Projeto Final ICD: Coronavirus Pandemic (COVID-19)

Letícia S. M. Alves[1], **Luiz P. P. Amaral**[1], **Othávio R. C. Araújo**[1], **Pedro G. B. Ribeiro**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

{leticiasma,luizphilippe,ruddacc,pdrgeovanni}@ufmg.br

1. Links do Projeto

Código fonte: https://github.com/LuizPPA/covid-19-spread-analysis
Apresentação em vídeo: https://youtu.be/IgW3NuSMdeA

2. Introdução

2.1. Motivação

O tema escolhido para o trabalho foi motivado pela notória proporção que a pandemia do COVID-19 tomou no mundo inteiro. Todos os dias, a toda hora, mais e mais dados da doença surgem nos meios de comunicação. Devido a isso, temos um grande volume de informações que pode ser ricamente explorado. Nosso objetivo com o trabalho é, portanto, realizar algumas das possíveis análises desses dados abordando questões interessantes que podem ser levantadas.

2.2. Nossas perguntas de pesquisa

1. Como o numero de habitantes impacta no número de infectados?
2. Como a densidade populacional impacta no número de infectados?
3. Como países desenvolvidos e países em desenvolvimento respodem à pandemia?
4. A prevalência de fumantes interfere na taxa de mortalidade por Covid-19?

3. Metodologia

3.1. A base de dados utilizada

O projeto foi realizado sobre a base pública: ”Coronavirus Pandemic (COVID-19) – the data”.

O banco fornece dados por países. Encontramos por exemplo: número total de casos e mortes, novos casos e mortes diariamente, número de testes realizados, taxa de testes que deram positivo para a doença etc. Também podemos filtrar esses dados pelo tamanho da população, densidade, idade média da população e outras variáveis. Além disso, essas taxas podem ser relacionadas a fatores como fumar, sexo (feminino e masculino), extrema pobreza, entre outros. Assim inicia nossa busca de como cada atributo de nosso banco influencia os dados registrados.

O banco: `https://ourworldindata.org/coronavirus-data`

3.2. Tratamento dos Dados

Dependendo da pergunta de interesse foi necessário escolher certas colunas em detrimento de outras a fim de manter o foco no que queríamos analisar. Além disso, limpamos os dados excluíndo linhas com valores faltantes quando necessário.

4. Resultados

4.1. Como o número de habitantes impacta no número de infectados?

4.1.1. Análise Exploratória

Para responder à nossa pergunta, foi plotado o número total de casos registrados de Covid-19 em alguns países (com diferentes números de habitantes) no decorrer dos dias da pandemia. Buscamos dessa forma enxergar um padrão sobre como a doença se comporta ao longo do tempo, e nos gráficos gerados (Figuras: 1, 2, 3, 4) é perceptível uma similaridade entre os seus formatos de curva.

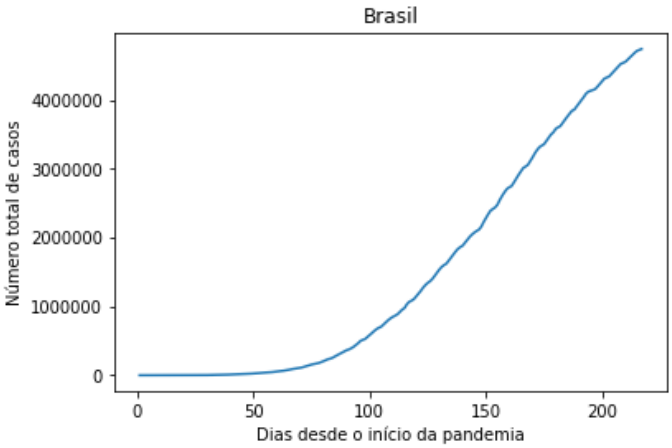

Figura 1. Casos registrados no Brasil com o passar dos dias de pandemia

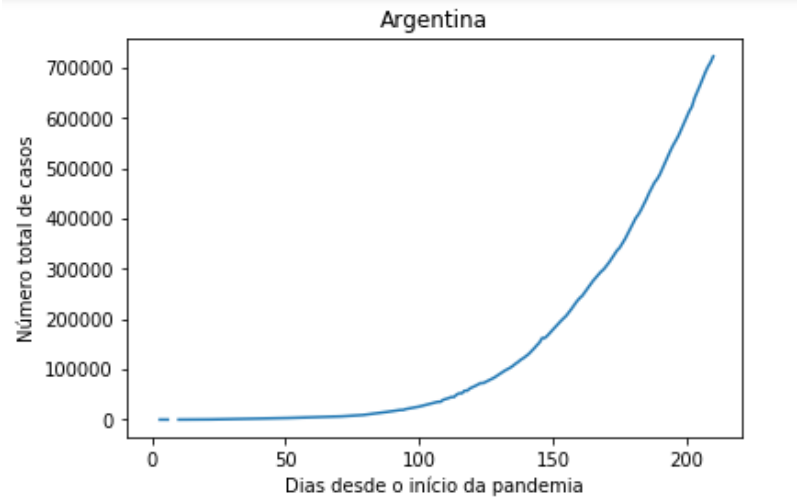

Figura 2. Casos registrados na Argentina com o passar dos dias de pandemia

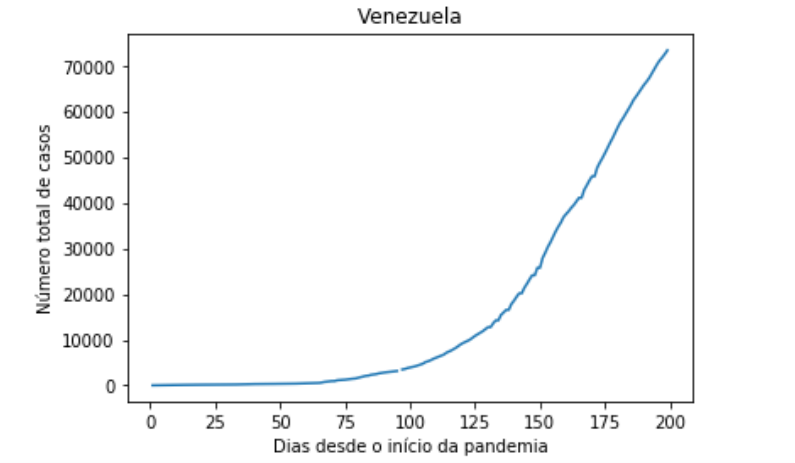

Figura 3. Casos registrados na Venezuela com o passar dos dias de pandemia

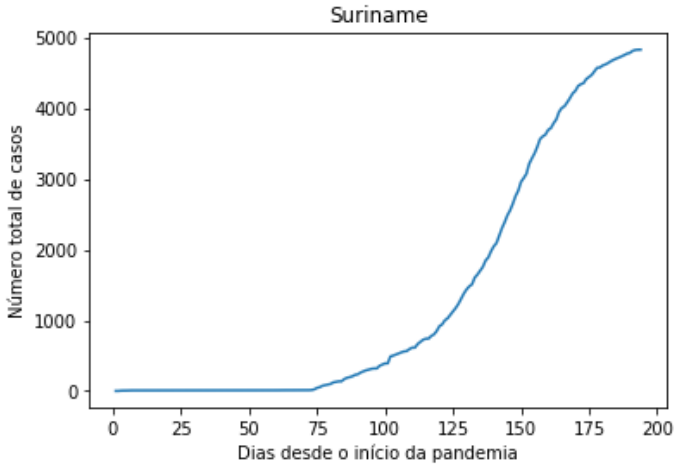

Figura 4. Casos registrados no Suriname com o passar dos dias de pandemia

Com essa observação, foi levantada a hipótese de que talvez o tamanho da população não influencia na propagação da Covid. Para trabalharmos em cima dessa ideia levantamos a seguinte hipótese nula: **"O tamanho da população influencia na quantidade de infectados pela Covid-19"**.

