# Modelo não supervisionado para classificação de blocos econômicos usando Pandemia do Coronavírus (COVID-19) entre 2020 e 2022.

Rodrigo Da Costa[1]*; Renato Máximo Sátiro[2]

[1] Georgia Institute of Technology. Mestrando em Data Analytics. Atlanta, Geórgia, Estados Unidos
[2] Universidade Federal de Goiás (UFG). Doutorando em Administração. Goiânia, Goiás, Brasil
*autor correspondente: costa@gatech.edu

**Modelo não supervisionado para classificação de blocos econômicos usando Pandemia do Coronavírus (COVID-19) entre 2020 e 2022.**

## Resumo

A pandemia ocasionada pelo vírus SARS-CoV-2 é a maior emergência em saúde pública desde a eclosão da gripe espanhola no início do século XX. A crise causada pela COVID-19 desencadeou uma série de mudanças em nível mundial, trazendo consequências para os mais variados campos da experiência humana. Neste contexto, a ciência de dados é um dos campos científicos que pode trazer informações sobre algumas questões relativas ao fenômeno. Foram utilizados os dados da pandemia do Coronavírus entre 2020 e 2022 gerenciados pelo 'Global Change Data Lab' em colaboração com a universidade de Oxford. Utilizando análise de decomposição do número total de casos acumulado para todos os países e a técnica de flutuação empírica dos resíduos com limite de $1.7 \times 10^{-4}$, identificam-se três datas chaves (Dezembro de 2020, Maio de 2021 e Janeiro de 2022). Com as três datas definidas, foi aplicada a análise de Cluster K-means sendo $k = 3$ (três grupos: verde, azul e laranja). Para cada um dos grupos, realizou-se a análise de componentes principais (PCA) e apresentamos cada um dos dados em um mapa coroplético. Por fim, por intermédio das conclusões observadas, constatou-se que podemos utilizar os blocos econômicos conhecidos como o BRICS e o G7 para interpretar a classificação e identificar os países entre (a) verde são os 'países não desenvolvidos', (b) azul são os 'países emergentes' e (c) laranja que são os 'países desenvolvidos'. Tais resultados mostram que podemos utilizar os dados do COVID-19 para representar também fatores sociais, geográficos, políticos e econômico e enriquecer às discussões já bastante complexas referentes ao instável contexto vivenciado atualmente.

**Palavras-chave:** k-means; processos de flutuação empírica; PCA (Análise do componente principal).

**Unsupervised model for classifying economic blocks using Coronavirus Pandemic (COVID-19) between 2020 and 2022.**

## Abstract

The pandemic caused by the SARS-CoV-2 virus is the biggest public health emergency since the outbreak of the Spanish flu at the beginning of the 20th century. The crisis caused by COVID-19 triggered a series of changes worldwide, bringing consequences for the most varied fields of human experience. In this context, data science is one of the scientific fields that can provide information about some issues related to the phenomenon. Using data from the Coronavirus pandemic between 2020 and 2022 managed by the 'Global Change Data Lab' in collaboration with the University of Oxford. Using decomposition analysis of the accumulated total number of cases for all countries and the empirical fluctuation for residue with a $1.7 \times 10^{-4}$ threshold, we identified three key dates (December 2020, May 2021 and January 2022). With the three dates defined, we applied the Cluster K-means analysis being $k = 3$ (three groups: green, blue and orange). For each of the groups, we performed principal component analysis (PCA) and presented each of the data in a choropleth map. Finally, through the observed conclusions, we demonstrated that we can use know economic blocks such as the BRICS and the G7 to interpret the classification from the artificial intelligence and identify the countries between (a) green being 'undeveloped countries', (b) blue being the 'emerging countries' and (c) orange being the 'developed countries'. These results show that we can use COVID-19 data to also represent social, geographic, political and economic factors and enrich the already quite complex discussions regarding the unstable context currently experienced.

