57º SBPO
SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
Gramado • RS
DE 5 A 9 DE OUTUBRO DE 2025

# PREDICTOR2: UM MODELO PARA PREDIÇÃO DOS ÍNDICES DE CO2 NA ATMOSFERA

**David Franklin Pessoa Ferreira**
Universidade Federal Rural do Semi-Árido
Av. Francisco Mota, 572, Bairro Costa e Silva - Mossoró/RN - 59625-900
davidfpessoa02@gmail.com

**Sebastião Emídio Alves Filho**
Universidade Estadual do Rio Grande do Norte
Rua Professor Antônio Campos, s/n, BR 110, km 48, Bairro Costa e Silva - Mossoró/RN - 59600-000
sebastiaoalves@uern.br

**Matheus da Silva Menezes**
Universidade Federal do Rio Grande do Norte
Campus Universitário Lagoa Nova - Natal/RN - 59078-900
matheus.menezes@ufrn.br

**Heitor Pereira Liberalino Carlos**
Universidade Estadual do Rio Grande do Norte
Rua Professor Antônio Campos, s/n, BR 110, km 48, Bairro Costa e Silva - Mossoró/RN - 59600-000
heitorliberalino@uern.br

**RESUMO**

O aquecimento global é uma das maiores problemáticas da atualidade. Ao longo das décadas, as ações antrópicas colaboram com as mudanças climáticas sofridas [Houghton e Woodwell, 1989], dentre eles, a elevação dos Gases do Efeito Estufa (GEE) em especial o $CO_2$, por meio da queima de combustíveis fósseis entre outros. O presente estudo tem como objetivo usar dados de séries temporais dos índices de $CO_2$ na atmosfera, disponibilizados pela NASA, e monitorados pelo NOAA, e usar modelos de IA para predizer índices futuros. Para isso, implementou-se um modelo de *Machine Learning* (ML) gerado para este trabalho, o **PrediCtOr2**, e comparou-se os resultados obtidos com a biblioteca *Scikit-learn*. Validou-se os modelos, através do Método dos Mínimos Quadrados (MMQ). Ao final do experimento, pode-se comprovar a coerência das predições dos modelos, constatando sua eficiência para tomada de decisão acerca desta temática.

**PALAVRAS CHAVE. mudanças climáticas, regressão linear, aprendizado de máquina, séries temporais.**

**3. AG&MA - PO na Agricultura, Meio Ambiente e Sustentabilidade; 10. IA - PO e IA**

**ABSTRACT**

Global warming stands as one of the most critical challenges of the modern era. Over the decades, anthropogenic actions have collaborated with the climate changes suffered [Houghton e Woodwell, 1989], including the rise in Greenhouse Gas (GHG) emissions, particularly $CO_2$,

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

driven by fossil fuel combustion, among other sources. This study aims to utilize time-series data on atmospheric $CO_2$ levels, provided by NASA and monitored by NOAA, alongside AI models to predict future indices. For this purpose, a custom Machine Learning (ML) model, **PrediCtOr2**, was developed for this work, and its results were compared with those obtained using the Scikit-learn library. The models were validated through the Least Squares Method (LSM). By the conclusion of the experiment, the consistency of the models' predictions was confirmed, validating their efficiency for decision-making regarding this pressing issue.

**KEYWORDS. Climate change. Linear regression. Machine Learning. Time series.**

**3. AG&MA - PO na Agricultura, Meio Ambiente e Sustentabilidade; 10. IA - PO e IA**

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

## 1. Introdução

O oxigênio é um dos bens mais preciosos para os seres vivos. Do primeiro respirar até o último, toda uma existência se firma. No entanto, nosso ar e atmosfera não são compostos unicamente por oxigênio respirável ($O_2$). Diversos outros elementos os compõem, alguns destes atuam como uma barreira protetora para os seres vivos. Uma destas barreiras é a Camada de Ozônio (composta de ozônio, $O_3$), localizada entre 25 a 30 km acima da superfície terrestre, na estratosfera. Essa camada nos protege dos raios solares, como um filtro para os raios ultravioletas (raios UV). Um dos fatores que contribuem para o enfraquecimento da Camada de Ozônio é a presença excessiva de gases danosos, os chamados Gases do Efeito Estufa (GEE), como Óxido Nítrico ($NO$), Óxido Nitroso ($N_2O$), Metano ($CH_4$) e o Dióxido de Carbono ($CO_2$).

