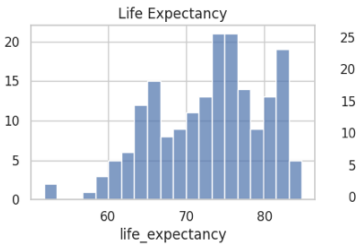

Figura 1. Expectativa de Vida

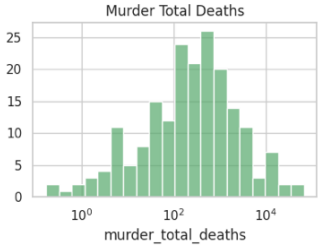

Figura 2. Total de assassinatos

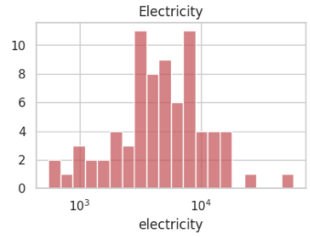

Figura 3. Energia elétrica

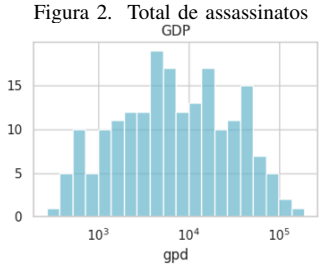

Figura 4. PIB *per capita*

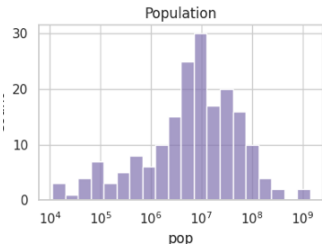

Figura 5. Número de Habitantes

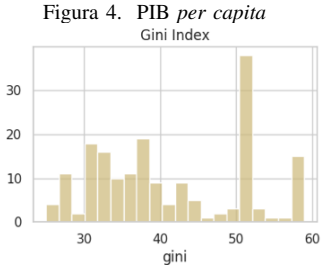

Figura 6. Coeficiente de Gini

Sobre a expectativa de vida, verificamos que essa se concentra, a maior parte, entre a faixa de 65 e 80 anos. Para o caso do número bruto de assassinatos, optou-se por utilizar uma escala logarítmica, já que alguns países possuem um número inferior a 100, enquanto outros possuem mais que 10 mil assassinatos para o ano de 2019, como é o caso do Brasil. Em virtude também da grande variação de número, a escala logarítmica também foi utilizada para a eletricidade gerada, para o PIB *per capita*, e para o número de habitantes.

Vê-se também que esse comportamento por meio dos *boxplots*, além de analisar-se que os dados referentes aos assassinatos brutos, PIB per capita e população possuem diversos *outliers*:

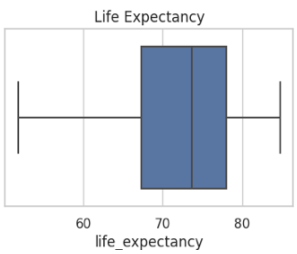

Figura 7. Expectativa de Vida

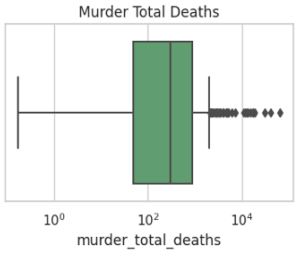

Figura 8. Total de assassinatos

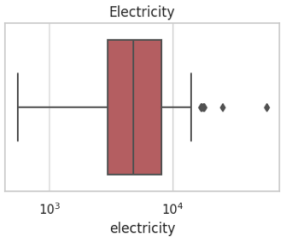

Figura 9. Energia elétrica

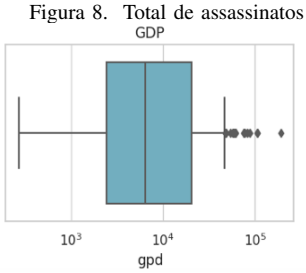

Figura 10. PIB *per capita*

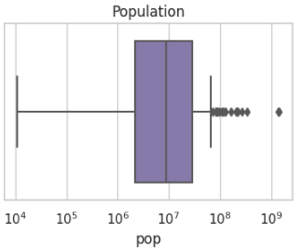

Figura 11. Número de Habitantes

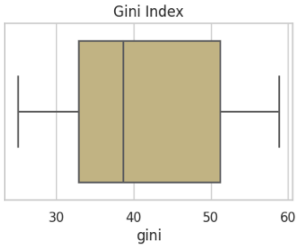

Figura 12. Coeficiente de Gini

Analisando os outros 5 preditores omitidos, também é possível fazer algumas observações. Para a mortalidade infantil, vê-se que a maior parte das amostras está inferior a um número de 40 para cada 1000 habitantes. Sobre a força de trabalho, vemos que nenhum país apresenta menos de 40 por cento ou mais de 90 por cento da força de trabalho composta por indivíduos entre 15 e 64 anos. Sobre os trabalhadores industriais, a porcentagem varia, majoritariamente, entre 10 e 30 por cento. Para o índice de saneamento, uma parte considerável dos países atinge o máximo possível. Por sua vez, no caso do IDH, nota-se que o índice está bem distribuído para os possíveis valores.

Verificando, nesse momento, os *box-plots* para os 5 últimos preditores, presentes no código disponibilizado: as considerações feitas sobre a mortalidade infantil e a força de trabalho são condizentes com a realidade. Em verdade, cerca de 50 por cento das amostras possui uma força de trabalho para a faixa etária escolhida entre 61 a 76 por cento da força de trabalho total. Enquanto isso, apenas metade das amostras possuem trabalhadores na indústria entre 15 e 25 por cento.

Parte-se agora para o cálculo da média μ_d, desvio padrão σ_d e assimetria γ_d, para $d = 1, \ldots, D$. Para tanto, utilizou-se as funções *mean*, *std* e *skew* na linguagem python. Os resultados foram consolidados e dispostos na tabela abaixo:

Tabela I
ANÁLISE MONOVARIADA INCONDICIONAL

	Preditores	Média	Desvio Padrão	Assimetria
0	life_expectancy	72.843	7.0368	-0.40705
1	murder	2011.60	6605.22	6.6189
2	electricity	6692.21	7581.48	4.5098
3	gpd	16793.01	24244.56	3.1328
4	pop	39893563.3	148982263.7	8.3164
5	gini	41.27	9.7267	0.24654
6	child	27.161	27.4698	1.4306
7	labour	68.30	11.1313	-0.6223
8	industry	19.59	7.9096	0.4122
9	hdi	0.7271	0.1502	-0.3129
10	sanitation	77.174	28.2703	-1.0833

Alguns detalhes merecem pontuação: A expectativa de vida, em geral, possui pouca assimetria e razoável desvio-padrão. Por sua vez, o número de assassinatos e a população possuem grandes assimetrias e altos desvios-padrão. Outros preditores possuem baixa assimetria, como o IDH, o Saneamento Básico, o Coeficiente de Gini e os relacionados a empregabilidade.

C. Análise monovariada condicional à classe

O procedimento da seção anterior será repetido, mas, dessa vez, considerando cada continente (ou classe) analisado. Por questões de brevidade, os histogramas e/ou *box-plots* estão em tamanho menor que na seção anterior, havendo a omissão de alguns. Para acesso completo, ver [5].

