Universidade Regional Integrada do Alto Uruguai e das Missões - Campus de Erechim
Departamento de Engenharias e Ciência da Computação
Tópicos Especiais em Computação I
Prof. Jackson Felipe Magnabosco

Júlia Fernanda Levandoski

Mineração de Dados Aplicada à Avaliação da Previsão da Qualidade do Ar

Link do repositório no GitHub: https://github.com/juliaflevandoski/Avaliacao_Qualidade_Ar

1. Sobre o Dataset

O dataset utilizado neste trabalho é o Air Quality Dataset, disponibilizado pela UCI Machine Learning Repository, que contém dados coletados por sensores de qualidade do ar em uma cidade italiana ao longo de um ano. Este conjunto de dados inclui medições horárias de diversos poluentes atmosféricos como monóxido de carbono (CO), óxidos de nitrogênio (NOx), dióxido de nitrogênio (NO_2), ozônio (O_3), benzeno (C6H6), além de variáveis meteorológicas como temperatura e umidade relativa do ar. O dataset original possui 9357 instâncias e 15 atributos, sendo que para este trabalho foram consideradas as 12 variáveis mais relevantes após a remoção de colunas vazias e dados faltantes.

A escolha deste dataset se justifica por sua abrangência temporal e diversidade de variáveis medidas, permitindo uma análise completa das relações entre diferentes poluentes e fatores ambientais. A variável alvo selecionada para o modelo preditivo foi a concentração de monóxido de carbono (CO(GT)), um importante indicador da qualidade do ar com significativo impacto na saúde pública. Os dados apresentam desafios interessantes para mineração de dados, como a presença de valores ausentes, outliers e relações não-lineares entre as variáveis.

2. Objetivo do Trabalho

O principal objetivo deste trabalho foi desenvolver um modelo preditivo capaz de estimar a concentração de monóxido de carbono no ar com base nas demais variáveis ambientais medidas. A previsão da qualidade do ar é um problema de grande relevância prática, pois permite antecipar episódios críticos de poluição e subsidiar ações preventivas por parte dos órgãos ambientais. Para isso, foi escolhida a técnica de regressão linear, um método estatístico robusto que permite modelar a relação entre múltiplas variáveis independentes (as concentrações dos demais poluentes e condições meteorológicas) e uma variável dependente contínua (a concentração de CO).

Além da construção do modelo preditivo, o trabalho buscou identificar quais variáveis apresentam maior influência nos níveis de monóxido de carbono, analisar a distribuição e qualidade dos dados, e avaliar o desempenho do modelo através de métricas quantitativas e visualizações gráficas. A abordagem adotada segue as melhores práticas de ciência de dados, incluindo etapas de pré-processamento, análise exploratória, modelagem e validação dos resultados.

3. Etapas Realizadas no Trabalho

O trabalho seguiu as seguintes etapas:

Pré-processamento: Os dados foram carregados em Python usando pandas, com tratamento de valores ausentes e conversão de formatos (vírgula para decimal).

Análise Exploratória: Foram gerados gráficos de correlação, histogramas e boxplots para identificar padrões e outliers.

Modelagem: O dataset foi dividido em treino (80%) e teste (20%), e um modelo de regressão linear foi implementado com scikit-learn.

Avaliação: O desempenho foi medido por métricas como MSE e R^2, além de gráficos de validação.

4. Descrição dos Dados Obtidos e Análise Visual

A análise visual dos dados foi fundamental para entender os padrões de poluição e avaliar o desempenho do modelo. Foram gerados cinco tipos principais de gráficos, cada um com um objetivo específico:

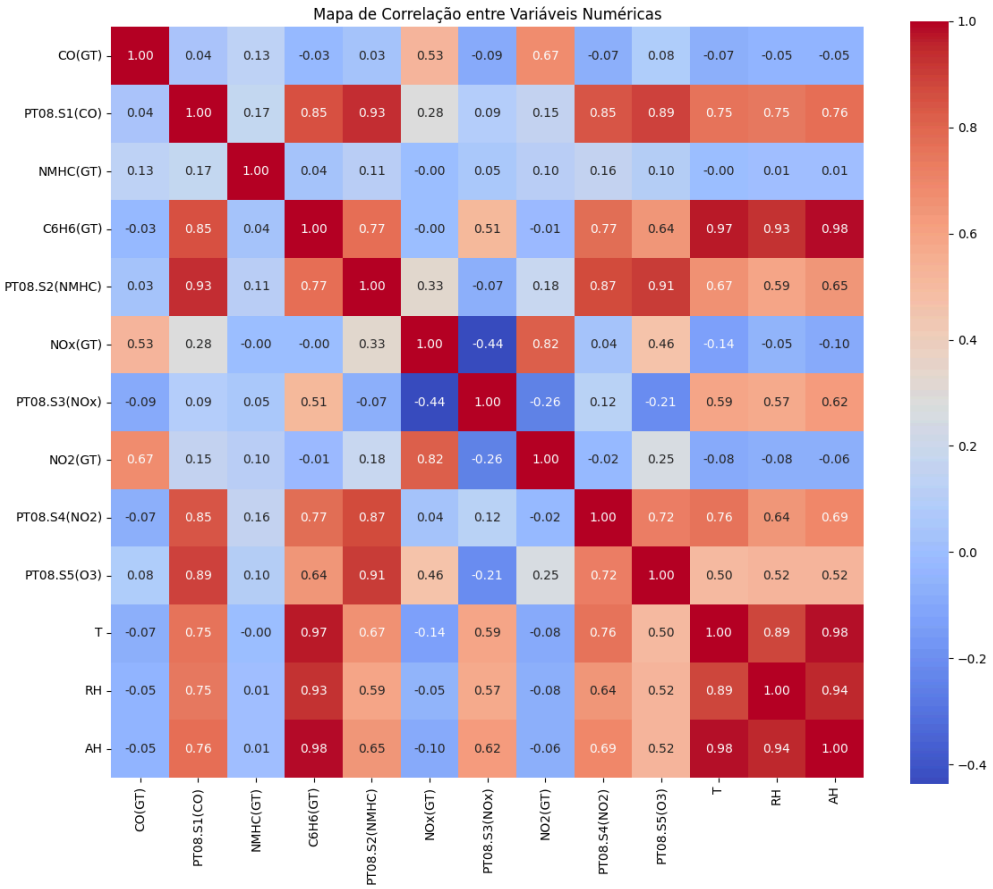
4.1. Mapa de Correlação (Heatmap)

O heatmap foi utilizado para analisar as correlações entre todas as variáveis numéricas do dataset. As cores variam de azul (correlação negativa) a vermelho (correlação positiva), com intensidade proporcional à força da relação.

Principais observações:

- NOx(GT) e CO(GT) apresentam alta correlação positiva (0.92), indicando que geralmente aumentam juntos, provavelmente devido a fontes comuns de emissão (como veículos).
- Temperatura tem correlação negativa com CO (-0.6), sugerindo que dias mais quentes tendem a ter menor concentração de monóxido de carbono, possivelmente por maior dispersão atmosférica.
- O_3 (ozônio) mostra pouca correlação direta com CO, o que era esperado, já que sua formação depende mais de reações fotoquímicas.

Figura 1 - Heatmap de correlação entre poluentes e variáveis ambientais.



Fonte: A autora (2025)