Segundo [Amarpuri et al., 2019] o $CO_2$ "*responde por 1% dos gases atmosféricos, mas é responsável por cerca de 81% das emissões totais, sendo um componente importante das emissões de Gases do Efeito Estufa (GEE)*", oriundas, entre outras fontes, da queima de combustíveis fósseis e da produção de cimento. Os autores ainda alertam que o $CO_2$ desempenha um papel crítico no aquecimento global. De forma a ilustrar da gravidade dessa situação, segundo [Nações Unidas, 2022], a Terra apresentou um aumento significativo na sua temperatura - cerca de 1,1ºC - se comparado com as temperaturas médias do final do século XIX. Algo que pode, potencialmente, causar desastres climáticos, como secas intensas, escassez de água, incêndios severos, elevação do nível do mar, inundações, derretimento das calotas polares, tempestades e declínio da biodiversidade [Nações Unidas, 2022] [Amarpuri et al., 2019]. Portanto, analisar dados de séries temporais com objetivo de monitor os GEE, em especial do $CO_2$, pode nos colocar na vanguarda na busca por soluções desta problemática.

Segundo [Box et al., 2015], *"Uma série temporal é uma sequência de observações coletadas sequencialmente ao longo do tempo".* As séries temporais são aplicadas em diversas áreas como economia, negócios (em especial o setor de mercado de ações), ciências naturais (em particular a geofísica e meteorologia) e ciências sociais. Em geral, a principal característica de uma série temporal é a dependência entre observações adjacentes. Esse aspecto de dependência é considerado de interesse prático. Para analisá-las, são necessárias técnicas que explorem essas dependências, exigindo o desenvolvimento de modelos estocásticos e dinâmicos, aplicados em contextos relevante [Box et al., 2015].

Para isso, é necessário o uso de modelos matemáticos capazes de prever o futuro olhando para o passado. Nesse contexto, os modelos de *Machine learning (ML)* têm demonstrado grande eficácia. Um modelo de ML devidamente treinado com dados fidedignos e com validação da sua acurácia - por meio da comparação entre predições geradas e os dados reais - pode se tornar uma ferramenta matemática de rápida resposta na predição dos próximos índices de $CO_2$ na atmosfera, com precisão e confiabilidade satisfatórias. Predizer os próximos níveis de $CO_2$ na atmosfera, mantendo-se a tendência atual, pode ajudar na compreensão do cenário futuro e no planejamento de ações mitigadoras.

Sendo assim, o objetivo deste trabalho é elaborar e treinar um modelo de ML (PrediCtOr2), utilizando como entrada dados compilados pela agência norte-americana NASA (*National Aeronautics and Space Administration* - "Administração Nacional da Aeronáutica e Espaço, em tradução direta") e monitorados pela NOAA (*National Oceanic and Atmospheric Administration* - "Administração Nacional Oceânica e Atmosférica", em tradução direta), referente ao período de 1958 até maio de 2024 [NOAA, 2022]. Utilizando o método de classificação por Regressão, foram considerados tanto o ajuste linear quanto os modelos não lineares logarítmico e exponencial, com intuito de predizer os próximos índices de emissão de $CO_2$ na atmosfera terrestre, assumindo a manutenção da tendência de comportamento observadas nas séries temporais.

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

## 2. Trabalhos Relacionados

Diversos estudos foram realizados utilizando séries temporais e modelos matemáticos ou de ML. Na China, [Li et al., 2017] utilizaram séries temporais de dados do período de 1998 a 2014, obtidas da *Wind Data*, do Anuário Estatístico da China e do Boletim Estatístico do Desenvolvimento Econômico e Social Nacional da China, ambos do Departamento de Estatística da China (*National Bureau of Statistics of China*), nos anos de 2015/16; com objetivo de realizar predições para os anos de 2015 a 2035. Eles utilizaram uma técnica de predição com base em um modelo de Impactos estocásticos por regressão à população, afluência e tecnologia (*Stochastic impacts by regression on population, affluence, and technology - STIRPAT*). Obtiveram resultados que, nos cenários baixo, médio e alto, no qual a China atingirá seu pico de emissões de carbono em 2024, 2027 e 2030, respectivamente, se mantidas as tendências da época. As métricas de desempenho utilizadas foram o RMSE (*Root mean squared error*), RIDGE (*Ridge regression*) e *Intercept*.

[Amarpuri et al., 2019] trabalharam com séries temporais com objetivo de predizer os índices de $CO_2$ para o ano de 2020 utilizando uma técnica de predição com base em um modelo híbrido de *Deep learning (DL)* de Rede Neural Convolutiva e Rede de Memória de Longo Prazo (*Convolution neural network and long short-term memory network - CNN-LSTM*).

[Kumar et al., 2020] trabalharam com séries de dados temporais coletados em tempo real por meio de sensores, entre os meses de julho a setembro de 2018, dos quais os sensores estavam localizados em três setores da Região Nacional da Capital, em Nova Delhi; O objetivo era predizer a qualidade do ar do dia seguinte. Para isso, utilizaram uma técnica de predição com base em um modelo de ML com Regressão linear e obtiveram resultados significativos. As métricas de desempenho utilizadas foram MAE (*Mean absolute error*), MSE (*Mean squared error*), RMSE e MAPE (*Mean absolute percentage error*).

