PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

João Carlos Valadares Ribeiro Filho

Impacto de indicadores socioeconômicos na quantidade de pessoas com AIDS

Belo Horizonte
2020

João Carlos Valadares Ribeiro Filho

Impacto de indicadores socioeconômicos na quantidade de pessoas com AIDS

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2020

SUMÁRIO

1. Introdução

1.1. Contextualização

O tema escolhido para esse trabalho de conclusão de curso de pós-graduação veio do interesse pela relação entre o desenvolvimento social e econômico e as áreas básicas da sociedade (segurança, educação e saúde). Durante anos, vários planos e diretrizes foram realizados a fim de aprimorar essas áreas. Isso refletiu na história e posicionou os países em diferentes graus de evolução socioeconômica.

Por exemplo, uma política de conscientização nacional contra uma epidemia pode focar em aprimorar a área da saúde através do aumento da qualidade da educação (seja valorizando os salários dos professores ou aumentando a quantidade desses nas instituições de ensino). Por outro lado, também é válido acreditar que melhores resultados serão adquiridos reduzindo a quantidade de desempregados ou aumentando o Produto Interno Bruto (PIB) do país.

Através da comparação de dados educacionais e econômicos, é possível analisar as medidas tomadas por diferentes países e compreender o resultado disso durante os anos. Logo, a relação entre a taxa de infectados por um vírus e o desenvolvimento educacional e econômico pode ser medida por indicadores valorados em uma linha de tempo. Esses indicadores abrangeriam dados ligados ao percentual de investimentos em áreas específicas, quantidade de professores, força de trabalho, quantidade de pessoas com acesso à informação e etc.

Dessa forma, com o objetivo de analisar e cruzar dados socioeconômicos de diferentes países, foram utilizados dados retirados das bases dos órgãos internacionais World Bank, OECD, ILOSTAT, UNICEF, UNESCO e UNAIDS.

- **World Bank:** ou Banco Mundial, foi criado após a Segunda Guerra Mundial com o objetivo de reajustar a economia mundial. Sua visão organizacional é "*reduzir a pobreza e gerar prosperidade compartilhada de uma maneira sustentável*". Além disso, o Banco Mundial realiza pesquisas e coletas de dados para expor uma perspectiva sobre a pobreza e as dificuldades que os seus países membros enfrentam.
- **OECD:** ou Organização para a Cooperação e o Desenvolvimento Econômico, "*constitui foro composto por 35 países dedicado à promoção de*

padrões convergentes em vários temas, como questões econômicas, financeiras, comerciais, sociais e ambientais", segundo o Itamaraty.

- **ILOSTAT:** ou Departamento de Estatística da Organização Internacional do Trabalho, objetiva prover estatísticas trabalhistas relevantes, oportunas e comparáveis.
- **UNICEF:** ou Fundo das Nações Unidas para a Infância, foca suas atividades na proteção dos direitos da criança.
- **UNESCO:** ou Organização das Nações Unidas para a Educação, a Ciência e a Cultura. Uma das formas da UNESCO conquistar seus objetivos (principalmente a erradicação do analfabetismo no mundo) é, também, focar em programas educacionais. Isso proporcionou a construção de uma base de dados sobre a situação da educação de vários países.
- **UNAIDS:** é uma parceria que encoraja, mobiliza e apoia países para alcançar o acesso universal à prevenção, ao tratamento e aos cuidados relacionados ao HIV.

Os dados provenientes dessas entidades já foram compilados e armazenados, pela UNESCO e pelo World Bank, em um conjunto de dados (*dataset*). Esse conjunto é composto por vários indicadores relevantes ao tema educacional e econômico. Os valores desses indicadores estão definidos em uma linha temporal. Há também o conjunto de dados provido pela UNAIDS que apresenta valores referentes ao número de pessoas com AIDS e também está distribuído em uma linha do tempo.

Todos os arquivos relacionados a esses *datasets*, *scripts*, arquivos intermediários de dados e dicionários estão armazenados no repositório que é informado no final deste documento (seção "**7. Links**").

1.2. O problema proposto

Esse trabalho propõe cruzar dados de 11 indicadores (temas da educação, ciência e economia) de um grupo de países e, a partir disso, realizar análises e classificações para concluir a relação, durante o período de 2010 a 2016, entre os indicadores socioeconômicos e sua possível influência na taxa de pessoas infectadas com HIV. Além disso, também serão aplicados, separadamente, os dados referentes ao Brasil e aos Estados Unidos da América para testar a eficiência da previsão desses resultados de classificação e previsão.

Perguntas propostas pelo 5-W:

- **Por que a predição de indicadores e o cruzamento dos seus dados é importante?**

  Uma ação a fim de atingir uma meta sanitária (definida em uma política nacional) de reduzir o número de pessoas atingidas por uma doença, pode focar em direcionar mais investimentos na educação fundamental e, assim, com o passar dos anos, isso se concretizar, por exemplo. Ou seja, ao comparar os dados dos indicadores socioeconômicos, será possível direcionar ações e investimentos para atingir objetivos ou sanar crises públicas, especificamente, o número de infectados por HIV.

  Além disso, ao analisar a linha de evolução dos valores desses indicadores, será possível afirmar, com um certo grau de confiança, o seu próximo resultado (também medido em outro indicador) e, assim, facilitar a tomada de decisões e a execução de diretrizes públicas.

- **Quais os dados analisados e suas respectivas fontes?**

  Os dados utilizados nesse trabalho são provenientes de um conjunto de dados da UNAIDS e de outro conjunto que agrega as bases de dados do World Bank, OECD, ILOSTAT, UNICEF e UNESCO.

  O primeiro *dataset*, o da UNAIDS, apresenta valores agrupados por países e distribuídos entre os anos de 2010 a 2019. Já o segundo, *dataset*, possui cerca de 4573 indicadores. Porém, a fim de solucionar o questionamento central desse trabalho, só foi necessário utilizar 11 indicadores. Além da objetividade da relação a ser analisada, muitos dos dados desses 4573 indicadores não estão preenchidos, já que dependem da publicação dos dados pelos países. Assim, esse grupo de 11 indicadores possui uma média alta de corretude e relação contextual com o tema desse trabalho.

- **Quais os objetivos a serem alcançados por essa análise?**

  O resultado dos processos de extração, transformação, classificação e análise, sobre os dados dos indicadores, objetiva criar uma relação entre os indicadores educacionais e econômicos e o aumento, ou não, do número de infectados com o vírus HIV. Com isso, será possível entender a influência

entre esses indicadores, mas também analisar o impacto que eles podem causar entre si.

Dessa forma, os possíveis relacionamentos entre os indicadores socioeconômicos serão expressos na hipótese que centra o tema desse trabalho: “Impacto de indicadores socioeconômicos na quantidade de pessoas com AIDS”. Isso será comprovado com a validação do resultado de previsões realizadas com os dados dos países Brasil e Estados Unidos da América (EUA).

- **Quais os aspectos geográficos relacionados à análise realizada?**

  Os conjuntos de dados utilizados são geograficamente plurais, já que são compostos por fontes de diferentes países de todos os continentes. Esses conjuntos foram criados por organizações internacionais que objetivam praticar ações globais através da análise de dados de várias nações.

- **Qual o período está sendo analisado?**

  O conjunto de dados da UNAIDS inicia em 2010 e vai até 2019. Já o conjunto do composto (proveniente de dados da UNESCO, World Bank e etc), inicia no ano de 1970 e vai até 2016. Porém, após realizar uma análise sobre a relevância dos indicadores para responder a hipótese proposta e o percentual de preenchimento desses dados, foi definido o período entre 2010 a 2016. O notebook “***1_extracao_faixa_temporal_melhor_distribuicao_valores_aids***” realiza esse processo.

2. Coleta de Dados

Os dados utilizados nesse trabalho foram obtidos do catálogo do World Bank e da UNAIDS. O endereço para adquirí-los, na fonte original, está presente no final deste documento. A fim de facilitar a obtenção desses dados, eles também estão armazenados no repositório desse trabalho (também presente na seção de “7. Links”).

O *dataset* proveniente da UNAIDS é composto por 32 colunas:

- **Nome do país** (“*Country*”);
- As outras 31 colunas que representam um **intervalo de anos entre 2010 a 2019**. Porém, aqui, cada ano pode ter três margens de valores. Por exemplo: o ano de 2010 é representado por três colunas: “*2010*”, “*2010_lower*” e “*2010_upper*”.

Esse *dataset* também possui 172 linhas (uma linha para cada país, continente ou conjunto de países). O nome do arquivo que o representa é “***Epidemic transition metrics_Trend of new HIV infections.csv***” (armazenado no repositório mencionado na seção “**7. Links**”). O tipo dos dados presentes nas colunas de datas pode ser inteiro ou texto e varia de acordo com o conteúdo. Para alcançar o objetivo de relacionar a quantidade de pessoas com AIDS e indicadores socioeconômicos (que poderiam influenciar positivamente na diminuição dessa taxa de infecção), será utilizado o período de anos de 2010 a 2016. Para chegar nesse período mencionado, foram aplicados processos de extração e análise (os passos desses processos estão presentes no *notebook* “***1_extracao_faixa_temporal_melhor_distribuicao_valores_aids***”) e um período foi definido de acordo com a média de dados preenchidos para os indicadores relevantes.

Descrição das colunas:

	Nome da coluna	Descrição	Tipo
1	Country	Nome do país.	Texto
2	2010	Valor referente ao indicador para o ano de 2010.	Texto ou inteiro
3	2011	Valor referente ao indicador para o ano de 2011.	Texto ou inteiro

4	2012	Valor referente ao indicador para o ano de 2012.	Texto ou inteiro
5	2013	Valor referente ao indicador para o ano de 2013.	Texto ou inteiro
6	2014	Valor referente ao indicador para o ano de 2014.	Texto ou inteiro
7	2015	Valor referente ao indicador para o ano de 2015.	Texto ou inteiro
8	2016	Valor referente ao indicador para o ano de 2016.	Texto ou inteiro

Tabela 2.1 – Colunas do dataset "Epidemic transition metrics_Trend of new HIV infections.csv"

Exemplo: o país Afeganistão, para o ano de 2010, apresentou a quantidade de pessoas com AIDS igual a 3288 ("<1000" fica igual a "999"; "<500", igual a 499; e esses dois somados, igual a 1800, chegando-se ao resultado de 3288). Seguem os dados presentes nas colunas dessa linha que permitiram chegar a essa conclusão:

Afghanistan,<1000,<500,1800

- Nome do país: ***Afghanistan***
- Valores para o ano de 2010: ***<1000 <500 1800***

O dataset proveniente do World Bank é composto por 69 colunas:

- **Nome do país** ("*Country Name*");
- **Código do país** ("*Country Code*");
- **Código do Indicador** ("*Indicator Code*"); e
- 65 colunas que representam um **intervalo de anos entre 1970 a 2100**.

Esse conjunto de dados está presente no arquivo "***EdStatsData.csv***" (armazenado no repositório, mencionado na seção "**7. Links**", dentro da pasta "data" em "*Edstats_csv.zip*"). Além disso, esse *dataset* também possui 836.930 linhas. Esse número existe porque cada país tem, em média, 4.573 indicadores (nem todos os países atribuíram valores a todos os indicadores). Logo, cada país pode atribuir valores aos 4.573 indicadores distribuídos nos anos de 1970 a 2016 (as colunas de 2017 a 2100 estão em branco).

A fim de solucionar a hipótese “Impacto de indicadores socioeconômicos na quantidade de pessoas com AIDS”, definida no capítulo anterior deste documento, será necessário lidar apenas com 11 indicadores.

Para chegar nesse número de 11 indicadores, foram previamente realizados processos de extração e análise através da execução dos passos presentes no notebook “***A1_extracao_massa_indicadores_aids***”. Complementar a isso, foram escolhidos os indicadores mais contextualmente relevantes com o tema deste trabalho.

Descrição das colunas:

	Nome da coluna	Descrição	Tipo
1	Country Name	Nome do país. Não apresenta valores nulos.	texto
2	Indicator Name	Nome do indicador. Não apresenta valores nulos.	texto
3	Indicator Code	Código do indicador. Não apresenta valores nulos.	texto
4	2010	Valor referente ao indicador para o ano de 2010.	float64
5	2011	Valor referente ao indicador para o ano de 2011.	float64
6	2012	Valor referente ao indicador para o ano de 2012.	float64
7	2013	Valor referente ao indicador para o ano de 2013.	float64
8	2014	Valor referente ao indicador para o ano de 2014.	float64
9	2015	Valor referente ao indicador para o ano de 2015.	float64
10	2016	Valor referente ao indicador para o ano de 2016.	float64

*Tabela 2.2 – Descrição das colunas do dataset “**massa_bruta_pais_por_indicadores_aids.csv**”*

Descrição dos indicadores presentes nas linhas e agrupados por país:

	Código do indicador	Descrição	Fonte
1	NY.GDP.MKTP.CD	PIB do país. “O PIB é a soma de todos os bens e serviços finais produzidos por um país, estado ou cidade, geralmente em um ano.” (IBGE)	World Bank e OECD
2	IT.NET.USER.P2	Quantidade de usuários de internet por 100 pessoas.	International Telecommun

		Usuários da Internet são pessoas que usaram a Internet (de qualquer local) nos últimos 3 meses. A Internet pode ser usada através de um computador, telefone celular, assistente digital pessoal, máquina de jogos, TV digital etc.	ication Union. World Telecommun ication/ICT.
3	SP.POP.GROW	Percentual do crescimento populacional.	Eurostat U.S. Census Bureau UNESCO Institute for Statistics
4	SL.UEM.TOTL. ZS	Percentual da força de trabalho desempregada.	International Labour Organization , ILOSTAT database.
5	SE.XPD.TOTL. GD.ZS	Despesa total do governo em educação (atual, capital e transferências), expressa como porcentagem do PIB.	UNESCO Institute for Statistics
6	UIS.LP.AG15T2 4	Número total da população entre 15 e 24 anos que não sabe ler e escrever.	UNESCO Institute for Statistics
7	UIS.ILLPOP.AG 25T64	Número total da população entre 25 e 64 anos que não sabe ler e escrever.	UNESCO Institute for Statistics
8	SE.PRM.TCHR	Quantidade de professores no ensino fundamental.	UNESCO Institute for Statistics
9	UIS.ROFST.1	Taxa de crianças fora da escola em idade escolar primária.	UNESCO Institute for Statistics
10	SE.PRM.ENRL. TC.ZS	Quantidade de professores por aluno no ensino fundamental.	UNESCO Institute for Statistics
11	SE.PRM.TENR	Número total de alunos da faixa etária oficial da escola primária que estão matriculados no ensino fundamental ou médio, expressos como porcentagem da população correspondente.	UNESCO Institute for Statistics

*Tabela 2.3 – Descrição dos indicadores presentes na coluna "Indicator Code" do dataset "**massa_bruta_pais_por_indicadores_aids.csv**"*

Exemplo: o indicador com código "*SE.XPD.TOTL.GD.ZS*" medido pelo país "Brasil", durante os anos de 2010 a 2016, está presente, como uma linha, no arquivo "*massa_bruta_pais_por_indicadores_aids.csv*". Segue a linha, os campos e sua respectiva descrição.

