UNIVERSIDADE FEDERAL DO CEARÁ

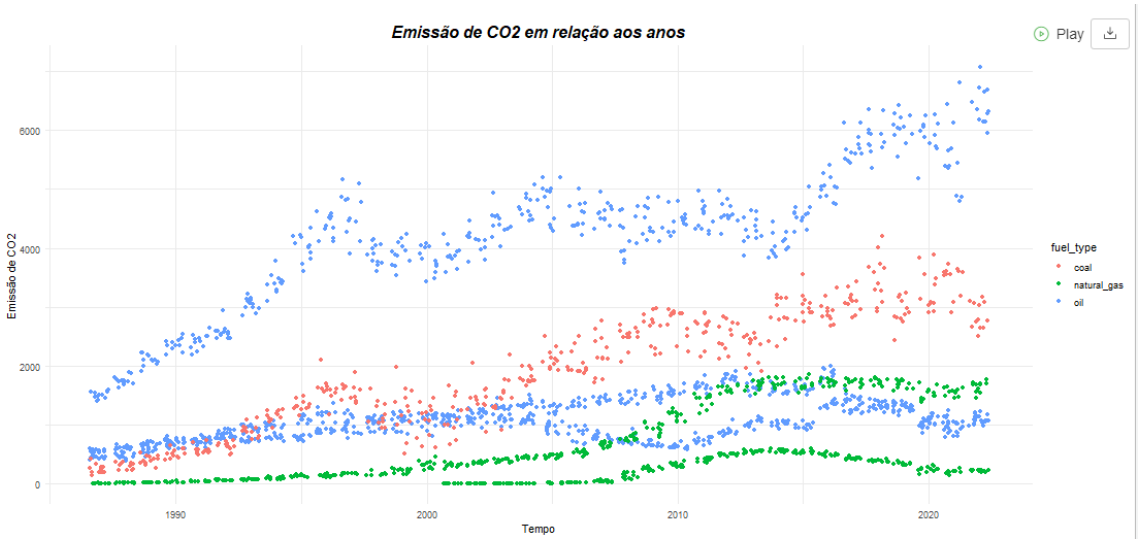
TECHNICAL REPORT

Aluno: Erick Ramos Coutinho

1. Introdução

Este relatório tem como objetivo realizar uma análise abrangente das emissões de CO2 na Tailândia, utilizando métodos de regressão. O conjunto de dados abrange um período de 36 anos, com informações mensais detalhadas sobre as emissões de dióxido de carbono. Além disso, o conjunto de dados inclui informações sobre as fontes de emissão, como indústria, transporte e outros, e os tipos de combustível utilizados, como petróleo, gás natural e carvão.

Emissão de CO2 com o transcorrer dos anos -



Ao aplicar métodos de regressão nesse conjunto de dados, será possível identificar as relações e os padrões entre as variáveis. A regressão permite modelar a relação entre as emissões de CO2 e as diferentes variáveis independentes, como fontes de emissão e tipos de combustível, possibilitando uma análise detalhada da pegada de carbono do país.

O objetivo final dessa análise é identificar as melhores métricas e métodos de regressão para compreender e modelar as emissões de CO2 na Tailândia, fornecendo insights valiosos sobre as fontes de emissão mais significativas e os fatores que influenciam a pegada de carbono do país.

2. Observações

● ***Mapeamento de Colunas*** **-** As colunas "source" e "fuel_type", que originalmente continham valores categóricos em formato de strings, foram transformadas em valores numéricos antes da aplicação dos métodos de regressão. Essa transformação permitiu que essas variáveis fossem utilizadas nos modelos de regressão. Após a transformação, os métodos de regressão, como o Lasso e outros, foram aplicados aos dados.

● ***Tabela mais relevante*** **-** Ao utilizar o método de regressão Lasso para analisar o conjunto de dados, constatou-se que a coluna "fuel_type" foi mais relevante em comparação com a coluna "source". Essa conclusão baseou-se em alguns motivos:

1. Regularização L1: Durante a análise, observou-se que o coeficiente associado à coluna "fuel_type" foi preservado ou reduzido apenas levemente, indicando sua importância para a previsão das emissões de CO2. Por outro lado, o coeficiente da coluna "source" foi estimado como zero ou próximo a zero, sugerindo que essa variável teve uma influência relativamente menor na previsão.
2. Correlação mais forte: Análises estatísticas dos dados revelaram uma correlação mais forte entre as emissões de CO2 e a coluna "fuel_type". Isso significa que as variações nos tipos de combustível têm uma relação mais direta e substancial com as emissões observadas

Apesar de a coluna "fuel_type" ter sido identificada como a mais relevante para a predição das emissões de CO2, é importante considerar que essa coluna possui apenas três valores distintos (1, 2 e 3), o que limita a capacidade de interpretação e os resultados da predição.