O trabalho de [Taheri e Razban, 2021] teve por objetivo predizer as concentrações de $CO_2$ em uma variedade de horizontes de previsão, para isso, utilizaram seis técnicas de predição com base em modelos de ML, são elas: *Support Vector Machine (SVM)*, *AdaBoost*, *Random Forest (RF)*, *Gradient Boosting (GB)*, *Logistic Regression (LR)* e *Multilayer Perceptron (MLP)*. Os resultados mostraram que o modelo MLP pode prever fortemente o comportamento volátil do $CO_2$ e também supera outros modelos em termos de precisão. Além disso, uma estratégia de controle capaz de modelar e detectar padrões dinâmicos do nível de $CO_2$ foi utilizada para modular a taxa de ventilação em tempo real e também reduzir o consumo de energia. As métricas utilizadas foram RMSE, MAE, MAPE e $R^2$.

[Silva, 2022] trabalhou com dados históricos sobre a elevação média do nível do mar, obtidos de medições no período de 1993 a 2022, coletados e disponibilizados pela NASA; O objetivo era predizer o nível do mar nas principais áreas do litoral piauiense suscetíveis a inundação devido ao avanço do mar para os próximos 78 anos (2023-2100). A técnica utilizada foi o Método dos Mínimos Quadrados (MMQ) com ajuste dos dados para as funções de comportamento linear, logarítmica, exponencial, potencial e polinomial de segundo grau (quadrática), buscando obter uma função que melhor represente o comportamento dos dados; esse tabelamento foi feito no *Excel 2020*; como resultado desse tabelamento, algumas funções apresentaram bons ajustes, ficando dentro de faixas estimadas pelo Sexto Relatório de Avaliação do IPCC (*Intergovernmental Panel on Climate Change*) 2022, os casos linear e logarítmico num cenário mais otimista e o caso e quadrático em um cenário mais pessimista. Diante dos resultados encontrados, foi utilizado o software *QGIS* para simular o impacto desta variação no nível do mar na costa do litoral piauiense; Assim, verificou-se que em ambos os cenários, as áreas de preservação ambiental, aquicultura, salicultura além do Porto Internacional do Piauí serão afetados em decorrência desse problema.

[Kumari e Singh, 2023] apresentaram os efeitos prejudiciais do $CO_2$ na Índia, e traba-

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

lharam com dados univariados de séries temporais dos anos de 1980 a 2019, tendo como fonte desses dados o CAIT, com objetivo de realizar predições para os anos de 2020 até 2030. Utilizaram seis técnicas de predição; Três com base em modelos estatísticos: *Autoregressive-integrated moving average (ARIMA)*, *Seasonal autoregressive-integrated moving average with exogenous factors (SARIMAX)* e o modelo *Holt-Winters*; Dois modelos de ML: *Linear Regression (LinR)* e RF; E um modelo DL: *Deep learning-based long short-term memory (DL-LSTM)*. O melhor modelo apresentado por eles foi o DL-LSTM. Foram utilizadas as seguintes métricas de desempenho MSE, RMSE, MSLE (*Mean squared log error*), MAE, MAPE, MedAE (*Median absolute error*), Max Error, $R^2$, EVS (*Explained variance score error*).

Com base nesta revisão de literatura, verificou-se poucos trabalhos utilizando Regressão linear (com os casos Logarítmico e Exponencial) e o MMQ. Com isso, este trabalho visa contribuir no estudo destes modelos aplicados no contexto da predição de dados de séries temporais. Trabalhar o conhecimento teórico por trás dos algoritmos e modelos matemáticos é primordial para se avançar em modelos melhores, mais robustos e confiáveis. Uma vez que o RMSE e o $R^2$ são métricas confiáveis e foram as mais usadas nos trabalhos encontrados, opou-se por elas neste trabalho.

## 3. Metodologia

Os códigos foram implementados em *Python*, utilizando o *Google Colaboratory* ou *Colab*, *API* em nuvem da *Google Research*. Outra informação relevante, é que este trabalho se baseou em um tripé, formado pelo modelo matemático do Método dos Mínimos Quadrados (MMQ), o modelo de ML PrediCtOr2 (com base na teoria da ML) e um modelo de regressão com *Scikit-learn - SKLearn* (biblioteca de código aberto). Com isso, o experimento pôde testar a capacidade e confiabilidade do PrediCtOr2 em comparação com os resultados obtidos via MMQ e *SKLearn*.