- Nome do país: ***Brazil***
- Código do indicador: ***SE.XPD.TOTL.GD.ZS***
- Valores para a faixa temporal de 2010 a 2016: ***5.64,5.75,5.8,5.99,0.0,0.0,0.0***

Os indicadores relacionados ao desenvolvimento econômico e os investimentos na educação seguem a ideia da questão: "um país em crescimento econômico e que investe cada vez mais na educação consegue prover um ambiente que diminui a chance de pessoas se contaminarem com AIDS?".

Já os indicadores de cunho social objetivam questionar: "A quantidade de pessoas que possuem acesso à informação pela internet juntamente às capacitadas a preencher um posto de trabalho são mais conscientes e possuem menor chance de se contaminarem com HIV?".

Por último, os indicadores educacionais foram selecionados para embasar o questionamento: "Quanto mais conhecimento, menor a chance de se expor ao HIV?".

Assim, esses indicadores mencionados podem ser classificados como possíveis "influenciadores" da taxa de contaminação com AIDS.

O dataset da UNAIDS foi obtido em 21/07/2020 pelo endereço "https://lawsandpolicies.unaids.org/". Já o do World Bank, foi obtido em 30/06/2020 pelo endereço "*http://datacatalog.worldbank.org/dataset/education-statistics*".

3. Processamento/Tratamento de Dados

Para realizar as atividades de processamento e tratamento dos dados, foi utilizada a linguagem Python através de *Jupyter Notebooks*, a fim de alcançar simplicidade e facilidade de compreensão e reprodução dos passos.

Para transformar os *datasets* "***EdStatsData.csv***" e "***Epidemic transition metrics_Trend of new HIV infections.csv***" em um conjunto de dados refinado e apto a ser utilizado na etapa de classificação (realizada através dos passos do notebook "***5_classificacao_aids_por_indicadores_socioeconomicos***"), foi necessário executar vários processos que envolveram limpeza, seleção e transformação dos dados presentes nos notebooks "***A1_extracao_massa_indicadores_aids***", "***1_extracao_faixa_temporal_melhor_distribuicao_valores_aids***", "***2_extracao_transformacao_massa_indicadores_todos_paises_aids***", "***3_extracao_transformacao_massa_aids***" e "***4_transformacao_unificacao_datasets***". Então, em seguida, os *datasets* principais (mencionados no início desse parágrafo) serão analisados e, posteriormente, esse processo de transformação será explicado com foco em cada *dataset* utilizado (principais e auxiliares).

1) *EdStatsData.csv*

- Quantidade de registros: 61.198.101 (69 colunas e 886.929 linhas).
- Quantidade de registros ausentes: 53.455.179
- Quantidade de registros duplicados: 0
- Filtragem de indicadores pouco preenchidos: esse *dataset* possui, aproximadamente, 87% do seu conteúdo em branco e isso serviu para embasar a ideia de selecionar indicadores com alta taxa de preenchimento. Essa e a filtragem de dados não preenchidos estão descritos no passo a passo do notebook "***2_extracao_transformacao_massa_indicadores_todos_paises_aids***".
- Tratamento dos dados: como mencionado, os dados ausentes foram usados para auxiliar na escolha dos possíveis indicadores socioeconômicos (selecionando os indicadores com menos dados

ausentes) que farão relação com o indicador de pessoas com AIDS (presente em outro *dataset*). Por fim, os dados dos indicadores e das colunas selecionados (2010 a 2016) foram substituídos por zero para, em seguida, serem trocados pelo valor anterior ou posterior válido. Ou seja, não haverão valores zerados nos indicadores socioeconômicos (os valores da coluna do indicador “AIDS” não serão alterados por essa lógica).

Os valores dos registros presentes na coluna “PIB” foram transformados para a unidade “bilhões” para melhorar a análise e classificação.

■ Seleção dos dados: os dados selecionados desse *dataset* são os referentes aos países listados em “***lista_paises_aids.csv***”, aos indicadores definidos em “***codigos_indicadores_relevantes_aids.csv***” e à faixa temporal de 2010 a 2016 especificada em “***lista_anos_aids.csv***”.

■ Resultado: a aplicação dos passos anteriores resultaram em um conjunto de dados mais refinado que foi armazenado em “***massa_bruta_pais_por_indicadores_aids.csv***”. Esse conjunto apresenta 21.483 registros (2.387 linhas e 9 colunas) e nenhum registro não preenchido.

2) *Epidemic transition metrics_Trend of new HIV infections.csv*

■ Quantidade de registros: 5.472 (32 colunas e 171 linhas).

■ Quantidade de registros ausentes: 1.626; nesse caso, é considerado um registro ausente aquele cujo valor seja igual a “…”.

■ Quantidade de registros duplicados: 0

■ Seleção dos dados: os dados selecionados desse *dataset* são os referentes aos países listados em “***lista_paises_aids.csv***”, aos indicadores definidos em “***codigos_indicadores_relevantes_aids.csv***” e à faixa temporal de 2010 a 2016 (especificada em “***lista_anos_aids.csv***”). Além disso, no *notebook* “***4_transformacao_unificacao_datasets***”, antes de exportar o conjunto de dados unificado, é realizada uma filtragem dos países presentes em “***massa_bruta_aids.csv***”, cujos valores são todos

zerados. Logo, uma lista de países (que possuam todos os valores preenchidos no *dataset* "***massa_bruta_aids.csv***") é usada para selecionar os registros dos indicadores socioeconômicos e das pessoas infectadas com AIDS.

- Tratamento dos dados: os dados presentes no *dataset* "***Epidemic transition metrics_Trend of new HIV infections.csv***" são numéricos e textuais. Os dados numéricos não sofrerão alteração, já os textuais sim. Esses últimos podem vir no formato, por exemplo, "…" ou "<100". Assim, os registros com valores iguais a "…" serão substituídos por zero (porque essa coluna será transformada em zero ou um), mas registros com valores iguais a "<100" serão transformados em "99", por exemplo (como há o sinal de "menor que" o valor "100" é subtraído um, se fosse "maior que" seria somado um).

- Resultado: a aplicação dos passos anteriores resultaram em um conjunto de dados mais refinado e que foi armazenado em "***massa_bruta_aids.csv***". Esse conjunto apresenta 1.192 registros (149 linhas e 8 colunas) e nenhum registro não preenchido.

Será apresentada, agora, uma descrição da sequência de passos, presentes nos *notebooks* armazenados no repositório descrito na seção "**7. Links**" desse documento, com foco nos *datasets* principais e nos que foram criados para auxiliar todos esses processos de extração e transformação.

Para orientar a execução das atividades de transformação foi necessário criar outros conjuntos de dados. Esses conjuntos possuem um volume menor, pois o seu objetivo é agregar dados comuns que guiem a unificação dos *datasets* principais: "***EdStatsData.csv***" e "***Epidemic transition metrics_Trend of new HIV infections.csv*"**. São esses os conjuntos auxiliares: "***brasil_relevancia_indicadores.csv***", "***codigos_indicadores_relevantes_aids.csv***", "***lista_anos_aids.csv***", "***lista_paises_aids.csv***", "***lista_suja_paises.csv***", "***massa_bruta_pais_por_indicadores_aids.csv***", "***massa_bruta_aids.csv***",

“***final_transformacao_dados_unificados_aids_indicadores.csv***” e “***final_transformacao_dados_unificados_aids_indicadores_br_eua.csv***”.

O conjunto de dados “***brasil_relevancia_indicadores.csv***” é criado no *notebook* “***A1_extracao_massa_indicadores_aids***”. Isso ocorre pela execução dos passos:

1) Carga completa do *dataset* “***EdStatsData.csv***”;

2) Os dados são filtrados para o país Brasil (“Brazil”) e armazenados no DataFrame (estrutura da biblioteca *Pandas*) “*df_brasil*”.

3) Esse DataFrame possui a lista de todos os indicadores com dados do Brasil (cada país tem dados para a mesma quantidade de indicadores). Assim, é calculada a quantidade de dados não preenchidos e, dessa forma, são selecionados os indicadores com maior taxa preenchimento. Esses indicadores foram armazenados no arquivo “***brasil_relevancia_indicadores.csv***”.

Os dados armazenados em “***brasil_relevancia_indicadores.csv***” serviram de base para uma posterior coleta e análise manual dos indicadores, a fim de encontrar os que possuam maior relação contextual e relevância com o tema desse trabalho. O resultado dessa análise foi a lista dos 11 indicadores, que acabou sendo armazenada em “***codigos_indicadores_relevantes_aids.csv***”, criada manualmente já que seu conteúdo dependeu da escolha contextual realizada anteriormente. Esse arquivo possui apenas uma coluna com os códigos de cada indicador.

O arquivo “***lista_anos_aids.csv***” possui a lista dos anos (2010 a 2016) a serem utilizados para auxiliar na extração e união dos dados oriundos dos dois *datasets* principais (presentes em “***EdStatsData.csv***” e “***Epidemic transition metrics_Trend of new HIV infections.csv***”). A obtenção dessa lista de anos ocorre através da execução dos passos presentes e descritos no notebook “***1_extracao_faixa_temporal_melhor_distribuicao_valores_aids***”, segue a ideia por trás disso:

1) Carga completa do dataset “***EdStatsData.csv***”.

2) Carga dos indicadores socioeconômicos listados em “***codigos_indicadores_relevantes_aids.csv***”.

3) Filtragem dos dados carregados no passo “1” de acordo com os dados presentes na lista de indicadores carregada no passo “2”. Isso é armazenado no DataFrame (estrutura da biblioteca *Pandas*) “*df_faixa_temporal_indicadores_relevantes*”.

4) Seleciona as colunas do DataFrame “*df_faixa_temporal_indicadores_relevantes*” cujos valores possuam boa taxa de preenchimento dos dados diferentes de zero (um passo anterior substituiu os campos vazios por zero). Isso resulta nas colunas presentes no intervalo dos anos de 1970 a 2016.

5) Carrega o *dataset* “***Epidemic transition metrics_Trend of new HIV infections.csv***” e constata-se que este inicia sua distribuição dos dados a partir do ano de 2010. Logo, será esse o primeiro ano da faixa temporal.

6) Para escolher o ano final dessa faixa temporal, foi necessário comparar a qualidade dos dados das colunas de 2015 e 2016 do *dataset* "***EdStatsData.csv***". Após isso, foi possível concluir que a coluna de 2016 possui uma média de dados preenchidos maior que a de 2015. Dessa forma, será o ano de 2016 que fechará a faixa temporal a ser exportada.

O conjunto de dados presente no arquivo “***lista_paises_aids.csv***” é encontrado a partir da remoção de repetições presentes em “***lista_suja_paises.csv***”. Esse último *dataset* foi criado manualmente e seus dados são os nomes de países (mas com algumas repetições). A sua origem foi a coluna “*Country Name*” do arquivo “***EdStatsData.csv***”, sendo retirado o nome de continentes. Isso está descrito no passo “*3. Organizando nome de países definidos em "lista_suja_paises*” do notebook “***2_extracao_transformacao_massa_indicadores_todos_paises_aids***”.

O arquivo “*massa_bruta_pais_por_indicadores_aids.csv*” é gerado também no notebook “***2_extracao_transformacao_massa_indicadores_todos_paises_aids***”

e isso ocorre no passo “*4. Extraindo e transformando dados referentes aos países definidos para "**massa_bruta_pais_por_indicadores.csv**"*”. Sendo a sequência disso:

1) Os dados provenientes de “***EdStatsData.csv***” foram previamente carregados em um DataFrame (estrutura da biblioteca *Pandas*) “*df_todos_paises*”.

2) A última coluna desse DataFrame não faz referência a um ano válido e nem possui dados, logo é retirada.

3) São selecionadas as linhas cujos dados da coluna de países (“*Country Name*”) estejam presentes em “***lista_paises_aids.csv***”.

4) Agora a filtragem das linhas ocorre pela coluna de indicadores (“*Indicator Code*”). Assim, as linhas cujos dados estejam presentes em “***codigos_indicadores_relevantes_aids.csv***” são mantidas.

5) Por fim, seleciona as colunas que estão presentes na faixa temporal definida em “***lista_anos_aids.csv***”, preenche os campos vazios de todas linhas com zero, remove as colunas “*Country Code*” e “*Indicator Name*” e finaliza com a exportação disso em “***massa_bruta_pais_por_indicadores_aids.csv***”.

Os dados presentes em “***massa_bruta_aids.csv***” foram resultado da sequência de passos presentes no notebook “***3_extracao_transformacao_massa_aids***”. Segue uma descrição desses passos:

1) São carregados os dados do arquivo “***Epidemic transition metrics_Trend of new HIV infections.csv***”.

2) São selecionadas os registros desse *dataset* cujos países estão presentes em ”***lista_paises_aids.csv***”.

3) É carregada a faixa temporal definida em “***lista_anos_aids.csv***” para guiar a seleção das colunas que estão dentro dessa faixa.

4) São realizadas transformações dos dados, pois alguns deles estão em um formato de aproximação e outros preenchidos com "...". Por exemplo, o primeiro registro da coluna "2010" possui o valor "<1000", logo esse será convertido para "999". Já registro da oitava linha da coluna "2010", será convertido para "0" possui o seu conteúdo é igual a "…".

5) Finalmente, as colunas que referenciam o mesmo ano serão consolidadas em uma só e seus dados, somados. Por exemplo, as colunas "2010", "2010_lower" e "2010_upper" resultarão em uma nova coluna com nome "2010" e com valores do resultado da soma das três. Isso é feito também para os anos de 2011, 2012, 2013, 2014, 2015 e 2016. O resultado disso é exportado para "***massa_bruta_aids.csv***".