**Keywords:** k-means; empirical fluctuation processes; PCA (Principal component analysis)

## Introdução

A pandemia ocasionada pelo vírus SARS-CoV-2 é a maior emergência em saúde pública desde a eclosão da gripe espanhola no início do século XX. A crise causada pela COVID-19 desencadeou uma série de mudanças em nível mundial, trazendo consequências para os mais variados campos da experiência humana. Neste contexto, a ciência de dados é um dos campos científicos que pode trazer informações sobre algumas questões relativas ao fenômeno.

Um caso específico é a classificação de blocos econômicos. Esta classificação é um processo importante para entendermos a dinâmica macroeconômica dos países e a pandemia do coronavírus trouxe uma oportunidade única de compararmos os países com uma única lente (mesmo padrão) onde 63 métricas foram contabilizadas de Janeiro de 2020 à Janeiro de 2022. Entre as métricas, temos informações que vão além da simples caracterização da doença e apresentam caraterísticas Políticas, Económicas, Sociais e Tecnológicas como por exemplo “life_expectancy”, “handwashing_facilities”, “human_development_index”, “gdp_per_capita”, “aged_65_older”, “stringency_index” entre outros.

Uma das formas mais comuns de avaliação e classificação macroeconômica é a utilização da técnica PEST ou PESTEL. Esta técnica porém faz parte do grupo de análises administrativas para mapear pontos fortes e fracos e como resultado pode ser bastante subjetiva.

Com base nos dados da pandemia COVID-19, podemos utilizar uma técnica de aprendizagem não supervisionada (k-means) para classificar os países em blocos econômicos de forma independente.

## Material e Métodos

Os dados da pandemia COVID-19 serão obtidos de [1]. Com base nos dados e após um processo de limpeza, vamos utilizar a função *‘seasonal_decompose’* da biblioteca de *‘python statsmodels’* para identificar os pontos de separação para dividir os dados em blocos (antes e depois do ponto de quebra). Dentro destes blocos, podemos selecionar uma data crítica e rodar a aprendizagem não supervisionada em k-means utilizando o pacote *‘sklearn.cluster/KMeans’* em python.

Além disto, também rodaremos uma análise PCA (Análise do componente principal) para reduzir a dimensionalidade dos dados e apresentar os resultados da classificação em duas dimensões.

## Coleta e Limpeza dos Dados

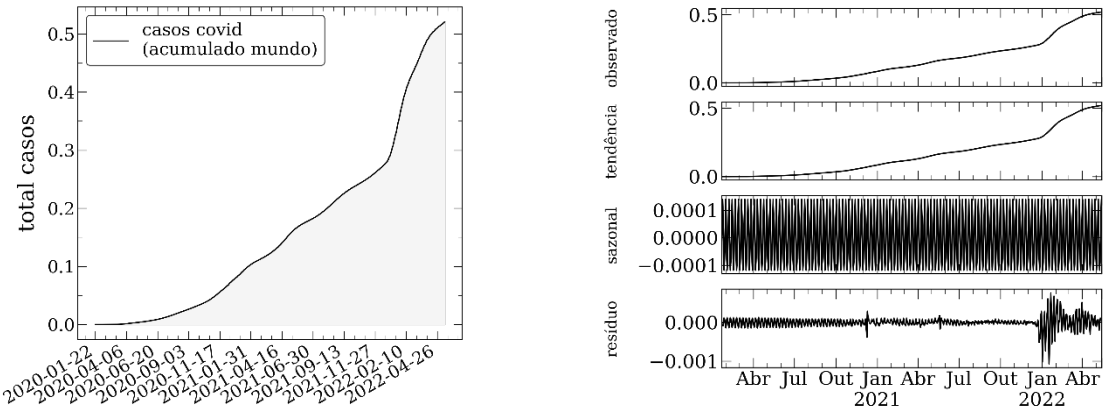
Os dados que utilizamos são curados pela equipe do '*Global Change Data Lab*' em colaboração com a universidade de Oxford. Todos os dados produzidos pela equipe são curados e mantidos pela instituição e são disponibilizados sob a licença CC '*creative commons*' onde qualquer pessoa pode utilizar os dados bastante atribuir o crédito para a instituição. O corpus atual dos dados contém 67 colunas onde 63 destas foram utilizadas para a nossa análise. Além disto, os dados apresentam 222 países com informação desde Janeiro de 2020 e com atualizações diárias. No nosso caso, utilizaremos os dados até Janeiro de 2022.