4.1.2. Teste de Hipótese e Intervalo de Confiança

Para realizarmos nosso teste de hipótese, dividimos o grupo de países em 2: países com uma grande população e países com uma pequena população. O parâmetro de divisão foi 20 milhões de habitantes. Foram realizados três testes distintos utilizando o bootstrap de 5000 amostras. O primeiro teste utiliza dados do 30º dia desde o início da pandemia, o segundo do 60º dia e o terceiro do 90º dia. Abaixo encontram-se os três histogramas (Figuras: 5, 6, 7) representando o bootstrap.

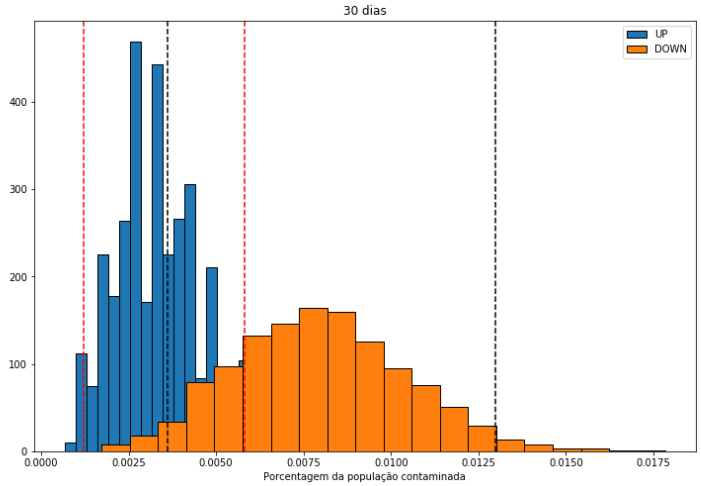

Figura 5. Bootstrap com dados do 30º dia de pandemia

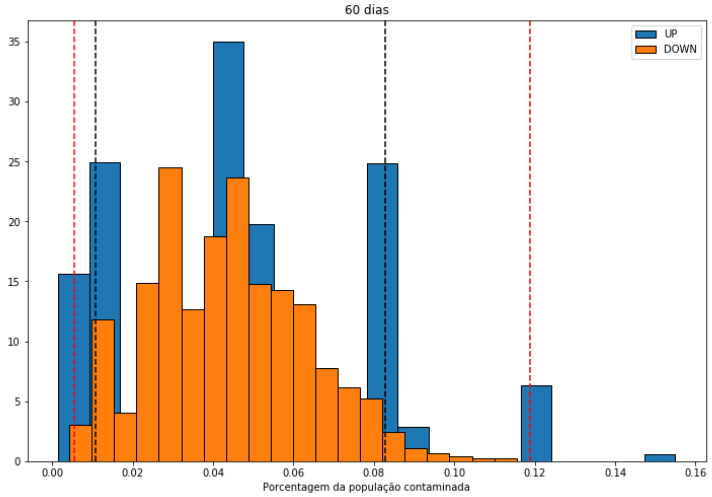

Figura 6. Bootstrap com dados do 60º dia de pandemia

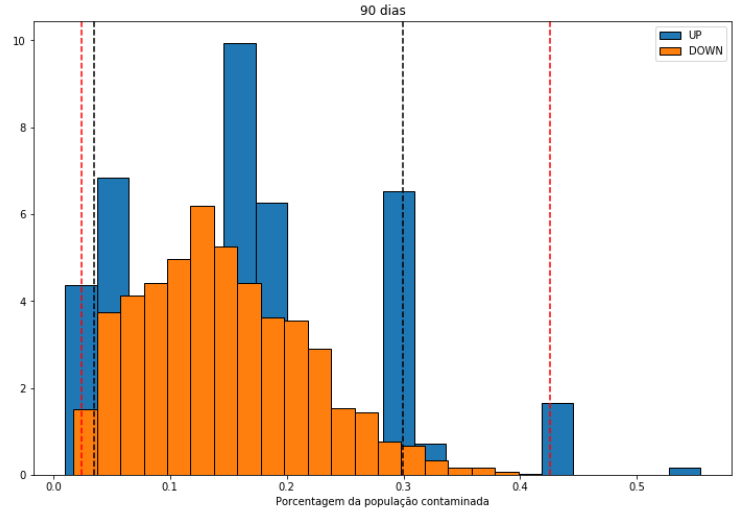

Figura 7. Bootstrap com dados do 90º dia de pandemia

Nos histogramas apresentados existem duas linhas pontilhadas vermelhas e pretas, que representam, respectivamente, o intervalo de confiança de **95%** do grupo de países com população superior a 20 milhões e o intervalo de confiança de **95%** do grupo de países com população inferior a 20 milhões. É perceptível que os dois intervalos de confiança se cruzam em todos os três histogramas plotados e isso não nos permite rejeitar a hipótese nula. Porém, ainda assim podemos inferir que o tamanho da população não influencia, determinantemente, na propagação da doença.

4.1.3. Conclusão

Os resultados obtidos com os três histogramas plotados nos levam a acreditar que a população de uma país exerce sim influência na propagação da doença. Mas, dada a análise exploratória inicial, podemos inferir que a população não afetará, obrigatoriamente de forma direta, a propagação. Isto é, podemos presumir que existem outros fatores distintos que tem um grande impacto na velocidade de propagação do vírus. Como exemplo as medidas tomadas pelo governo no início da pandemia e o comprometimento da população com as orientações da Organização Mundial da Saúde, que provavelmente geram um maior impacto maior no número de infectados em comparação ao tamanho da população.

E para demonstrar os resultados obtidos temos um exemplo da vida real: dois países com uma grande diferença no número total de habitantes, China e Estados Unidos, sendo que o primeiro possui uma população aproximadamente 4.25 vezes maior que a do segundo porém o número total de casos registrados pelo primeiro, até o dia 31/10/2020, é de 85.973, enquanto o segundo registrou 9.078.485 casos.

4.2. Como a densidade populacional impacta no número de infectados?

4.2.1. Análise Exploratória

Para responder à nossa pergunta, primeiramente temos na Figura 8 uma ideia visual da relação entre os dados: Densidade Populacional e Total de casos (por milhão). Observando esse gráfico, aparentemente não há quase nenhuma correlação entre essas duas variáveis.