Por fim, os arquivos "***final_transformacao_dados_unificados_aids_indicadores.csv***" e "***final_transformacao_dados_unificados_aids_indicadores_br_eua.csv***" armazenam o resultado dos processos de seleção, extração e transformação de dados aplicados anteriormente envolvendo, direta ou indiretamente, os outros *datasets*. O primeiro arquivo possui o *dataset* utilizado para uma classificação geral dos modelos, já o segundo é utilizado para validar o resultado de previsões realizadas com os dados específicos do Brasil e dos Estados Unidos da América (EUA).

O *notebook* que gera esses arquivos é o "***4_transformacao_unificacao_datasets***".

1) Os arquivos "***massa_bruta_pais_por_indicadores_aids.csv***" e "***massa_bruta_aids.csv***" são, respectivamente, carregados nos DataFrames (estrutura da biblioteca *Pandas*) "*massa_bruta_pais_por_indicadores*" e "*massa_bruta_aids*".

2) São definidas listas com os dados referenciais úteis para a extração dos registros das colunas.

3) Os registros do DataFrame "*massa_bruta_pais_por_indicadores*" são filtrados pelas listas de indicadores presentes em uma das listas com dados referenciais (esse conjunto de indicadores foi escolhido de acordo com a sua

relevância para o tema desse trabalho) e pelos países presentes no DataFrame “*massa_bruta_aids*”.

4) Os dados resultantes do passo anterior são unidos através da coluna dos países, pois os dois possuem registros com nomes de países e esse será o ponto de interseção.

5) Os dados do indicador “AIDS” serão, agora, transformados em “0” ou “1”. O valor “0” representará a diminuição ou estagnação do número de pessoas infectadas pelo vírus HIV; já “1”, o aumento desse número. Logo, para chegar a isso, as sete colunas de datas (2010 a 2016) resultarão em seis, uma para a diferença entre 2011 e 2010, outra para a diferença entre 2012 e 2011 e etc. Por exemplo, o registro da coluna “2011” para o indicador “AIDS” da linha 12 (país “*Afghanistan*”) é, aproximadamente, “3,5” e o da coluna “2010” é “3,3”, assim o valor resultante será “1”, pois houve um aumento do número de infectados de 2010 para 2011, e isso será armazenado em uma coluna “2010” de uma nova estrutura.

5) Depois, os dados referentes ao países Brasil e EUA serão transferidos desse DataFrame (“*df_resultado_transformacao*”) para um segundo (“*df_paises_testes*”) que ficará apenas com os dados dos indicadores desses países, mas ambos na mesma faixa temporal.

6) Após isso, os dados zerados de todas as colunas, dos dois DataFrames, exceto a do indicador “AIDS”, serão preenchidos pelo valor anterior ou posterior para manter uma variação estável e não afetada por ausência de preenchimento do *dataset* original.

7) Em seguida, os dados, dos dois DataFrames, dos registros de cada país serão armazenados em uma estrutura de pilha que representará a coluna da estrutura final a ser exportada. Por exemplo, os dados da segunda linha dos anos de 2010 a 2015, do indicador “AIDS” e do país “Afghanistan” são, aproximadamente, “0”, “2.95”, “0”, “0”, “0” e “0” que representarão os primeiros registros da coluna “AIDS” (nome do indicador) da nova estrutura a ser criada como resultado disso.

8) Antes da exportação, os dados da coluna “PIB”, dos dois *DataFrames*, são convertidos para a quantidade de bilhões.

9) Por fim, essas novas estruturas formadas pelos dados dos indicadores distribuídos em colunas serão exportadas, respectivamente, para

"***final_transformacao_dados_unificados_aids_indicadores.csv***" e "***final_transformacao_dados_unificados_aids_indicadores_br_eua.csv***". Antes disso, as colunas dos dois *DataFrames* foram renomeadas para auxiliar na compreensão dos resultados da classificação, pois antes estavam como códigos oriundos do *dataset* "***EdStatsData.csv***". Logo, segue uma tabela com o código do indicador e seu respectivo nome:

Código Indicador	Novo nome
NY.GDP.MKTP.CD	PIB
IT.NET.USER.P2	Usuários Internet
SP.POP.GROW	Crescimento Populacional
SL.UEM.TOTL.ZS	Desemprego
SE.XPD.TOTL.GD.ZS	Investimentos Educação
UIS.LP.AG15T24	Analfabetismo 15 e 24
UIS.ILLPOP.AG25T64	Analfabetismo 25 e 64
SE.PRM.TCHR	Qtd Professores EF
UIS.ROFST.1	Crianças Fora Escola
SE.PRM.ENRL.TC.ZS	Qtd Alunos por professor
SE.PRM.TENR	Qtd Alunos EF

Tabela 3.1 – Dicionário dos códigos dos indicadores e seus respectivos nomes.

Dessa forma, é possível concluir que as diferenças entre os *DataFrames* "*df_indicadores_em_colunas_testes*" e "*df_indicadores_em_colunas*" são:

- O DataFrame "*df_indicadores_em_colunas_testes*" possui apenas dados dos países Brasil e EUA, logo possui apenas 12 linhas. Além disso, possui uma coluna com o nome "Pais Periodo" com o conteúdo igual ao nome do país e o período da faixa temporal que a linha pertence, por exemplo, o valor dessa coluna na primeira linha do DataFrame é igual a "*Brasil 2010-2011*".

- Já o DataFrame “*df_indicadores_em_colunas*” possui registros de todos os países, menos os do Brasil e do Estados Unidos da América. E como esses dados serão usados para treinar os modelos, não será necessário criar uma coluna similar à mencionada acima.

4. Análise e Exploração dos Dados

O *notebook* “***A2_analise_dataset_unificado***” possui as informações técnicas relacionadas às descrições a serem explicitadas. Neste capítulo, será realizada uma exploração dos dados resultantes dos processos executados e explicados nos capítulos anteriores. Os dados mencionados serão os presentes nos *datasets* representados pelos arquivos “***final_transformacao_dados_unificados_aids_indicadores.csv***”, “***final_transformacao_dados_unificados_aids_indicadores_br_eua.csv***” e “***massa_bruta_pais_por_indicadores_aids.csv***”.

O conjunto de dados a serem analisados em “***final_transformacao_dados_unificados_aids_indicadores.csv***” é o presente na coluna “AIDS”. Essa coluna é categórica e seus dados estão com valores "0" ou "1". De acordo com a descrição dos valores dessa coluna, a quantidade de registros na categoria "0" prevalece sobre a categoria "1", especificamente 81.19% dos registros são da categoria "0". Isso permite concluir que os países apresentaram, entre 2010 e 2016, uma diminuição ou estagnação da taxa de pessoas infectadas com AIDS.

```
In [34]: df_analise['AIDS'].value_counts()

Out[34]: 0.0    492
         1.0    114
         Name: AIDS, dtype: int64
```

Figura 4.1 – Dados coluna “AIDS”.

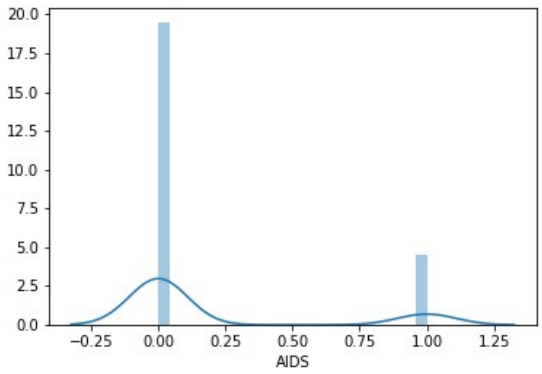

Figura 4.2 – Gráfico dados coluna “AIDS”

O “***final_transformacao_dados_unificados_aids_indicadores_br_eua.csv***” representa um *dataset* a ser utilizado para validar o resultado das classificações a serem realizadas no capítulo 5. Esse conjunto de dados possui 12 linhas: seis para o Brasil e as outras seis para os dados dos Estados Unidos da América. Os dados dos dois países representam os valores dos indicadores para os períodos de 2010 a 2016. O Brasil apresenta quatro situações de aumento da quantidade de pessoas com AIDS (de 2010 a 2011, 2012 a 2013, 2013 a 2014 e 2014 a 2015) e duas sem aumento. Já para os Estados Unidos da América, há apenas um período em que houve aumento da quantidade de pessoas com AIDS: de 2013 a 2014.

Os próximos dados a serem analisados serão os provenientes do *dataset* “***massa_bruta_pais_por_indicadores_aids.csv***”, pois o resultado final da transformação (“***final_transformacao_dados_unificados_aids_indicadores.csv***”) agrupou e mesclou os dados dos países e dos indicadores em colunas e isso impede detalhar a análise.

Analisando a coluna do PIB dos países ("*NY.GDP.MKTP.CD*") do *dataset* mencionado, é possível entender que, entre 2010 e 2016, 6% dos países não tiveram o seu PIB registrado e 14.71% ficaram acima da média. Segue a lista dos países donos dos 10 maiores PIBs (média de valores entre 2010 a 2016): Estados Unidos da América, China, Japão, Alemanha, Reino Unido, França, Brasil, Itália, Índia e Rússia. Os dados dessa coluna mostram-se como um conjunto desbalanceado de valores e é possível visualizar que três países possuem PIBs muito fora do normal, ou seja, fica evidente que a diferença de poder econômico entre os países é enorme.

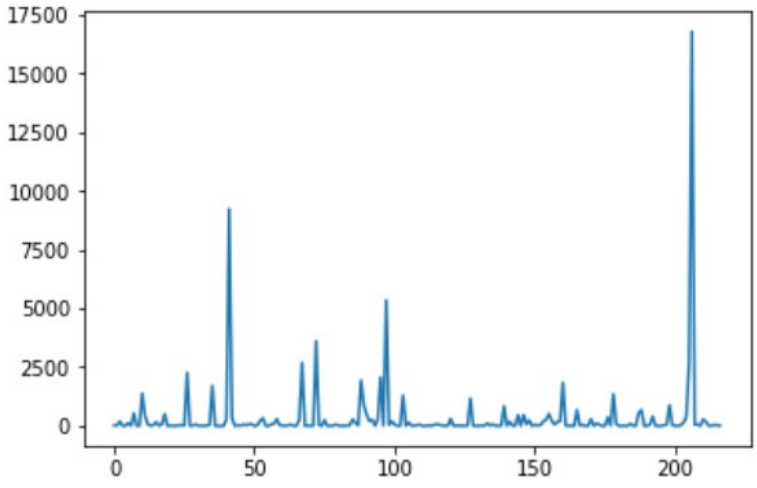

Figura 4.3 – Gráfico coluna "PIB"

Os dados referentes à quantidade de usuários de internet (distribuídos na coluna "*IT.NET.USER.P2*") originam-se de uma média balanceada da quantidade de usuários de internet a cada 100 pessoas. Esse indicador dá uma ideia geral da popularização do acesso ao serviço de telecomunicação e, consequentemente, do acesso à informação. De acordo com a distribuição desses dados, há alta variabilidade dos valores, nenhum valor discrepante, simetria exata dos dados e apenas 6,45% dos dados não estão preenchidos.

```
Out[176]: count    217.00
          mean      40.22
          std       29.13
          min        0.00
          25%       13.37
          50%       39.40
          75%       64.77
          max       96.51
          Name: Media, dtype: float64
```

Figura 4.4 – Descrição dos dados referentes à quantidade de usuários de internet

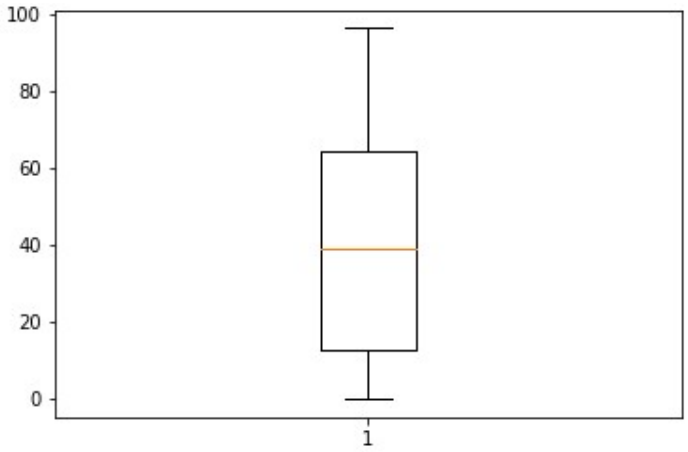

Figura 4.5 – Distribuição dos dados referentes à quantidade de usuários de internet

Os dados referentes ao crescimento populacional ("*SP.POP.GROW*") representam uma média ponderada anual do país (em termos percentuais). Esse indicador considera como "população" todos as pessoas residentes no país, independentemente da situação legal de cidadania. O resultado apresentado na

descrição dos dados e no *boxplot* abaixo mostra uma leve assimetria positiva, poucos valores discrepantes e uma pequena variabilidade. Além disso, há apenas 0.92% de registros sem valores e, inclusive, é possível concluir que 25 países apresentaram valores negativos (diminuição da população) para essa taxa, média do período entre 2010 a 2016, são alguns desses países: Albânia, Japão, Polônia, Portugal, Porto Rico, Ucrânia, Ilhas Virgens e etc.

```
Out[179]: count    217.00
          mean       1.35
          std        1.32
          min       -1.75
          25%        0.44
          50%        1.17
          75%        2.18
          max        6.85
          Name: Media, dtype: float64
```

Figura 4.6 – Descrição dos dados referentes ao crescimento populacional

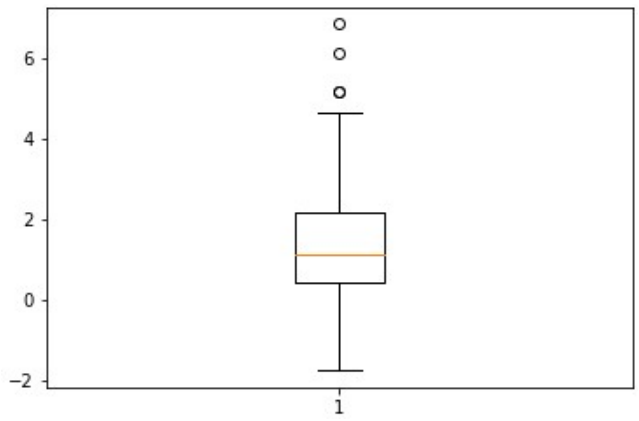

Figura 4.7 – Distribuição dos dados referentes ao crescimento populacional

O próximo indicador a ser analisado é a taxa de desemprego (“*SL.UEM.TOTL.ZS*”). O desemprego aqui é representado por um indicador baseado em um percentual médio balanceado medido anualmente e embasado no total da força de trabalho do país. A descrição do resultado dos dados apresenta a presença de valores zerados (14,29% do total de países), vários registros com dados discrepantes, uma leve assimetria positiva e uma pequena variabilidade. Segue a lista de alguns países com as maiores taxas de desemprego nesse período: Ilhas Salomão, Gâmbia, Macedônia, Moçambique, Espanha e etc.

```
Out[43]: count    217.000000
         mean       7.715405
         std        6.685136
         min        0.000000
         25%        3.100000
         50%        6.185714
         75%       10.700000
         max       31.057143
         Name: Media, dtype: float64
```

