Quaestum

DOI: https://doi.org/10.22167/2675-441X-2024741
Quaestum 2024; 5:e2675741
ISSN: 2675-441X

Artigo

Caracterização do perfil global de emissões de gases de efeito estufa utilizando "machine learning"

Characterizing the global greenhouse gases emissions using machine learning

Luis Felipe Alves Frutuoso[1]*; William Barbosa[2]

Recebido: mai. 10, 2023
Aceito: jan. 17, 2024

[1]Doutor em Engenharia Química. Rua Marques de Herval, 90, Sede UO-BS, Valongo, Santos, São Paulo, 11010-310, Brasil

[2]Doutor em Economia Aplicada. Rua Correia de Lemos, 780, Chácara Inglesa, São Paulo, São Paulo, 04140-000, Brasil

*Autor correspondente: felipelfaf@gmail.com

Resumo: Considerando o contexto atual, em que se busca um crescimento econômico sustentável com ênfase em políticas e incentivos associados a questões ambientais, este estudo investigou o grau de importância relativo de determinantes socioeconômicos no entendimento do perfil de emissões de gases de efeito estufa a partir de uma abordagem de "machine learning". Foi estimado um modelo do tipo "random forest" a partir de dados sobre a capacidade produtiva econômica e a quantidade de emissões de gases de efeito estufa no período entre 1990 e 2018. A amostra estudada consistiu em países que representavam as maiores e menores economias globais, selecionados a partir do seu nível de atividade econômica no período. Inicialmente, identificaram-se as variáveis mais relevantes a partir da técnica de eliminação recursiva de variáveis; em seguida, o modelo foi treinado empregando a técnica de "cross validation"; e, por fim, foi validado com os dados selecionados para teste. As métricas de desempenho não indicaram problemas de "overfitting", e os resíduos das estimativas se comportaram de acordo com a distribuição normal. A partir do modelo estimado neste trabalho, observou-se que o perfil de emissões de gases de efeito estufa foi influenciado de maneira distinta dependendo do país analisado, de forma que os fatores mais ou menos relevantes indicaram estar associados com o nível de atividade econômica. Assim, as discussões e a modelagem apresentadas no presente trabalho se propuseram a incentivar políticas de incentivo e medidas de controle direcionadas aos setores mais relevantes, que pudessem contribuir para um crescimento econômico sustentável.

Palavras-chave: capacidade produtiva econômica; crescimento econômico sustentável; emissões; "random forest".

Abstract: Considering the current context, in which sustainable economic growth is sought with an emphasis on policies and incentives associated with environmental issues, this study investigated the degree of relative importance of global socioeconomic factors in understanding the profile of greenhouse gas emission based on a machine learning approach. A random forest algorithm was estimated from data on economic productive capacity and the amount of greenhouse gas emissions in the period between 1990 and 2018. The database was composed of countries that represented the major and minor global economies, selected by their economic activity level in the period. Initially, the most important variables of the study were identified using the recursive variable elimination technique; then, the model was trained using cross validation technique; and, finally, it was validated with the data selected for testing. The performance metrics did not indicate overfitting issues, and the residuals of the estimates behaved in accordance with the normal distribution. From the model estimated in this study, it was observed that the global greenhouse gas emission profile was influenced differently depending on the country analyzed, so that the more or less relevant factors indicated to be associated with the level of economic activity. Thus, the discussions and modeling presented in this study aimed to encourage incentive policies and control measures aimed at the most relevant sectors, which could contribute to sustainable economic growth.

Keywords: economic growth; sustainable growth; emissions; random forest.

1. Introdução

O aumento da capacidade de produção promovido pela industrialização e as novas tecnologias têm impulsionado o crescimento econômico mundial e causado diversas mudanças na sociedade, gerando empregos e melhorando o padrão de vida das pessoas. Entretanto, essa modernização tem gerado consequências negativas ao meio ambiente, principalmente pelo aumento da emissão de gases e geração de efluentes[1].

A atividade humana interfere na mudança climática de várias formas. O uso inadequado da terra (desmatamento), o aumento da atividade agropecuária e industrial e a queima de combustíveis fósseis podem ser citados como exemplos. Os efeitos negativos dessa interferência são perceptíveis. Os relatos de catástrofes climáticas como inundações, derretimento de geleiras, aumento do nível do mar, secas e ondas de calor têm se tornado frequentes[2].

Um dos maiores problemas associados à intensificação dessas atividades é o aumento da emissão de gases de efeito estufa (GEE), que recebem essa denominação por sua característica de absorver a radiação terrestre. O vapor d'água, o dióxido de carbono (CO_2), o metano (CH_4) e o óxido nitroso (N_2O) são as principais substâncias que afetam a taxa de aquecimento do planeta[3]. Os diversos setores da atividade humana contribuem para o aumento da emissão de GEE de formas distintas. O setor de agropecuária, por exemplo, é responsável pela emissão de uma grande quantidade de metano pelo rebanho, enquanto a queima de combustíveis fósseis resulta na emissão de CO_2 e vapor d'água diretamente na atmosfera[4].

Para que se possa quantificar as emissões de um país, é preciso considerar o volume bruto de emissões e o tipo de gás que é lançado na atmosfera, pois os gases responsáveis pelo efeito estufa apresentam potencial de aquecimento distinto. Em linhas gerais, o potencial de aquecimento representa uma medida de comparação entre a habilidade de absorção de calor de um gás em relação ao CO_2, adotado como padrão de referência[5]. O gás metano e o óxido nitroso, por exemplo, apresentam potencial de aquecimento superior ao do CO_2 em cerca de 21 e 310 vezes, respectivamente[6]. A utilização do potencial de aquecimento em conjunto com o volume bruto de emissões resulta na métrica conhecida como equivalência em dióxido de carbono (CO_2eq).

Com base nessas informações, nota-se que o perfil de emissões de uma nação pode ser influenciado por diversos fatores, os quais muitas vezes não se manifestam de forma direta. Assim, é possível formular diversas hipóteses para a compreensão do perfil de emissões global. A título de exemplo, pode-se esperar que nações de intensa atividade econômica apresentem as maiores taxas de emissão de GEE, porém o estilo de vida da população destes países tem potencial para estabelecer tendências não esperadas.

O objetivo deste estudo foi estimar um modelo a partir do "random forest", utilizando dados sobre a capacidade produtiva econômica e a quantidade de emissões de gases de efeito estufa no período entre 1990 e 2018, a fim de identificar o grau de importância relativo de determinantes socioeconômicos como variáveis preditoras para auxiliar o entendimento do perfil de emissões de GEE das nações. Esperou-se que os resultados obtidos nesta pesquisa pudessem subsidiar ações e fomentar políticas públicas mais eficientes em prol do desenvolvimento sustentável.