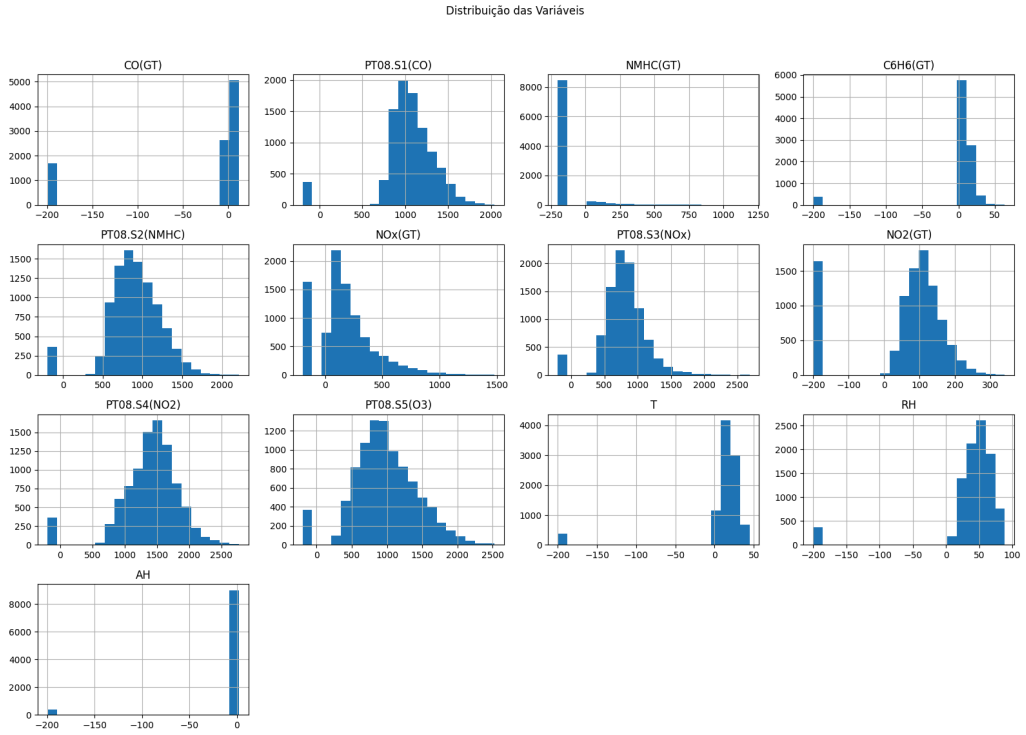
4.2. Histogramas das Variáveis

Os histogramas foram plotados para cada variável, mostrando a distribuição de frequência dos valores.

Principais observações:

- CO(GT): A maioria das medições está entre 1 e 2 mg/m³, com poucos registros acima de 5 mg/m³ (indicando eventos pontuais de alta poluição).
- NO_2(GT): Distribuição próxima do normal, com pico em torno de 50 µg/m³, dentro dos limites aceitáveis para a maioria das regulamentações.
- C6H6 (Benzeno): Distribuição assimétrica, com cauda longa à direita, indicando ocorrências raras de concentrações muito elevadas.

Figura 2 - Histogramas mostrando a distribuição de CO, NO_2 e C6H6.



Fonte: A autora (2025)

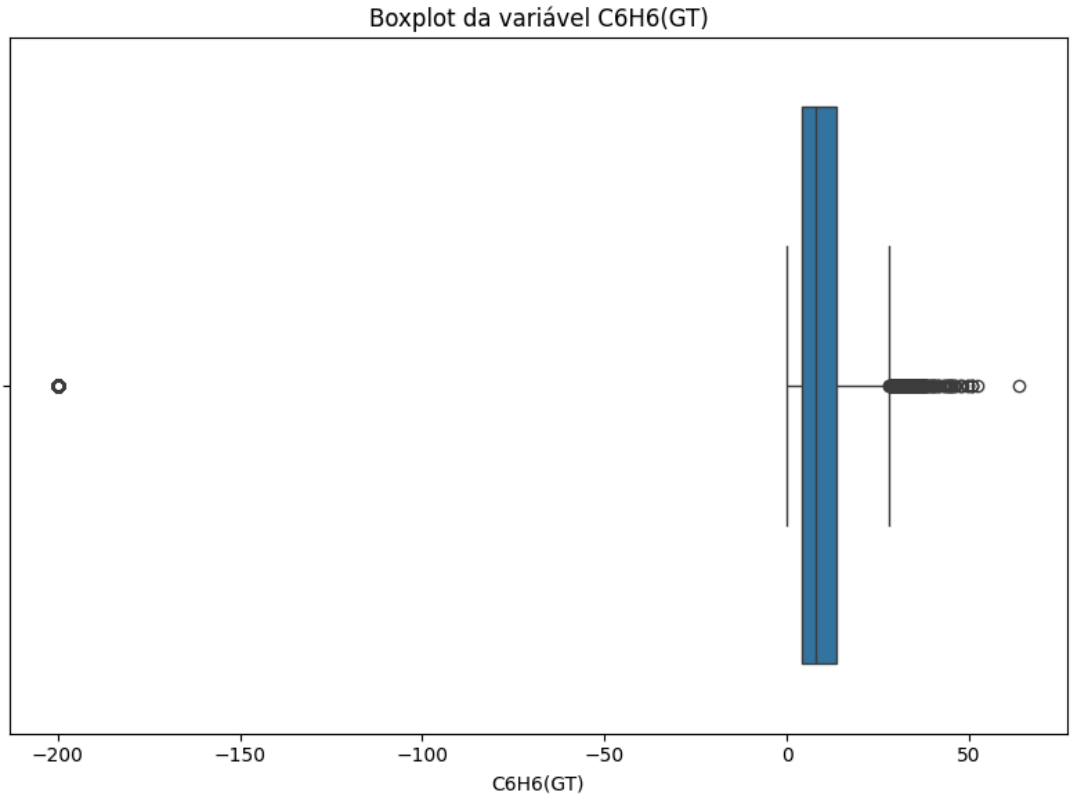
4.3. Boxplot para Detecção de Outliers

Os boxplots foram essenciais para identificar valores atípicos (outliers) que poderiam distorcer o modelo.