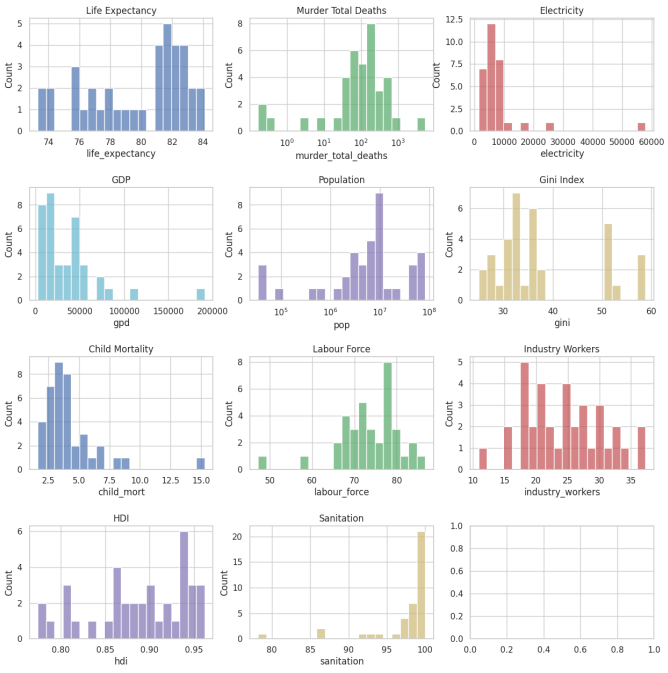

Figura 13. Europa - Histogramas

Iniciando pela Europa, vemos que o IDH se encontra em níveis bastante elevados, chegando a casa dos 0.95. Além disso o número total de mortes se concentra, em sua maioria, na ordem de 10^2. O coeficiente de Gini se concentra, em maior

parte, abaixo da média geral encontrada na seção anterior. A mortalidade infantil é extremamente baixa, e os índices de saneamento básico se encontram próximos de 100 por cento. A Europa possui a maior média de expectativa de vida, mas não o menor número de assassinatos, que pertence à Oceania. Possui a maior renda per capita, um coeficiente de Gini baixo, além de ínfima mortalidade infantil.

No caso da Ásia, vê-se que a expectativa de vida é superior a da África, apesar de a população possuir um número bastante superior:

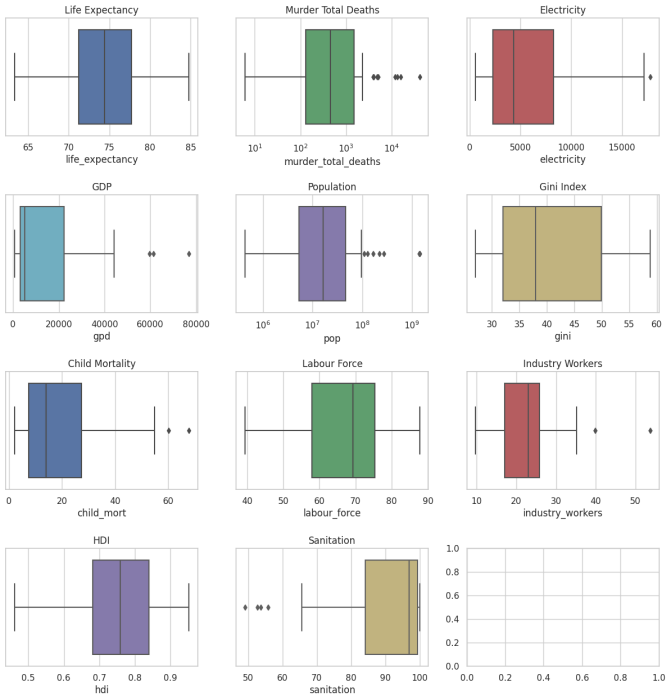

Figura 14. Ásia - Box-plots

Para o continente africano, vemos que há uma maior heterogeneidade nas distribuições, a exemplo da força de trabalho.

