Projeto de Formatura – 2023

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia de Computação

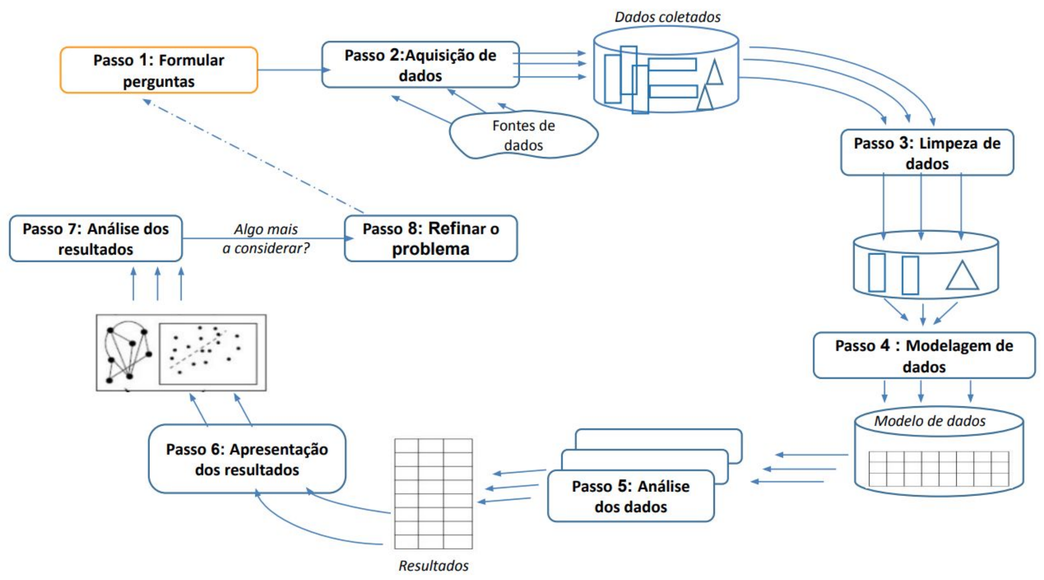
Tema: **Aplicação de modelos de Aprendizado de Máquina na estimação da coluna troposférica de NO2 do TROPOMI no estado do Pará**

Objetivo

A Amazônia desempenha um papel crítico no sistema climático global e o NO2, um importante poluente atmosférico, afeta a qualidade do ar. Devido às frequentes nuvens na região, a obtenção de dados por sensoriamento remoto é desafiadora. Assim, esse estudo tem como objetivo:

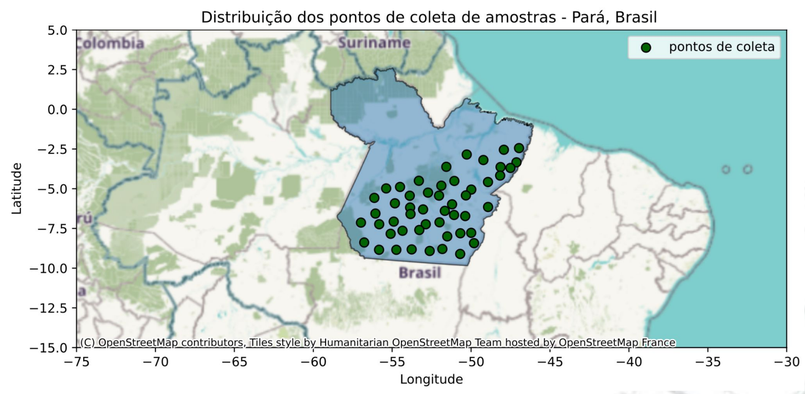
- Desenvolver um modelo para estimar a concentração de NO2 na Amazônia
- Gerar um método de imputação de dados para a região.

O projeto foi desenvolvido seguindo um ciclo de projeto de ciência de dados em oito etapas.