Figura 4.8 – Descrição dos dados referentes à taxa de desemprego

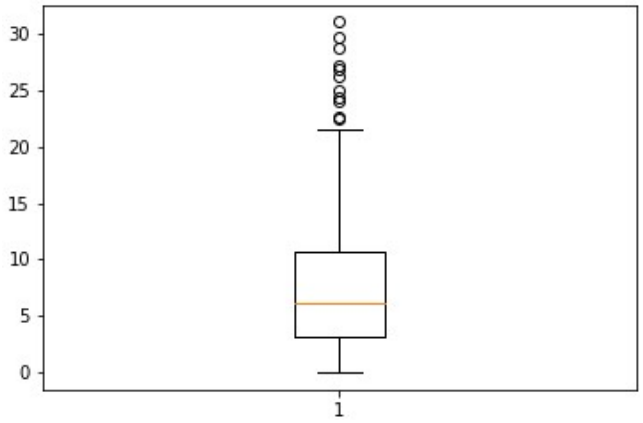

Figura 4.9 – Distribuição dos dados referentes à taxa de desemprego

Ainda no *dataset* "*massa_bruta_pais_por_indicadores_aids.csv*", o indicador "*SE.XPD.TOTL.GD.ZS*" representa o percentual do PIB investido em educação. Os valores desse indicador mostram que 27,65% deles não foram preenchidos. Por outro lado, não existem valores discrepantes. Porém, há uma quantidade razoável de variabilidade dos valores desses registros e a presença de assimetria negativa regular. Não há existência de valores negativos. Lista dos países que mais investem em educação de forma proporcional ao seu PIB: Cuba, Estados Federados da Micronésia, Ilhas Salomão, Namíbia, Dinamarca, Islândia, Suécia e Malta.

```
Out[188]: count    217.00
          mean       3.36
          std        2.62
          min        0.00
          25%        0.00
          50%        3.54
          75%        5.19
          max       12.84
          Name: Media, dtype: float64
```

Figura 4.10 – Descrição dos dados referentes à taxa de investimento na educação

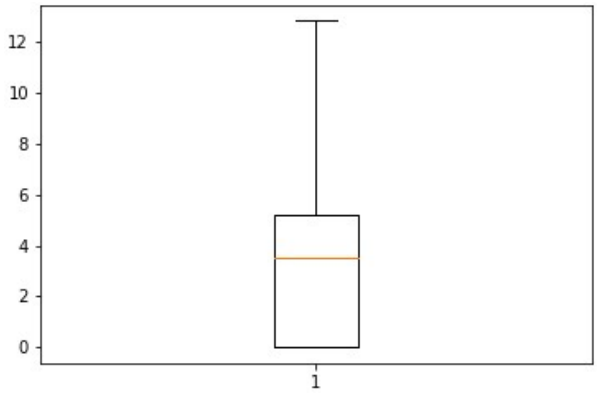

Figura 4.11 – Distribuição dos dados referentes à taxa de investimento na educação

O indicador de analfabetismo mede a quantidade de jovens (com idade entre 15 e 24 anos) que não sabem ler e escrever (não possuem, na sua vida cotidiana, a compreensão de declarações curtas e simples). Esse indicador é referenciado pelo código "*UIS.LP.AG15T24*" no mesmo *dataset* mencionado acima. Como resultado da descrição dos seus dados, esse *dataset* apresenta 45,16% de campos não preenchidos e isso reflete um conjunto de dados que pode induzir falsas conclusões. Há também a existência de três valores discrepantes (referentes aos países Índia, Paquistão e Bangladesh), como mostra o gráfico abaixo.

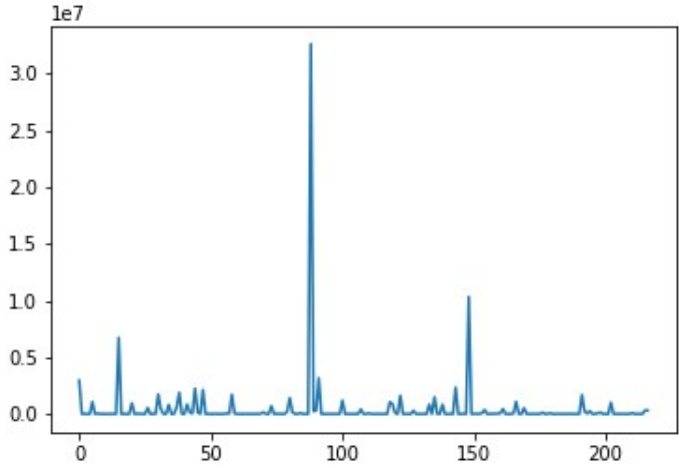

Figura 4.12 – Gráfico de dados da taxa de analfabetismo entre 15 e 24 anos

No mesmo contexto, há outro indicador de analfabetismo, mas para adultos e idosos (25 a 64 anos). Esse indicador é definido pelo código "*UIS.ILLPOP.AG25T64*" e também possui a diretriz de considerar analfabetos as pessoas que não possuem compreensão de declarações curtas e simples. A descrição da análise dos seus dados apresenta 45,62% de valores não preenchidos e cerca de 15,25% países com taxa acima da média. Segue a lista de alguns países com a maior quantidade de adultos e idosos analfabetos: Índia, Paquistão, Bangladesh, China, Egito, Brasil, Iraque, Afeganistão e etc. Percebe-se que Índia, Paquistão e Bangladesh apresentam as mesmas posições nas duas listagens de analfabetismo. Mas também destacam-se China e Brasil, que juntos com a Índia, ocupam, respectivamente, as posições de segundo, sétimo e nono lugar na lista de países mais economicamente poderosos.

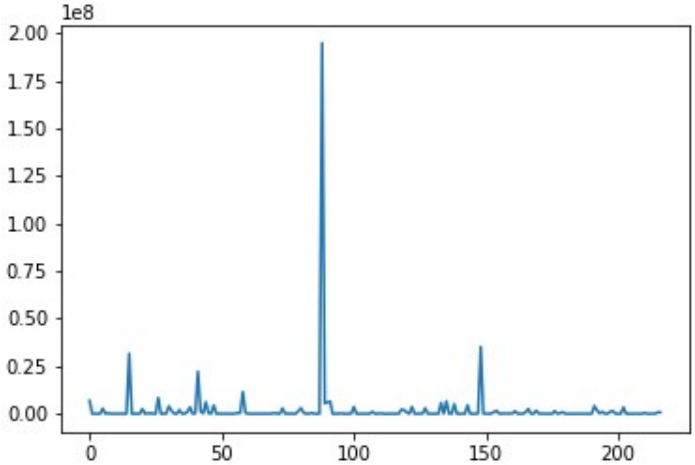

Figura 4.13 – Gráfico de dados da taxa de analfabetismo entre 25 e 64 anos

Nesse conjunto de dados, há outro indicador da área de educação: a quantidade de professores no ensino fundamental cujo código é “*SE.PRM.TCHR*”. Esse indicador objetiva identificar quantos professores, seja de escola pública ou privada, atuam no ensino fundamental (acredita-se que a qualidade do ensino fundamental tem grande influência na vida profissional dos trabalhadores e em outros fatores da vida do cidadão, como por exemplo, conscientização para prevenção de doenças). O resultado da descrição dos dados desse indicador mostra a presença de 14,29% de países com dados não preenchidos e 20,43% países com valores acima da média. Segue a lista dos 10 países com maior número de professores do ensino fundamental: China, Índia, Estados Unidos da América, Indonésia, Brasil, Nigéria, México, Filipinas, Bangladesh e Egito. Interessante notar a presença de países com alto índice de analfabetismo (Índia, China, Brasil, Nigéria, Bangladesh e Egito) aqui nessa lista. Isso pode estar mostrando a reação do governo desses países na tentativa de combater deficiências do setor da educação no seu país. Segue o gráfico com a ilustração dos dados desse indicador, destaque para a alta quantidade de professores dos três países: Índia, China e Brasil.

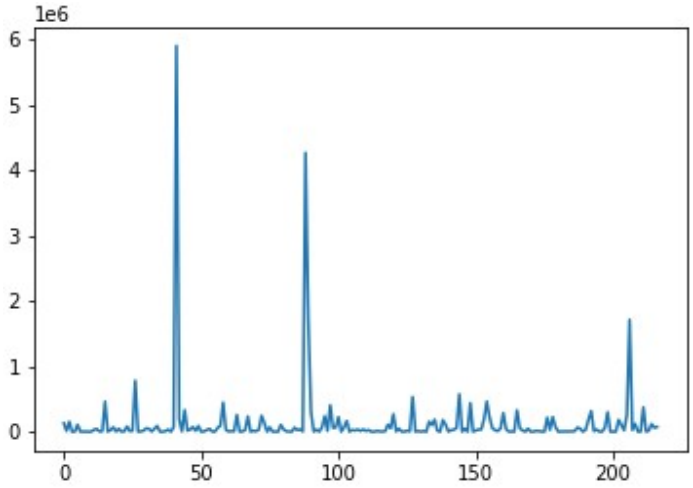

Figura 4.14 – Gráfico da quantidade de professores no ensino fundamental

O indicador “*UIS.ROFST.1*” representa o número de crianças que não estão matriculadas no ensino fundamental. Esse número é expresso como um percentual de crianças que estão aptas a estudar no ensino fundamental, mas, por algum motivo, simplesmente não estão. O conjunto de dados desse indicador mostra que 19,35% dos países não possuem valores preenchidos e 30,29% estão acima da

média. A descrição desse mesmo conjunto apresenta alta quantidade de dados discrepantes, baixa variabilidade e um nível regular de assimetria positiva. A lista dos 5 países com maior quantidade de crianças que não estão estudando no ensino fundamental: Sudão do Sul, Libéria, Eritreia, Guiné Equatorial e Sudão.

```
Out[22]: count    217.00
         mean       7.44
         std       11.20
         min        0.00
         25%        0.64
         50%        3.16
         75%        8.91
         max       64.18
         Name: Media, dtype: float64
```

Figura 4.15 – Descrição dos dados referentes à quantidade de crianças fora da escola do nível fundamental

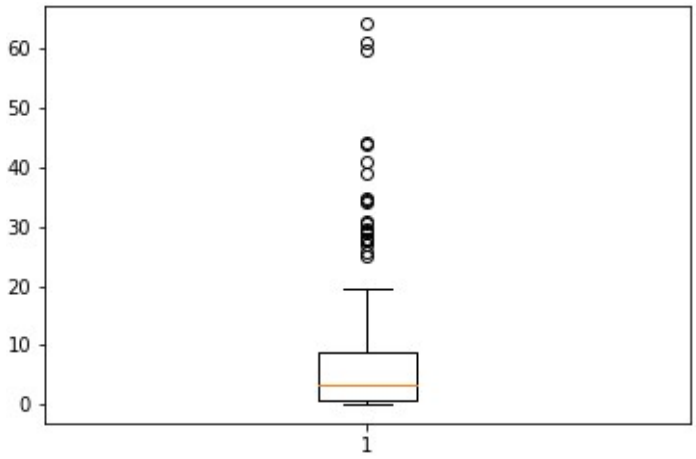

Figura 4.16 – Distribuição da quantidade de crianças fora da escola do nível fundamental

Mais um indicador da área de educação será analisado, o "*SE.PRM.ENRL.TC.ZS*", que representa uma média da quantidade de alunos por professor no ensino fundamental. Mesmo havendo similaridade com outros indicadores, a análise deste é importante, pois trará a distribuição de alunos pela população de estudantes do ensino fundamental. Os dados relacionados a esse

indicador apresentam apenas 14,29% dos dados preenchidos, mas 39,25% de valores acima da média. A descrição desses dados ilustra alguns valores discrepantes, pequena variabilidade e a uma leve assimetria positiva. Segue a lista dos 10 países com a maior média de alunos por professor: República Centro-Africana, Malawi, Chade, Ruanda, Moçambique, Etiópia, Guiné-Bissau, Zâmbia, Sudão do Sul e Uganda. Vale observar que todos os países dessa lista são do continente africano.

```
Out[26]: count    217.00
         mean     20.49
         std      15.10
         min       0.00
         25%      11.28
         50%      17.03
         75%      28.11
         max      81.92
         Name: Media, dtype: float64
```

Figura 4.17 – Descrição dos dados referentes à quantidade de crianças por professor do nível fundamental

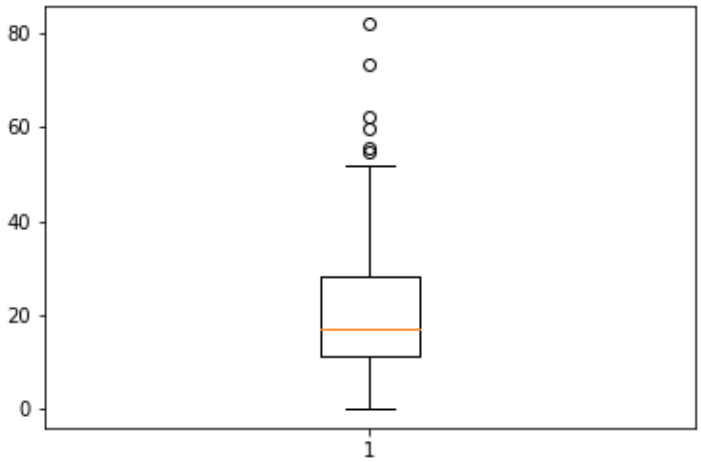

Figura 4.18 – Distribuição da quantidade de crianças por professor do nível fundamental

Por fim, o último indicador a ser analisado do *dataset* "*massa_bruta_pais_por_indicadores_aids.csv*" será um referente à taxa de alunos que ingressaram (estão matriculados) no ensino fundamental. Esse indicador está expresso como porcentagem da população correspondente, o seu código é "*SE.PRM.TENR*". Os dados desse indicador possuem 19,35% de valores não preenchidos e 69,71% de valores acima da média. Isso resulta em uma regular variabilidade dos dados, uma acentuada assimetria positiva e um grupo de registros com valor zerado que aparecem como discrepante. É perceptível que cerca de 70% dos países possuem grande quantidade de crianças matriculadas no ensino fundamental.

```
Out[32]: count    217.00
         mean      73.20
         std       37.47
         min        0.00
         25%       70.45
         50%       93.24
         75%       97.42
         max       99.95
         Name: Media, dtype: float64
```

Figura 4.19 – Descrição dos dados referentes à quantidade de crianças matriculadas no nível fundamental

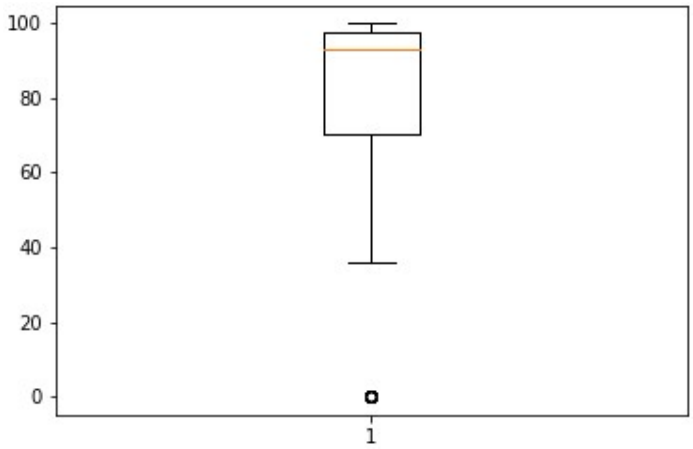

Figura 4.20 – Distribuição da quantidade de crianças matriculadas no nível fundamental

5. Criação de Modelos de Machine Learning

A fim de acompanhamento, abrir o *Jupyter Notebook* "***5_classificacao_aids_por_indicadores_socioeconomicos***", "***5_1_classificacao_aids_por_indicadores_socioeconomicos_brasil***" e "***5_2_classificacao_aids_por_indicadores_socioeconomicos_eua***".

