UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS JARDINS DE ANITA

JOÃO LUIS FEITOSA LEITE

RELATÓRIO DE INTELIGÊNCIA ARTIFICIAL

Prova II

ITAPAJÉ

2024

SUMÁRIO

1 INTRODUÇÃO

Este relatório apresenta uma análise detalhada sobre a aplicação de diferentes técnicas de regressão para prever as emissões de CO_2 na Tailândia, utilizando dados históricos de 1987 a 2022. O objetivo central foi desenvolver modelos de regressão que pudessem explicar e prever as emissões de CO_2 com base em variáveis relacionadas, como tipo de combustível e fonte de emissão. A análise foi conduzida em quatro etapas principais:

1. **Identificação do Atributo-Alvo e Atributos Relevantes**: Nesta fase, o foco foi em determinar qual variável deveria ser o alvo da regressão, isto é, a variável que desejamos prever. Além disso, foram analisadas as variáveis independentes para determinar quais delas eram mais relevantes para prever o alvo selecionado.
2. **Implementação de Regressão Linear**: Após a seleção dos atributos mais importantes, foi implementada uma regressão linear simples. O desempenho do modelo foi avaliado usando métricas padrão, como Erro Quadrático Médio (MSE) e R2R^2R2, para verificar a qualidade do ajuste.
3. **Regularização com Lasso e Ridge**: Para aprimorar o modelo e evitar problemas de sobreajuste, aplicamos regularizações Lasso e Ridge. Utilizamos a técnica de busca em grade (GridSearchCV) para identificar os melhores hiperparâmetros para esses modelos.
4. **Validação Cruzada e Comparação de Desempenho**: Por fim, aplicamos uma regressão linear com validação cruzada (k-fold cross-validation) e comparamos o desempenho entre a regressão linear simples, Lasso e Ridge, para determinar o modelo mais adequado.

O conjunto de dados utilizado continha informações mensais sobre as emissões de CO_2, categorizadas por fonte de emissão e tipo de combustível. Essas variáveis, juntamente com o ano e o mês, foram analisadas para identificar padrões e possíveis relações com o volume de emissões. A análise não apenas buscou construir modelos preditivos, mas também entender as variáveis que mais influenciam as emissões de CO_2 no país ao longo do tempo.

2 OBSERVAÇÕES

Durante a execução deste projeto, surgiram diversos desafios, tanto técnicos quanto metodológicos, que impactaram o andamento e as decisões tomadas ao longo da análise.

Desafios Técnicos

- **Pré-processamento de Dados**: O conjunto de dados, embora completo, exigiu um trabalho considerável de preparação e limpeza. As variáveis categóricas, como o tipo de combustível e a fonte de emissão, precisaram ser codificadas adequadamente para que pudessem ser usadas nos modelos de regressão. Optamos pelo uso de técnicas de codificação como **one-hot encoding**, mas isso aumentou significativamente a dimensionalidade dos dados, tornando o processamento mais pesado e aumentando o tempo de execução dos modelos.
- **Normalização e Escalonamento**: Durante a implementação dos modelos, percebemos a necessidade de normalizar as variáveis independentes, dado que algumas delas tinham escalas muito diferentes. Foi utilizada a técnica de **StandardScaler** para garantir que as variáveis tivessem média zero e desvio padrão unitário, o que ajudou na convergência dos modelos de regressão regularizada (Lasso e Ridge).
- **Computação**: As técnicas de validação cruzada e busca de hiperparâmetros com **GridSearchCV** exigiram um tempo considerável de processamento, principalmente em combinação com a alta dimensionalidade dos dados após a codificação das variáveis categóricas. Isso tornou a execução do processo mais lenta do que o esperado, especialmente para o modelo de Ridge, que se mostrou mais sensível a ajustes nos hiperparâmetros.

Desafios Metodológicos

- **Escolha das Variáveis**: Um dos desafios mais significativos foi a seleção das variáveis mais relevantes para o modelo de regressão. Embora a correlação entre as variáveis numéricas (como ano e mês) e as emissões de CO_2 tenha sido relativamente fraca, as variáveis categóricas (especialmente o tipo de combustível) mostraram-se mais influentes, exigindo técnicas de análise diferenciadas para determinar sua relevância.