2. Material e métodos

Dados

As análises apresentadas neste trabalho foram baseadas em informações coletadas a partir de duas bases de dados públicas e independentes, a Penn World Table (PWT)[7] e a Climate Analysis Indicators Tool (CAIT)[8].

A PWT consiste em uma base de dados contendo informações acerca da atividade econômica mundial. A gestão dos dados é feita pela Universidade da Califórnia em Davis, nos Estados Unidos, e pela Universidade de Groningen, nos Países Baixos. Há cerca de quatro décadas a PWT tem sido adotada como referência para obtenção de dados econômicos mundiais, permitindo realizar comparações da capacidade produtiva e até do estilo de vida da população entre as nações[7].

Existem diversas versões da PWT disponíveis para consulta, e a principal diferença entre elas é o período de coleta dos dados, os quais são atualizados constantemente. Contudo, há períodos em que a lista de países e a metodologia de cálculo dos indicadores são ligeiramente distintas. As alterações praticadas no versionamento buscam que os dados possibilitem a comparação do desempenho econômico entre os países ao longo do tempo[9]. As discussões deste trabalho se fundamentaram na versão 10.0 da PWT, atualizada em 18 de junho de 2021, a qual contém o histórico de dados de 183 países no período compreendido entre os anos de 1950 e 2019.

A principal variável da tabela é o Produto Interno Bruto (PIB), indicador bastante utilizado em macroeconomia, que representa a soma de todos os bens e serviços finais produzidos por um país, estado ou cidade[10]. Os países

geralmente calculam o PIB nas suas respectivas moedas, e para que seja possível compará-los na mesma base é preciso considerar a métrica da paridade do poder de compra (PPC).

A PWT apresenta o PIB real dos países de duas formas: sob a ótica do dispêndio ("expenditure") e sob a ótica da produção ("output"). Feenstra et al.[7] afirmaram que o PIB calculado sob a ótica do dispêndio permite uma comparação do padrão de vida entre os países, enquanto o PIB sob a ótica da produção transmite uma noção da produtividade econômica. Além disso, a tabela contém dados baseados em padrões de referência adotados pelo Programa de Comparação Internacional (ICP) e informações fornecidas pela contabilidade nacional dos países ("national accounts").

Além do valor final do PIB, a PWT contém informações detalhadas das parcelas que compõem este indicador. É possível consultar a parcela relativa ao consumo das famílias e gastos dos governos, assim como a parcela relacionada com investimentos produtivos. Adicionalmente, há dados relacionados com a produtividade nacional, como o expresso pela variável "total factor productivity" (TFP). Esse conjunto de informações permite extrair ideias e formular hipóteses mais aprofundadas sobre a atividade dos países.

Com relação à emissão de GEE, as informações consideradas neste estudo foram obtidas a partir da Climate Watch, a qual consiste num repositório de dados públicos que conta com um conjunto de ferramentas para visualização e análise de informações climáticas[8]. Os dados avaliados tiveram referência na CAIT, ferramenta desenvolvida e atualizada pelo World Resources Institute (WRI). Essa base de dados considera as emissões dos principais gases de efeito estufa reconhecidos pelo protocolo de Kyoto, principalmente metano, dióxido de carbono, óxido nitroso e os gases fluorados. Nela são contabilizadas as emissões de diversos setores econômicos, tais como agropecuária, transporte e eletricidade, processos industriais, resíduos (aterros), mudança do uso da terra, desmatamento etc. O inventário de emissões extraído contemplou dados entre 1990 e 2018, sendo expresso em termos de equivalência de CO_2 (CO_2eq), o que permitiu incorporar a contribuição de setores econômicos distintos. Dessa forma, a análise do perfil de emissões global se deu com base nos dados agregados da PWT e da CAIT.

Tratamento e análise estatística

Com base no conjunto de dados agregados da PWT e da CAIT, realizou-se uma análise descritiva utilizando medidas de tendência central e de dispersão para organizar os dados e selecionar uma amostra para aplicação dos algoritmos de "machine learning". Inicialmente determinou-se o PIB per capita e calculou-se sua média para o intervalo dos dados agregados. Os países foram organizados em ordem decrescente e, em seguida, foram selecionadas as dez maiores e as dez menores economias para compor a amostra de estudo, constituindo-se, portanto, uma amostra com 20 observações.

O conjunto original de dados continha cerca de 52 variáveis com informações sobre os indicadores econômicos e sobre as características e origem dos dados. As variáveis "i_cig" e "i_xm", por exemplo, indicaram se os dados sobre o consumo das famílias e balança comercial de determinado país, respectivamente, foram extrapolados ou interpolados. Dessa forma, entendeu-se que era possível selecionar apenas uma parte destas variáveis para a modelagem do problema. A identificação e a descrição das variáveis utilizadas neste estudo foram do tipo numérico, permitindo a aplicação das técnicas e métodos da estatística descritiva (Tabela 1).

Outro cuidado importante adotado no tratamento dos dados foi a verificação da quantidade de dados não disponíveis (NA) dentro do conjunto selecionado. Verificou-se que todos os dados para os países que compunham a amostra final estavam disponíveis, não tendo sido necessário, portanto, estimar valores desconhecidos usando média ou mediana, por exemplo.

Tabela 1. Variáveis selecionadas para modelagem do perfil de emissões de gases de efeito estufa (GEE)

Nome da variável	Tipo	Descrição
pop	Explicativa	População (milhões)
emp	Explicativa	Número de pessoas trabalhando
h_idex	Explicativa	Índice de capital humano
consumo	Explicativa	Consumo das famílias e gastos do governo, considerando a paridade do poder de compra
absorcao	Explicativa	Consumo das famílias e gastos do governo acrescidos da parcela de investimento produtivo considerando a paridade do poder de compra
PIB_percapita	Explicativa	PIB pela ótica do dispêndio, considerando a paridade do poder de compra
MtCO2eq	Resposta	Emissões de GEE expressas em milhões de toneladas equivalentes de CO_2

Fonte: Groningen Growth and Development Centre[11]
Nota: A variável MtCO2eq foi obtida a partir dos dados disponibilizados pelo Climate Watch[8]; PIB: Produto Interno Bruto; GEE: gases do efeito estufa

Modelagem “machine learning”: “random forest”

Os algoritmos de “machine learning” podem ser empregados na modelagem de diversos tipos de problemas, de forma que não há um algoritmo universal que funcione bem para todos os casos. É preciso que o usuário reconheça o algoritmo mais apropriado para o problema sendo investigado, com base na natureza dos dados e no tipo de problema em questão.

Dentre as possíveis aplicações, destacam-se os problemas de regressão e classificação de dados. A regressão é um método no qual o algoritmo é usado para estimar um valor baseado em uma série de variáveis. Problemas de classificação, por sua vez, buscam identificar a classe a que determinada observação pertence com base em seus atributos, isto é, as variáveis que compõem a base de dados[12].