Nessa etapa, serão aplicados algoritmos de classificação (*Dummy*, *SVC*, *Gaussian Naive Bayes* e *Decision Tree*) para tratar a hipótese "Impacto de indicadores socioeconômicos na quantidade de pessoas com AIDS" e validar a relação entre os indicadores explicitados através dos testes realizados com os dados dos países selecionados nas etapas anteriores: Brasil e Estados Unidos da América. Todos os modelos foram configurados para definir a ordem dos números aleatórios a fim de retirar a aleatoriedade da separação dos dados de treino e teste. Além disso, houve também a preocupação em separar proporcionalmente os dados de acordo com os dados da coluna "AIDS".

Os conjuntos de dados a serem utilizados para treinamento serão "***final_transformacao_dados_unificados_aids_indicadores.csv***" e "***final_transformacao_dados_unificados_aids_indicadores_br_eua.csv***", pois eles foram o resultado das etapas que envolveram processos de análise, seleção, tratamento e transformação. Cada bloco de dados é representado por um indicador e cada indicador tem a sua forma de registrar os seus valores. Já a coluna com a taxa de infectados com AIDS está registrando os seus valores de forma binária, ou seja, "0" quando a taxa diminuiu ou estagnou e "1" quando aumentou (ambos os valores são para o intervalo de um ano).

A solução da hipótese envolve a seleção de um grupo de indicadores e a análise do resultado de cada algoritmo de classificação para concluir a efetividade dos modelos aplicados (isso está descrito no *notebook* que lida com cada grupo de indicadores). Além disso, há também os resultados das previsões para validar a efetividade dessa classificação.

A seguir serão exibidos resultados dos modelos de classificação que trabalharam com a união dos dados dos dois *datasets* "***final_transformacao_dados_unificados_aids_indicadores.csv***" e "***final_transformacao_dados_unificados_aids_indicadores_br_eua.csv***". Em

seguida, serão realizadas previsões com os dados do segundo *dataset* em um modelo treinado com os dados do primeiro.

Segue a hipótese e suas respectivas classificações e análises para dados: **“Impacto de indicadores socioeconômicos na quantidade de pessoas com AIDS”.**

- Análise e conclusão:
  - O resultado da classificação pelo modelo ***Dummy***, por ser randômico, será útil como parâmetro de comparação com o resultado dos outros modelos, ou seja, será usado como uma "*baseline*". Os outros modelos utilizados foram: ***SVC***, ***Gaussian Naive Bayes*** (GNB) e ***Decision Tree Classifier*** (DTC).

- Notebook: “***5_classificacao_aids_por_indicadores_socioeconomicos***”.
- Objetivo: medir a relação entre os indicadores socioeconômicos e a alteração na quantidade de pessoas com AIDS.
- Indicadores utilizados na classificação:
  - PIB (coluna ”PIB” do *dataset*);
  - Quantidade de usuários de internet, a fim de ter noção do acesso à informação (coluna “*Usuários Internet*” do *dataset*);
  - Crescimento demográfico (coluna “*Crescimento Populacional*” do *dataset*);
  - Taxa de desemprego (coluna “*Desemprego*” do *dataset*);
  - Percentual do PIB que o governo direciona para a área da educação no seu país (coluna “*Investimentos Educação*” do *dataset*);
  - Quantidade de pessoas analfabetas com idade entre 15 e 24 anos (coluna “*Analfabetismo 15 e 24*” do *dataset*);
  - Quantidade de pessoas analfabetas com idade entre 25 e 64 anos (coluna “*Analfabetismo 25 e 64*” do *dataset*);
  - Quantidade de professores no ensino fundamental (coluna “*Qtd Professores EF*” do *dataset*);
  - Taxa de crianças fora da escola no ensino fundamental (coluna “*Crianças Fora Escola*” do *dataset*);
  - Proporção de alunos por professor no ensino fundamental (coluna “*Qtd Alunos por professor*” do *dataset*);

- Taxa de alunos que ingressaram no ensino fundamental (coluna “*Qtd Alunos EF*” do *dataset*).

- Resultados de cada modelo:

1) *Dummy:* seu resultado servirá como parâmetro de comparação para os outros modelos. As acurácias de treinamento e previsão estão próximas, mas os resultados das métricas de classificação para a classe 1 são baixos.

```
*********  Dummy  *********
Acurácia treinamento: 68.06%
Acurácia previsão: 72.04%

<< Matriz de confusão >>
Predito  0.0  1.0  All
Real
0.0      129   21  150
1.0       31    5   36
All      160   26  186

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.81      0.86      0.83       150
         1.0       0.19      0.14      0.16        36

    accuracy                           0.72       186
   macro avg       0.50      0.50      0.50       186
weighted avg       0.69      0.72      0.70       186
```

Figura 5.1 – Resultado classificação dos indicadores socioeconômicos pelo Dummy.

- Precisão: por causa da presença de mais valores para a classe 0 do que para a classe 1, o modelo apresentou maior precisão para classificar uma situação de diminuição ou estagnação da quantidade de pessoas com AIDS, do que de aumento.
- Revocação: quando o modelo tentou prever aumento na taxa de infecção, teve apenas 14% de sucesso. Porém, 86% quando tentou prever a diminuição ou estagnação dessa mesma taxa.

2) SVC: as acurácias de treinamento e previsão são praticamente iguais. O SVC teve uma acurácia de previsão aproximadamente 10% maior do que a do *Dummy*.

```
*********  SVC  *********
Acurácia treinamento: 82.64%
Acurácia previsão: 82.80%

<< Matriz de confusão >>
Predito  0.0  1.0  All
Real
0.0      150    0  150
1.0       32    4   36
All      182    4  186

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.82      1.00      0.90       150
         1.0       1.00      0.11      0.20        36

    accuracy                           0.83       186
   macro avg       0.91      0.56      0.55       186
weighted avg       0.86      0.83      0.77       186
```

Figura 5.2 – Resultado classificação dos indicadores socioeconômicos pelo SVC.

- Precisão: esse classificador conseguiu classificar todos os casos em que houve aumento de pessoas infectadas com AIDS e 82% dos casos de estagnação ou diminuição.
- Revocação: aqui houve um ótimo resultado no momento de prever os casos de estagnação ou diminuição do número de pessoas com AIDS, porém um resultado baixíssimo (11%) quando foi prever casos de aumento dessa quantidade. A diferença do número de casos de aumento ou não influenciou nessa diferença.

- **Gaussian Naive Bayes (GNB):** os valores da acurácia de treinamento e previsão também estão bem próximas. Esses valores também são aproximadamente 10% maiores que os do Dummy. Os resultados de classificação e previsão estão parecidos com os do SVC.

```
*********  Gaussian Naive Bayes  *********
Acurácia treinamento: 81.25%
Acurácia previsão: 81.18%

<< Matriz de confusão >>
Predito  0.0  1.0  All
Real
0.0      144    6  150
1.0       29    7   36
All      173   13  186

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.83      0.96      0.89       150
         1.0       0.54      0.19      0.29        36

    accuracy                           0.81       186
   macro avg       0.69      0.58      0.59       186
weighted avg       0.78      0.81      0.77       186
```

Figura 5.3 – Resultado classificação dos indicadores socioeconômicos pelo GNB.

- Precisão: o GNB conseguiu uma precisão na classificação de 83% dos casos em que houve diminuição ou estagnação da quantidade de pessoas com AIDS. Por outro lado, apenas 54% para os casos do aumento dessa mesma taxa.
- Revocação: aqui houve um ótimo resultado (96%) no momento de prever os casos de estagnação ou diminuição do número de pessoas com AIDS, mas um resultado baixo (19%) quando foi prever casos de aumento dessa quantidade.

- **Decision Tree Classifier (DTC):** o resultado dessa classificação apresentou 16% de diferença entre as acurácias de treinamento e de previsão. Porém, quando comparadas ao Dummy apresentam resultados melhores. Pelos dados apresentados abaixo, esse algoritmo teve um resultado melhor quando foi prever casos de aumento do número de pessoas infectadas com AIDS.

```
*********  DecisionTreeClassifier  *********
Acurácia treinamento: 99.07%
Acurácia previsão: 83.87%

<< Matriz de confusão >>
Predito  0.0  1.0  All
Real
0.0      136   14  150
1.0       16   20   36
All      152   34  186

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.89      0.91      0.90       150
         1.0       0.59      0.56      0.57        36

    accuracy                           0.84       186
   macro avg       0.74      0.73      0.74       186
weighted avg       0.84      0.84      0.84       186
```

Figura 5.4 – Resultado classificação dos indicadores socioeconômicos pelo DTC.

- Precisão: esse modelo apresentou uma precisão de 89% para classificar uma situação de diminuição ou estagnação da quantidade de pessoas com AIDS e 59% para os casos de aumento dessa quantidade.
- Revocação: quando o modelo tentou prever aumento na quantidade de pessoas com AIDS teve 56% de sucesso e 91% quando tentou prever a diminuição ou estagnação dessa mesma quantidade.

A figura 5.5 apresenta a lógica utilizada pelo DTC para classificar uma situação em que houve influência dos indicadores socioeconômicos na quantidade de pessoas infectadas pelo HIV ou não.

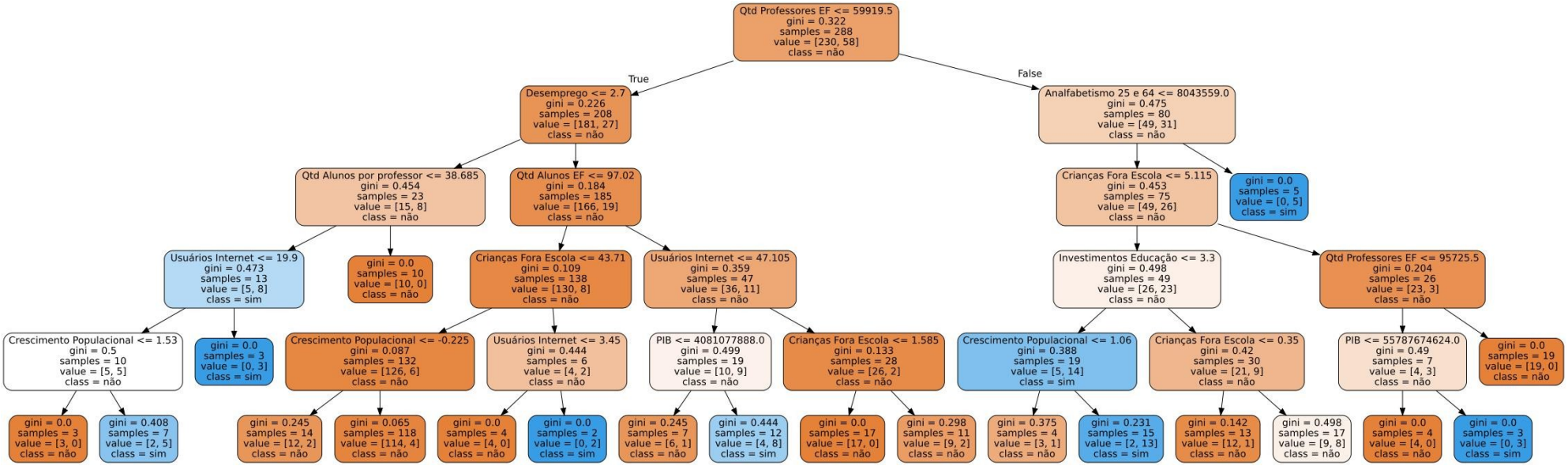

Figura 5.5 – Resultado gráfico da classificação dos indicadores socioeconomicos pelo DTC

Todos os modelos (SVC, *Gaussian Naive Bayes* e *Decision Tree Classifier*) tiveram, em média, acurácias de treinamento e precisão melhores do que as apresentadas pelo *Dummy*, segue uma tabela comparativa:

	Dummy	***SVC***	***GNB***	***DTC***
Acurácia treinamento	68.06%	82.64%	81.25%	99.07%
Acurácia previsão	72.04%	82.80%	81.18%	83.87%

Tabela 5.1 – Acurácias resultantes da aplicação dos modelos

De acordo com os resultados acima, todos os três modelos, quando comparados ao *Dummy*, apresentaram um aumento médio de 10% da acurácia de previsão. Por outro lado, também apresentaram acurácias de treinamento maiores que a do modelo de *baseline*, mas destaca-se o percentual de 99,07% do modelo *Decision Tree Classifier* (DTC).