Para um bom resultado na execução da nossa análise, decidimos remover todas as colunas de informação categórica (texto) mantendo apenas um identificador do país. Para os dados que apresentam valores nulos (NaN), a decisão foi transformar estes valores em zero ao invés de remover a linha. Todos os valores também foram normalizados utilizando a biblioteca de '*sklearn*' do python (MinMaxScaler).

## Processos de Flutuação Empírica

O primeiro passo da análise foi selecionar datas chaves para a análise. Para isto, utilizamos uma análise de resíduo com decomposição de tendência, sazonalidade e resíduo.

***Figura 1.*** *Casos de COVID (número total) acumulado para todos os países e análise de decomposição*




*(a)* *(b)*

Com a avaliação do resíduo, traçamos uma linha de limite para identificar quais datas apresentaram maio variação. Utilizando o limite de $1.7 \times 10^{-4}$, conseguimos identificar três datas críticas:

| Data | Resíduo |
|---|---|
| 2020-12-11 | $2.53 \times 10^{-4}$ |
| 2021-05-19 | $1.81 \times 10^{-4}$ |
| 2022-01-21 | $7.61 \times 10^{-4}$ |

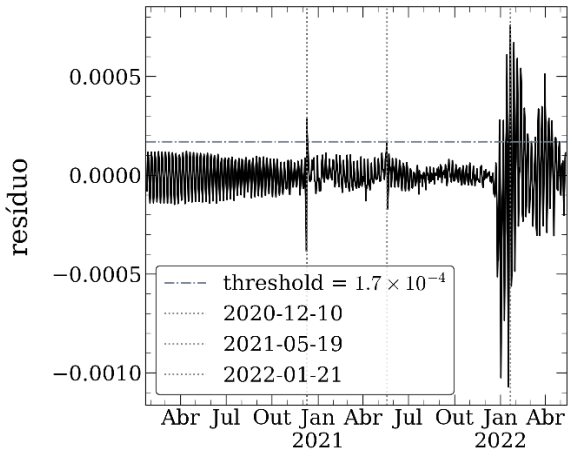
***Tabela 1.*** *Identificação da data por resíduo (flutuação) e limite.*

Foi executada a decomposição da série temporal em quatro fatores principais utilizando o modelo aditivo onde:

$$y_t = O_t + T_t + S_t + R_t \quad (1)$$

Onde: $O_t$ é o valor médio observado; $T_t$ é o componente da tendência; $S_t$ é o componente da sazonalidade e $R_t$ é o resíduo do componente aleatório da série temporal. Com isto, foram definidas as datas foram também validadas visualmente através do gráfico de resíduos onde vemos claramente que para as datas marcados temos fortes evidências de mudanças de comportamento nos dados:

***Figura 2.*** *Análise de resíduo mostrando os limites de datas*




Na Figura 2 podemos observar que o período a partir de Dezembro de 2021 é uma região de alto resíduo. Por este móvito, a decisão foi utilizar uma data próxima ao centro de massa do cluster de resíduos no mês de Janeiro e Fevereiro de 2022 (21 de Janeiro de 2022 para o este estudo).

PCA (Análise do componente principal)

Após a coleta de dados, também precisamos utilizar a análise de componente principal para reduzir os 63 componentes iniciais para apenas 2. Se levarmos em consideração um

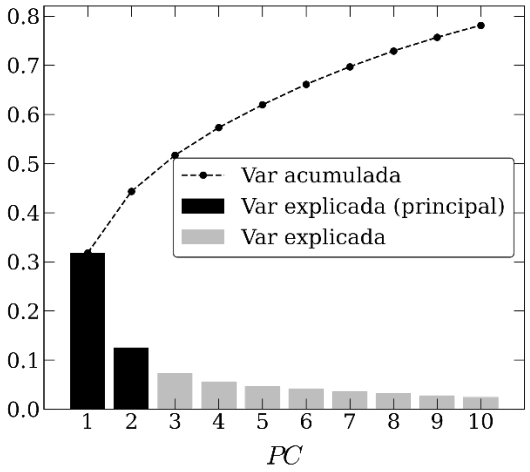
vetor com $m$ pontos de tal forma que $\{x^1, x^2, x^3, \dots x^m\}$. O PCA pode ser estimado como os autovetores da matrix de covariância de média central $\boldsymbol{X} = x - \mu$.