(Observação: Foi necessário utilizar uma escala logarítmica para os eixos x e y para melhor visualização da tendência dos dados)

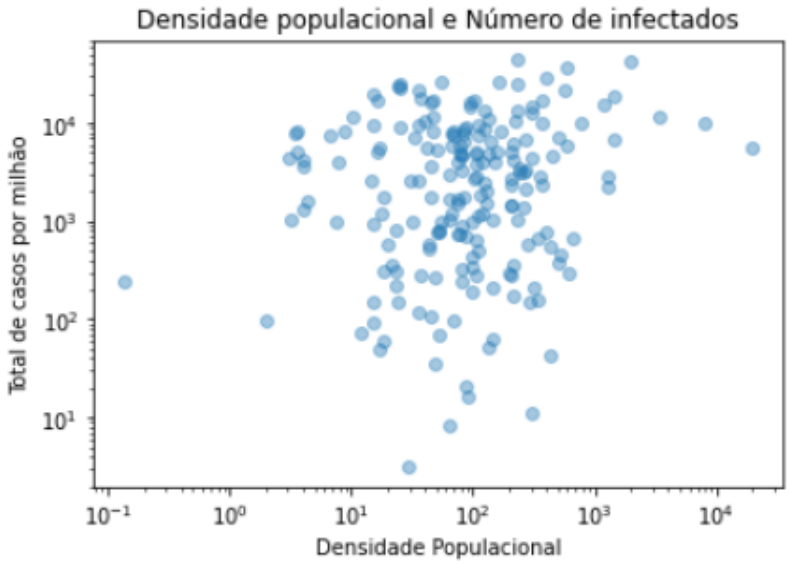

Figura 8. Densidade populacional e respectivos número de infectados (Escala logarítmica)

Pudemos de fato constatar essa baixa correlação calculando a Correlação de Pearson que foi de, aproximadamente, **0.071**.

Levantamos então a Hipótese Nula: **"A densidade populacional não influencia no número de casos da doença"**.

Tentaremos rejeitá-la a seguir.

4.2.2. Teste de Hipótese e Intervalo de Confiança

Consideramos como alta densidade populacional valores acima de 200. Para o teste de hipótese foi realizado um Teste de Permutação para 5000 amostras.

Abaixo segue o histograma da Figura 9 com as médias encontradas nas amostras para a variável Total de Casos (por milhão) para populações com densidades maiores que 200.

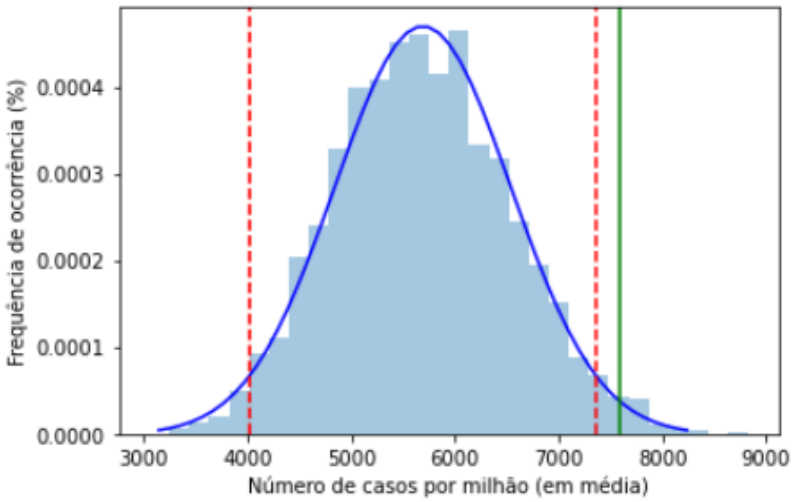

Figura 9. Histograma da média de casos totais (por milhão) das amostras

Junto ao histograma plotamos a curva de Distribuição Normal que se ajusta a esses dados. As linhas tracejadas em vermelho são o intervalo de confiança de **95%** para as médias. Tal intervalo vai de, aproximadamente, **4023.3** à **7352.8**.

Já a linha vertical em verde, é o valor médio real do número de casos para populações de alta densidade. Essa média real é de, aproximadamente **7595**. Porém, como esse valor está fora do intervalo de confiança calculado, podemos então rejeitar a hipótese nula. Ou seja, concluímos que: **A densidade populacional influencia sim no número total de casos da doença**.

4.2.3. Conclusão

Dadas as evidências, podemos dizer que, muito provavelmente, a densidade populacional de fato influencia o número total de casos da Covid-19 (uma vez que a hipótese nula foi rejeitada). Porém, como vimos anteriormente, a correlação entre esses dois fatores foi baixa. O que podemos então tentar argumentar é que outros fatores externos estão exercendo influência nesses resultados.

Intuitivamente, pensamos que quanto maior a densidade da população, maior seria o número de casos totais. Porém, isso não parece ser uma regra de tendência geral. Assim, pode ser que, por exemplo, fatores econômicos, políticos, índices de desenvolvimento ou medidas restritivas dos países façam com que, mesmo com a alta densidade de pessoas o número de casos não seja tão elevado quanto seria de se esperar.

Na Figura 10 plotamos algumas outras variáveis que poderiam estar correlacionadas e inluenciando a Densidade Populacional e o Total de Casos (por milhão).