Tabela II
ÁFRICA

	Preditores	Média	Desvio Padrão	Assimetria
0	life_expectancy	65.787	5.3616	-0.04028
1	murder	1543.97	3237.29	4.1594
2	electricity	2325.0	1382.71	1.4794
3	gpd	2608.45	3132.70	2.6742
4	pop	24628056.6	35828169.6	3.0777
5	gini	44.098	9.8490	-0.1760
6	child	58.679	28.6094	0.3585
7	labour_force	65.875	11.7977	-0.1182
8	industry	14.112	7.6033	0.4794
9	hdi	0.5679	0.1062	0.5713
10	sanitation	42.747	26.2080	0.7629

As expectativas de vida possuem níveis inferiores, além de o PIB per capita ser bem menor. Os menores índices de Saneamento Básico estão presentes nesta classe. Possui um IDH baixo, bem como índice de Saneamento Básico inferior às outras classes. A geração de energia elétrica é bem inferior à apontada pela Europa, mas é semelhante à média da América do Sul.

Tabela III
AMÉRICA DO SUL

	Preditores	Média	Desvio Padrão	Assimetria
0	life_expectancy	75.883	3.8181	-0.9246
1	murder	8541.73	18813.68	3.0351
2	electricity	2661.43	1025.40	0.6251
3	gpd	8607.27	3969.81	0.8333
4	pop	35679916.7	57963332.3	2.9783
5	gini	44.958	9.5655	-0.0913
6	child	16.539	7.3555	0.4485
7	labour_force	69.808	7.2424	-0.5901
8	industry	19.808	3.1164	0.2018
9	hdi	0.7701	0.0509	0.7156
10	sanitation	88.609	10.1427	-1.5960

A América do Sul possui uma média de mortalidade infantil mais elevada que a Europa, mas inferior à África. Possui uma média populacional bem superior à América do Norte e Central, além de semelhantes números de forças de trabalho e de porcentagem de trabalhadores na indústria. Em geral, percebe-se que muitos dos preditores da América do Norte e Central são semelhantes aos da América do Sul, o que demonstra que, possivelmente, a América Central tenha muito em comum com a América Latina como um todo, em que Estados Unidos e Canadá despontam como *outliers*.

Tabela IV
AMÉRICA DO NORTE

	Preditores	Média	Desvio Padrão	Assimetria
0	life_expectancy	75.030	3.7284	-0.7593
1	murder	2633.06	6642.33	3.5849
2	electricity	10477.5	6460.70	-0.1434
3	gpd	22707.10	24358.30	2.0325
4	pop	25444878.3	72173341.6	3.9925
5	gini	42.722	8.9249	-0.0723
6	child	16.670	12.4466	2.6235
7	labour_force	70.276	6.7926	-0.9875
8	industry	18.190	4.2095	-0.4539
9	hdi	0.7600	0.0903	-0.2790
10	sanitation	88.543	14.4575	-2.2975

Percebe-se que a Ásia possui a maior média populacional, o que pode ser a razão de não alcançar uma renda *per capita* tão alta. Entretanto, seu coeficiente de Gini demonstra valores razoáveis. Possui em média 22 por cento dos trabalhadores na indústria e uma baixa mortalidade infantil.

No caso da Oceania, vê-se que os histogramas estão bem distribuídos e heterogêneos. Isso pode se dar em virtude de ser um continente com diversas ilhas com espaço amostral de população diminuto, o que certamente pode influenciar na magnitude de alguns dos valore.

Com base em todas as análises realizadas, bem como estudo dos diversos gráficos e tabelas, pode-se chegar a conclusão de que a expectativa de vida ao nascer e a renda per capita são dois bons preditores para separar as classes, apesar de uma

divisão que leve em conta apenas essas duas características apresente limitações. Alguns índices como mortalidade infantil levantam indícios de estarem correlacionados com expectativa de vida, o que será aprofundado no tópico seguinte. Além disso, o IDH, por ser um preditor que possui influencia dos outros, também é uma boa variável para tentar resumir as classes, apesar de possíveis falhas.

D. Análise bivariada incondicional

Na análise bivariada incondicional, o intuito é de verificar como os preditores se relacionam. De início, uma visualização da correlação por meio da função *corr* do Python.