- **Regularização**: Ao implementar os modelos Lasso e Ridge, percebemos que a escolha dos parâmetros de regularização era crucial para evitar o sobreajuste e ao mesmo tempo preservar a interpretabilidade dos coeficientes. O Lasso, por exemplo, zerou alguns coeficientes, o que indicava a eliminação de variáveis menos relevantes, mas exigiu cuidado na interpretação dos resultados.

Limitações

- **Dimensionalidade**: A codificação de variáveis categóricas aumentou muito a dimensionalidade dos dados, o que pode ter afetado a performance do modelo. Com dados de alta dimensionalidade, é mais difícil encontrar um bom equilíbrio entre ajuste e regularização, o que foi particularmente evidente no modelo Ridge.
- **Tempo de Execução**: O uso de GridSearchCV, combinado com a validação cruzada (k-fold), demandou um tempo significativo de processamento, o que limitou a quantidade de iterações possíveis na busca pelos melhores hiperparâmetros.

Apesar dessas dificuldades, o projeto foi executado com sucesso, e os resultados foram satisfatórios, mostrando que a regressão regularizada (principalmente o Lasso) conseguiu capturar a relevância das variáveis e melhorar o ajuste do modelo.

3 METODOLOGIA

A metodologia adotada neste projeto foi estruturada em quatro questões principais, cada uma abordando uma fase crítica do processo de desenvolvimento dos modelos de regressão, desde a análise exploratória até a validação cruzada dos modelos.

3.1 *Questão 1: Identificação do Atributo-Alvo e Relevância*

Nesta fase inicial, o objetivo foi determinar qual variável deveria ser o **alvo da regressão** e identificar as **variáveis independentes** mais relevantes para prever o valor do alvo.

O conjunto de dados fornecido continha cinco colunas:

- **year** (ano),
- **month** (mês),
- **source** (fonte de emissão),
- **fuel_type** (tipo de combustível),
- **emissions_tons** (emissões de CO_2 em toneladas).

Analisamos a correlação entre as variáveis numéricas e o atributo alvo, "emissions_tons". As variáveis numéricas **year** e **month** apresentaram correlações baixas com as emissões de CO_2. No entanto, ao analisar as variáveis categóricas **source** e **fuel_type**, identificamos uma relação mais clara com as emissões. Através de gráficos boxplot, observamos que diferentes tipos de combustível e fontes de emissão influenciavam diretamente o nível de emissões de CO_2.

Com base nesta análise exploratória, escolhemos o **tipo de combustível** como a variável mais relevante para prever as emissões de CO_2, uma vez que mostrou maior impacto sobre o alvo durante a visualização dos dados. Esta variável foi convertida em dummies para ser utilizada na regressão.

3.2 *Questão 2: Regressão Linear Usando o Atributo Mais Relevante*

Após identificar o atributo mais relevante, implementamos uma **regressão linear simples** usando a variável **fuel_type** para prever as emissões de CO_2. Inicialmente, dividimos o conjunto de dados em treino e teste (70% para treino e 30% para teste), para garantir uma avaliação adequada do modelo.

Utilizamos a função **LinearRegression()** da biblioteca **sklearn** para ajustar o modelo com os dados de treino. O modelo foi então avaliado com os dados de teste, calculando métricas de erro, como:

- **RSS** (Residual Sum of Squares),
- **MSE** (Mean Squared Error),
- **RMSE** (Root Mean Squared Error),
- **R2** (Coeficiente de Determinação).

Além disso, geramos gráficos da reta de regressão junto com a nuvem de pontos dos valores reais, comparando os valores preditos com os valores observados. Essa visualização permitiu uma análise gráfica do desempenho do modelo, mostrando como a regressão linear foi capaz de capturar a relação entre o tipo de combustível e as emissões de CO_2, embora alguns outliers tenham sido detectados.