### 3.1. O Ambiente para Compilação - *Google Colab*

No *Colab*, foram necessárias importações de algumas bibliotecas, em suas versões disponíveis à época, listadas em diante: A biblioteca *Pandas*, para trabalhar com os dados; a biblioteca *MatPlotLib*, para criar gráficos e visualizar dados; a biblioteca *Stats* do *software open-source SciPy*, para lidar com os parâmetros estatísticos; a biblioteca *SeaBorn* para plotagem de gráficos; a biblioteca *NumPy*, para trabalhar com os *arrays* e transformações numéricas; o pacote *PyPlot* da biblioteca *MatPlotLib*, para gerar figuras, criar uma área de plotagem em figura e formatar linhas, rótulos, dentre outras funções do tipo na área de plotagem; a biblioteca *StatsModels*, para se estimar modelos estatísticos e realizar testes estatísticos; a biblioteca *SKLearn*, para gerar um modelo de ML pronto para usar. Os pacotes do *SKLearn* utilizados e são detalhados na **Subsubseção 3.3.2**.

### 3.2. Método dos Mínimos Quadrados - MMQ

Como forma a validar os resultados obtidos pelo PrediCtOr2, usou-se com o MMQ em conjunto com um modelo gerado do *SKLearn*. Com isso, trabalhou-se com os ajustes conforme conceituam [Almeida, 2019] [Montgomery e Runger, 2010], por meio dos dados, para estabelecer uma *Função* $G(x)$ cujo o gráfico se ajuste aos pontos observados da melhor forma, sem que ocorra *overfitting* ou *underfitting*.

Nesta seção, apresenta-se a metodologia de aplicação do MMQ para análise de regressão linear, empregada neste estudo. Com base no que dizem [Almeida, 2019] e [Montgomery e Runger, 2010], entende-se que o MMQ é uma técnica de otimização matemática usada para encontrar a curva com comportamento pré-determinado, de melhor ajuste aos dados. Isso minimiza a soma das diferenças quadradas entre os pontos dos dados observados e seus valores previstos correspondentes.

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

### 3.2.1. Caso Linear

Para se trabalhar com os dados no caso linear pelo MMQ, aplicou-se a *Função 1* de forma a determinar os valores de $a$ e $b$.

$$y = ax + b. \tag{1}$$

No qual, $y$ é a variável dependente (valor observado), $x$ é a variável independente (valor de entrada), $a$ é a interceptação da linha de regressão e $b$ é a inclinação da linha de regressão.

### 3.2.2. Estimativa dos Coeficientes *a* e *b*

Para estimar os coeficientes $a$ e $b$, usou-se as fórmulas derivadas do MMQ:

$$a = \frac{n.(\sum x_i.y_i) - \sum x_i.\sum y_i}{n.\sum x_i^2 - (\sum x_i)^2}. \tag{2}$$

$$b = \frac{\sum x_i.(\sum x_i.y_i) - \sum y_i.\sum x_i^2}{(\sum x_i)^2 - n.\sum x_i^2}. \tag{3}$$

### 3.2.3. Casos Não Lineares

Em alguns casos, pode ser necessário lidar com funções comparativas não lineares. A luz de [Almeida, 2019] e [Montgomery e Runger, 2010] pode-se chegar a conclusão de que, funções não lineares podem retomar ao caso linear aplicando transformações, ou funções, matemáticas aos dados originais, tornando-os linearmente relacionados. Ainda segundo os autores, transformações comuns incluem funções logarítmicas, exponenciais ou de potência aplicadas aos valores $x$ ou $y$. Neste estudo serão aplicadas as funções logarítmicas e exponenciais.

Após a linearização, a análise de regressão linear pode ser realizada utilizando o MMQ conforme descrito anteriormente. Os coeficientes $a$ e $b$ são, então, calculados com base nos dados, agora, linearizados.

### 3.2.4. Função Logarítmica

[Almeida, 2019] e [Montgomery e Runger, 2010] nos leva a compreensão de que a transformação logarítmica é utilizada quando a relação entre variáveis é melhor representada por uma curva logarítmica. Isso, é útil quando os dados exibem crescimento ou decaimento exponencial. [Montgomery e Runger, 2010] nos dá a compreensão que a transformação logarítmica é aplicada para linearizar a relação e permitir a aplicação de métodos de regressão linear. A equação para a transformação logarítmica é:

$$y' = \log(y). \tag{4}$$

No qual, $y'$ é o valor transformado da variável $y$.

A função logarítmica é frequentemente usada quando a variável dependente apresenta uma relação logarítmica com a variável independente. Assim, a equação para a função logarítmica é:

$$y = a.\ln(x) + b. \tag{5}$$

No qual, $y$ é a variável dependente, $x$ é a variável independente, $a$ é o intercepto, $b$ é a inclinação, e $\log(x)$ representa o logaritmo natural de $x$.

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

### 3.2.5. Função Exponencial

Para [Almeida, 2019] e [Montgomery e Runger, 2010], a transformação exponencial é utilizada quando a relação entre as variáveis é melhor representada por uma curva exponencial. Isso é útil quando os dados apresentam crescimento ou decaimento acelerado. Assim, pelo que expõe [Montgomery e Runger, 2010], a transformação exponencial é aplicada para linearizar a relação e permitir a aplicação de métodos de regressão linear.