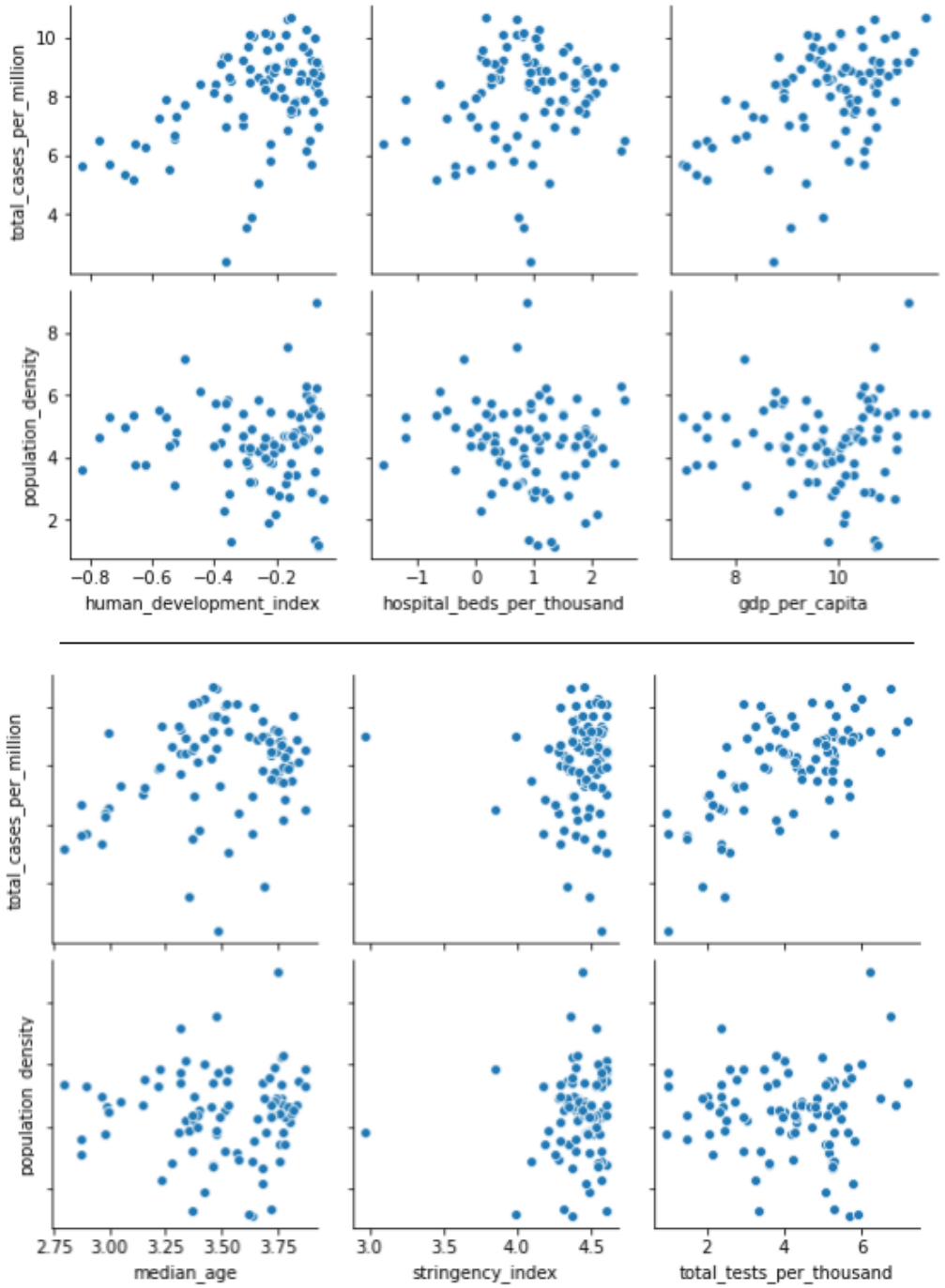

Figura 10. Outros fatores possívelmente relacionados à densidade populacional e ao número de infectados (Escalas logarítmicas)

As variáveis que aparentemente parecem estar mais relacionadas ao Total de Casos são o Índice de Desenvolvimento Humano, o GDP/PIB Per Capita, a Idade Média da População e o Total de Testes (por mil).

4.3. Como países desenvolvidos e países em desenvolvimento respondem a pandemia?

Para responder a essa pergunta, primeiramente precisamos definir uma métrica para considerar um país como ”desenvolvido”ou ”em desenvolvimento”. No nosso caso, adotaremos a ideia de que um país é desenvolvido quando seu índice de desenvolvimento humano (IDH) é igual ou superior à 0,7. Tentaremos relacionar diferentes estatísticas a respeito da evolução da pandemia em um país ao seu IDH e discorrer a causa dos resultados obtidos.

4.3.1. Análise Exploratória

Para entender e visualizar os dados à nossa disposição para estudos, inicialmente foi feito um plot de pares entre fatores socioeconômico e dados sobre o Covid, como IDH, número de casos, número de mortes, densidade populacional, número de leitos, porcentagem da população em condição de extrema pobreza, etc. Imediatamente fomos capazes de notar uma relação entre o IDH e o número de casos e mortes por milhão de habitantes, no entanto, ao contrário do que o senso comum pode sugerir, países com um IDH mais elevado apresentavam valores mais elevados tanto para o número de casos (Figura 11) quanto para o número de óbitos (Figura 12). Observando as demais relações, foram formuladas hipóteses para explicar as observações.

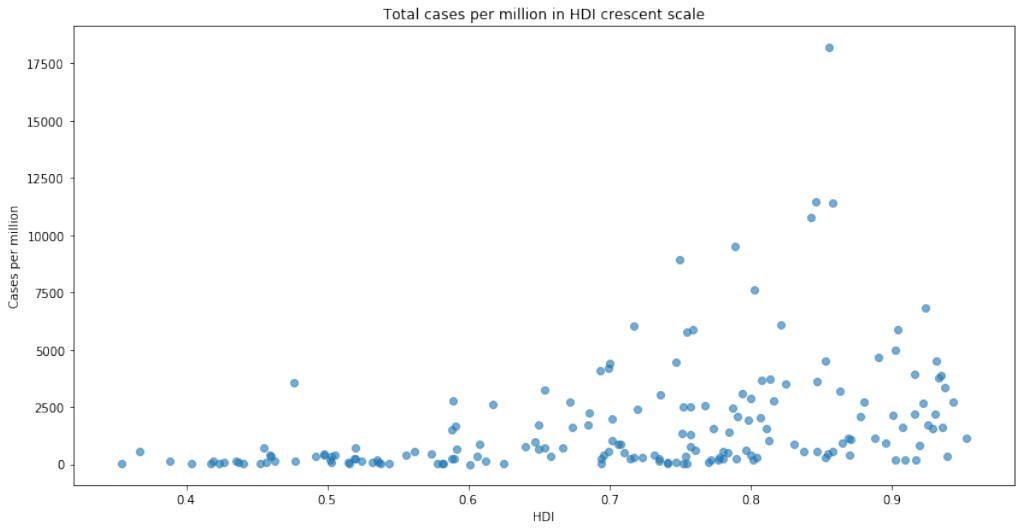

Figura 11. Gráfico de HDI e número de casos por milhões de habitantes.

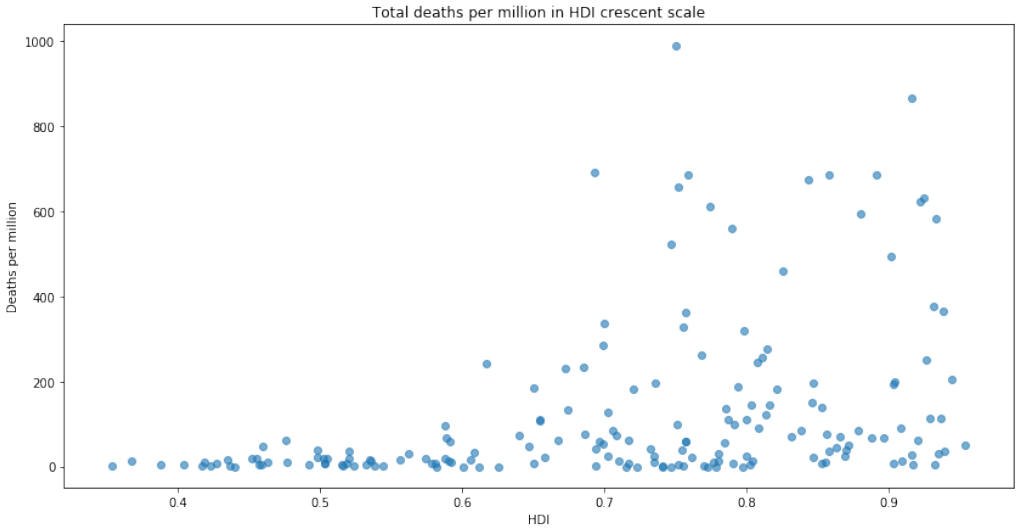

Figura 12. Gráfico de HDI e número de óbitos por milhões de habitantes.

4.3.2. IDH e número de casos

Foi realizado um teste de hipótese para que pudéssemos reconhecer formalmente a relação entre IDH e o número de casos registrados. Assumimos como hipótese nula: **"O IDH de um país não influencia no seu número de casos"**. Realizamos então um teste de permutação e os resultados são apresentados na Figura 13. Em azul, temos a distribuição obtida com as permutações realizadas (foram feitas mil repetições), em vermelho temos os limites do intervalo de confiança da distribuição e em verde temos a medida real obtida. Podemos notar que a média real está claramente fora do intervalo de confiaça.