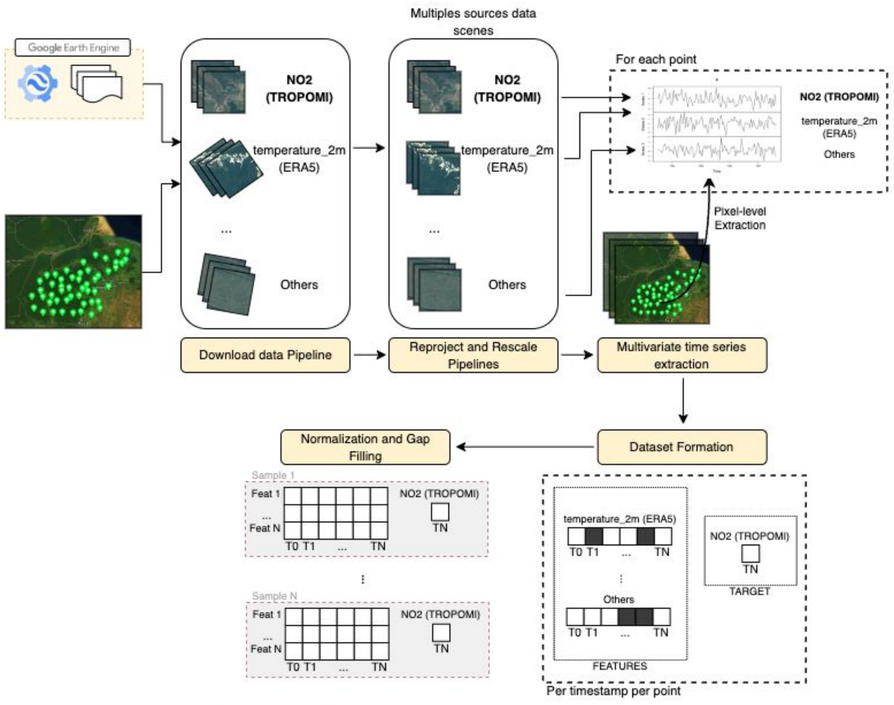


Métodos e Procedimentos

- **Estudo de caso**: 50 pontos aleatórios no Pará, Brasil, utilizados para a seleção de variáveis de sensoriamento remoto.



- Uso da API do Google Earth Engine em Python para adquirir os dados.
- Período de coleta de dados: Julho de 2017 a Abril de 2023.



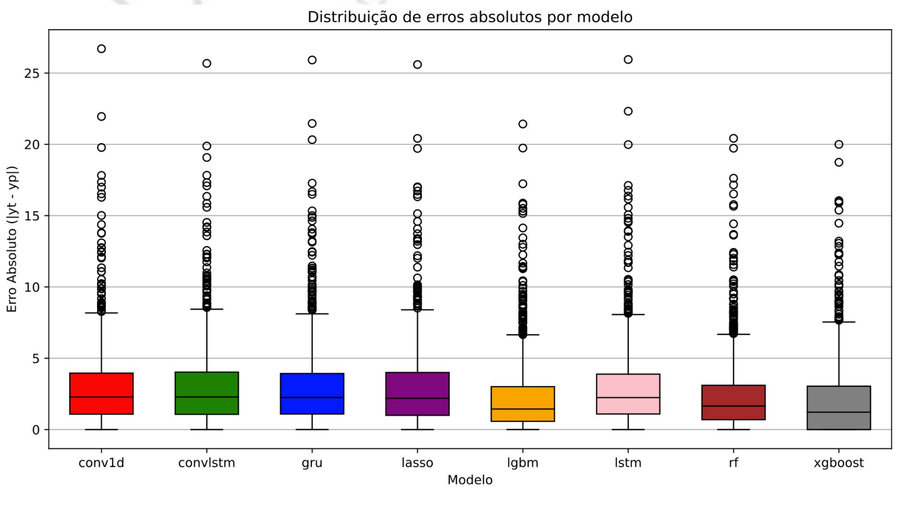
Pipeline de aquisição e pré-processamento dos dados:

- Uniformização dos dados para resolução de 11 km.
- Construção do conjunto de dados com séries temporais multivariadas.
- Imputação de valores ausentes pela média.
- Normalização das variáveis independentes.

Resultados

A tabela a seguir mostra os resultados obtidos no conjunto de dados de teste.

Modelo	R^2	r	MSE	RMSE	MAE
Lasso	0.23	0.48	14.70	3.83	2.84
RF	0.43	0.66	10.97	3.31	2.30
XGBoost	**0.47**	**0.68**	**10.17**	**3.19**	**2.00**
LightGBM	0.44	0.66	10.66	3.26	2.20
Conv1D	0.21	0.46	15.16	3.89	2.90
GRU	0.21	0.45	15.15	3.89	2.88
LSTM	0.22	0.47	14.90	3.86	2.86
ConvLSTM	0.19	0.43	15.53	3.94	2.92