A função exponencial é utilizada quando a variável dependente mostra uma relação exponencial com a variável independente. A equação para a função exponencial é:

$$y = b.\exp^{a.x}. \tag{6}$$

No qual, $y$ é a variável dependente, $x$ é a variável independente, $b$ é o intercepto, $a$ é a inclinação, e $\exp(x)$ representa a função exponencial de $x$.

Para estimar os coeficientes $a$ e $b$ na função exponencial, utilizamos os seguintes dados linearizados:

$$\ln(y) = ax + \ln(b) \tag{7}$$

## 3.3. Modelos de *Machine Learning - ML*

Foram implementados dois modelos de ML. Primeiro, utilizou-se o conceito básico (teoria de ML) para gerar o modelo proposto, *PrediCtOr2*; Segundo, usou-se a biblioteca *SKLearn*, como forma de testar a biblioteca para o experimento em questão, bem como, comparar com os resultados do *PrediCtOr2*.

### 3.3.1. PrediCtOr2

A teoria de ML prevê o modelo básico de Regressão Linear *($wx+b$, que retorna a Função $g(x) = ax + b$)*. É sob essa base que o modelo *PrediCtOr2* foi implementado.

Para isso, foram implementados 4 algorítimos sendo:

*Algoritmo 1*, que define um modelo de regressão linear com base na *Função 1*.

O *Algoritmo 2*, é a função de perda da Diferença de variância média ou Erro Quadrático Médio *(Mean Square Error – MSE)*, expressa:

$$MSE = \frac{1}{n}.\sum_{i=1}^{n}.(y_i - \hat{y}_i)^2. \tag{8}$$

No qual, $y_i$ são os valores reais, $\hat{y}_i$ são os valores previstos, e $n$ é o número de observações. O MSE penaliza fortemente grandes erros, uma vez que os erros são elevados ao quadrado. Isso pode ser útil quando grandes desvios são indesejáveis.

O *Algoritmo 3*, é uma função de otimização com base no método de Gradiente descendente. No contexto de aprendizado de máquina, essa função geralmente é a função de custo ou perda, que mede a diferença entre os valores previstos pelo modelo e os valores reais observados [Bishop e Nasrabadi, 2006].

Por fim, o *Algoritmo 4* é uma função de iteração que retorna $a$ e $b$, após receber em cada iteração o *Algoritmo 3*, com base no número de épocas estipuladas. Após esses preparativos, se inicializa os parâmetros $a$ e $b$ usando as *Equações 9* e 10.

$$a = \frac{y_i - \hat{y}}{x_i - \hat{x}}. \tag{9}$$

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

$$b = \frac{(\hat{y}.x_i) - (y_i.\hat{x})}{x_i - \hat{x}}. \tag{10}$$

No qual, $y_i$ são os valores da Média mês (índices mensais de $CO_2$), $\hat{y}$ é o valor observado da Média mês, $xi$ são os valores dos Dados decimais (referência numérica para mês/ano da coleta), $\hat{y}$ é o valor observado dos Dados decimais. Em seguida, define-se o valor de $Lr$ (*Learning rate*). Com isso, pode-se iniciar as iterações de $a$ e $b$, que recebem o *Algoritmo 4*. Define-se a variável $prediction$ (predição) que utiliza o *Algoritmo 1*, a variável local $loss$ (perda) que recebe o *Algoritmo 2*. Rescreve-se os parâmetros $a$ e $b$, bem como a variável $loss$; e, finalmente, exibe-se os resultados ($x$, $y$ e $prediction$).

### 3.3.2. Usando o *Scikit-learn (SKLearn)*

O *SKLearn* é uma biblioteca de código aberto, disponibilizada e mantida, entre outras empresas, pela *Google*.

Para nosso experimento, utilizou-se os pacotes *RidgeCV*, *LassoCV*, *LinearRegression*, *ElasticNet* da biblioteca *Linear Model* para trabalhar com os modelos de ajuste linear; *StandardScaler* da biblioteca *Preprocessing*, para normalizar os dados; *train_test_split* da biblioteca *model_select*, para separar os dados de Treino e Teste; *GridSearchCV* da biblioteca *Model_Selection*, para se realizar uma busca sobre valores de parâmetro especificados para um estimador; *R2_Score* da biblioteca *Metrics*, para se definir o coeficiente de determinação na regressão linear; *SVR* da biblioteca *SVM*, que é a forma de regressão do *SVM*; *RandomForestRegressor*, para se ter um meta estimador que ajuste uma série de árvores de decisão de classificação em várias subamostras do conjunto de dados e usar a média para melhorar a predição e controlar o ajuste excessivo, e *GradientBoostingRegressor* para construir um modelo aditivo de uma maneira progressiva, que permite a otimização de funções de perda diferenciáveis arbitrárias, ambos da biblioteca *Ensemble*.