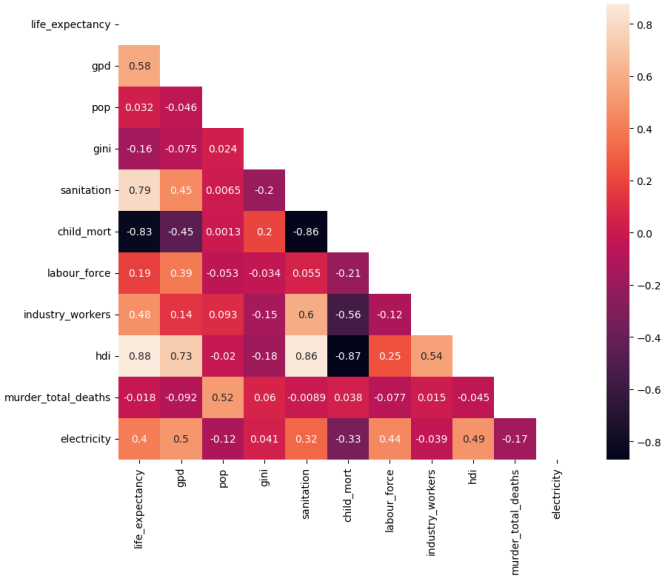

Figura 15. Tabela de Correlação entre os Preditores

Iniciando pelas maiores correlações, conforme o esperado e indicado nas seções anteriores, a mortalidade infantil possui uma grande correlação negativa com a expectativa de vida. O saneamento básico de um país também está fortemente ligado aos índices de mortalidade infantil. E, além disso, a mortalidade infantil se demonstra como um grande sugestivo de baixo desenvolvimento, já que possui grande correlação negativa com o Índice de Desenvolvimento Humano de uma amostra.

Sobre a renda *per capita*, há uma baixa correlação com o número de habitantes, mas uma correlação considerável com expectativa de vida. Uma maior renda per capita não está correlacionada com baixo número de assassinatos, mas está relacionada de forma considerável com uma baixa mortalidade infantil e alto IDH, além de força de trabalho mais concentrada entre 15 e 64 anos.

Sobre o coeficiente de Gini, há pouca correlação com os outros preditores, notando-se, de forma previsível, uma certa correlação com o IDH de forma negativa. Vale ressaltar, também, que o coeficiente de gini está levemente atrelado ao saneamento e mortalidade infantil, bem como número de trabalhadores na indústria. Apesar de essa correlação possuir baixa magnitude, é interessante notar que a natureza do índice, por si só, não permite grandes variações em seus valores, sendo mais interessante uma análise relativa de variação do índice que uma análise absoluta.

Interessante destacar que a força de trabalho maior concentrada entre 15 e 64 anos está razoavelmente relacionada ao IDH. Enquanto isso, a força de trabalho na indústria possui uma correlação de 0.54 com o desenvolvimento humano.

De modo curioso, o IDH possui mínima correlação com o número de habitantes. É possível concluir, de certo modo, que o grande volume de habitantes em um país não é indicativo ou causa de subdesenvolvimento.

Em seguida, foram montadas as tabelas de dispersão, escolhendo-se os melhores pares para a confecção das tabelas [1].

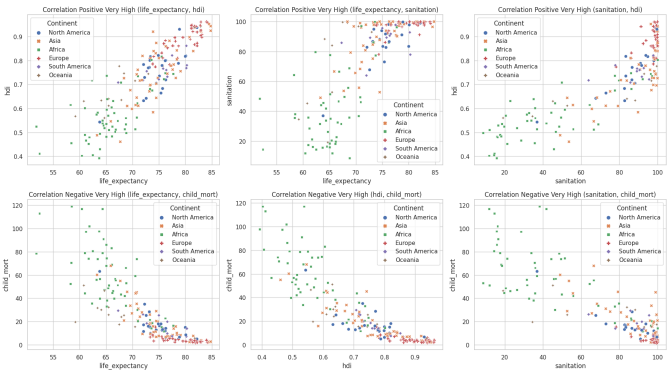

Figura 16. Tabela de dispersão entre os Preditores

Para sustentar a análise, vemos uma correlação muito alta entre os seguintes pares:

- IDH e Expectativa de Vida
- Saneamento Básico e Expectativa de Vida
- IDH e Saneamento Básico
- Mortalidade Infantil e Expectativa de Vida (negativo)
- Mortalidade Infantil e IDH (negativo)
- Mortalidade Infantil e Saneamento Básico (negativo)

Ou seja, esses fatores (IDH, expectativa de vida, saneamento básico e mortalidade infantil), parecem um bom norte a ser tomado quanto à proporção de políticas públicas. Apesar de serem um tanto abstratos quando pensa-se em medidas concretas, o gráfico permite concluir que são áreas interessantes para os gestores estudarem ao gerir uma determinada população.

Pela análise das tabelas de outras tabelas de dispersão aqui omitidas, também se verificou uma alta correlação entre os seguintes preditores, que corroboram a argumentação feita sobre a tabela de correlação:

- PIB *per capita* e expectativa de vida
- População e número absoluto de assassinatos
- PIB *per capita* e IDH
- IDH e porcentagem de trabalhadores na indústria
- Saneamento básico e porcentagem de trabalhadores na indústria
- Mortalidade infantil e porcentagem de trabalhadores na indústria (negativa)

Importante ressaltar que, para alguns desses preditores, escalas logarítimicas foram adotadas, pelos motivos já expostos na Análise monovariada incondicional.

E. Análise multivariada incondicional

A partir do gráfico de dispersão a seguir, pode-se observar que há muitas classes sobrepostas. Isso indica que apesar de PC1 e PC2 serem os Componentes Principais, ou seja, aqueles que possuem os maiores autovalores, eles não representam bem a separação de classes. Algo que é visto também através do gráfico da figura 17, pois os PC1 e PC2 somam menos de 60% da variação dos dados.

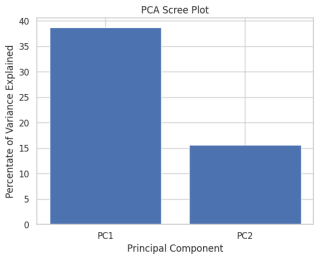

Figura 17. Percentual dos Componentes Principais

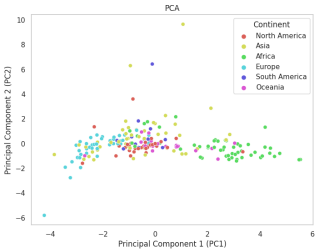

Figura 18. Gráfico de dispersão

Observa-se também que os continentes mais difíceis de serem discriminados são a América do Norte, América do Sul, Asia e Oceania, que estão na área mais central do gráfico. Já os continentes África e Europa encontram-se mais afastados do centro do gráfico e com menor sobreposição, de modo que são mais fáceis de serem diferenciados entre si e visualmente apresentam uma fronteira linear. Além disso, América do Norte e Europa também podem ser vistos dessa forma, pois apesar de estarem mais próximos apresentam uma fronteira visualmente linear.

A fim de poder testar os dados e as funções utilizadas, o código está disponibilizado de forma *open-source* no sítio [5].

REFERÊNCIAS

[1] Ahemaitihali, A.; Dong, Z. Spatiotemporal Characteristics Analysis and Driving Forces Assessment of Flash Floods in Altay. Water2022,14,331. https://doi.org/ 10.3390/w14030331

[2] Gapminder Foundation, ”Gapminder World Data,”[Online]. Available: https://www.gapminder.org. [Accessed: 06/09/2023].

[3] James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

[4] Kuhn, M. and Johnson, K., 2013. Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

[5] Pré-processamento global de dados: Uma análise do desenvolvimento humano - Código em Python. Disponível em: Google Colab