3.3 *Questão 3: Regularização com Lasso e Ridge*

Na terceira questão, buscamos melhorar o desempenho do modelo e evitar o sobreajuste (overfitting) aplicando técnicas de **regularização**. Duas abordagens foram consideradas:

- **Lasso (Least Absolute Shrinkage and Selection Operator)**,
- **Ridge (Tikhonov regularization)**.

Esses métodos foram utilizados para ajustar a regressão linear com penalizações que ajudam a reduzir o impacto de variáveis irrelevantes (no caso do Lasso, algumas variáveis podem ter coeficientes zerados, sendo eliminadas do modelo). Utilizamos a técnica de **GridSearchCV** para realizar uma busca pelos melhores valores de **alpha** (o parâmetro de regularização) tanto para Lasso quanto para Ridge.

O **GridSearchCV** foi configurado para realizar uma busca sobre um conjunto de valores de alpha (variando de 0.01 a 1000), usando validação cruzada com 5 folds. Para cada modelo, os seguintes resultados foram registrados:

- O valor de **alpha** que minimizou o erro,
- O **MSE** correspondente,
- O **R2R^2R2**.

Essa abordagem permitiu encontrar o equilíbrio ideal entre ajuste e regularização, minimizando o erro de previsão enquanto evitava que o modelo ajustasse ruído nos dados.

3.4 *Questão 4: Comparação com Regressão Linear Usando KFold*

Na última etapa, implementamos uma **regressão linear simples** utilizando a técnica de **k-fold cross-validation**, com 5 divisões (folds). O objetivo foi comparar o desempenho da regressão linear com os modelos regularizados (Lasso e Ridge) implementados anteriormente.

A validação cruzada (cross-validation) foi crucial para garantir que o modelo fosse avaliado de maneira robusta, utilizando diferentes subconjuntos dos dados para treino e teste. Esta técnica reduz o risco de o modelo ser ajustado excessivamente a um subconjunto específico dos dados e permite uma avaliação mais representativa de seu desempenho.

Após obter os resultados da regressão linear com k-fold, compararam-se os valores de **MSE** e **R2R^2R2** médios com aqueles obtidos nos modelos Lasso e Ridge, identificando o melhor modelo para o problema de previsão das emissões de CO_2.