## 3.4. Os Dados Analisados

Os dados são disponibilizados pela NASA e mantidos e monitorados pela NOAA [NOAA, 2022] periodicamente, desde março de 1958, sendo realizadas várias coletas por mês. Dessas coletas mensais, é feito uma média mensal, que totalizam 12 médias por ano. A base de dados compilados utilizado neste trabalho continham 795 registros, no momento da realização dos testes (maio de 2024). Cabe destacar que, para efeito de reprodutividade, trabalhou-se diretamente com os dados disponíveis em [NOAA, 2022]. Como supramencionado, mensalmente o site recebe um novo índice coletado. Logo, a base de dados altera-se periodicamente.

## 3.5. O Coeficiente de Determinação ($R^2$)

[Almeida, 2019] e [Montgomery e Runger, 2010] informam que o coeficiente de determinação ($R^2$) é uma medida estatística que indica a proporção da variância na variável dependente que é explicada pela(s) variável(eis) independente(s) em um modelo de regressão linear. A equação para $R^2$ é:

$$R^2 = \frac{SQreq}{SQtotal}. \tag{11}$$

No qual:

$$SQreq = \sum [G(x_i) - \bar{y}]^2. \tag{12}$$

$$SQtotal = \sum (y_i - \bar{y})^2. \tag{13}$$

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

$$\bar{y} = \frac{1}{n}.\sum y. \tag{14}$$

Com isso, $(0 <= R^2 <= 1)$, no qual $G(x_i)$ é o valor estimado no ponto (observado), $y_i$ é o valor tabelado, $\bar{y}$ é a média dos valores observados, e $n$ é o tamanho da amostra. Segundo [Montgomery e Runger, 2010] um valor de $R^2$ próximo de 1 indica que o modelo explica uma grande proporção da variância na variável dependente e, portanto, é considerado um bom ajuste. Por outro lado, um valor de $R^2$ próximo de 0 indica que o modelo não explica muita variância na variável dependente, sendo considerado um ajuste ruim.

### 3.6. Raiz do Erro Quadrático Médio (*Root Mean Squared Error - RMSE*)

Além do $R^2$, optou-se por outra forma confiável de aferição de modelos preditivos para mensurar a qualidade dos modelos empregados no caso, o RMSE. Este, é a uma das métricas mais amplamente usadas para aferição de acurácia de modelos preditivos de ML. Tal métrica parte da premissa simples, fazer a média dos erros (no qual subtrai-se o valor real de $y$ do valor predito de $y$) elevado ao quadrado, e por fim, tira-se a raiz da soma.

Pode-se melhor entender tal métrica ao expressá-la (Equação 15):

$$RMSE = \sqrt{\frac{1}{n}.\sum_{i=1}^{n}.(y_i - \hat{y}_i)^2}. \tag{15}$$

Segundo [Kumar et al., 2020] *"Enquanto a (Mean Squared Error - MSE) trabalha sobre viés e variância para previsões, o (Root Mean Square Error - RMSE) traz a unidade de volta à unidade original ao tomar a raiz quadrada do MSE. O RMSE é utilizado para medir a análise da discrepância entre os valores reais e estimados."*

Essa medida também tende a superestimar erros grandes, ao elevar os desvios ao quadrado para impedir que os desvios positivos e negativos se cancelem, o que pode ajudar a eliminar os métodos com esses erros. Nessa métrica, quanto menor for seu resultado, melhor a qualidade do modelo.

## 4. Resultados e Discussões

Nesta secção, apresenta-se os resultados obtidos neste estudo. Cabe salientar que, como forma de melhor apresentar os resultados obtidos, esses serão expostos por meio de tabelas, referente a cada um dos casos trabalhados. No repositório do *GitHub* do projeto [Ferreira, 2025] constam o tabelamento completo, gráficos e outros artefatos produzidos.

As métricas usadas nos modelos propostos foram o RMSE, para medir a análise da discrepância entre os valores reais e estimados [Kumar et al., 2020], e o $R^2$.

### 4.1. Resultados Obtidos

Os resultados obtidos a partir dos experimentos realizados estão apresentados a seguir. A **Tabela 1** mostra os resultados obtidos em todos os casos (regressão linear, logarítmica e exponencial). A **Tabela 2** apresenta os resultados aplicados ao conjunto de teste.

Observa-se que o modelo de regressão exponencial foi o que apresentou melhor ajuste, sugerindo que os níveis de $CO_2$ estão aumentando exponencialmente. Tais indicadores representam uma ameça aos seres vivos. Segundo [Pinto et al., 2010] e [Masson-Delmotte et al., 2019] a elevação da temperatura global em 2C$^o$ representa um ameça a biodiversidade. [Pinto et al., 2010] destaca a necessidade de manter as concentrações de $CO_2$ entre 350-400ppmv. Valores entre 540-970ppmv podem acarretar riscos severos de aumento da temperatura. Nas predições do PrediCtO2, a partir de fevereiro de 2059 a concentração de $CO_2$ ultrapassará 500ppmv.