Algumas bases de dados exibem peculiaridades que dificultam a parametrização dos algoritmos. Nesse sentido, verifica-se amplo destaque para os métodos de “ensemble”, cujas estimativas são obtidas por meio da combinação das previsões de vários modelos, buscando um melhor desempenho. Um dos algoritmos dessa classe de métodos que se destaca é o conhecido por “random forest”, que combina as estimativas individuais de uma grande quantidade de modelos baseados em árvores de decisão, as quais são parametrizadas a partir de diferentes conjuntos de dados formados por amostras aleatórias dos dados originais, procedimento conhecido como “bagging”. Além disso, o usuário deve ajustar os hiperparâmetros do algoritmo para que as estimativas sejam mais precisas[13],[14].

Foi aplicado o algoritmo “random forest” nos dados agregados da PWT e da CAIT entre 1990 e 2018 utilizando, na modelagem, as variáveis apresentadas na Tabela 1, sendo a emissão de GEE — representada pela variável MtCO2eq — a variável resposta deste estudo. Por se tratar de uma variável do tipo contínua, as métricas de desempenho adotadas neste trabalho foram a raiz do erro quadrático médio (RMSE), o erro médio absoluto (MAE) e o coeficiente de determinação (R^2). Os principais hiperparâmetros avaliados neste estudo foram o número de árvores construídas antes de uma predição — ntree —, o número de variáveis aleatoriamente selecionadas para construção de cada árvore — mtry — e, por fim, o número mínimo de folhas em cada árvore — nodesize[15],[16]. A base de dados foi dividida em base de treino e teste antes da aplicação do algoritmo “random forest” nos dados. Usou-se a base de treino para construir o modelo, e a base de testes foi utilizada para avaliá-lo. O ajuste dos hiperparâmetros e o desempenho do modelo foram avaliados com a técnica de “cross validation”.

3. Resultados e discussão

Análise estatística

A amostra final com os 20 países selecionados foi organizada em ordem crescente segundo o PIB per capita médio (US$) entre 1990 e 2018 (Tabela 2).

Verificou-se que os dez países de menor PIB per capita se concentravam no continente africano, enquanto os dez países de maior PIB per capita se distribuíam pela Europa, Ásia e América (devido à presença dos Estados Unidos). O PIB per capita médio dos dez países de menor desenvolvimento econômico valia US$ 1.001,21; o dos dez países de maior economia foi igual a US$ 62.855,45, aproximadamente 63 vezes superior. De acordo com Coelho et al.[17], o baixo desempenho econômico das nações africanas se deve à dependência da comercialização de produtos básicos não industrializados (“commodities”) e aos problemas históricos de baixa governabilidade, como a falta de transparência e de austeridade, por exemplo.

Percebeu-se que algumas variáveis da amostra apresentavam natureza e intervalos diferentes. Por exemplo, os valores mínimos e máximos observados para o índice de capital humano (variável h_idex) foram de 1,06 e 4,15, nesta ordem, enquanto o intervalo de variação do PIB per capita foi de US$ 244 a US$ 166.520. Desta forma, com base na segmentação dos dados em dois “clusters”, separando as maiores e menores economias entre 1990 e 2018, realizou-se o reescalonamento das variáveis por meio da técnica de normalização Z-score. Esse procedimento foi importante para evitar que a escala das variáveis influenciasse a parametrização do modelo. Antes de prosseguir com o tratamento, verificou-se ainda que não havia dados desconhecidos (NA) no conjunto de variáveis selecionadas neste estudo, não sendo necessário, portanto, estimar valores desconhecidos em nenhuma das variáveis.

Apresentam-se, nas Figuras 1 e 2, respectivamente, os gráficos “boxplot” do PIB per capita (US$) do grupo com as dez menores e com as dez maiores economias. Em linhas gerais, os países apresentam uma distribuição do PIB per capita bastante similar, isto é, sem grandes diferenças entre as observações de cada grupo. Entretanto, algumas observações merecem destaque. Ruanda, por exemplo, foi o país com maior amplitude do PIB per capita no período. Esse comportamento pode ser atribuído principalmente pelas recentes mudanças políticas que o país atravessou, superando uma era marcada pelo genocídio e atraindo cada vez mais investimento estrangeiro[18].

Com relação ao grupo das dez maiores economias, a amplitude do PIB per capita do Catar e a presença de

Luxemburgo merecem destaque. O Catar era um país extremamente pobre, com sua economia baseada na pesca; contudo, a recente descoberta de grandes reservas de óleo e gás trouxeram uma nova perspectiva, enriquecendo o país e posicionando-o como uma das maiores economias mundiais[19].

Tabela 2. Observações selecionadas e organizadas segundo o PIB per capita (US$)

País	Continente	PIB per capita (US$)			
		Média	Desvio-padrão	Máximo	Mínimo
Burundi	África	743	76,3	657	929
Libéria	África	773	303	245	1.498
RDC	África	816	264	533	1.511
Etiópia	África	911	455	515	2.097
Moçambique	África	941	240	547	1.271
RCA	África	1.012	114	770	1.228
Malawi	África	1.074	114	844	1.410
Nigéria	África	1.090	56	1000	1.202
Ruanda	África	1.280	413	527	2.047
Madagascar	África	1.371	183	1.147	1.656
Irlanda	Europa	45.076	19.239	19.627	90.300
Noruega	Europa	49.722	14.003	28.446	67.758
Estados Unidos	América	50.750	6.72	38.627	62.274
Suíça	Europa	53.501	10.357	41.114	71.616
Singapura	Ásia	55.028	233.31	21.527	89.339
Kuwait	Ásia	55.031	21.077	13.214	92.274
Brunei	Ásia	66.486	17.020	41.315	98.537
Luxemburgo	Europa	79.569	21.411	45.944	111.704
Catar	Ásia	82.677	49.378	23.639	166.520
Emirados Árabes	Ásia	90.714	17.433	66.166	124.115

Fonte: Resultados originais da pesquisa
Nota: PIB: Produto Interno Bruto; RDC: República Democrática do Congo; RCA: República Centro-Africana