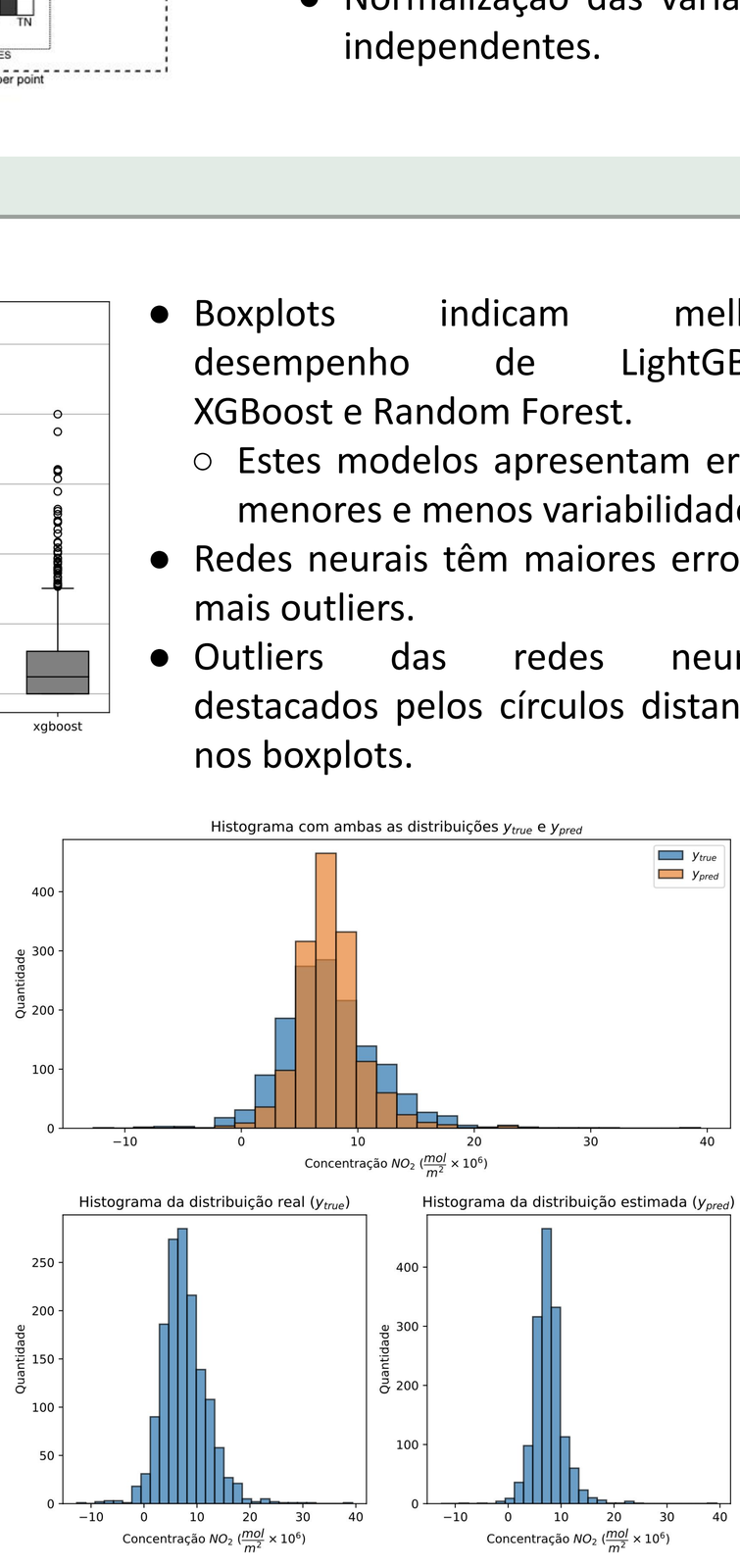


- Boxplots indicam melhor desempenho de LightGBM, XGBoost e Random Forest.
  - Estes modelos apresentam erros menores e menos variabilidade.
- Redes neurais têm maiores erros e mais outliers.
- Outliers das redes neurais destacados pelos círculos distantes nos boxplots.

- Modelos de árvore de decisão (Random Forest, XGBoost, LightGBM) superaram modelos lineares e redes neurais.
- Avaliação baseada em métricas R^2, r, MSE, RMSE, e MAE.
- XGBoost com melhor R^2 e menores pontuações em r, MSE, RMSE, e MAE.

Histograma do melhor modelo (XGBoost)

- Histograma superior mostra sobreposição das distribuições reais (azul) e estimadas (laranja) de NO2.
- Formas das distribuições são similares, indicando previsões alinhadas com os dados reais.
- Maioria dos dados preditos concentrados entre 5 e 15, com a distribuição real mostrando maior dispersão.



Integrantes: Wesley Pereira de Almeida

Professor Orientador: Pedro Luiz Pizzigatti Corrêa

Coorientadora: Luciana Varanda Rizzo