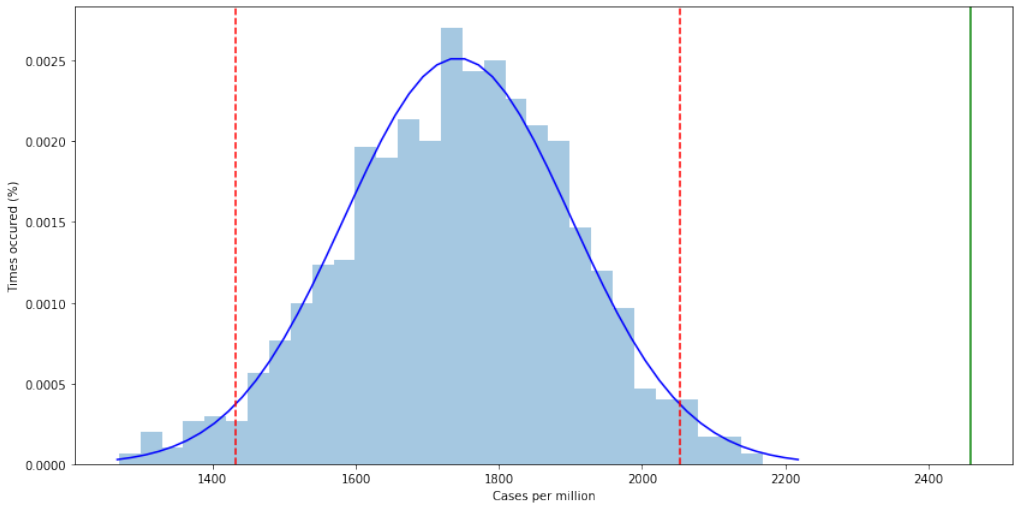

Figura 13. Teste de permutação realizado sobre os valores de número de casos em relação ao IDH.

Uma possível expliação para o motivo desta tendência é o número de testes realizados. Conseguimos verificar que países desenvolvidos apresentam um número de testes

realizados bem mais elevados do que os países em desenvolvimento. Além disso, países desenvolvidos tinham uma taxa de testes positivos consideravelmente menor do que aqueles em desenvolvimento, o que pode indicar uma subnotificação de casos em locais com menor IDH. Outro teste de hipótese foi realizado comprovando a relação entre o IDH e a taxa de resultados positivos (Figura 14).

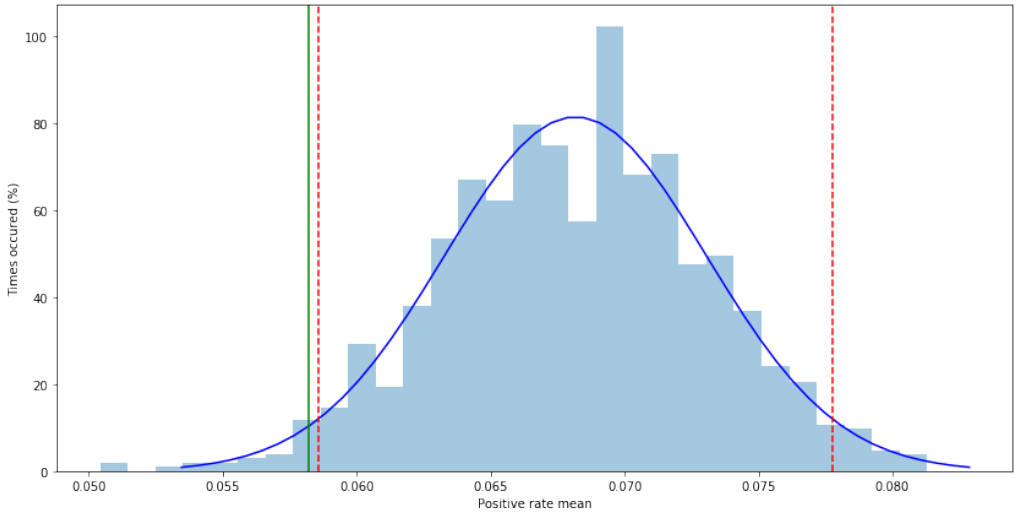

Figura 14. Teste de permutação realizado sobre os valores de taxa de testes positivos em relação ao IDH.

Uma observação interessante ao observar o boxplot dos dados de taxa de testes positivos foi a presença de oito *outliers* entre os dados, oito países com IDH superior a 0.7 com taxas de testes positivos muito superiores as dos demais países com IDH semelhantes. O boxplot e a tabela com os países são apresentados abaixo.

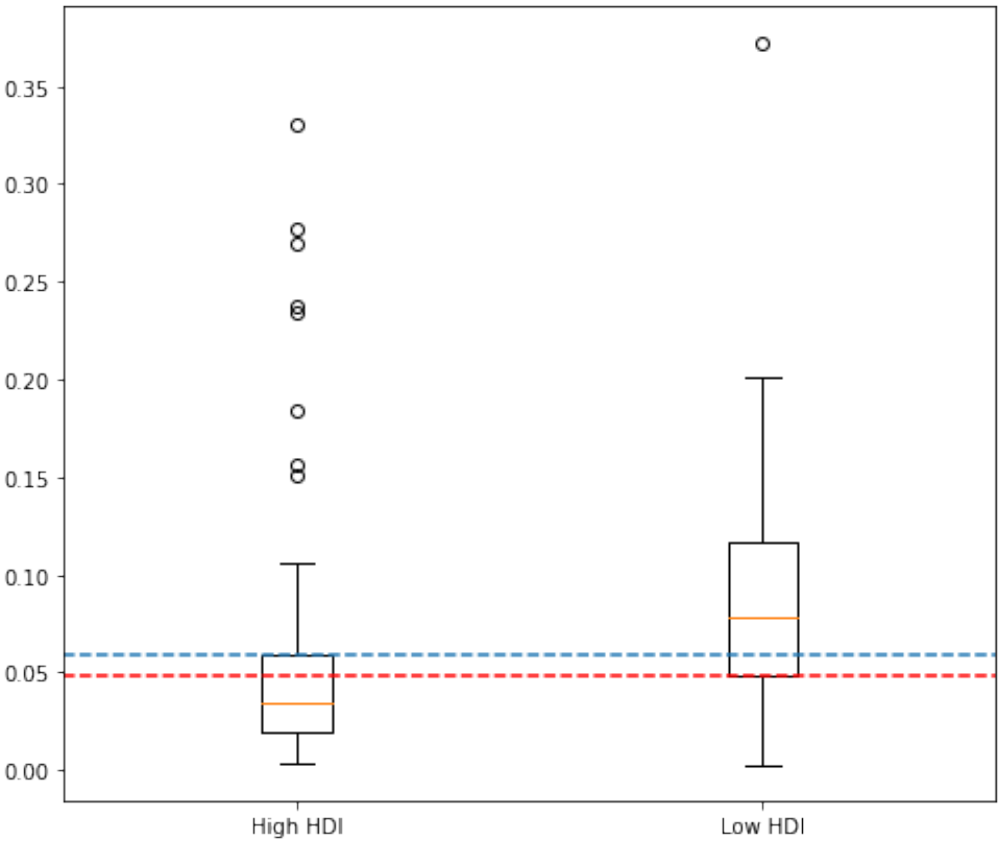

Figura 15. Box plot de taxas de testes positivos para países desenvolvidos (à esquerda) e em desenvolvimento (à direita).

	human_development_index	positive_rate
location		
Argentina	0.825	0.269115
Chile	0.843	0.150424
Dominican Republic	0.736	0.234042
Kuwait	0.803	0.184037
Mexico	0.774	0.277225
Oman	0.821	0.330039
Panama	0.789	0.237934
Qatar	0.856	0.156365

Figura 16. Países com alto IDH com taxa de testes positivos superior à 15%.