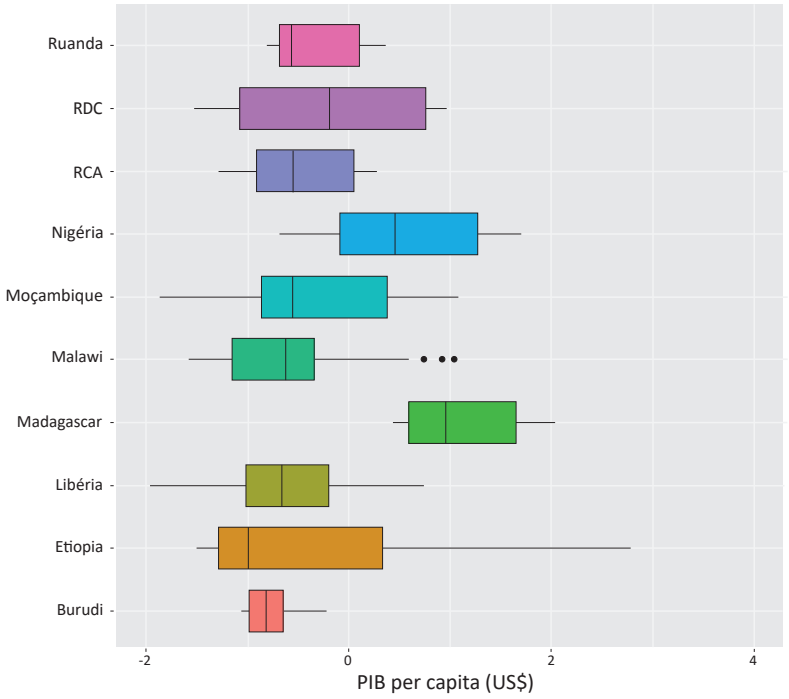

Figura 1. "Boxplot" do Produto Interno Bruto (PIB) per capita (US$) das dez menores economias
Fonte: Resultados originais da pesquisa
Nota: RDC: República Democrática do Congo; RCA: República Centro-Africana

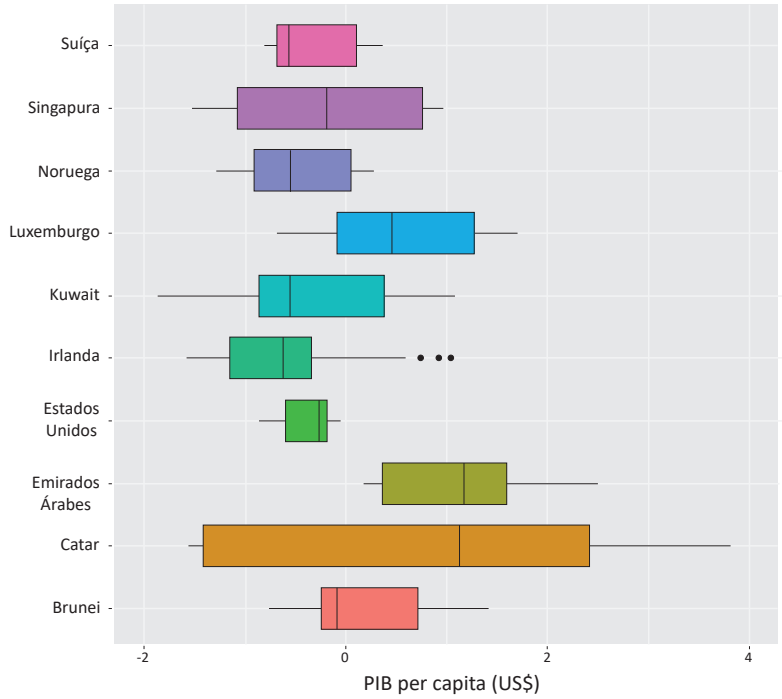

Figura 2. "Boxplot" do Produto Interno Bruto (PIB) per capita (US$) das dez maiores economias
Fonte: Resultados originais da pesquisa

A posição de Luxemburgo como uma das maiores economias mundiais com base no PIB per capita se deve à alta taxa de trabalhadores expatriados — reflexo do investimento estrangeiro no país —, os quais contribuem para a riqueza nacional, porém não são contabilizados como residentes oficiais[20].

A taxa média de emissões de GEE entre 1990 e 2018 dos países que compunham a amostra é apresentada na Figura 3.

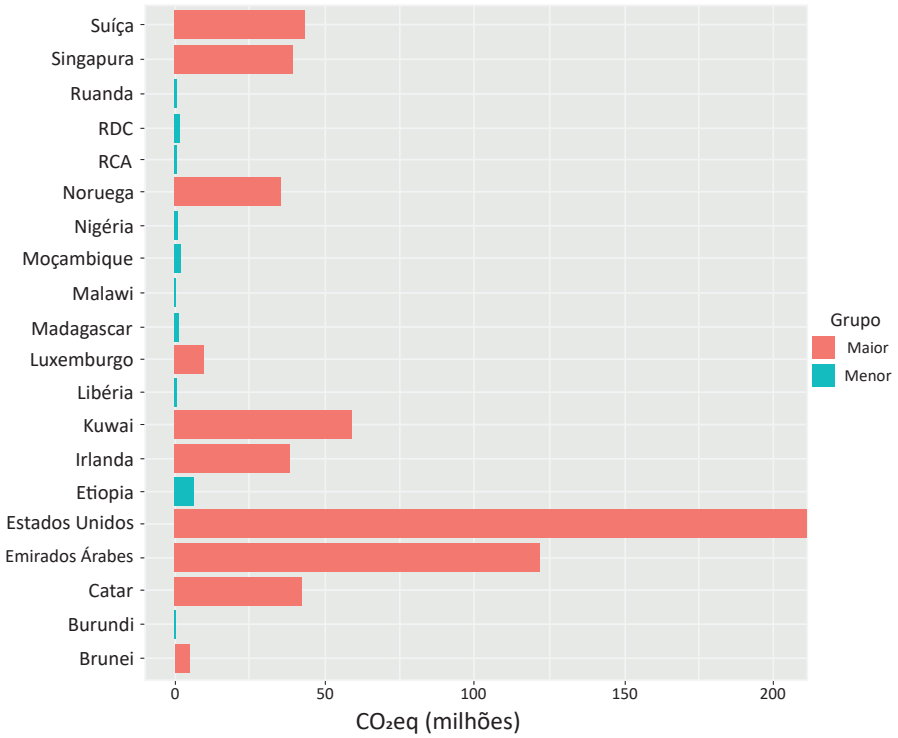

Figura 3. Emissão média de gases de efeito estufa (CO_2eq) entre 1990 e 2018
Fonte: Resultados originais da pesquisa

Os países que compunham o grupo das maiores economias apresentaram as maiores taxas de emissões de GEE. Esse resultado pode ser justificado pelo maior nível de industrialização e pela maior população desses países, que implicariam uma maior demanda e consumo de bens e produtos, resultando em maiores emissões.