$$C = \frac{1}{m}\sum_{i=1}^{m}(x^i - \mu)(x^i - \mu)^T \quad (2)$$

Onde $\mu = \frac{1}{m}\sum_{i=1}^{m} x^i$. Cada um dos autovetores $\boldsymbol{w}$ da matrix de covariância corresponde a um autovalor $\lambda_n$ de forma que:

$$\boldsymbol{w^T C w} = \boldsymbol{w^T}\,\lambda\,\boldsymbol{w} = \lambda \quad (3)$$

***Figura 3.*** *Análise de PCA apresentando o valor da variância acumulada para os 10 primeiros componentes.*

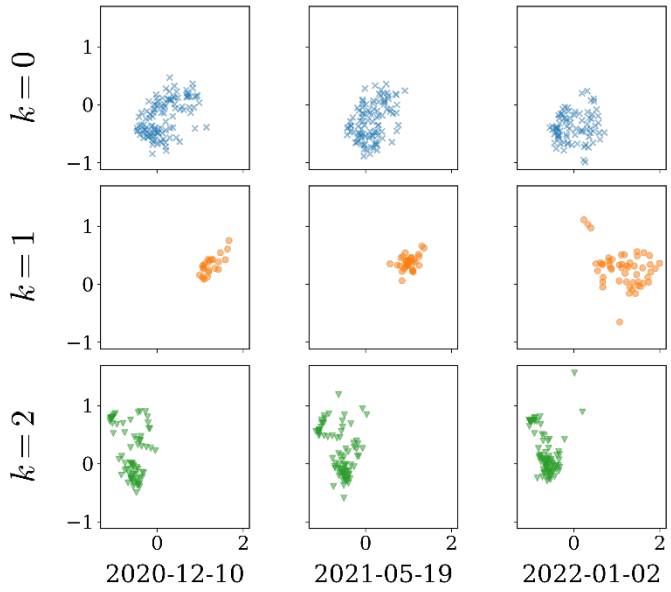



Como mostra a figura 3. Se considerarmos os dois primeiros componentes, vemos que a variância explicada é da ordem de 45%. ($PC_1 \approx 31\%\ e\ PC_2 \approx 14\%$). A análise de componente principal é importante para conseguirmos visualizar os dados em duas dimensões.

Análise de Cluster K-means

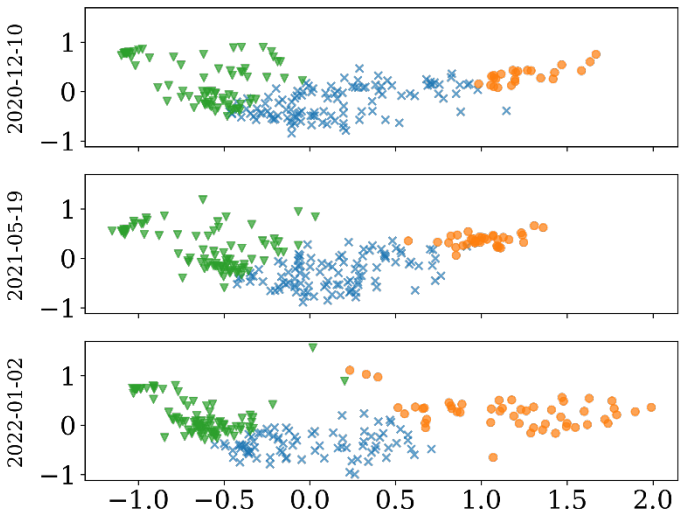
Com a definição das três datas que vão ser processadas, como modelo de classificação utilizaremos o modelo de agrupamento de k-means com valor do cluster sendo $k = 3$ (três grupos). O modelo faz parte do grupo de modelos com aprendizado de máquina não supervisionado e foi escolhido uma vez que não queremos criar bias humano na classificação.