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

Tabela 1: Resultados obtidos

| | **PrediCtO2** | ***SKLearn*** | **MMQ** |
|---|---|---|---|
| | | *Regressão Linear* | |
| **a** | 1.6742 | 1.6361 | 1.6361 |
| **b** | -2974.6782 | -2898.7762 | -2898.7762 |
| **R²** | 0.9757 | **0.9762** | **0.9762** |
| **RMSE** | 4.9418 | **4.8868** | **4.8868** |
| | | *Caso Logarítmico* | |
| **a** | 3345.5921 | 3255.7118 | 3255.7118 |
| **b** | -25043.7916 | -24373.8686 | -24373.6202 |
| **R²** | 0.8342 | **0.9750** | **0.9750** |
| **RMSE** | 12.8949 | **5.0067** | **5.0067** |
| | | *Caso Exponencial* | |
| **a** | 0.00293 | 0.00453 | 0.00453 |
| **b** | 0.04184 | 0.04347 | 0.04347 |
| **R²** | 0.8625463 | **0.9849** | **0.9849** |
| **RMSE** | 0.03235 | **0.0107** | 3.1792 |

Fonte: Autoria própria (2024).

### 4.2. Conjunto Teste

Como forma de demonstrar a eficiência do experimento, compara-se o valor real mensurado pela NOAA com as estimativas dos modelos empregados.

Ao final, realizou-se as predições propostas pelos modelos dos meses de janeiro, fevereiro, março e abril do ano de 2025, sublinhados na **Tabela 2**.

Como visto, apesar do PrediCtOr2 ter seus coeficientes diferentes dos outros dois modelos, ele foi o que melhor se aproximou dos valores reais nos últimos anos, com uma variação em média menor que 1ppmv (2020-2022) em seus resultados. Apesar de ter apresentado altas variações entre os anos de 1965 a 2015, chegando 9.31ppmv em 1995 (março de 1995 o valor real foi de 361.98ppmv, contra a estimativa de 371.29ppmv). O tabelamento completo pode ser encontrado no repositório do projeto [Ferreira, 2025].

Já com os modelos MMQ e *SKLearn*, houve inversão. Eles começaram fazendo estimativas baixas, aproximaram-se do valor real no período entre 1970 a 1995 (quando voltar a ficar com variações acima de 2ppmv) e voltaram a se distanciar dos valores reais a partir de 2010, tendo suas estimativas para o último período tabelado, março de 2023, com uma variação de -7.62ppmv (valor real 420.99ppmv contra a estimativa de 413.37ppmv, de ambos os modelos). As melhores estimativas, no tabelamento, se deu no período de 1985 (março), cuja a variação foi de apenas 0.12ppmv (valor real foi de 347.91ppmv, contra a estimativa de 348.03ppmv) para MMQ e 0.05ppmv (valor real foi de 347.91ppmv, contra a estimativa de 347.96ppmv) no *SKLearn*.

## 5. Considerações Finais

Neste trabalho, foram utilizados, considerando os dados de $CO_2$ disponíveis, técnicas para a modelagem dos dados e predição dos mesmos, considerado o MMQ e dois modelos de *Machine learning*, sendo uma aplicação implementada via teoria (PrediCtOr2) e outro por meio de biblioteca *SKLearn*.

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

Tabela 2: Conjunto teste

| **Período** | **Valor real** | *Caso exponencial* | | |
|---|---|---|---|---|
| | | **PrediCtOr2** | **SKLearn** | **MMQ** |
| **03/1958** | 315.70 | **315.70** | 308.22 | 307.97 |
| **03/1960** | 317.58 | **318.48** | 311.01 | 310.78 |
| **03/1965** | 320.89 | 325.53 | **318.07** | 317.89 |
| **03/1970** | 326.93 | 332.75 | **325.29** | 325.17 |
| **03/1975** | 331.94 | 340.13 | 332.68 | **332.62** |
| **03/1985** | 347.91 | 355.37 | **347.96** | 348.03 |
| **03/1995** | 361.98 | 371.29 | **363.94** | 364.15 |
| **03/2015** | 401.74 | 405.31 | 398.14 | **398.66** |
| **03/2020** | 414.74 | **414.29** | 407.18 | 407.79 |
| **03/2022** | 418.81 | **417.94** | 410.85 | 411.50 |
| **03/2023** | 420.99 | **424.63** | 413.37 | 413.37 |
| **PREDIÇÕES ESTIMADAS** | | | | |
| **01/2025** | *426.65* | *428.19* | *416.81* | *416.81* |
| **02/2025** | *427.09* | *428.35* | *416.97* | *416.97* |
| **03/2025** | *428.15* | *428.51* | *417.13* | *417.13* |
| **04/2025** | *NC*[1] | *428.68* | *417.29* | *417.29* |

Fonte: Autoria própria (2024).