Ao comparar os percentuais de precisão de classificação e previsão, cada modelo apresentou resultados diferentes, ou seja, nenhum modelo obteve resultados acima dos outros em todos os cenários. Assim, o DTC foi o melhor em classificar os casos de diminuição ou estagnação da quantidade de pessoas com AIDS, mas o SVC foi mais preciso com os casos de aumento dessa quantidade. Já nos cenários de previsão, o resultado foi o inverso do anterior, o SVC conseguiu prever todos os cenários de diminuição ou estagnação da quantidade de pessoas com AIDS, mas o DTC apresentou melhor resultado para os cenários de aumento. Segue uma tabela comparativa dos resultados apresentados nos relatórios individuais de cada modelo.

	Quantidade de pessoas com AIDS	**Dummy**	**SVC**	**GNB**	**DTC**
Precisão de classificação	Diminuição ou estagnação	81%	82%	83%	89%
	Aumento	19%	100%	54%	59%
Previsão	Diminuição ou estagnação	86%	100%	96%	91%
	Aumento	14%	11%	19%	56%

Tabela 5.2 – Resultados de classificação e previsão dos modelos

Agora serão realizadas as previsões com os modelos treinados pelos dados do *dataset* “***final_transformacao_dados_unificados_aids_indicadores.csv***”.

A primeira previsão acontecerá com os dados do Brasil. A lógica que gerou esses resultados está no *notebook* “***5_1_classificacao_aids_por_indicadores_socioeconomicos_brasil***”.

- Resultados de cada modelo:

1) *Dummy:* as acurácias de treinamento e previsão estão muito diferentes e os resultados das métricas de classificação para as duas classes foram baixos.

```
*********  Dummy  *********
Acurácia treinamento: 70.46%
Acurácia previsão: 16.67%

<< Matriz de confusão >>
Predito  0.0  1.0  All
Real
0.0        1    1    2
1.0        4    0    4
All        5    1    6

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.20      0.50      0.29         2
         1.0       0.00      0.00      0.00         4

    accuracy                           0.17         6
   macro avg       0.10      0.25      0.14         6
weighted avg       0.07      0.17      0.10         6
```

Figura 5.6 – Resultado previsão dos indicadores socioeconômicos do Brasil pelo Dummy

- Precisão: o resultado foi muito baixo para as duas classes. Só conseguiu classificar uma situação de estagnação ou diminuição da quantidade de pessoas com AIDS de forma correta.
- Revocação: o modelo não conseguiu prever as situações de aumento da quantidade de pessoas com AIDS, mas conseguiu prever 50% dos casos contrários.

2) SVC: as acurácias de previsão estão muito abaixo da de treinamento. O SVC teve uma acurácia de previsão aproximadamente 16% a mais do que a do *Dummy*.

```
*********  SVC  *********
Acurácia treinamento: 83.17%
Acurácia previsão: 33.33%

<< Matriz de confusão >>
Predito  0.0  All
Real
0.0        2    2
1.0        4    4
All        6    6

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.33      1.00      0.50         2
         1.0       0.00      0.00      0.00         4

    accuracy                           0.33         6
   macro avg       0.17      0.50      0.25         6
weighted avg       0.11      0.33      0.17         6
```

Figura 5.7 – Resultado previsão dos indicadores socioeconômicos do Brasil pelo SVC

- Precisão: esse modelo conseguiu classificar apenas 33% dos casos em que houve estagnação ou diminuição da quantidade de pessoas com AIDS, porém nenhum caso de aumento.
- Revocação: a previsão foi ótima para os casos de diminuição ou estagnação, mas sem sucesso para os de aumento. Logo, sua média de efetividade fica em 50%.

3) Gaussian Naive Bayes (GNB): a acurácia de treinamento foi ótima e a de previsão foi boa.

```
*********  Gaussian Naive Bayes  *********
Acurácia treinamento: 82.18%
Acurácia previsão: 66.67%

<< Matriz de confusão >>
Predito  1.0  All
Real
0.0        2    2
1.0        4    4
All        6    6

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00         2
         1.0       0.67      1.00      0.80         4

    accuracy                           0.67         6
   macro avg       0.33      0.50      0.40         6
weighted avg       0.44      0.67      0.53         6
```

Figura 5.8 – Resultado previsão dos indicadores socioeconômicos do Brasil pelo GNB

- Precisão: o GNB não conseguiu classificar os casos de diminuição ou estagnação da quantidade de pessoas infectadas com AIDS. Porém, teve 67% de sucesso na classificação dos casos de aumento dessa mesma quantidade.
- Revocação: esse modelo previu 100% dos casos de aumento da quantidade de pessoas com AIDS, mas foi incapaz de prever o contrário.

4) Decision Tree Classifier (DTC): o resultado dessa classificação apresentou ótimo resultado para a acurácia de treinamento, e uma de apenas 33% de previsão.

```
*********  DecisionTreeClassifier  *********
Acurácia treinamento: 99.67%
Acurácia previsão: 33.33%

<< Matriz de confusão >>
Predito  0.0  All
Real
0.0        2    2
1.0        4    4
All        6    6

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.33      1.00      0.50         2
         1.0       0.00      0.00      0.00         4

    accuracy                           0.33         6
   macro avg       0.17      0.50      0.25         6
weighted avg       0.11      0.33      0.17         6
```

Figura 5.9 – Resultado previsão dos indicadores socioeconômicos do Brasil pelo DTC

■ Precisão: o valor da precisão para a diminuição ou estagnação da quantidade de pessoas com AIDS foi de 33%, já para o aumento foi de 0%.

■ Revocação: as previsões para a classe 1 (aumento da quantidade de pessoas com AIDS) foi 0% de sucesso, mas para a classe 0 (estagnação ou diminuição) foi de 100%.

Nenhum modelo teve uma acurácia de previsão elevada, mesmo possuindo bons resultados de treinamento. Isso se deve pelo fato de que 66,67% dos casos do Brasil foram de aumento da quantidade de pessoas com AIDS e os registros da massa de treinamento são, em sua maioria, de casos de diminuição ou estagnação dessa quantidade. Logo, o desbalanceamento dos dados de treinamento deixou os modelos com baixa capacidade de prever situações de crescimento da quantidade de pessoas com AIDS em um país (durante 2010 a 2016).

	Dummy	***SVC***	***GNB***	***DTC***
Acurácia treinamento	70,46%	83,17%	82,18%	99,67%
Acurácia previsão	16,67%	33,33%	66,67%	33,33%

Tabela 5.3 – Acurácias resultantes da aplicação dos modelos para os dados do Brasil

Os dados acima mostram que mesmo o modelo de *Gaussian Naive Bayes* tendo uma acurácia de treinamento menor que o modelo de *Decision Tree Classifier*, a sua acurácia de previsão foi 30% maior.

Comparando os resultados da tabela abaixo, os modelos SVC e DTC tiveram a mesma precisão de classificação dos casos de diminuição ou estagnação da quantidade de pessoas com AIDS, já o GNB foi o único que conseguiu classificar casos de aumento.

Ainda nessa mesma tabela, os modelos SVC e DTC tiveram 100% de sucesso na previsão de casos de diminuição ou estagnação da quantidade de pessoas com AIDS. Porém, apenas o GNB conseguiu prever casos de aumento dessa medida, mas além de conseguir prever, o seu resultado foi de 100% de sucesso nisso.

	Quantidade de pessoas com AIDS	**Dummy**	**SVC**	**GNB**	**DTC**
Precisão de classificação	Diminuição ou estagnação	20%	33%	0%	33%
	Aumento	0%	0%	67%	0%
Previsão	Diminuição ou estagnação	50%	100%	0%	100%
	Aumento	0%	0%	100%	0%

Tabela 5.4 – Resultados de classificação e previsão dos modelos para os dados do Brasil

A segunda previsão acontecerá com os dados dos Estados Unidos da América. A lógica que gerou esses resultados está no *notebook* "***5_2_classificacao_aids_por_indicadores_socioeconomicos_eua***".

- Resultados de cada modelo:

  1) *Dummy:* as acurácias de treinamento e previsão estão próximas e a média dos resultados das métricas de classificação para as duas classes foram baixas.

```
*********  Dummy  *********
Acurácia treinamento: 70.46%
Acurácia previsão: 66.67%

<< Matriz de confusão >>
Predito  0.0  1.0  All
Real
0.0        4    1    5
1.0        1    0    1
All        5    1    6

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.80      0.80      0.80         5
         1.0       0.00      0.00      0.00         1

    accuracy                           0.67         6
   macro avg       0.40      0.40      0.40         6
weighted avg       0.67      0.67      0.67         6
```

Figura 5.10 – Resultado previsão dos indicadores socioeconômicos dos EUA pelo Dummy

  ■ Precisão: o resultado tanto para classificar casos de estagnação ou diminuição da quantidade de pessoas com AIDS foi de 80% de sucesso, mas 0% na situação de aumento.

  ■ Revocação: o modelo não conseguiu prever as situações de aumento da quantidade de pessoas com AIDS, mas conseguiu prever 80% dos casos contrários.

2) SVC: os percentuais das acurácias de treinamento e previsão foram praticamente iguais a 83%.

```
*********  SVC  *********
Acurácia treinamento: 83.17%
Acurácia previsão: 83.33%

<< Matriz de confusão >>
Predito  0.0  All
Real
0.0        5    5
1.0        1    1
All        6    6

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.83      1.00      0.91         5
         1.0       0.00      0.00      0.00         1

    accuracy                           0.83         6
   macro avg       0.42      0.50      0.45         6
weighted avg       0.69      0.83      0.76         6
```

Figura 5.11 – Resultado previsão dos indicadores socioeconômicos dos EUA pelo SVC

■ Precisão: esse modelo conseguiu classificar 83% dos casos de estagnação ou diminuição da quantidade de pessoas com AIDS, mas nenhum caso de aumento.

■ Revocação: a previsão foi de 100% para os casos de diminuição ou estagnação, mas sem sucesso para os de aumento. Logo, sua média de efetividade fica de 50%.

3) Gaussian Naive Bayes (GNB): a acurácia de treinamento foi de 82,18% e a de previsão foi de apenas 16,67%.

```
*********  Gaussian Naive Bayes  *********
Acurácia treinamento: 82.18%
Acurácia previsão: 16.67%

<< Matriz de confusão >>
Predito  1.0  All
Real
0.0        5    5
1.0        1    1
All        6    6

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00         5
         1.0       0.17      1.00      0.29         1

    accuracy                           0.17         6
   macro avg       0.08      0.50      0.14         6
weighted avg       0.03      0.17      0.05         6
```

Figura 5.12 – Resultado previsão dos indicadores socioeconômicos dos EUA pelo GNB

- Precisão: o GNB não conseguiu classificar os casos de diminuição ou estagnação da quantidade de pessoas infectadas com AIDS e teve apenas 17% de sucesso na classificação dos casos de aumento dessa mesma quantidade.
- Revocação: esse modelo previu 100% dos casos de aumento da quantidade de pessoas com AIDS, mas foi incapaz de prever o contrário.

4) Decision Tree Classifier (DTC): o resultado dessa classificação apresentou praticamente 100% de sucesso para a acurácia de treinamento e conseguiu uma taxa 83,3% de acurácia de previsão.

```
*********  DecisionTreeClassifier  *********
Acurácia treinamento: 99.67%
Acurácia previsão: 83.33%

<< Matriz de confusão >>
Predito  0.0  All
Real
0.0        5    5
1.0        1    1
All        6    6

<< Relatório de classificação >>
              precision    recall  f1-score   support

         0.0       0.83      1.00      0.91         5
         1.0       0.00      0.00      0.00         1

    accuracy                           0.83         6
   macro avg       0.42      0.50      0.45         6
weighted avg       0.69      0.83      0.76         6
```

Figura 5.13 – Resultado previsão dos indicadores socioeconômicos dos EUA pelo DTC

- Precisão: o modelo foi incapaz de classificar os casos de aumento da quantidade de pessoas com AIDS, mas conseguiu 83% de precisão na classificação de diminuição ou estagnação dessa mesma quantidade.
- Revocação: a precisão para prever casos da classe 0 (estagnação ou diminuição) foi de 100%, mas foi incapaz de prever os casos da classe 1 (aumento da quantidade de pessoas com AIDS).

Diferente dos resultados com os dados do Brasil, aqui houve ótimos resultados para as acurácias. Em especial o modelo *Decision Tree Classifier* apresentou uma acurácia de treinamento de 99,67% e uma de previsão de 83,33%. Apresentando assim as melhores médias de acurácias.

	Dummy	***SVC***	***GNB***	***DTC***
Acurácia treinamento	70,46%	83,17%	82,18%	99,67%
Acurácia previsão	66,67%	83,33%	16,67%	83,33%

Tabela 5.5 – Acurácias resultantes da aplicação dos modelos para os dados dos EUA

Já a tabela abaixo apresenta os percentuais de acerto para as previsões das classes 0 e 1. Assim, segundo os dados da tabela 5.6, os modelos SVC e DTC, igual ao resultado com os dados do Brasil, tiveram a mesma precisão de classificação dos casos de diminuição ou estagnação da quantidade de pessoas com AIDS, já o GNB, com apenas 17% de precisão, foi o único que conseguiu classificar casos de aumento.

Além disso, os modelos SVC e DTC tiveram 100% de sucesso na previsão de casos de diminuição ou estagnação da quantidade de pessoas com AIDS, mas apenas o GNB conseguiu prever casos de aumento dessa medida, com 100% de sucesso nisso.

	Quantidade de pessoas com AIDS	**Dummy**	**SVC**	**GNB**	**DTC**
Precisão de classificação	Diminuição ou estagnação	80%	83%	0%	83%
	Aumento	0%	0%	17%	0%
Previsão	Diminuição ou estagnação	80%	100%	0%	100%
	Aumento	0%	0%	100%	0%

Tabela 5.6 – Resultados de classificação e previsão dos modelos para os dados dos EUA

Dessa forma, os modelos treinados com todos os dados (menos os do Brasil e dos EUA) tiveram melhores acurácias de previsão para os cenários do EUA. O

modelo *Decision Tree Classifier* conseguiu uma acurácia média de previsão de 83,33%. Porém, os percentuais específicos de previsão foram muito próximos, logo em ambos os casos seria preciso utilizar um modelo para prever cada situação, como: o DTC ou SVC para situações em que a hipótese vincule a ideia de diminuição ou estagnação da quantidade de pessoas infectadas com AIDS e o GNB para o caso de aumento dessa mesma quantidade.

6. Apresentação dos Resultados

Segundo a orientação desse capítulo, segue o *workflow* motivador desse estudo através do modelo canvas proposto por Vasandani.