Entretanto, a Etiópia, país listado no grupo das menores economias, apresentou taxa de emissões de GEE de ordem de grandeza similar à de países do grupo das maiores economias. De acordo com Engdaw[21], mesmo com o aumento recente da parcela de contribuição dos processos industriais e de geração de energia, o perfil de emissões de GEE na Etiópia ainda é fortemente influenciado pelo nível de atividade agrícola e desmatamento, caracterizando-se como um país com perfil de emissões distinto ao do grupo das economias mais desenvolvidas.

Portanto, foi realizada a modelagem dos dados por meio do algoritmo "random forest" com o intuito de identificar as variáveis que caracterizavam o perfil de emissões dos países a partir do ajuste dos hiperparâmetros.

Modelagem

O principal objetivo de aplicar o algoritmo "random forest" aos dados foi construir um modelo de regressão de natureza descritiva com o intuito de identificar padrões e tendências. A partir da amostra dividida entre países de menores e maiores economias, separou-se a base de dados em dois conjuntos, treino e teste, seguindo uma proporção de 70/30, respectivamente.

É possível que nem todos os atributos do conjunto de dados sejam relevantes para a modelagem do perfil de emissões. Nesse sentido, a técnica de eliminação recursiva de atributos foi aplicada — "recursive feature elimination" (RFE) em inglês —, com um estimador do tipo "random forest" para determinar a importância relativa dos atributos. Essa função avaliou a contribuição relativa dos atributos a partir da construção de sucessivos modelos "random forest" treinados com múltiplas versões da base de dados de teste obtidas por "cross validation", considerando três métricas de desempenho: raiz do erro quadrático médio (RMSE), erro médio absoluto (MAE) e coeficiente de determinação (R^2).

Calculou-se então a raiz do erro quadrático médio em função do número de atributos utilizados na construção de modelos "random forest" para os conjuntos de dados das maiores e menores economias (Figura 4). Observou-se que o desempenho do modelo melhorou à medida que se adicionaram novos atributos para ambos os conjuntos de dados; contudo, o número ideal de atributos foi diferente dependendo do conjunto de países considerado.

Identificou-se que um modelo construído a partir de seis atributos apresentava a menor raiz do erro quadrático médio para ambos os conjuntos. Contudo, notou-se que a utilização de apenas três atributos resultou em erros médios similares aos reportados para o modelo com seis atributos para o conjunto das menores economias, com uma diferença de 6%. Assim, foi definido que três atributos seriam suficientes para modelar o conjunto de dados das menores economias, a partir de um resumo das métricas de desempenho para os modelos construídos (Tabela 3).

Tabela 3. Métricas de desempenho dos modelos "random forest" construídos para as observações das maiores e menores economias por meio da técnica "recursive feature elimination" (RFE)

Métrica	Maiores economias	Menores economias
RMSE	0,0265	0,2187
MAE	0,0109	0,1316
R^2	0,9993	0,9438

Fonte: Resultados originais da pesquisa
Nota: RMSE: raiz do erro quadrático médio; MAE: erro médio absoluto; R^2: coeficiente de determinação

A redução praticada no número de atributos simplificou o modelo, facilitando sua interpretação sem perder representatividade, visto o elevado valor do coeficiente de determinação e a pequena alteração nas demais métricas de desempenho.

Nem todas as variáveis selecionadas contribuíram para a modelagem dos dados da mesma forma. Foi possível avaliar a contribuição relativa de cada uma delas através da determinação de sua importância ("feature importance"). A Figura 5 exibe a importância de cada variável na modelagem dos dados utilizando "random forest". A variável "emp" (quantidade de pessoas empregadas) foi a que exerceu maior influência em ambos os modelos. Esta variável apresentou relação direta com o padrão de vida da população, dado que uma maior quantidade de pessoas empregadas poderia indicar uma nação que apresentava maior nível de consumo de bens e serviços, resultando em uma maior emissão de GEE.

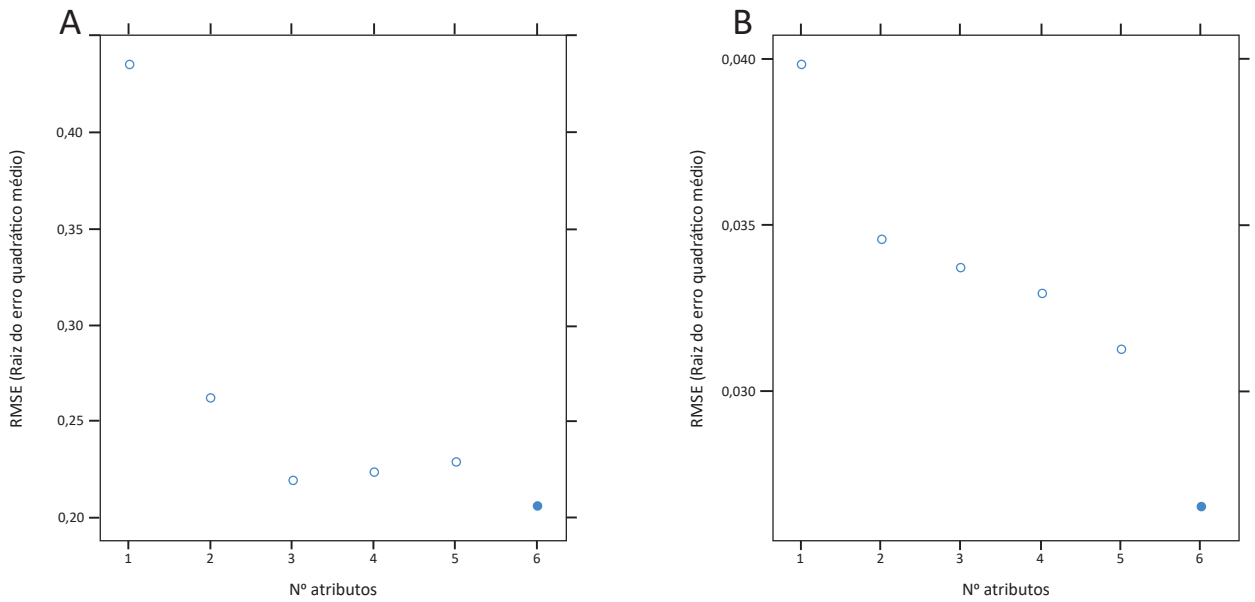

Figura 4. Raiz do erro quadrático médio de modelos "random forest" construídos para o conjunto de dados das menores (A) e maiores (B) economias utilizando quantidades diferentes de atributos
Fonte: Resultados originais da pesquisa

A segunda variável de maior relevância se alterou de acordo com o modelo considerado. Para os países de menor economia, verificou-se que a parcela referente à absorção — consumo das famílias acrescido do investimento produtivo — exerceu maior influência no perfil de emissões de GEE, enquanto para os países de maior economia o consumo das famílias foi a variável mais relevante na ótica do PIB. Os países de menor economia encontravam-se em um estágio de desenvolvimento em que se incentivava o investimento produtivo das empresas, visando a um maior crescimento econômico, o que influenciou de forma mais intensa o perfil de emissões. Em contrapartida, os países que constituíam o grupo das maiores economias exibiram um perfil de emissões de GEE impulsionado principalmente pelo consumo das famílias e pelo tamanho da sua população. É importante frisar que essa diferença na relevância relativa entre as variáveis foi verificada com base nas variáveis normalizadas, visando reduzir o efeito da magnitude dos valores. Ou seja, essa análise não indicou que o investimento produtivo de um determinado grupo fosse maior ou menor que o de outro, porém sugeriu uma importância relativa distinta em sua comparação.