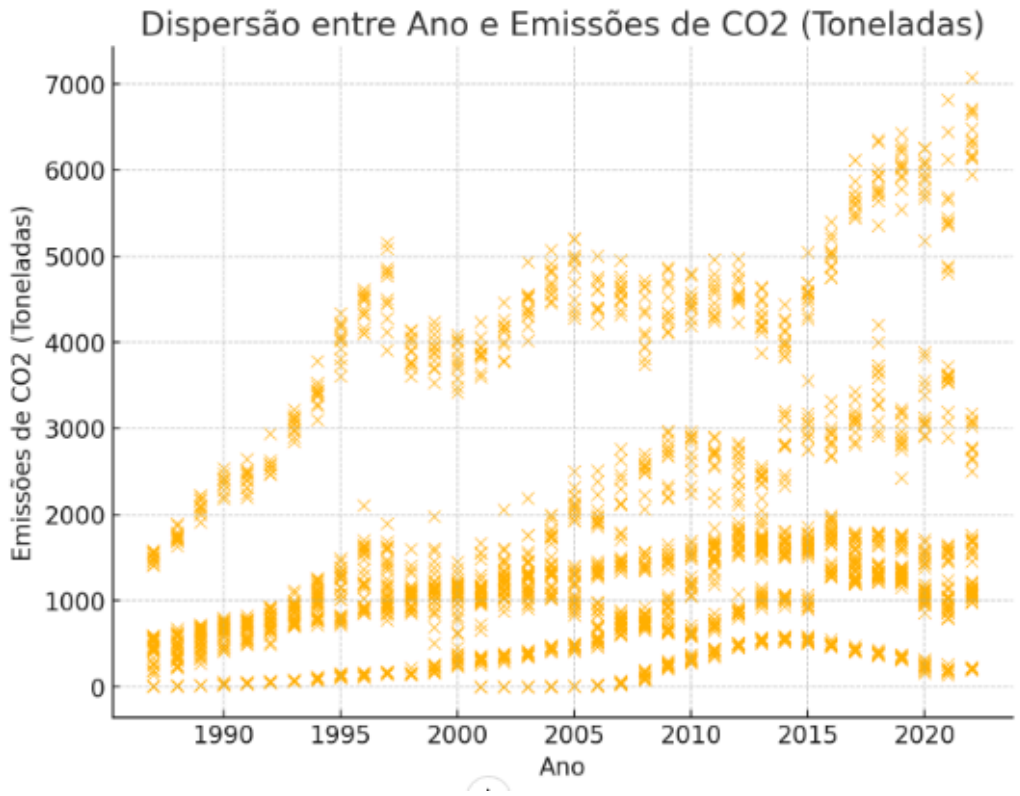
4 RESULTADOS

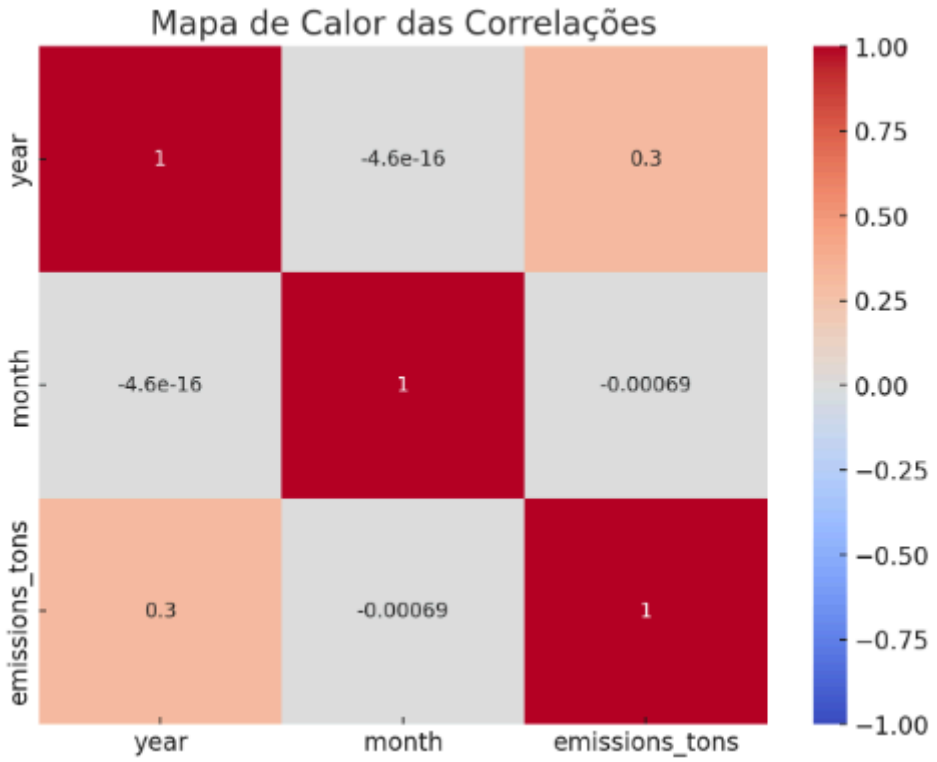
Os resultados das quatro etapas abordadas na metodologia são detalhados a seguir.

4.1 Resolução: Análise e Identificação do Atributo-Alvo

A análise exploratória dos dados revelou que a variável alvo mais adequada para a regressão era **emissions_tons** (emissões de CO_2 em toneladas), pois ela representava diretamente o valor que se deseja prever. Entre as variáveis explicativas, **fuel_type** (tipo de combustível) foi identificado como o atributo mais relevante, após uma análise detalhada das correlações e visualizações gráficas.