Portanto, embora os dados possam sugerir que os países desenvolvidos têm mais casos de COVID, é possível que isso seja devido a existência de subnotificação nos países em desenvolvimento, dessa forma, é possível dizer que países com maior IDH são mais eficientes na identificação da doença, o que auxilia no combate ao vírus.

4.3.3. IDH e número de óbitos

Baseado na conclusão acima, poderíamos esperar que países desenvolvidos apresentassem um menor número de óbitos, já que a doença pode ser fatal diagnosticada ou não, e com um diagnóstico precoce e mais recursos para tratamento de doentes, seria natural que mais mortes fossem evitadas. No entanto, os dados também vão contra essa noção inicialmente. Uma justifica que aceitaria todas as observações realizadas, é que países com maior IDH, em geral, têm uma maior idade média da população, e como o Covid apresenta maior risco a pessoas com 65 anos ou mais, é natural que um país com uma maior faixa de população idosa apresente um maior número de óbitos. Os dados de idade média corroboram com essa justificativa, vide a figura 17.

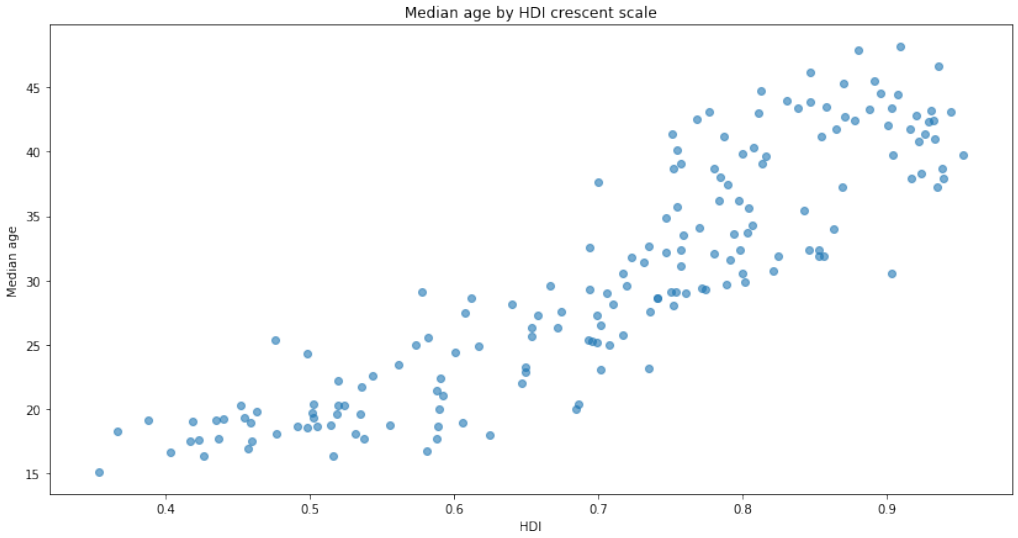

Figura 17. Gráfico de idade média da população e IDH de diferentes países. É possível observar a presença de uma tendência diretamente proporcional.

Para analisar como a tendência do crescimento da idade média da população pode estar relacionada à tendência do crescimento de número de óbitos, realizamos duas regressões polinomiais cujos resultados são exibidos nas figuras 18 e 19.

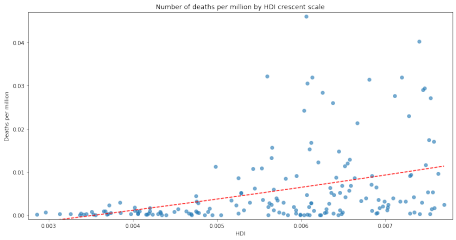

Figura 18

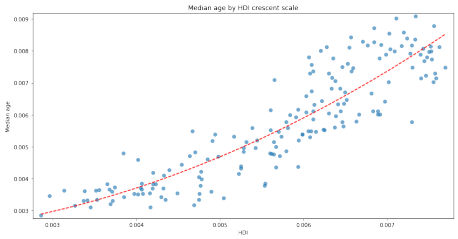

Figura 19

O resultado pode ser melhor visualizado na figura 20 onde plotamos ambas as curvas normalizadas sobre o mesmo plano e sem os pontos dos dados. É possível notar que as tendências se acompanham, por mais que não se trate de uma relação um para um, o que já era esperado, visto que um maior número de pacientes em estado crítico implica em mais pacientes internados e menos pacientes com quadros menos críticos recebendo tratamento adequado.

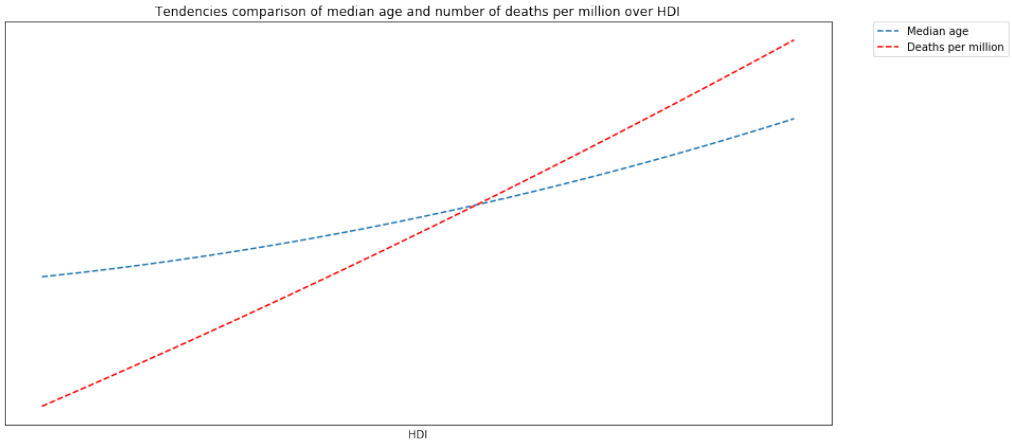

Figura 20. Tendências de idade média e número de óbitos relacionadas ao IDH de diferentes países. Em azul a tendência da idade média da população, em vermelho, a tendência do número de óbitos por milhões de habitantes.

4.3.4. Conclusão

O senso comum pode sugerir que países desenvolvidos são menos impactados pela pandemia do Covid-19, no entanto, os dados mostram por muitas vezes o contrário. É seguro dizer no entanto, que mesmo que tais países sejam tão afetados (ou mais afetados) que os demais, eles apresentam uma maior facilidade de lidar com os impactos que são consequência dessa crise.

Notamos que países com alto IDH apresentam uma maior eficiência na identificação de casos, por isso, é incerto dizer que esses países apresentam um maior número de casos, embora isso seja sugerido pelos dados. Esses países apresentam uma maior taxa de mortalidade por Covid, mas é possível relacionar essa tendência com fatores mais etários do que socioeconômicos.