***Figura 4.*** *Análise de K-means com k=3 (três grupos). Utilizando os dois componentes principais do PCA para visualização em 2D.*




Como mostra a figura 4 – a análise de PCA é importante para visualizarmos que todas as cores (agrupamentos do modelo de k-means) estão equivalentes.

***Figura 5.*** *Resultado do modelo preditivo treinado para $k = 3$ (três grupos).*



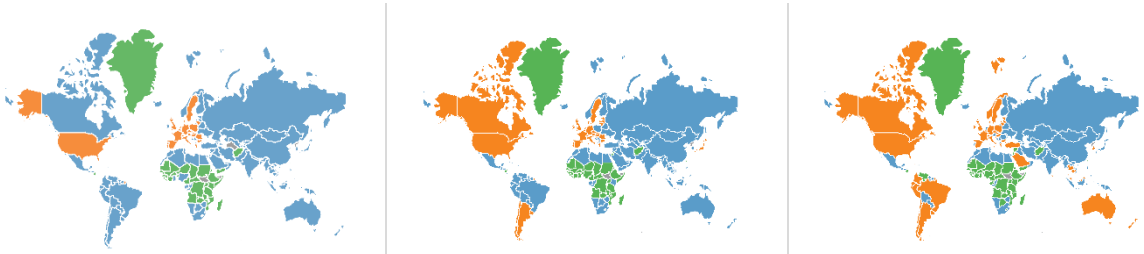

## Resultados e Discussão

Para visualizarmos o modelo de predição, apresentamos os resultados da análise do componente principal (PCA) que utiliza passo de decomposição de valor singular (SVD). Com

isto, conseguimos reduzir os dados de 63 componentes para apenas duas componentes (duas dimensões do gráfico, eixo x e y) onde apresentamos na figura 5 a classificação dos dados em três grupos (verde, azul e laranja) onde cada ponto no gráfico de resultados é um país e sua cor é definida com base no modelo de k-means.

Com base nos resultados do modelo de classificação k-means, utilizamos a biblioteca gráfica D3 para gerar o mapa coroplético onde tomamos o cuidado de manter as mesmas cores apresentadas na análise de PCA do resultado do modelo preditivo de classificação.

***Figura 6.*** *Evolução dos grupos macroeconômicos da esquerda para a direita partindo de Dezembro de 2020, Maio de 2021 e finalmente Janeiro de 2022.*

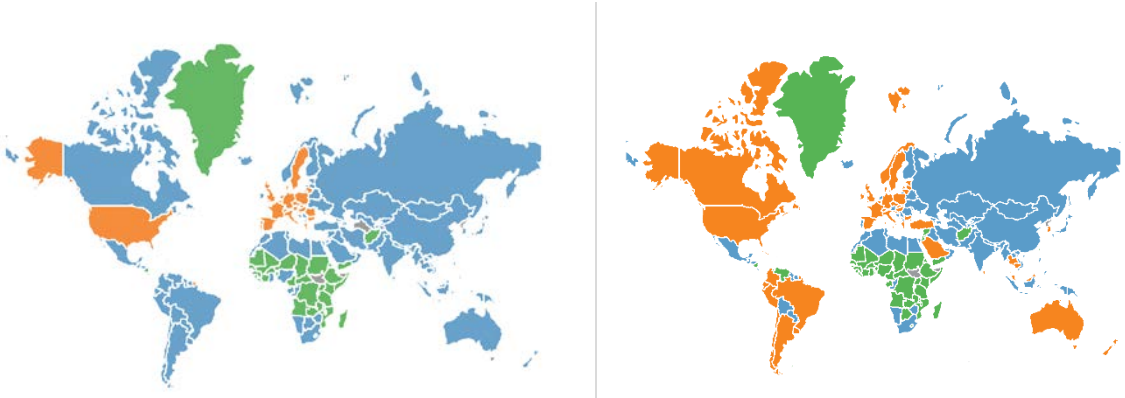


*(a)* *(b)* *(c)*

As três datas que escolhemos baseados na nossa análise de flutuação empírica dos resíduos são importantes para entender a evolução da pandemia. A primeira data (Dezembro de 2020) foi quando ainda não tínhamos nenhuma vacina e a pandemia atingia ponto crítico – esta data foi quando tivemos os Estados Unidos autorizando de forma emergencial o início da vacinação.