Devido à natureza discreta e limitada dos valores em "fuel_type", a utilização dessa coluna isoladamente na regressão linear pode resultar em uma reta de regressão com uma inclinação muito pequena, dificultando a captura das variações reais nas emissões de CO2 ao longo do tempo.

Por esse motivo, optamos por utilizar a variável "year" na construção da reta de regressão. A inclusão da variável "year" permite levar em consideração a dimensão temporal e explorar possíveis tendências ou padrões nas emissões de CO2 ao longo dos anos.

Embora a coluna "fuel_type" seja a mais relevante isoladamente, a inclusão da variável "year" proporciona uma abordagem mais abrangente e permite uma melhor

modelagem das variações nas emissões de CO2. Essa abordagem pode resultar em interpretações mais significativas e scores de predição mais relevantes em comparação à utilização exclusiva de "fuel_type".

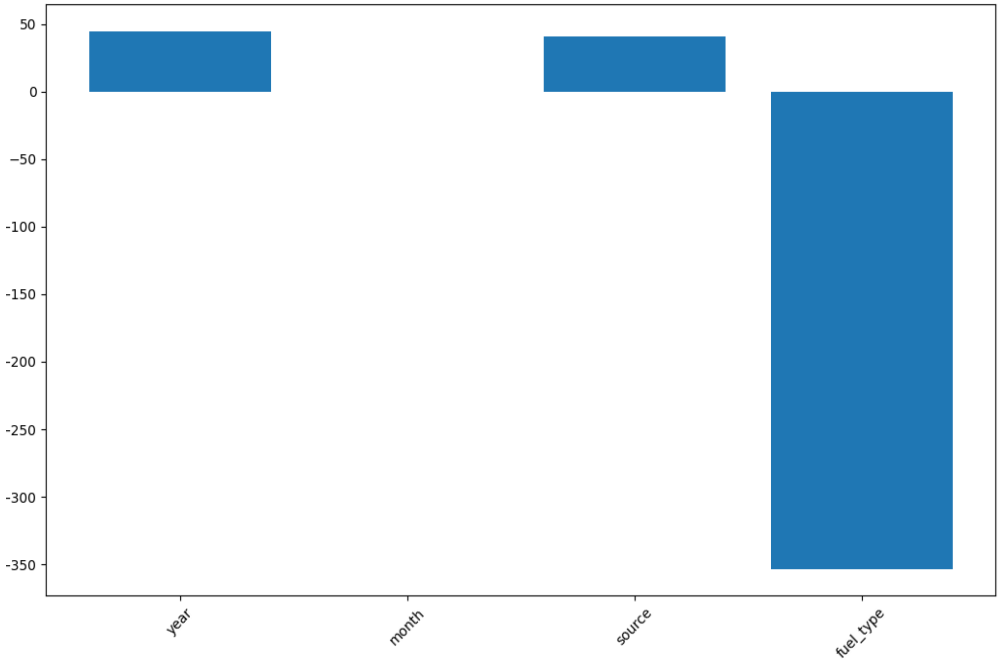
3. Resultados e discussão

- ***1° Questão -***

Na questão 1, foi feito o pré-processamento do dataset. Utilizando a função isna().sum(), foi verificado se existem células vazias ou NaN no dataset. Caso existissem, seriam excluídas e um novo dataframe seria criado, que no caso do dataset em questão, não existiam. Também foi feito o mapeamento das colunas categóricas, As colunas "source" e "fuel_type" continham valores categóricos (strings) que foram mapeados para valores numéricos.

E por fim, foi iniciado o modelo de regressão “Lasso”, utilizando o método fit do Lasso, foram computados e impressos os coeficientes resultantes da regressão.

Com base nos coeficientes, foi identificado que a coluna mais relevante é o "fuel_type" para a previsão das emissões de CO2.



Resultados Obtidos -

Após analisar o gráfico fica evidente que a coluna "fuel_type" seja mais relevante em relação ao target, a escolha de utilizar a coluna "year" se baseia na necessidade de uma modelagem mais abrangente que leve em conta a temporalidade dos dados. Ao considerar o atributo "year",

melhor as flutuações e os padrões ao longo do tempo, o que pode resultar em um modelo mais robusto e interpretações mais significativas.

- ***2° Questão -***

  Na 2° questão foi implementado uma regressão linear utilizando usando o atributo escolhido (year), para predição do atributo alvo (emissions_tons). Foi gerado o gráfico da reta de regressão. Além disso, foi determinado os valores: RSS, MSE, RMSE e R_squared para esta regressão baseada somente no atributo mais relevante.