4.4. A prevalência de fumantes interfere na taxa de mortalidade por Covid-19?

Neste tópico a vertente principal é analisar o número de fumantes em uma região e assim, de acordo com o número de mortes daquele país, determinar se, de alguma forma, essas razões se relacionam.

4.4.1. Análise exploratória

A visualização do comportamento dos dados, como nas Figura 21 e 22, pode mostrar como aparentemente os dados são afetados pela falta de padronização na disponibilidade de dados de cada país.

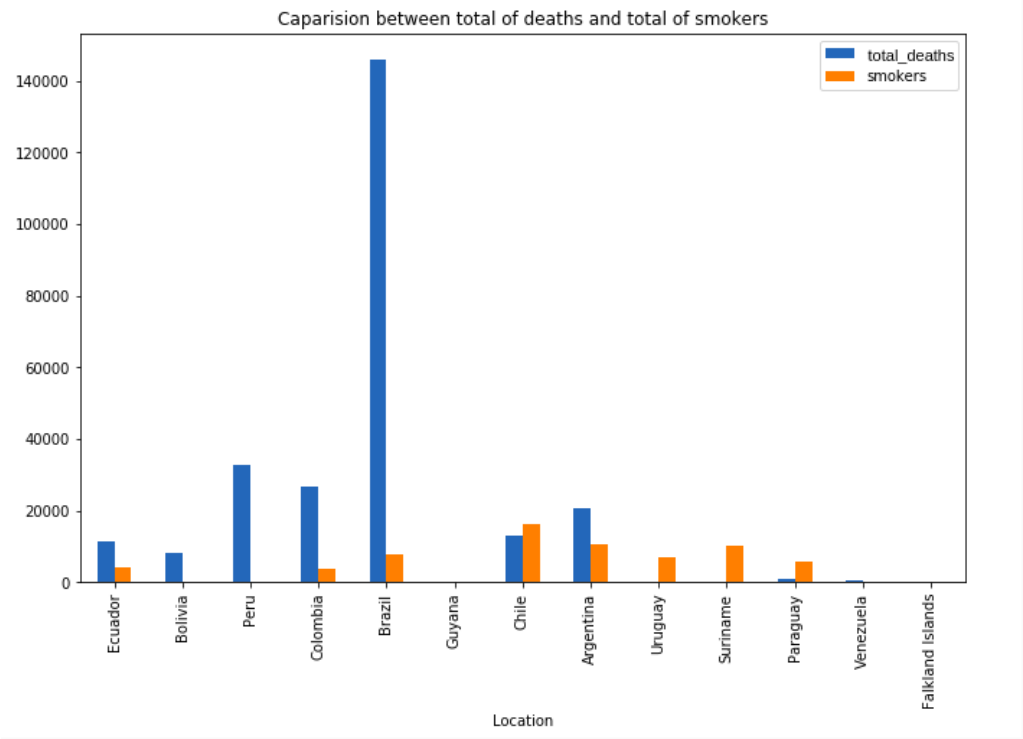

Figura 21. Relação entre o número total de mortes e o número total de fumantes positivos em cada país.

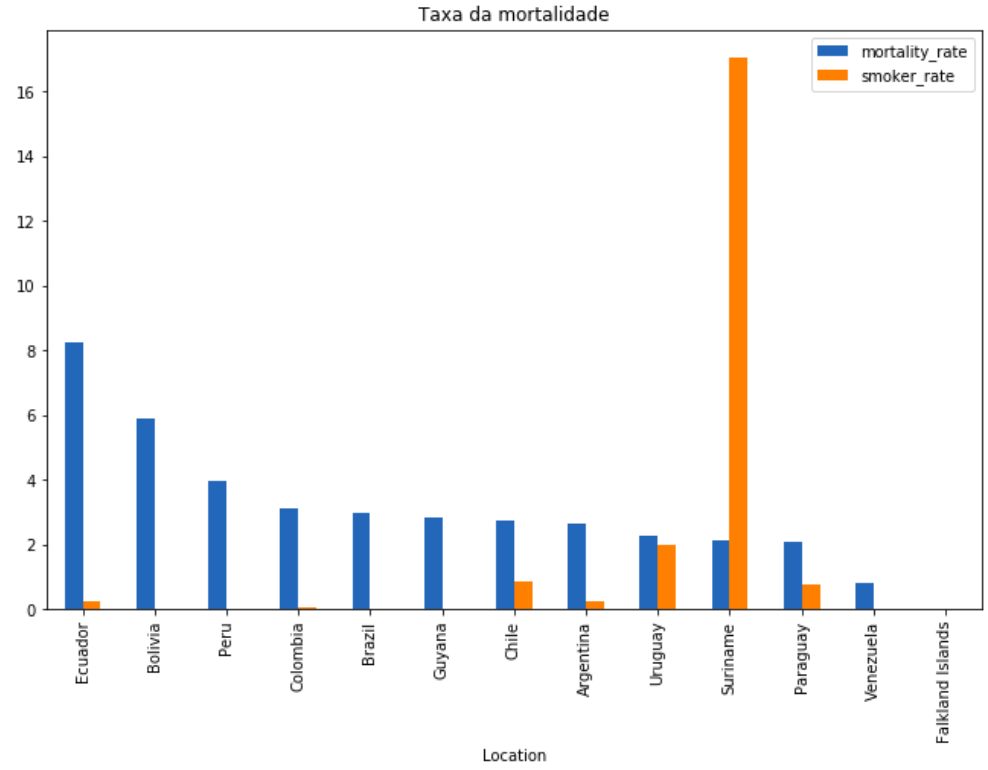

Figura 22. Relação entre a razão de mortes e a razão de fumantes na população.

Depois de abordado que os dados são desbalanceados e não tem uma boa representatividade do mundo real, nossa proposta pra esse tópico é tentar encontrar alguma análise estatística que nos mostre alguma relação entre total de mortes e número de fumantes.

4.4.2. Regressão linear

Neste seção, nós avaliamos uma análise não supervisionada utilizando, afim de balancear os dados, o conjunto de observações no mundo inteiro. Em uma primeira tentativa de realizar uma medida estatística para relacionar as razões abordadas neste tópico, tentamos prever com um regressor linear a quantidade de fumantes em uma determinada amostra como na Figura 23. Depois disso, na Figura 24, avaliamos um outro modelo, Knn, para ter métricas de comparação.