Os resultados obtidos com os métodos foram, em geral, satisfatórios. Nesse contexto, pode-se destacar o caso exponencial em todos os modelos trabalhados. O PrediCtOr2, demonstrou-se eficiente em suas predições atuais, tendo coeficiente de determinação ($R^2$) mais baixo do que os demais modelos, porém obteve excelente índice $RMSE$ no caso exponencial.

De forma geral o experimento se mostrou satisfatório, haja vista, os três modelos propostos tiveram bons resultados na predição das estimativas. Os modelos conseguiram mostrar sua eficácia em predições. O que prova que tais modelos são eficientes para tomada de decisões, sendo esse nosso objeto de estudo.

Com isso, espera-se que esse estudo sirva de base para novas pesquisas e validações dos modelos aqui empregados, bem como outros de mesma eficiência.

## Referências

Almeida, C. G. d. (2019). *Cálculo Numérico*. UFU. URL `https://repositorio.ufu.br/bitstream/123456789/25218/1/Calculo%20Numerico.pdf`.

Amarpuri, L., Yadav, N., Kumar, G., e Agrawal, S. (2019). Prediction of co2 emissions using deep learning hybrid approach: A case study in indian context. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, p. 1–6.

Bishop, C. M. e Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer, 1 edition.

Box, G. E., Jenkins, G. M., Reinsel, G. C., e Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons, 5 edition. ISBN 978-1-118-67502-1.

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br

Ferreira, D. F. P. (2025). Predictor2. GitHub repository. URL https://github.com/David-invictus/PrediCtOr2.git. Repositório de software no GitHub.

Houghton, R. A. e Woodwell, G. M. (1989). Global climatic change. *Scientific American*, 260(4): 36–47. ISSN 00368733, 19467087. URL http://www.jstor.org/stable/24987210.

Kumar, R., Kumar, P., e Kumar, Y. (2020). Time series data prediction using iot and machine learning technique. *Procedia Computer Science*, 167:373–381. ISSN 1877-0509. URL https://www.sciencedirect.com/science/article/pii/S1877050920307067. International Conference on Computational Intelligence and Data Science.

Kumari, S. e Singh, S. K. (2023). Machine learning-based time series models for effective co2 emission prediction in india. *Environmental Science and Pollution Research*, 30(55):116601–116616. URL https://doi.org/10.1007/s11356-022-21723-8.

Li, L., Lei, Y., He, C., Wu, S., e Chen, J. (2017). Prediction on the peak of the c [o. sub. 2] emissions in china using the stirpat model. *Advances in Meteorology*, 2017. URL https://doi.org/10.1155/2016/5213623.

Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P. R., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Connors, S., Matthews, J. B. R., Chen, Y., Zhou, X., Gomis, M. I., Lonnoy, E., Maycock, T., Tignor, M., e Waterfield, T. (2019). Aquecimento global de 1,5°c: Sumário para formuladores de políticas. Relatório Especial do IPCC sobre o Aquecimento Global de 1,5°C, Versão em Português. URL https://www.ipcc.ch/site/assets/uploads/2019/07/SPM-Portuguese-version.pdf. Tradução não oficial para o português. Relatório original publicado em 2018.

Montgomery, D. C. e Runger, G. C. (2010). *Applied statistics and probability for engineers*. John wiley & sons, 5 edition. ISBN 13: 978-0-470-05304-1.

Nações Unidas (2022). O que são as mudanças climáticas? URL https://brasil.un.org/pt-br/175180-o-que-sao-mudancas-climaticas.

NOAA (2022). Co2 measurements. Noaa.gov. URL https://gml.noaa.gov/webdata/ccgg/trends/co2/co2_mm_mlo.txt.

Pinto, E. d. P. P., Moutinho, P., Stella, O., Castro, I., Mazer, S., Rettmann, R., e Moreira, P. F. (2010). Perguntas e respostas sobre aquecimento global. Technical report, Instituto de Pesquisa Ambiental da Amazônia, Belém/PA. URL https://cmsdespoluir.cnt.org.br/Documents/PDFs/Perguntas%20e%20respostas%20sobre%20AQUECIMENTO%20GLOBAL.pdf.

Silva, H. A. P. D. (2022). Simulação da elevação do nível médio do mar utilizando métodos numéricos e ferramentas de geoprocessamento: estudo de caso. Master's thesis, Universidade Federal Rural do Semi-Árido (UFERSA), Mossoró/RN. URL https://repositorio.ufersa.edu.br/handle/prefix/9072.

Taheri, S. e Razban, A. (2021). Learning-based co2 concentration prediction: Application to indoor air quality control using demand-controlled ventilation. *Building and Environment*, 205:108164. ISSN 0360-1323. URL https://www.sciencedirect.com/science/article/pii/S0360132321005655.

**DOI:** 10.59254/sbpo-2025-212656
https://proceedings.science/p/212656?lang=pt-br