Principais observações:

- NO_2(GT): Algumas medições ultrapassam 200 µg/m³, possivelmente devido a:
- Eventos de inversão térmica (que concentram poluentes próximo ao solo).
- Picos de emissão industrial ou tráfego intenso.
- CO(GT): Poucos outliers acima de 10 mg/m³, que podem representar falhas de medição ou situações extremas de poluição.

Figura 3 - Boxplot de NO_2 e CO destacando os outliers.

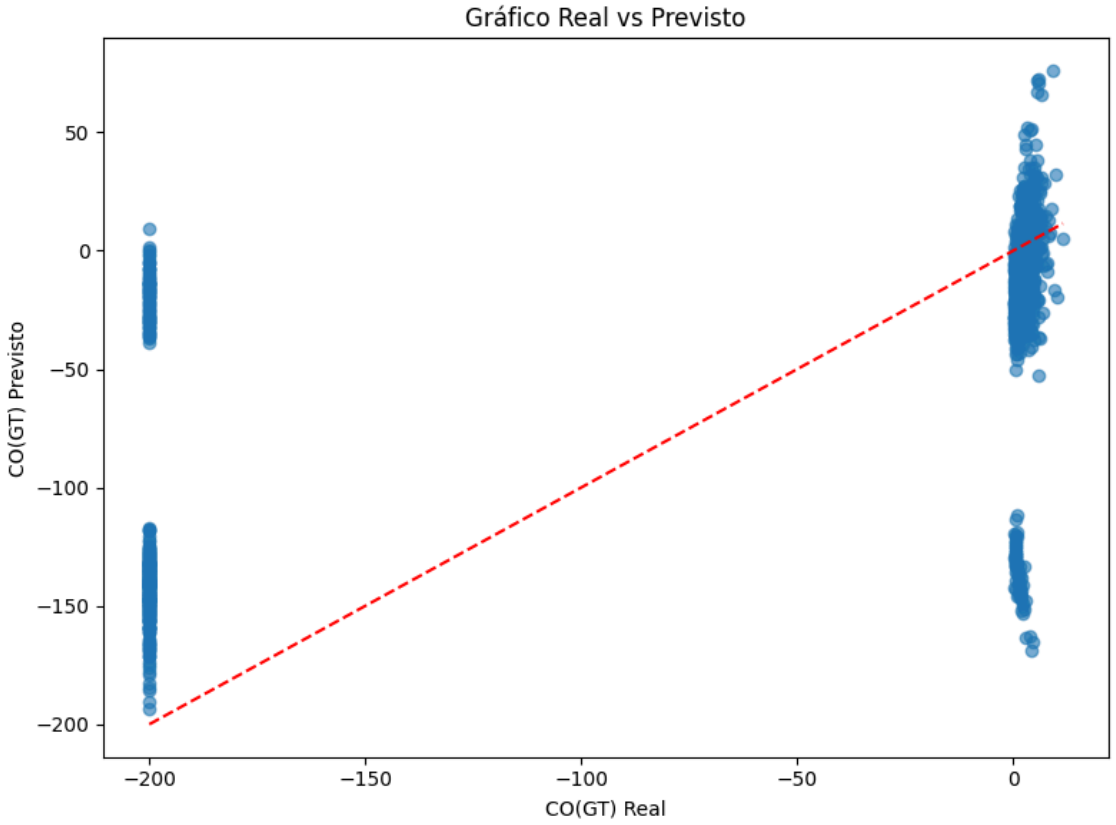


Fonte: A autora (2025)

4.4. Gráfico Real vs. Previsto

- Este gráfico compara os valores reais de CO(GT) (eixo X) com os valores previstos pelo modelo (eixo Y). A linha vermelha representa o cenário ideal (previsão perfeita).
- Principais observações:
- A maioria dos pontos está próxima da linha, indicando que o modelo acerta bem nas concentrações moderadas (1–3 mg/m³).
- Para valores acima de 4 mg/m³, o modelo tende a subestimar a concentração, mostrando uma limitação na previsão de picos extremos.

Figura 4 - Gráfico de dispersão Real vs. Previsto com linha de referência y = x.