<table>
<tr><td colspan="3">Título: “Impacto de indicadores socioeconômicos na quantidade de pessoas com AIDS”</td></tr>
<tr><td>Definição do problema: analisar conjuntos de dados sociais, econômicos e educacionais para relacionar com o aumento ou não da quantidade de pessoas com AIDS.</td><td>Resultados e previsões: relacionar dados de indicadores das entidades mundiais e prever cenários de aumento ou diminuição de pessoas com AIDS para os países Brasil e Estados Unidos da América.

Hipótese: Impacto de indicadores socioeconômicos na quantidade de pessoas com AIDS.</td><td>Aquisição de dados: os conjuntos de dados foram obtidos no repositórios do Banco Mundial e da UNAIDS.</td></tr>
<tr><td>Modelagem: são aplicados vários processos de transformação e análise para chegar em um dataset a ser utilizado pelos modelos de classificação da biblioteca sklearn.</td><td>Avaliação do modelo: os resultados dos modelos são avaliados através do relatório de classificação e da matriz de confusão.</td><td>Preparação dos dados: os dados que não vieram preenchidos foram tratados através da substituição de valores do registro vizinho ou preenchidos com zero, a depender do caso.</td></tr>
</table>

O trabalho realizado com os registros de diferentes *datasets* tornou possível conhecer o resultado da coleta de dados que diferentes organizações mundiais realizaram e como isso pode auxiliar diretrizes governamentais de vários países espalhados pelo mundo.

O conjunto de dados fornecido pelo Banco Mundial, por ser o resultado da união de outros conjuntos, apresentou uma grande quantidade de valores em branco, mas também mostrou que existem muitos indicadores que possibilitam extrair conclusões sobre o estado das áreas da educação, economia e saúde de um país.

Por outro lado, o *dataset* fornecido pela UNAIDS apresentou mais registros com dados preenchidos, mas em uma faixa temporal mais curta (2010 a 2019). Além disso, 81.19% dos casos presentes nesse *dataset* foi de estabilização ou diminuição da quantidade de pessoas com AIDS e isso atrapalhou os modelos de classificação a identificarem um caso de aumento dessa quantidade.

A fim de explorar o resultado dessas classificações, foram realizadas duas execuções de teste separadas: uma com os dados do Brasil e outra com os dos Estados Unidos da América. Os dados desses dois países foram divididos em seis períodos:

- 2010 a 2011;
- 2011 a 2012;
- 2012 a 2013;
- 2013 a 2014;
- 2014 a 2015; e
- 2015 a 2016.

De acordo com os dados do Brasil, apenas nos períodos entre 2011 a 2012 e 2015 a 2016 houve estagnação ou diminuição na quantidade de pessoas com AIDS. Já os dados dos Estados Unidos da América apresentaram outro cenário: apenas entre 2013 a 2014, houve aumento dessa quantidade de pessoas.

O resultado dos percentuais de classificação e previsão variaram muito entre os algoritmos utilizados. Isso tornou possível concluir que para medir o aumento da quantidade de pessoas com AIDS, o algoritmo que apresentou melhores resultados foi o Gaussian Naive Bayes, já para a situação inversa foram os modelos *SVC* e *Decision Tree Classifier*.

Além disso, a análise dos dados relacionados com os indicadores socioeconômicos, entre 2010 e 2016, permitiu concluir que:

- Vinte e cinco países tiveram decréscimo demográfico.
- Os dez países com a maior quantidade de estudantes por professor (no ensino fundamental) são do continente africano (ver *figura 6.1*).

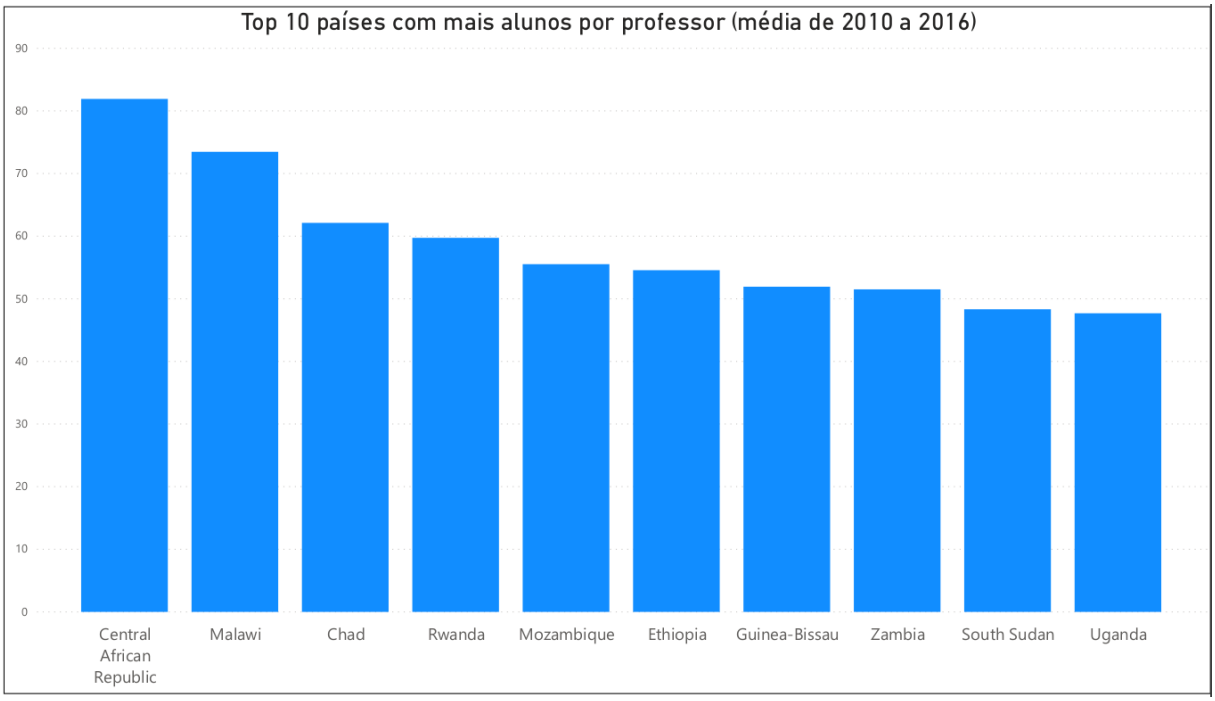

Figura 6.1 – Lista dos 10 países com maior quantidade de alunos por professor (média de 2010 a 2016).

- A Espanha possui um PIB médio de 1,361 trilhões de dólares e está entre os 10 países com maiores taxas de desemprego (ver *figura 6.2*).

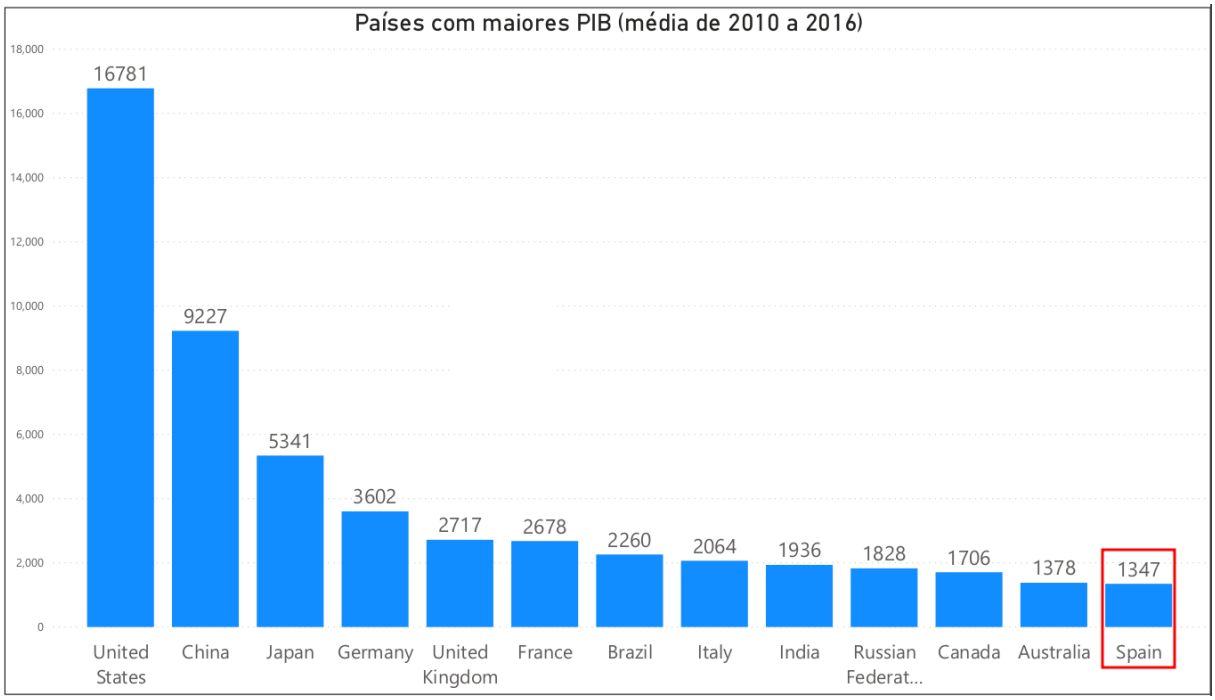

Figura 6.2 – Lista dos 10 países com maiores PIB (média de 2010 a 2016).

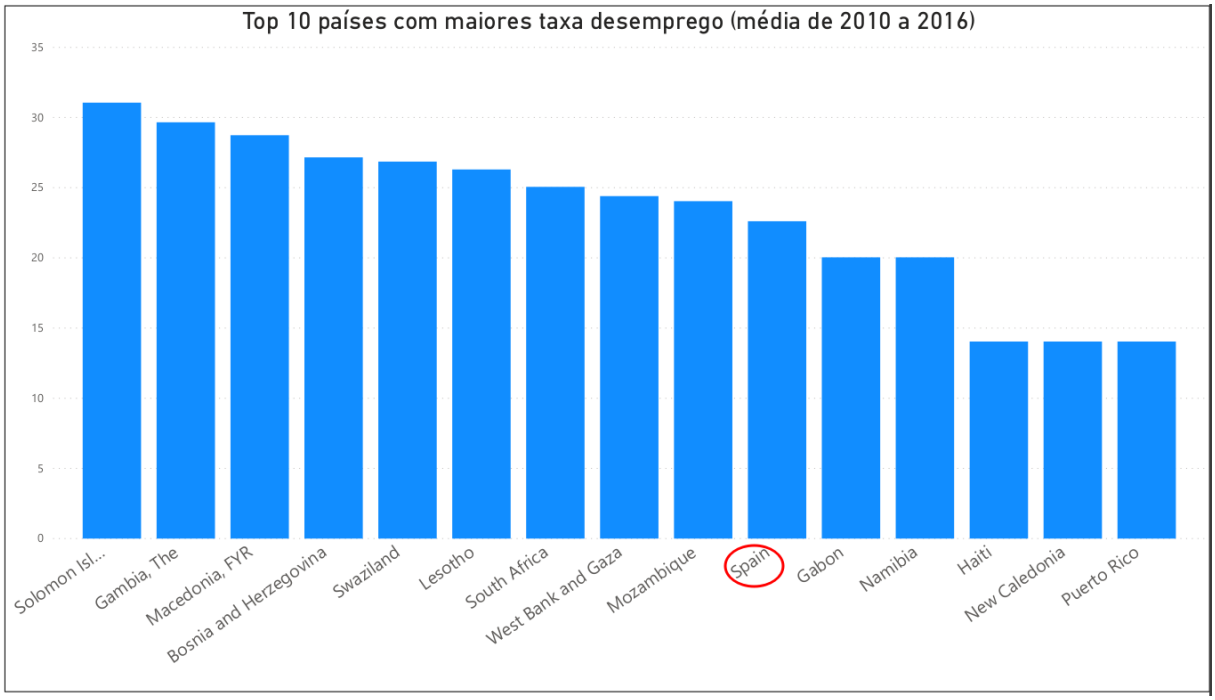

Figura 6.3 – Lista dos 10 países com maiores taxas de desemprego (média de 2010 a 2016).

- Nenhum dos 10 países (ver *figura 6.2*) com maiores PIB está entre os 10 países que mais investem em educação (ver *figura 6.4*).

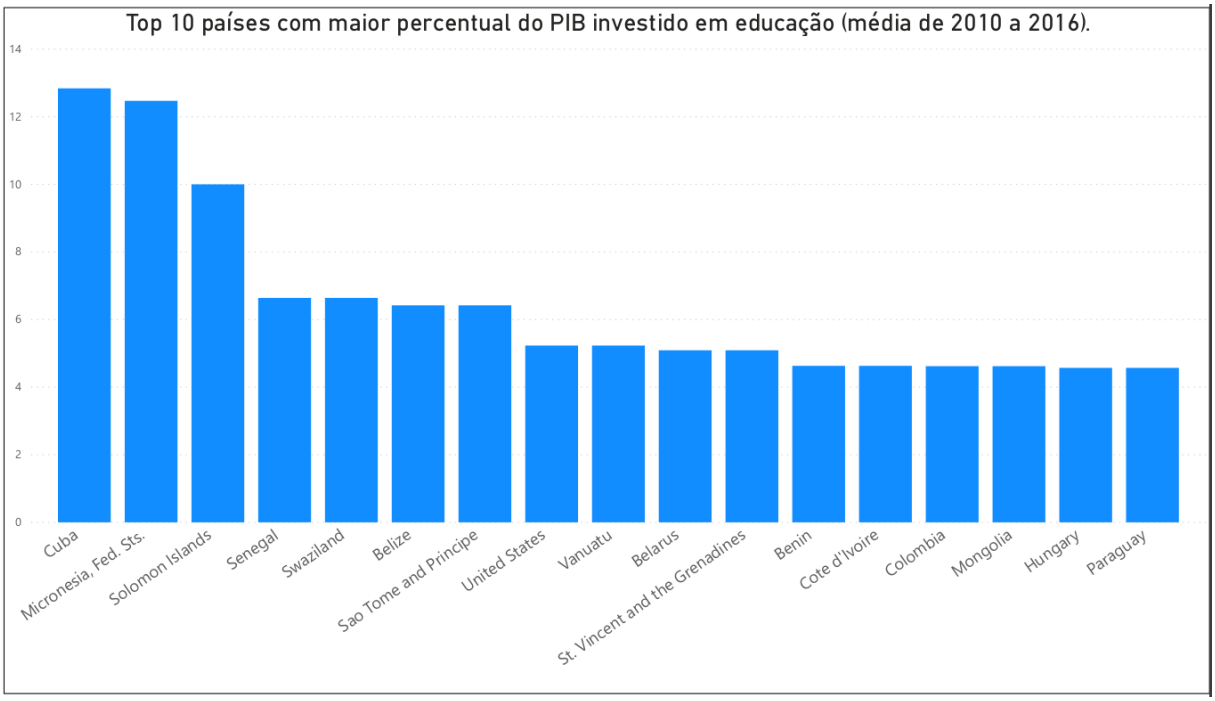

Figura 6.4 – Lista dos 10 países com maior percentual do PIB investido em educação (média de 2010 a 2016).