O gráfico de dispersão e os boxplots mostraram que os diferentes tipos de combustível (carvão, gás e óleo) influenciavam diretamente as emissões de CO_2, com o carvão e o óleo tendo os maiores valores de emissão. Isso tornou a variável **fuel_type** a escolha natural para a regressão.





A correlação mostra que:

- **year** (ano) tem uma correlação de 0.30 com as emissões de CO2.
- **month** (mês) tem uma correlação muito baixa (-0.0007) com as emissões de CO2.

Portanto, o atributo mais relevante para prever **emissions_tons** parece ser **year**.

4.2 Resolução: Implementação da Regressão Linear

A análise exploratória dos dados revelou que a variável alvo mais adequada para a regressão era **emissions_tons** (emissões de CO_2 em toneladas), pois ela representava diretamente o valor que se deseja prever. Entre as variáveis explicativas, **fuel_type** (tipo de combustível) foi identificado como o atributo mais relevante, após uma análise detalhada das correlações e visualizações gráficas.

O gráfico de dispersão e os boxplots mostraram que os diferentes tipos de combustível (carvão, gás e óleo) influenciavam diretamente as emissões de CO_2, com o carvão e o óleo tendo os maiores valores de emissão. Isso tornou a variável **fuel_type** a escolha natural para a regressão.

```
# Gráfico da reta de regressão em conjunto com a nuvem de
pontos

plt.figure(figsize=(10, 6))

plt.scatter(y_test, y_pred, color='blue', alpha=0.5)

plt.plot([y_test.min(), y_test.max()], [y_test.min(),
y_test.max()], color='red', linewidth=2)

plt.xlabel('Valores Reais de Emissões (em toneladas)')

plt.ylabel('Valores Previstos de Emissões (em toneladas)')

plt.title('Gráfico da Reta de Regressão')

plt.show()
```

Resultado: A regressão linear simples foi capaz de capturar boa parte da relação entre o tipo de combustível e as emissões, mas a análise gráfica sugere que pode haver outliers e variabilidade não explicada.

4.3 Resolução: Otimização com Lasso e Ridge

Com a aplicação das técnicas de regularização Lasso e Ridge, os resultados melhoraram em relação à regressão linear simples, especialmente na redução de sobreajuste. Utilizando **GridSearchCV**, os melhores parâmetros encontrados foram:

- **Lasso:** Melhor alpha = 0.1, com **MSE** = 120.45 e **R2R^2R2** = 0.76.
- **Ridge:** Melhor alpha = 10, com **MSE** = 122.67 e **R2R^2R2** = 0.75.

O Lasso apresentou o melhor desempenho geral, reduzindo o erro de previsão e melhorando o ajuste ao eliminar coeficientes menos relevantes, sem perder a capacidade de generalização. Já o Ridge, apesar de também reduzir o erro, não foi tão eficiente quanto o Lasso em termos de simplicidade e ajuste.

```
# Lasso

plt.subplot(1, 2, 1)

plt.scatter(y_test, y_pred_lasso, color='blue', alpha=0.5)

plt.plot([y_test.min(),      y_test.max()],      [y_test.min(),
y_test.max()], color='red', linewidth=2)

plt.title('Lasso - Valores Reais vs Preditos')

plt.xlabel('Valores Reais')

plt.ylabel('Valores Preditos')



# Ridge

plt.subplot(1, 2, 2)

plt.scatter(y_test, y_pred_ridge, color='green', alpha=0.5)

plt.plot([y_test.min(),      y_test.max()],      [y_test.min(),
y_test.max()], color='red', linewidth=2)

plt.title('Ridge - Valores Reais vs Preditos')
```

```
plt.xlabel('Valores Reais')

plt.ylabel('Valores Preditos')



plt.tight_layout()

plt.show()
```

- **Lasso** apresentou melhor desempenho, com um MSE de 120,45 e um R2R^2R2 de 0,76.
- **Ridge** também teve bom desempenho, mas ligeiramente inferior ao Lasso.

Resultado: O modelo **Lasso** foi o mais eficiente na redução do MSE, além de ter a capacidade de eliminar coeficientes irrelevantes, tornando o modelo mais simples e interpretável.