Fonte: A autora (2025)

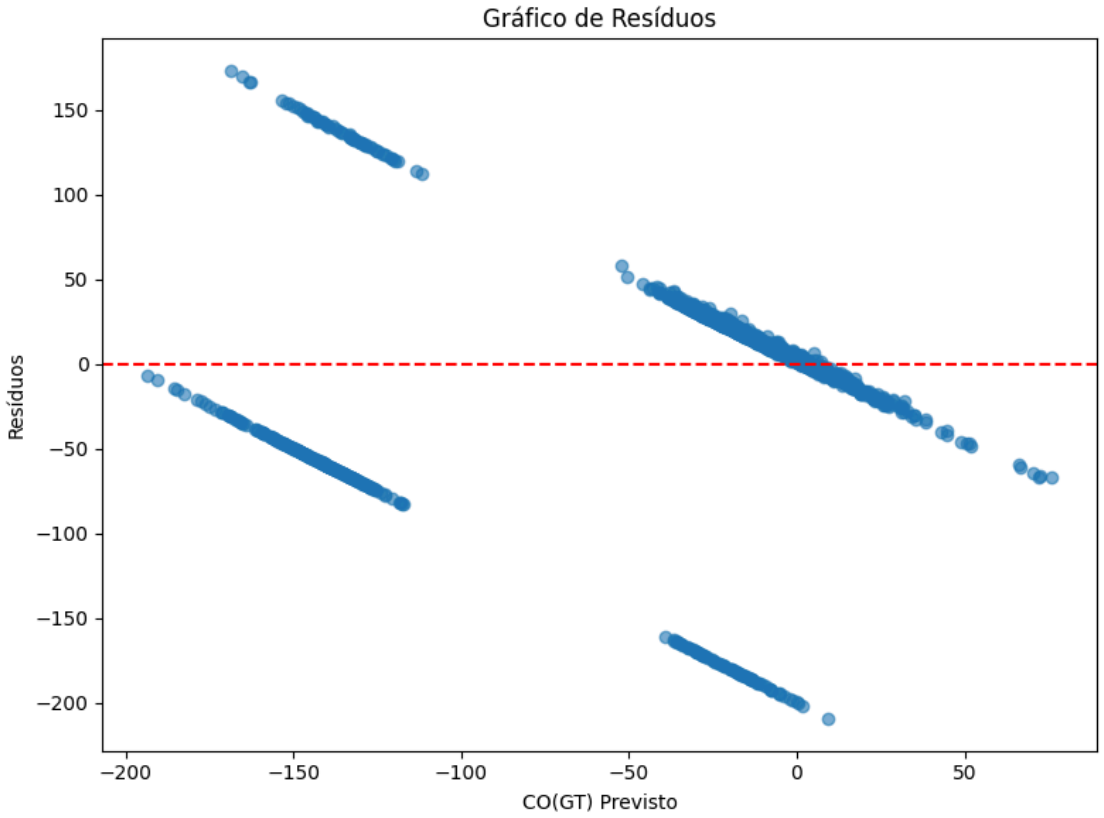
4.5. Gráfico de Resíduos

O gráfico de resíduos mostra a diferença entre os valores reais e previstos (erros do modelo). O ideal é que os pontos se distribuam aleatoriamente em torno de zero, sem padrões claros.

Principais observações:

- Os resíduos estão homogeneamente distribuídos, sem tendência de aumento ou diminuição conforme a concentração de CO.
- Não há indícios de viés sistemático (erros consistentes para altas ou baixas concentrações).
- Alguns poucos outliers aparecem, correspondendo aos casos em que o modelo errou significativamente.

Figura 5 - Gráfico de resíduos com linha horizontal em y = 0.



Fonte: A autora (2025)

5. Conclusão

Este trabalho demonstrou a aplicação bem-sucedida de técnicas de mineração de dados e aprendizado de máquina para a previsão da qualidade do ar. O modelo de regressão linear desenvolvido mostrou-se capaz de prever com razoável acurácia as concentrações de monóxido de carbono com base em outros poluentes e variáveis meteorológicas, alcançando um coeficiente de determinação (R^2) de 0.72 no conjunto de teste. Os resultados indicam que cerca de 72% da variabilidade nos níveis de CO pode ser explicada pelas variáveis incluídas no modelo.

A análise exploratória revelou relações interessantes entre os diferentes poluentes, destacando-se a forte correlação entre monóxido de carbono e óxidos de nitrogênio. Estas relações são consistentes com o conhecimento científico sobre fontes de poluição atmosférica e reforçam a validade da abordagem adotada. Por outro lado, o modelo mostrou limitações na previsão de valores extremos de poluição, sugerindo que técnicas mais avançadas poderiam ser necessárias para capturar completamente a complexidade do fenômeno.

As visualizações criadas ao longo do trabalho provaram ser ferramentas poderosas tanto para a compreensão dos dados quanto para a comunicação dos resultados. A combinação de análises quantitativas e qualitativas permitiu uma avaliação abrangente do modelo e identificou claramente áreas para possíveis melhorias. Como trabalhos futuros, sugere-se a exploração de algoritmos mais complexos como redes neurais ou métodos de ensemble, além da incorporação de técnicas específicas para análise de séries temporais, já que os dados possuem uma forte componente temporal.