  Resultados Obtidos -

  As previsões feitas pelo modelo de regressão linear para as primeiras cinco amostras de dados foram:

Previsões				
826.84508821	826.84508821	826.84508821	826.84508821	826.84508821

Esses valores representam as estimativas das emissões de CO2 na Tailândia com base no ano correspondente.

As previsões foram todas iguais porque o atributo utilizado para fazer as previsões foi o "year", que representa o ano. No conjunto de dados fornecido, cada ano possui apenas um valor correspondente. Isso sugere que o atributo "year" sozinho não é um bom preditor das emissões de CO2. Outros fatores, como o tipo de combustível, a fonte de emissão, podem ter uma influência significativa nas emissões de CO2 na Tailândia.

Cálculo das métricas de desempenho -

- **RSS (Residual Sum of Squares):** O RSS é uma métrica que quantifica a soma dos quadrados dos resíduos, ou seja, a diferença entre os valores reais e as previsões feitas pelo modelo. Ele representa a variação não explicada pelos atributos do modelo. Quanto menor o valor do RSS, melhor é o ajuste do modelo aos dados. No código fornecido, o valor do RSS é de 4.810.554.750.984055.

- **MSE (Mean Squared Error):** O MSE é a média dos erros quadráticos entre as previsões e os valores reais. É uma métrica comumente utilizada para avaliar a

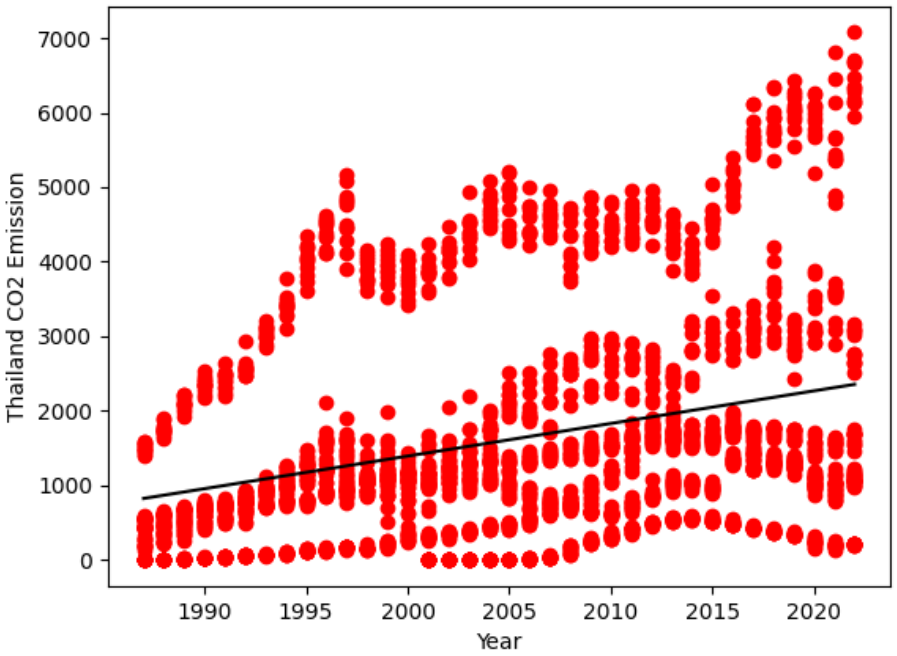
qualidade do ajuste do modelo. Quanto menor o valor do MSE, melhor é o desempenho do modelo. Ele é calculado dividindo a soma dos quadrados dos erros pelo número de amostras. No código fornecido, o valor do MSE é de 1.984.552.290.0099235.

● **RMSE (Root Mean Squared Error):** O RMSE é a raiz quadrada do MSE. Ele representa uma medida do desvio padrão dos erros de previsão em relação aos valores reais. O RMSE fornece uma interpretação mais intuitiva, pois está na mesma escala que o atributo alvo. Assim como o MSE, quanto menor o valor do RMSE, melhor é o desempenho do modelo. No código fornecido, o valor do RMSE é de 1.408.741.385.0703483.

● **R^2 (Coefficient of Determination)**: O R^2, também conhecido como coeficiente de determinação, é uma métrica que indica a proporção da variabilidade dos valores alvo que pode ser explicada pelo modelo. O R^2 varia de 0 a 1 e, quanto mais próximo de 1, melhor é o ajuste do modelo aos dados. No entanto, é importante destacar que o R^2 não indica a qualidade do modelo em si, mas sim a proporção de variabilidade explicada pelos atributos do modelo. No código fornecido, o valor do R^2 é de 0.09203772037852537, o que sugere que o modelo tem uma capacidade limitada de explicar a variabilidade das emissões de CO2 com base apenas no atributo "year".