Uma comparação da ordem de relevância das variáveis entre os dois grupos mostrou que o índice de capital humano (h_idex) — que sugeria uma indicação da qualidade com que os recursos econômicos eram empregados na melhoria da saúde e da educação da população — foi mais relevante para as nações de menor economia. O indicador do índice de capital humano do grupo que representava as menores economias apresentou uma importância relativa de 4,81%, frente a 0,21% de importância do grupo das maiores economias. Com relação à atividade econômica, observou-se que o perfil de emissões de GEE das nações de menor economia foi mais influenciado pelo nível de importações do que de exportações, ao passo que o perfil de emissões dos países de maior economia apresentou tendência inversa, ou seja, com as exportações exercendo maior influência.

A análise apresentada indicou a quantidade de atributos necessários para a modelagem dos dados, tornando a parametrização mais ágil e reduzindo o custo computacional. Os modelos "random forest" de cada conjunto de dados foram ajustados a partir de uma base de dados de treino composta por 70% das observações dos dados originais selecionados aleatoriamente. Os 30% remanescentes formaram a base de dados de teste, utilizada para avaliar o desempenho dos modelos. A variável resposta adotada foi a quantidade de emissões de GEE, expressa em milhões de toneladas equivalentes de CO_2.

Realizou-se o treinamento dos modelos com a técnica de validação cruzada "k-fold cross validation", considerando cinco grupos ("folds"). O diferencial da aplicação da técnica se baseou na obtenção da estimativa do modelo a partir da média das estimativas de múltiplos modelos individuais, no caso cinco. O principal benefício desta abordagem foi o melhor desempenho do modelo, reduzindo o erro (RMSE e MAE) e melhorando a representatividade dos dados (R^2).

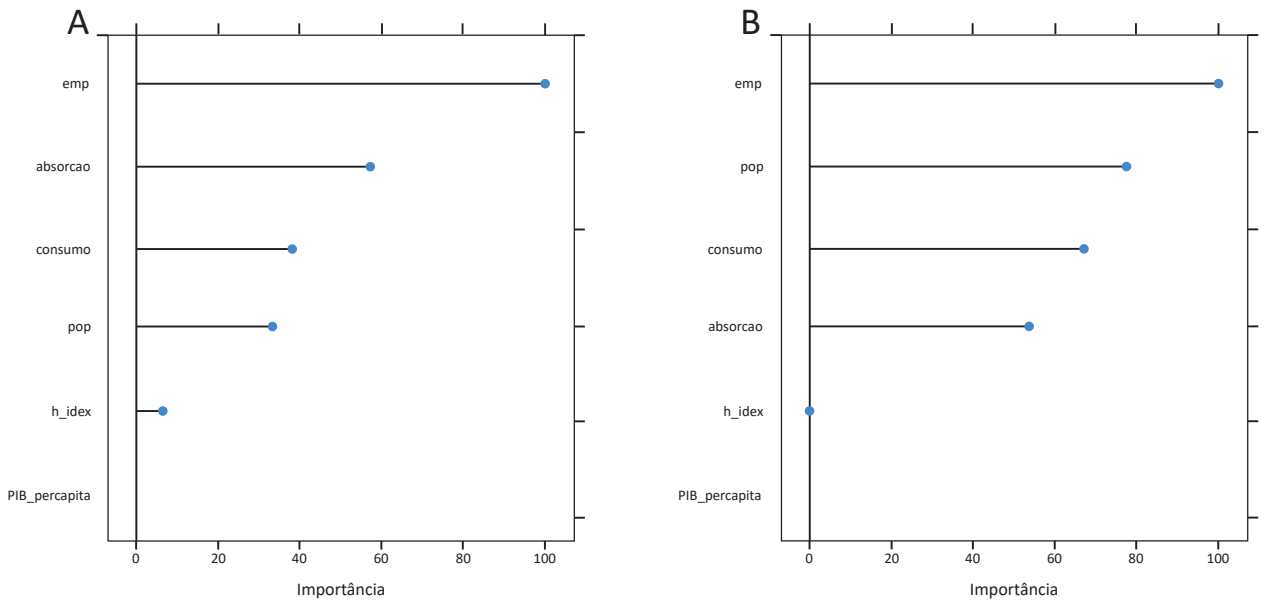

Figura 5. Importância relativa das variáveis do conjunto de dados das menores (A) e maiores (B) economias considerando um modelo "random forest"
Fonte: Resultados originais da pesquisa
Nota: emp: número de pessoas trabalhando; absorcao: consumo das famílias e gastos do governo acrescidos da parcela de investimento produtivo considerando a paridade do poder de compra; consumo: consumo das famílias e gastos do governo, considerando a paridade do poder de compra; pop: população (milhões); h_idex: índice de capital humano; PIB_percapita: PIB per capita pela ótica do dispêndio, considerando a paridade do poder de compra

Em seguida, empregaram-se os modelos parametrizados no treinamento para avaliar as estimativas utilizando os dados de teste. As métricas de desempenho dos modelos "random forest" na etapa de treinamento e nos dados de teste são apresentadas na Tabela 4.

Tabela 4. Métricas de desempenho dos modelos "random forest"

Métrica	Maiores economias		Menores economias	
	Treino	Teste	Treino	Teste
RMSE	0,0369	0,0258	0,1938	0,1659
MAE	0,0137	0,0076	0,1039	0,0804
R^2	0,9989	0,9993	0,9684	0,9553

Fonte: Resultados originais da pesquisa
Nota: RMSE: raiz do erro quadrático médio; MAE: erro médio absoluto; R^2: coeficiente de determinação

Verificou-se que as métricas de desempenho obtidas na etapa de treinamento ficaram próximas dos resultados obtidos com a técnica de RFE, corroborando com o benefício do uso da validação cruzada. Outro aspecto positivo correspondeu ao melhor desempenho do modelo nos dados de teste quando comparados com os dados de treino, indicando não haver "overfitting" e sugerindo boa capacidade de generalização.