A segunda data (Maio de 2021) já apresenta um momento diference onde vários países já tinham administrado a segunda dose da vacina. Por fim, a terceira data (Janeiro de 2022) traz uma foto mais atual da classificação dos países.

***Figura 7.*** *Comparativo lado a lado com o começo da análise (Dezembro de 2020) com o final da análise (Janeiro de 2022)*



*(a)* *(b)*

A figura 6 (a, b e c) apresenta a classificação dos dados nos três grupos (verde, azul e laranja) onde cada um dos pontos da análise de k-means é um país e sua cor definida pelo algoritmo de inteligência artificial não supervisionada.

Observando os dados da figura 6.a. (Dezembro de 2020), notamos uma possível correlação entre o grupo da cor azul e os países dos Mercados Emergentes. Pela literatura, sabemos que um dos melhores exemplos dos países Emergentes é bloco conhecido como BRICS formado pelo acrônimo de Brasil, Rússia, Índia, China e África do Sul. Estes cinco países juntos representam a organização internacional independente conhecida como BRICS que encoraja a aproximação comercial, política e cultural na cooperação entre as noções e o termo foi originalmente cunhado em 2001 para representar o fato de que estas nações apresentam um rápido potencial de crescimento comparado com outras nações.

Um outro bloco econômico bastante conhecido é o bloco chamado de G7. Este bloco é formado pelas nações: Canada, França, Alemanha, Itália, Japão, Reino Unido e Estados Unidos da América e representam as maiores economias, IMF '*International Monetary Fund*', do mundo. O bloco do G7 são os países mais ricos que apresentam democracias liberais e são geralmente o modelo de Países Desenvolvidos.

Utilizando estes dois blocos econômicos (BRICS e G7) como Mercados Emergentes e Mercados Desenvolvidos, podemos então associar o modelo de classificação aqui apresentado da seguinte forma:

| Data | Cor | k | Interpretação (por dominância) |
|---|---|---|---|
| 2020-12-11 | Azul | 0 | Emergentes |
| | Laranja | 1 | Desenvolvidos |
| | Verde | 2 | Subdesenvolvidos |
| 2021-05-19 | Azul | 0 | Emergentes |
| | Laranja | 1 | Desenvolvidos |
| | Verde | 2 | Subdesenvolvidos |
| 2022-01-21 | Azul | 0 | Emergentes |
| | Laranja | 1 | Desenvolvidos |
| | Verde | 2 | Subdesenvolvidos |

***Tabela 2.*** *Interpretação do modelo K-means utilizando BRICS e G7 como referência para identificar Mercados Emergentes, Mercados Desenvolvidos e Mercados subdesenvolvidos.*

A análise feita aqui só considera três datas (três pontos na linha do tempo) e embora suficiente como comparação não representa o período crítico. Uma melhor análise aqui poderia considerar ao invés de uma única data um período de tempo e aplicar a técnica do voto majoritário das classificações do período. Infelizmente, uma análise deste tipo exige um grande esforço já que uma das características do modelo de k-means é a alocações de grupos aleatoriamente. Como a classificação é aleatória, processando a mesma data, teríamos os mesmos eixos do PCA mas as cores (valores de k) podem ser diferentes em cada regressão.