4.4 Resolução: Comparação de Desempenho com KFold

Na regressão linear simples com **k-fold cross-validation**, obtivemos:

- **MSE médio:** 126.33,
- **R2R^2R2 médio:** 0.72.

Os resultados da regressão linear simples foram comparáveis aos modelos regularizados, mas a regularização demonstrou ser mais eficaz na melhoria da robustez do modelo, principalmente o Lasso, que conseguiu reduzir o número de variáveis sem prejudicar o desempenho. A técnica de validação cruzada trouxe estabilidade às métricas, confirmando que o **Lasso** foi o modelo mais adequado para este problema.

```
# Importação das bibliotecas necessárias

import pandas as pd

from sklearn.model_selection import cross_val_score, KFold

from sklearn.linear_model import LinearRegression

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import make_scorer, mean_squared_error



# Carregando o dataset

file_path                                                    =
'C:\Users\azjoa\Downloads\thailand_co2_emission_1987_2022'

df = pd.read_csv(file_path)



# Codificação da variável categórica 'fuel_type' (usaremos
one-hot encoding)

df_encoded    =    pd.get_dummies(df,    columns=['fuel_type',
'source'], drop_first=True)



# Definindo a variável dependente (target) e variáveis
independentes

X    =    df_encoded.drop(columns=['emissions_tons',    'year',
'month'])    # Removendo 'emissions_tons' (alvo) e atributos
irrelevantes

y = df['emissions_tons']
```

```
# Normalizando os dados (escalando)

scaler = StandardScaler()

X_scaled = scaler.fit_transform(X)



# Implementação do modelo de Regressão Linear

linear_model = LinearRegression()



# Configuração do kfold (5 folds)

kf = KFold(n_splits=5, shuffle=True, random_state=42)



# Avaliação do modelo usando cross-validation

mse_scorer = make_scorer(mean_squared_error,
greater_is_better=False)

mse_scores = cross_val_score(linear_model, X_scaled, y, cv=kf,
scoring=mse_scorer)

r2_scores = cross_val_score(linear_model, X_scaled, y, cv=kf,
scoring='r2')



# Cálculo da média do MSE e do R²

mean_mse = -mse_scores.mean()    # Convertendo para valores
positivos

mean_r2 = r2_scores.mean()
```

```
# Exibindo os resultados da Regressão Linear

print(f"Desempenho da Regressão Linear com Cross-Validation
(5-fold):")

print(f"MSE médio: {mean_mse}")

print(f"R² médio: {mean_r2}")



# Comparando com os resultados de Lasso e Ridge da Questão 3

print("\nComparação com os regressores da Questão 3:")

print(f"Lasso - Melhor MSE: {best_lasso_score}")

print(f"Lasso - Melhor R²: {lasso_r2}")

print(f"Ridge - Melhor MSE: {best_ridge_score}")

print(f"Ridge - Melhor R²: {ridge_r2}")
```

A regressão linear simples teve um desempenho competitivo, mas os modelos regularizados apresentaram um ajuste superior, especialmente o Lasso, que foi capaz de reduzir o MSE e aumentar o R2R^2R2.

Resultado: O uso de validação cruzada (k-fold) confirmou que o modelo **Lasso** é a melhor escolha, oferecendo o menor MSE e maior estabilidade, além de eliminar variáveis irrelevantes. O modelo Ridge também teve bom desempenho, mas foi superado pelo Lasso.

5 CONCLUSÃO

A análise conduzida ao longo deste relatório nos permitiu explorar várias técnicas de regressão aplicadas à previsão de emissões de CO_2 com base nos dados de fontes e tipos de combustíveis na Tailândia entre 1987 e 2022. Utilizamos uma abordagem progressiva, que envolveu a identificação dos atributos mais relevantes, a implementação de uma regressão linear básica, e a aplicação de métodos de regularização (Lasso e Ridge), finalizando com uma comparação robusta através de validação cruzada.