Adicionalmente, analisou-se o padrão das estimativas por meio da determinação dos resíduos, os quais foram calculados pela diferença entre o valor estimado pelo modelo e o valor presente nos dados. A distribuição dos resíduos versus os valores esperados é ilustrada pela Figura 6.

Os resíduos calculados para os dados das maiores e menores economias encontravam-se próximos de zero (Figura 6). Os resíduos do modelo que descrevia os dados das menores economias (Figura 6A) estavam distribuídos aleatoriamente, indicando apresentar heterocedasticidade. Por outro lado, o modelo que representava as maiores economias (Figura 6B) mostrou uma concentração dos resíduos em torno de zero, sugerindo possível homocedasticidade. Vale destacar, contudo, que o perfil de emissões destes países foi bastante similar: baixa amplitude nos dados e baixas emissões, as quais, associadas às boas estimativas apresentadas pelo modelo, resultaram em baixos resíduos.

Os resultados obtidos por meio da modelagem "random forest" subsidiam a avaliação inicial acerca das variáveis socioeconômicas mais relevantes quanto ao perfil de emissões de GEE. Ao comparar o perfil de emissões entre os grupos das menores e maiores economias, notou-se que o perfil de emissões das nações de menor economia

sofria maior influência do investimento produtivo e do nível de importações do que o perfil das nações de maior economia.

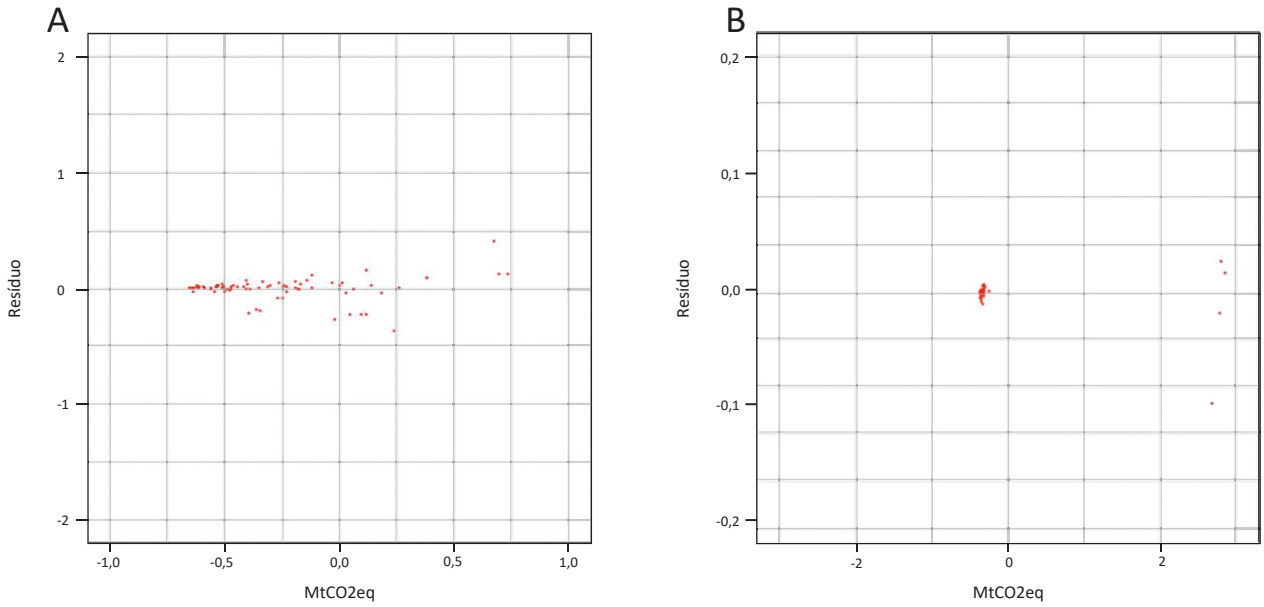

Figura 6. Resíduos das estimativas obtidas a partir de modelos "random forest" em função dos valores esperados de emissões de gases de efeito estufa para as observações das menores (A) e maiores (B) economias
Fonte: Resultados originais da pesquisa
Nota: MtCO2eq: emissões de GEE expressas em milhões de toneladas equivalentes de CO_2

Segundo os dados apresentados na Tabela 2, verificou-se que o grupo das menores economias estava concentrado no continente africano. Grisotto[22] destacou a tendência crescente de investimentos em tal continente, impulsionada pelas inúmeras lacunas e oportunidades que as nações dele oferecem. Além de possuir territórios extensos e pouco povoados, a África conta com uma população jovem e culturalmente inventiva, receptiva à tecnologia; contudo, tem dificuldade de acesso a serviços. Outro aspecto positivo é a proximidade geográfica de mercados maduros como a China, por exemplo. Esses fatores têm atraído investimento estrangeiro ao continente, justificando a maior relevância das variáveis de absorção e importações no perfil de emissões dessas nações. Por outro lado, as mesmas oportunidades que estimulam as importações e o investimento produtivo no continente acabam por impulsionar as exportações nos países que representam as maiores economias.

4. Conclusão

O presente estudo apresentou uma avaliação descritiva da condição socioeconômica de uma amostra de países, selecionados a partir do seu nível de atividade econômica, em conjunto com a aplicação de uma modelagem "machine learning" que empregou um modelo de "random forest" como um regressor, buscando aprofundar a compreensão acerca dos fatores que influenciavam o nível de GEE. A modelagem permitiu verificar que os fatores que caracterizavam o nível de emissões de GEE dos países de maior e menor atividade econômica eram distintos. Considerando a importância relativa das variáveis estudadas, observou-se que o perfil de emissões de gases de efeito estufa dos países de maior economia foi mais influenciado pelo padrão de consumo das famílias e gastos do governo, ao passo que os países de menor economia apresentaram um perfil fortemente influenciado pelo padrão de consumo das famílias e gastos do governo acrescido da parcela de investimento produtivo, ambos considerando a paridade do poder de compra.

A análise apresentada foi consolidada a partir de duas bases de dados independentes, o que reduziu possíveis vieses de amostragem e pôde prover informações relevantes para subsidiar ações de incentivo e desenvolvimento econômico. Por exemplo, os países de maior economia podem identificar os setores e produtos de maior consumo e implementar políticas públicas de sustentabilidade ambiental, social e de governança [do inglês "environmental, social and governance" (ESG)] ou determinar medidas que busquem um consumo sustentável; enquanto isso, os países de menor economia podem estabelecer medidas de incentivo, como desoneração fiscal ou baixo juro de

captação de recursos, para atrair empresas que priorizam ações ambientalmente conscientes.

Por fim, é importante frisar que as reflexões apresentadas neste estudo visam fomentar ideias para equacionar o problema das emissões de gases de efeito estufa associados ao constante crescimento econômico mundial. Isso não significa dizer que os países devam valorizar um fator em detrimento de outro; espera-se, no entanto, que essas reflexões sirvam como um direcionador para priorizar recursos e esforços, os quais muitas vezes são limitados e escassos.