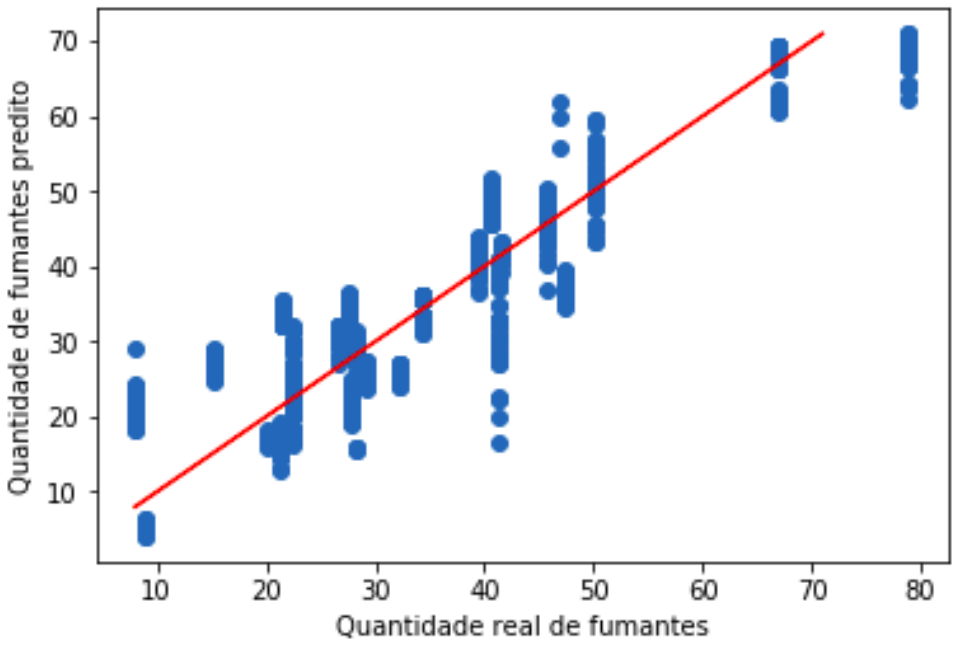

Figura 23. Predição de fumantes em um conjunto de dados com erro médio quadrado igual a 51.48.

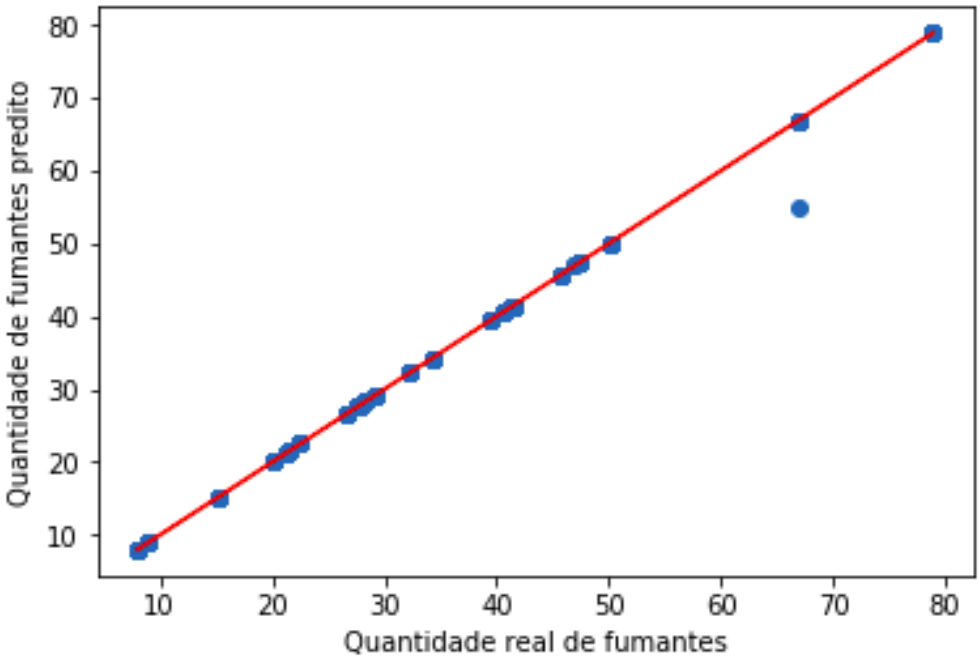

Figura 24. Predição de fumantes em um conjunto de dados com erro médio quadrado igual a 0.12.

Apesar dos dados apresentarem uma quantidade desbalanceada de informações em relação ao número de fumantes, algoritmos baseado em máximos locais próximos se saíram bem ao tentar prever a quantidade de fumantes em uma amostra selecionada. Os modelos de Regressão Linear e Knn obtiveram r2 igual a 0.82 e 1.0, respectivamente.

4.4.3. Classificação

Ainda com os dados do mundo inteiro e com a finalidade de achar uma relação entre a taxa de mortalidade e a taxa de fumantes, nessa seção selecionamos apenas as classes *smoker rate* e *mortality rate* e 4 continentes para realizar, de acordo com essas classes, rotulagem por região.

Dessa vez, o algoritmo usado para classificação foi o *Random Forest* baseado em seleção randômica de dados para fazer uma rotulação supervisionada em múltiplas casses. Com isso, na Figura 25 é possível ver o reporte dessa análise.

	precision	recall	f1-score	support
Africa	1.00	1.00	1.00	488
Asia	1.00	1.00	1.00	438
North America	1.00	0.99	1.00	168
South America	1.00	1.00	1.00	65
accuracy			1.00	1159
macro avg	1.00	1.00	1.00	1159
weighted avg	1.00	1.00	1.00	1159

Figura 25. Médias do algoritmo Random Forest.

Dado essas métricas quase perfeitas, queremos finalmente entender qual foi a razão mais relevante para essa classificação e com isso, utilizando o mesmo *Random Forest*, exploramos a importância de cada razão na Figura 26.

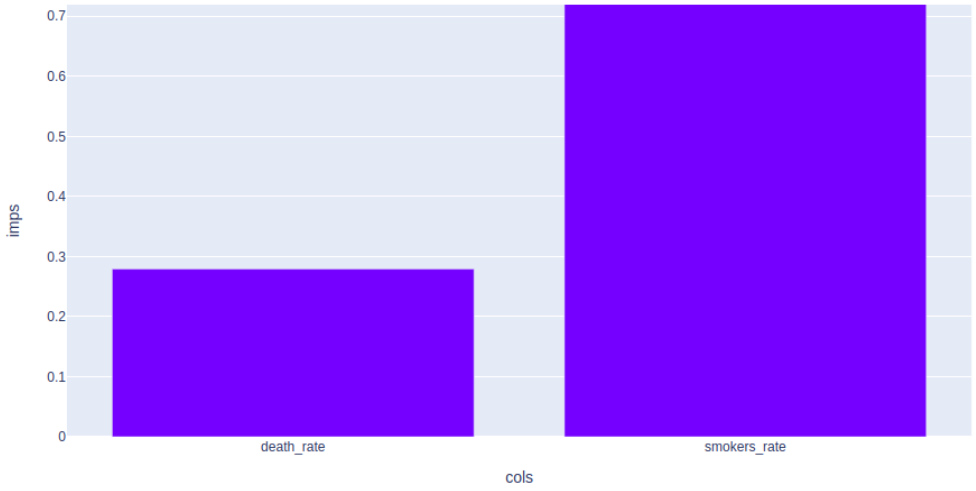

Figura 26. Feature importance com Random Forest.

4.4.4. Conclusão

Apesar dos dados se mostrarem extremamente desbalanceados, por causa da baixa padronização de informação disponível, e a pergunta inicial não ser respondida, nesta seção, conseguimos desenvolver muitos modelos estatísticos e podemos, ao longo dessa análise, aprender um pouco com os dados disponibilizados. Vimos que para classificação, a taxa de fumentes influencia o algoritmo *Random Forest* bem mais que a taxa de mortes.

Com estes fatos, podemos concluir que mesmo com os dados disponíveis a taxa de mortalidade é influenciada pela taxa de fumantes em uma população.