- Os países Índia, Paquistão e Bangladesh estão nas mesmas posições das listas de países com maiores taxas de analfabetismo de 15 a 24 anos e de 25 a 64 anos.

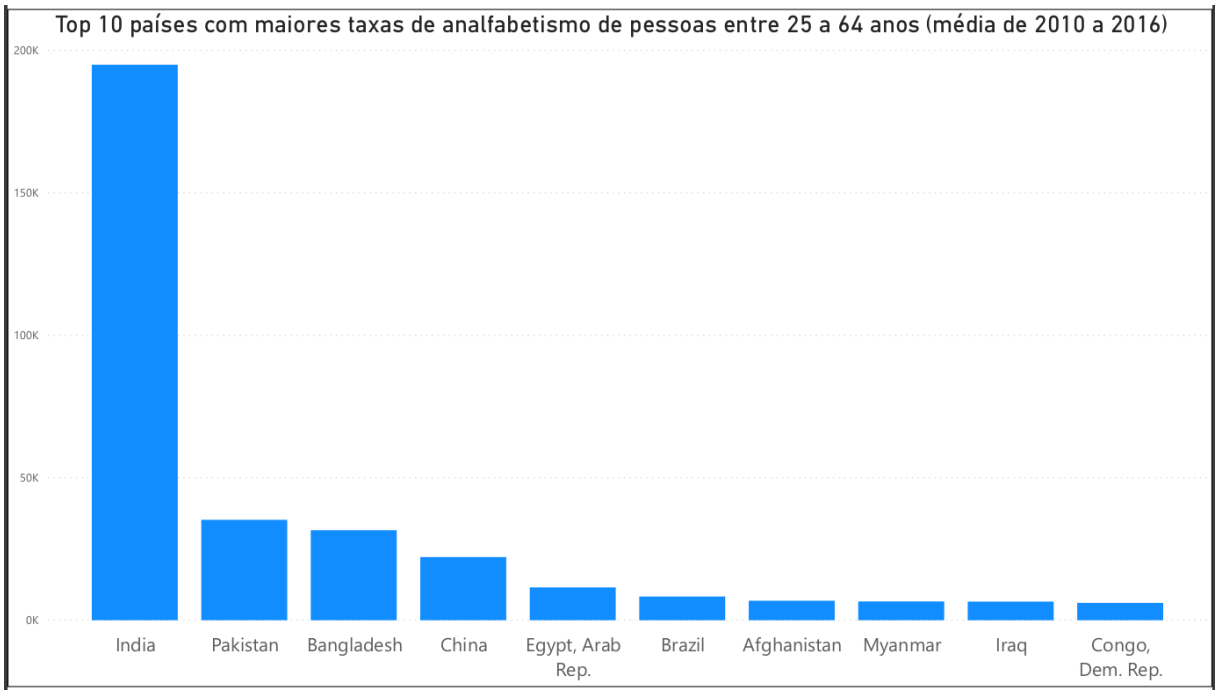

Figura 6.5 – Lista de 10 países com maiores taxas de analfabetismo entre 25 a 64 anos (média de 2010 a 2016).

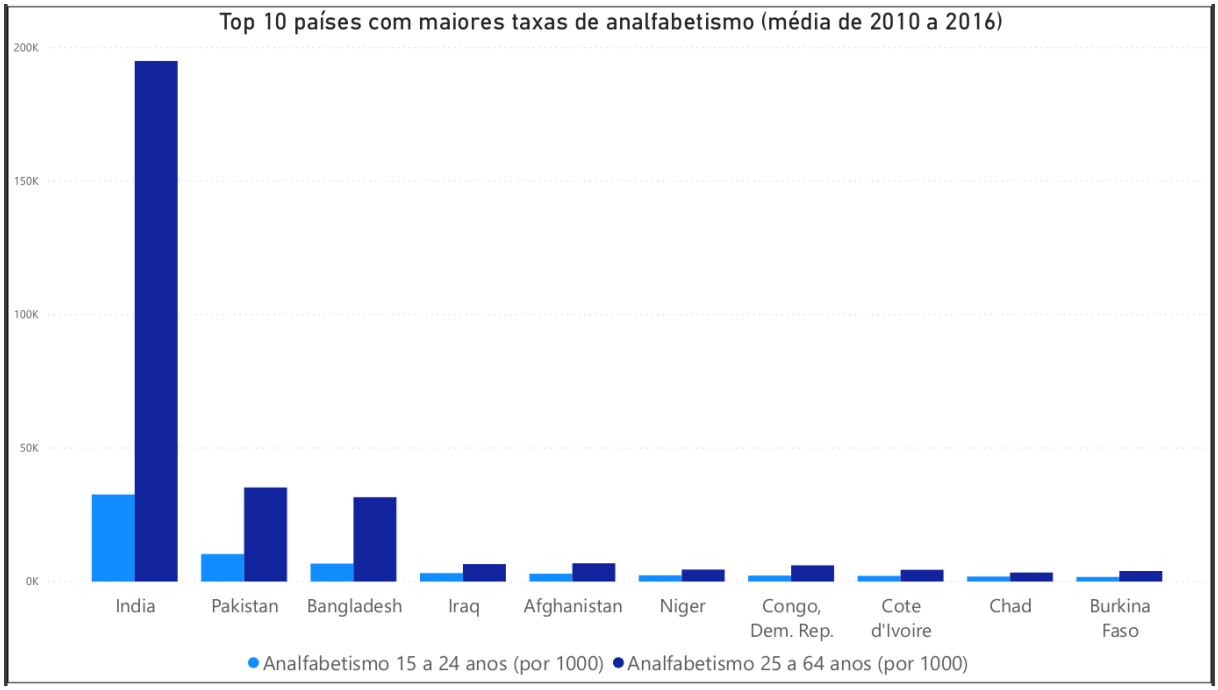

Figura 6.6 – Lista de 10 países com maiores taxas de analfabetismo entre 15 a 24 anos e 25 a 64 anos (média de 2010 a 2016).

- A *figura 6.5* mostra o impacto do desenvolvimento econômico dos países emergentes na taxa de analfabetismo. Destacam-se China e Brasil que, juntos com a Índia, ocupam, respectivamente, as posições de segundo, sétimo e nono lugar na lista de países com maiores médias de PIB entre 2010 a 2016 (ver *figura 6.2*), mas também estão entre os países com maiores taxas de analfabetismos de pessoas entre 25 e 64 anos (ver *figura 6.6*).
- Mais de 70% dos países possuem uma quantidade acima da média de alunos matriculados no ensino fundamental.
- Os 5 países com maior quantidade de crianças que deveriam estudar (no ensino fundamental), mas não estão, são do continente africano (Sudão do Sul, Libéria, Eritreia, Guiné Equatorial e Sudão).

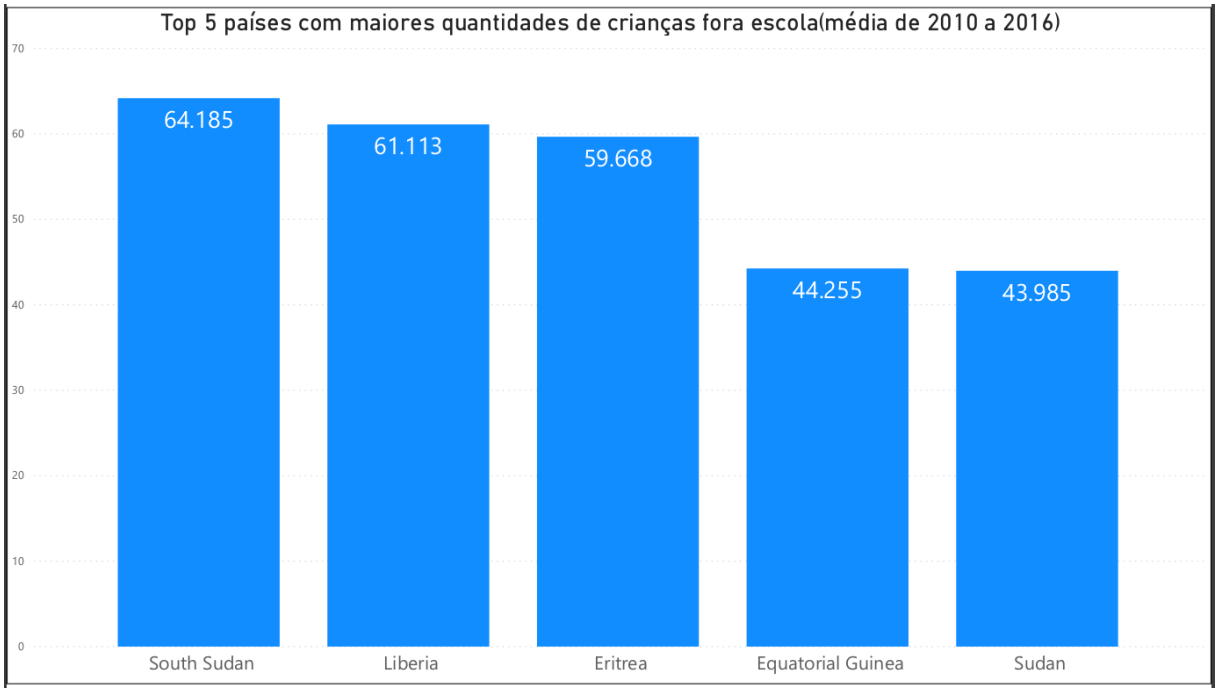

Figura 6.7 – Lista de 5 países com maiores quantidades de crianças fora da escola (média de 2010 a 2016).

- Entre 2010 e 2016, houve mais ocorrências de diminuição ou estagnação da quantidade de pessoas com AIDS, em vez de aumento.

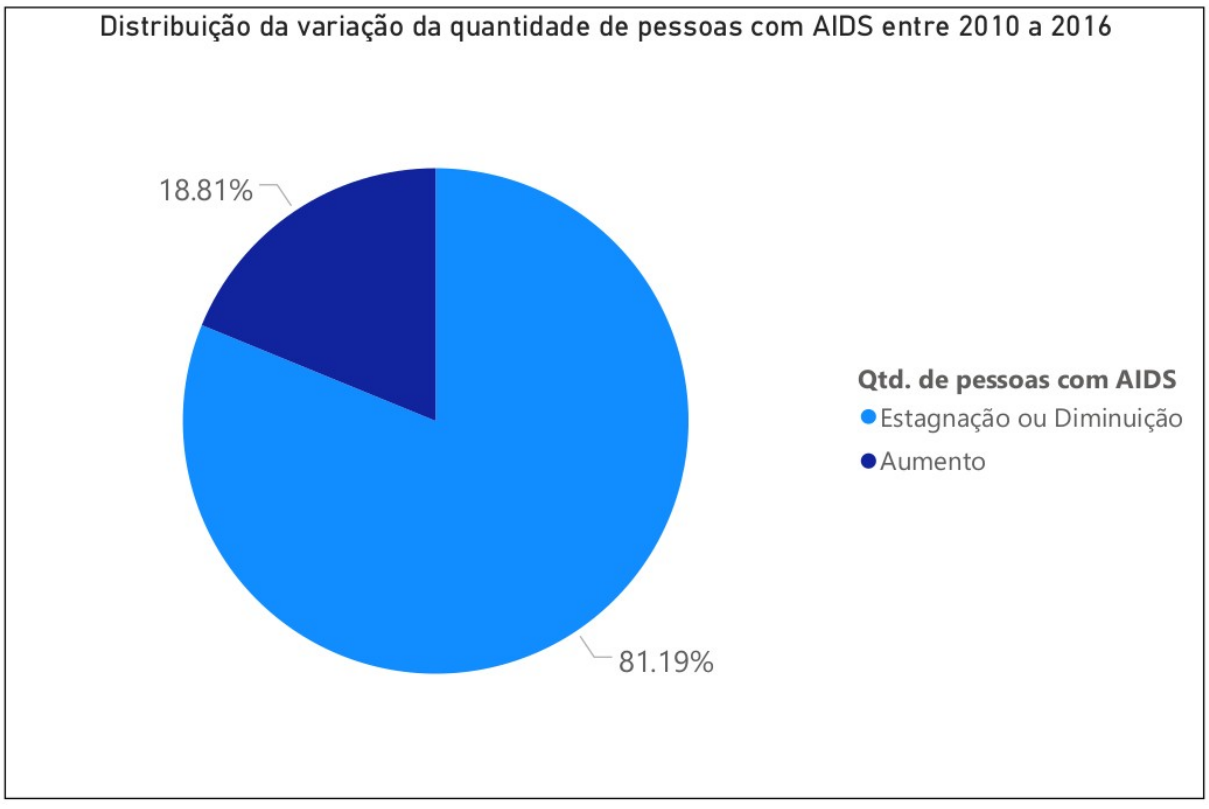

Figura 6.8 – Distribuição da variação da quantidade de pessoas com AIDS (média de 2010 a 2016).

7. Links

Link para o vídeo:

- youtu.be/LaORwovB9Rk
- https://drive.google.com/drive/folders/1FqUE6EhjP9rytREQyedB0spyTbz3lkqb?usp=sharing

Link para o repositório:

- https://github.com/joaocvr/puc-minas-tcc
- https://drive.google.com/drive/folders/1FqUE6EhjP9rytREQyedB0spyTbz3lkqb?usp=sharing

Links para os datasets:

- datacatalog.worldbank.org/dataset/education-statistics
- https://lawsandpolicies.unaids.org/

*Esses datasets (”*Edstats_csv.zip*” e “*Epidemic transition metrics_Trend of new HIV infections.csv*”) foram armazenados, a fim de *backup*, na pasta “*data*” do repositório https://github.com/joaocvr/puc-minas-tcc

LISTA DE FIGURAS

Figura 6.8 – Distribuição da variação da quantidade de pessoas com AIDS (média de 2010 a 2016).

LISTA DE TABELAS