Contribuições dos autores: Frutuoso, L.F.A.: Conceitualização; Aquisição de Dados; Análise de Dados; Definição da Metodologia; Escrita e Edição. Barbosa, W.: Conceitualização; Aquisição de Dados.

Como citar: Frutuoso, L.F.A.; Barbosa, W. 2024. Caracterização do perfil global de emissões de gases de efeito estufa utilizando "machine learning". Quaestum 5: e2675741.

Referências

[1] United States Environmental Protection Agency (EPA). 2022. Inventory of U.S. greenhouse gas emissions and sinks: 1990–2020. Disponível em: <https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks-1990-2020>. Acesso em: 21 set. 2022.
[2] Vicente, M.C.P. 2016. Cadernos Adenauer XVII, nº 2 — Mudanças climáticas: o desafio do século. Fundação Konrad Adenauer, Rio de Janeiro, RJ, Brasil.
[3] Yoro, K.O.; Daramola, M.O. 2020. CO_2 emission sources, greenhouse gases, and the global warming effect. p. 3-28. In: Rahimpour, M.R.; Farsi, M.; Makarem, M.A. (eds.). Advances in Carbon Capture: Methods, Technologies and Applications. Woodhead Publishing, Cambridge, UK. https://doi.org/10.1016/B978-0-12-819657-1.00001-3.
[4] European Environment Agency (EEA). 2023. EMEP/EEA air pollutant emission inventory guidebook 2023. Disponível em: <https://www.eea.europa.eu/publications/emep-eea-guidebook-2023>. Acesso em: 01 mar. 2024.
[5] Environmental Protection Agency (EPA). Understanding global warming potentials. Disponível em: <https://www.epa.gov/ghgemissions/understanding-global-warming-potentials>. Acesso em: 12 out. 2023.
[6] Foster, P.; Ramaswamy, V.; Artaxo, P.; Berntsen, T.; Betts, R.; Fahey, D.W.; Haywood, J.; Lean, J.; Lowe, D.C.; Myhre, G.; Nganga, J.; Prinn, R.; Raga, G.; Schulz, M.; Van Dorland, R. 2007. Changes in atmospheric constituents and in radiative forcing. In: Solomon, S.; Qin, D.; Manning, M.; Chen, Z.; Marquis, M.; Averyt, K.B.; Tignor, M.; Miller, H.L. (eds.). Climate Change 2007: the physical science basis. Contribution of working group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK. Disponível em: <http://www.ipcc.ch/pdf/assessment-report/ar4/wg1/ar4-wg1-chapter2.pdf>. Acesso em: 21 set. 2022.
[7] Feenstra, R.C.; Inklaar, R.; Timmer, M.P. 2015. The next generation of the Penn World Table. American Economic Review 105(10): 3150-3182. https://doi.org/10.1257/aer.20130954.
[8] Climate Watch. 2022. GHG Emissions. World Resources Institute. Disponível em: <https://www.climatewatchdata.org/ghg-emissions>. Acesso em: 07 nov. 2022.
[9] Dowrick, S. 2005. The Penn World Table: a review. The Australian Economic Review 38(2): 223-228. https://doi.org/10.1111/j.1467-8462.2005.00369.x.
[10] Instituto Brasileiro de Geografia e Estatística (IBGE). 2022. Produto Interno Bruto – PIB. Disponível em: <https://www.ibge.gov.br/explica/pib.php#:~:text=O%20PIB%20%C3%A9%20a%20soma,R%24%208%2C7%20trilh%C3%B5es.>. Acesso em: 05 nov. 2022.
[11] Groningen Growth and Development Centre. 2021. PWT 10.0 Penn World Table version 10.0. https://doi.org/10.15141/S5Q94M.
[12] Crisci, C.; Ghattas, B.; Perera, G. 2012. A review of supervised machine learning algorithms and their applications to ecological data. Ecological Modelling 240: 113-122. https://doi.org/10.1016/j.ecolmodel.2012.03.001.
[13] Boateng, E.Y.; Otoo, J.; Abaye, D.A. 2020. Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: a review. Journal of Data Analysis and Information Processing 8(4): 341-357.
[14] Breiman, L. 2001. Random forests. Machine Learning 45: 5-32. Disponível em: <https://link.springer.com/article/10.1023/A:1010933404324>. Acesso em: 13 nov. 2022.
[15] Liaw, A.; Wiener, M. 2002. Classification and Regression by randomForest. R News 2(3): 18-22.
[16] Silva, S.H.G.; Teixeira, A.F.S.; Menezes, M.D.; Guilherme, L.R.G.; Moreira, F.M.S.; Curi, N. 2017. Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF). Ciência e Agrotecnologia 41(6): 648-664. https://doi.org/10.1590/1413-70542017416010317.
[17] Coelho, S.T.; Pereira, A.S.; Bouille, D.H.; Mani, S.K.; Recalde, M.Y.; Savino, A.A.; Stafford, W.H.L. 2020. Overview of Developing Countries. p. 9-61. In: Coelho, S.T.; Pereira, A.S.; Bouille, D.H.; Mani, S.K.; Recalde, M.Y.; Savino, A.A.; Stafford, W.H.L. Municipal solid waste energy conversion in developing countries: technologies, best practices, challenges and policy. Elsevier, Oxford, UK.
[18] Riveira, C. 2017. Mbaku, da Brookings: Ruanda cresce, mas a que preço? Revista Exame. Disponível em: <https://exame.com/mundo/mbaku-da-brookings-ruanda-cresce-mas-a-que-preco/>. Acesso em: 13 nov. 2022.
[19] Ashghal. 2022. Qatar Past, Present and Future. Disponível em: <https://www.ashghal.gov.qa/en/AboutQatar/Pages/Qatar-Today.aspx>. Acesso em: 13 nov. 2022.
[20] Eurostat. 2022. GDP at regional level. Disponível em: <https://ec.europa.eu/eurostat/statistics-explained/index.php?title=GDP_at_regional_level>. Acesso em: 13 nov. 2022.
[21] Engdaw, B.D. 2020. Assessment of the trends of greenhouse gas emission in Ethiopia. Geography, Environment, Sustainability 13(2): 135-146. https://doi.org/10.24057/2071-9388-2018-61.
[22] Grisotto, R. 2018. Empreendedorismo na África: novos negócios atraem investidores e gigantes de tecnologia. Disponível em: <https://epocanegocios.globo.com/Empreendedorismo/noticia/2018/05/empreendedorismo-na-africa-novos-negocios-atraem-investidores-e-gigantes-de-tecnologia.html>. Acesso em: 12 jan. 2023.