## Conclusão

Conseguimos classificar os países em clusters utilizando o modelo de k-means com $k = 3$ (três grupos) para três datas diference: Dezembro de 2020, Maio de 2021 e Janeiro de 2022. Para Dezembro de 2020 notamos que o modelo de classificação corretamente classificou todos os países emergentes como $k = 0$ (azul) e classificou 6 dos 7 países do G7 como fazendo parte dos países desenvolvidos como $k = 1$ (laranja). O único país classificado como algo não esperado do G7 foi Japão que apresentou características similares aos países Emergentes para Dezembro de 2020. O caso do Japão é interessante uma vez o mesmo fenômeno que identificamos em Dezembro de 2020 (figura 7.a) se repetiu em Janeiro de 2022 (figura 7.b) onde o Japão novamente foi o único país do G7 classificado como um mercado emergente ao invés de Desenvolvido.

Este resultado é interessante e exige uma maior investigação uma vez que o Japão é historicamente visto como um país desenvolvido, mas este trabalho apresenta evidências que a influência dos países da Ásia-Pacífico (que em geral são emergentes) pode ser significante no Japão.

Um outro caso interessante é o do Brasil. O Brasil é o único país do BRICS que apresenta uma mudança na classificação de Mercados Emergentes, $k = 0$ (azul), para Mercados Desenvolvidos com $k = 1$ (laranja). Esta mudança aconteceu apenas agora em Janeiro de 2022 e pode ser interpretado como um cenário positivo de melhoria dos casos de pandemia do país. O caso do Brasil é outro interessante e que exige um maior estudo uma vez que a sua classificação é bastante sensível neste modelo e dependendo das datas escolhidas pode ser classificado até como um país subdesenvolvido dependendo. Esta sensibilidade indica que o Brasil está bem perto de uma linha de transição de fase do modelo e um trabalho mais detalhado com aplicação de um algoritmo de '*Random Forest*' pode nos ajudar a identificar quais são os fatores principais dos 63 mapeados que contribuem para a sua classificação (ou sensibilidade). É importante ressaltar que o modelo k-means é um modelo de aprendizado de máquina não supervisionado (onde os grupos são formados de forma automática por suas proximidades nos dados) e os resultados são bastante sensíveis em geral.

Por fim, concluímos que a utilização de inteligência artificial para a classificação de blocos econômicos é hoje uma realidade e muitas outras fontes de dados podem ser utilizadas aplicando as mesmas técnicas aqui descritas. Isto traz uma perspectiva de melhoria significativa ao processo atual com mais escala onde a classificação pode ser atualizada até mesmo em tempo real algo que seria inimaginável no passado.

## Referências

ALVES, H. J. d. P., et al. **A pandemia da COVID-19 no Brasil: Uma aplicação do método de clusterização k-means**. Research, Society and Development, 9(10), 2020.

ARTHUR, D., VASSILVITSKII, S.: **k-means++: The advantages of careful seeding**. Tech. rep., Stanford, 2006

BOSTOCK, Mike. D3.js - **Data-Driven Documents**. Disponível em: 'http://d3js.org/'. Acesso em: 11 de Maio 2022.

BROWN R.L., DURBIN J., EVANS J.M., **Techniques for testing constancy of regression relationships over time**, Journal of the Royal Statistal Society, B, 37, p. 149-163, 1975.

RITCHIE, Hannah et al. **Coronavirus Pandemic (COVID-19)**. Disponível em: https://ourworldindata.org/coronavirus. Acesso em: 11 de Maio 2022.

COSTA, Rodrigo. **MBA-USP-TCC**. Disponível em: 'https://github.com/costargc/MBA-USP-TCC/'. Acesso em: 11 de Maio 2022.

KUAN C.-M., Chen, **Implementing the fluctuation and moving estimates tests in dynamic econometric models**, Economics Letters, n.44, p. 235-239, 1994.

O'SULLIVAN, A., & SHEFFRIN, S. M. **Economics: principles in action**. Needham, Mass, Prentice Hall, 2003.

ZUBAIR M., *et al* **An Efficient K-Means Clustering Algorithm for Analysing COVID-19**. Hybrid Intelligent Systems. vol 1375. Springer, Cham, 2021.

ZARIKAS, V., et al. **Clustering analysis of countries using the covid-19 cases dataset**. Data in Brief, 31, 1-8. 2020.