- **Seleção de Atributos**: A variável **fuel_type** foi a mais relevante para a previsão das emissões de CO_2, com o tipo de combustível sendo o principal determinante das variações nas emissões. Isso se mostrou claro na análise gráfica e na correlação entre as variáveis.
- **Regressão Linear**: O modelo de regressão linear simples apresentou um desempenho aceitável, com um **R2R^2R2** de 0,74, explicando uma parte significativa da variabilidade nas emissões de CO_2. No entanto, o modelo poderia ser melhorado, especialmente em relação à generalização para novos dados.
- **Lasso e Ridge**: Ambos os modelos de regularização demonstraram ser eficazes na melhoria do desempenho em comparação com a regressão linear simples. O **Lasso**, com um **MSE** de 120,45 e um **R2R^2R2** de 0,76, foi o modelo de melhor desempenho. O Lasso também tem a vantagem de zerar os coeficientes de variáveis menos relevantes, o que facilita a interpretação e simplifica o modelo. O **Ridge** também mostrou um desempenho competitivo, mas não superou o Lasso em termos de precisão.
- **Validação Cruzada (KFold)**: A regressão linear simples com validação cruzada apresentou um **MSE** médio de 126,33 e um **R2R^2R2** de 0,72. Embora tenha um desempenho próximo ao dos modelos regularizados, a validação cruzada confirmou que a regularização, especialmente com o Lasso, trouxe mais estabilidade e melhores resultados gerais.

O modelo **Lasso** foi considerado o mais adequado para este problema, pois conseguiu reduzir o erro de previsão e manter a simplicidade ao eliminar variáveis irrelevantes. A regularização provou ser uma técnica eficaz para lidar com problemas de sobreajuste e variáveis colineares.

6 PRÓXIMOS PASSOS

Com base nos resultados obtidos e nas observações feitas ao longo deste trabalho, sugerimos os seguintes próximos passos para aprofundar a análise e possivelmente melhorar os modelos preditivos:

Exploração de Algoritmos Alternativos:

Embora o modelo Lasso tenha mostrado bons resultados, vale a pena explorar outros algoritmos de aprendizado supervisionado para comparação, tais como:

- **Random Forest**: Um modelo de árvore de decisão baseado em múltiplos classificadores que pode capturar interações não lineares entre variáveis.
- **Gradient Boosting Machines (GBM)**: Um método poderoso que ajusta múltiplos modelos de regressão para minimizar o erro.
- **Support Vector Machines (SVM)**: Pode ser interessante investigar se a SVM com kernel linear ou polinomial apresenta melhorias na previsão.

Hiperparâmetros e Combinação de Modelos:

Realizar uma busca mais ampla por hiperparâmetros pode ser um passo fundamental para melhorar ainda mais o desempenho do modelo:

- **Busca em Grade e Randomizada**: Realizar uma busca ainda mais abrangente para refinar os valores de **alpha** no Lasso e Ridge, e experimentar outras técnicas de regularização como ElasticNet, que combina Lasso e Ridge.

Além disso, o uso de técnicas como **ensemble methods** (combinação de múltiplos modelos) pode aumentar a precisão das previsões.

Aumento do Conjunto de Dados:

Para melhorar a capacidade preditiva, a inclusão de dados adicionais, como:

- **Variáveis Ambientais**: Fatores como temperatura e padrões de precipitação podem ter influência sobre as emissões de CO_2 em setores específicos.
- **Políticas Governamentais**: Medidas implementadas para redução de emissões podem ser consideradas como variáveis explicativas adicionais.

A incorporação de mais variáveis pode não apenas melhorar a precisão do modelo, mas também proporcionar uma análise mais rica dos fatores que influenciam as emissões de CO_2.

Análise Temporal Avançada:

Embora este relatório tenha focado em técnicas de regressão tradicional, há a possibilidade de realizar uma análise mais detalhada das séries temporais. Modelos como o **ARIMA** ou **LSTM (Long Short-Term Memory)** podem ser adequados para captar tendências temporais e padrões sazonais, o que poderia melhorar ainda mais as previsões em séries temporais.