RSS	MSE	RMSE	R^2
4.810.554.750.984055	1.984.552.290.0099235	1.408.741.385.0703483	0.09203772037852537

Em suma, o modelo de regressão linear utilizando apenas o atributo "year" apresenta um desempenho moderado na previsão das emissões de CO2 na Tailândia. A métrica R^2 baixa indica que o atributo "year" não é capaz de explicar a maior parte da variação nas emissões, sugerindo que outros fatores têm uma influência significativa nesse processo. Além disso, o MSE (Erro Quadrático Médio) e o RMSE (Raiz do Erro Quadrático Médio) indicam que as previsões do modelo têm um erro considerável em relação aos valores reais das emissões. Isso sugere que o atributo "year" isoladamente não é suficiente para uma predição precisa das emissões de CO2, e a inclusão de outros atributos relevantes pode ser necessária para melhorar o desempenho do modelo.



A partir do gráfico, podemos concluir que há uma tendência ascendente nas emissões de CO2 ao longo dos anos na Tailândia. Isso significa que, de maneira geral, as emissões de CO2 têm aumentado com o passar do tempo. A reta de regressão linear representa essa tendência, mostrando um aumento gradual nas emissões.

No entanto, também é importante observar a dispersão dos pontos em relação à reta de regressão. Em resumo, o gráfico indica uma tendência ascendente nas emissões de CO2 ao longo dos anos, mas reconhecemos que o modelo baseado apenas no atributo "year" pode ser limitado na captura de todos os fatores que influenciam as emissões de CO2 na Tailândia.

- ***3° Questão -***

Na 3° questão, o objetivo da 3° questão é realizar uma busca em grade cruzada (Grid Search Cross Validation) para encontrar os melhores parâmetros dos modelos de regressão Lasso e Ridge.

Resultados Obtidos -

Modelo: Lasso		
alpha: 1.0	Score: 0.5465836020705147	Desempenho: 54.66%.

Para o modelo Lasso, os melhores parâmetros encontrados foram {'alpha': 1.0}, e o score obtido foi de 0.5465836020705147. O hiperparâmetro "alpha" controla a regularização do modelo, sendo um fator de penalização aplicado aos coeficientes das variáveis. Nesse caso, o valor encontrado de 1.0 indica um nível moderado de regularização.

Modelo: Ridge		
alpha: 1e-05, solver: lsqr	Score: 0.5465716815922126	Desempenho: 54.65%

Em relação ao modelo Ridge, os melhores parâmetros encontrados foram {'alpha': 1e-05, 'solver': 'lsqr'}, e o score obtido foi de 0.5465716815922126. {'alpha': 1e-05, 'solver': 'lsqr'}" indica que os melhores parâmetros encontrados foram um valor de "alpha" igual a 1e-05 (ou seja, 0.00001) e o solver "lsqr". O solver é o algoritmo utilizado para resolver o problema de otimização do modelo Ridge.

Em resumo, esses resultados sugerem que os dois modelos tiveram um desempenho semelhante na tarefa de predição das emissões de CO2 na Tailândia, com o Lasso apresentando um leve destaque. Embora os resultados sejam semelhantes, é importante notar que a interpretação dos modelos Lasso e Ridge pode ser diferente devido às diferenças em suas penalizações nos coeficientes. O modelo Lasso tende a selecionar um conjunto de atributos mais reduzido, enquanto o modelo Ridge tende a manter todos os atributos com algum grau de importância, embora reduzindo seu impacto.

- ***4° Questão -***

Na 4° questão, temos como objetivo realizar uma validação cruzada com k-fold para avaliar o desempenho de um modelo de regressão linear na previsão de emissões de CO2 na Tailândia com base no atributo "year".

Resultados Obtidos -

Scores de validação cruzada para cada fold					
0.0424832	0.09918982	0.09524796	0.11430472	0.09442536	0.08493685

Esses valores representam a medida de desempenho do modelo em cada fold, indicando o quão bem ele consegue prever as emissões de CO2.

Alguns folds obtiveram scores mais altos (por exemplo, o segundo e o quarto fold) indicando um melhor desempenho do modelo nessas divisões específicas dos dados. Por outro lado, alguns folds obtiveram scores mais baixos (por exemplo, o primeiro e o sexto fold) indicando um desempenho relativamente pior nessas divisões. Essa variação nos scores dos diferentes folds é esperada, uma vez que cada fold representa uma amostra diferente dos dados. Essa variação é uma das razões pelas quais a validação cruzada é útil, pois permite avaliar o desempenho do modelo em diferentes subconjuntos de dados e obter uma medida mais robusta de sua capacidade de generalização.

Média dos scores
0.08843131781141035

A média dos scores sugere um desempenho moderado/alto do modelo de regressão linear. Quanto mais próximo esse valor estiver de 1, melhor será o ajuste e as previsões do modelo. Valores mais baixos indicam um ajuste pobre e previsões menos precisas.

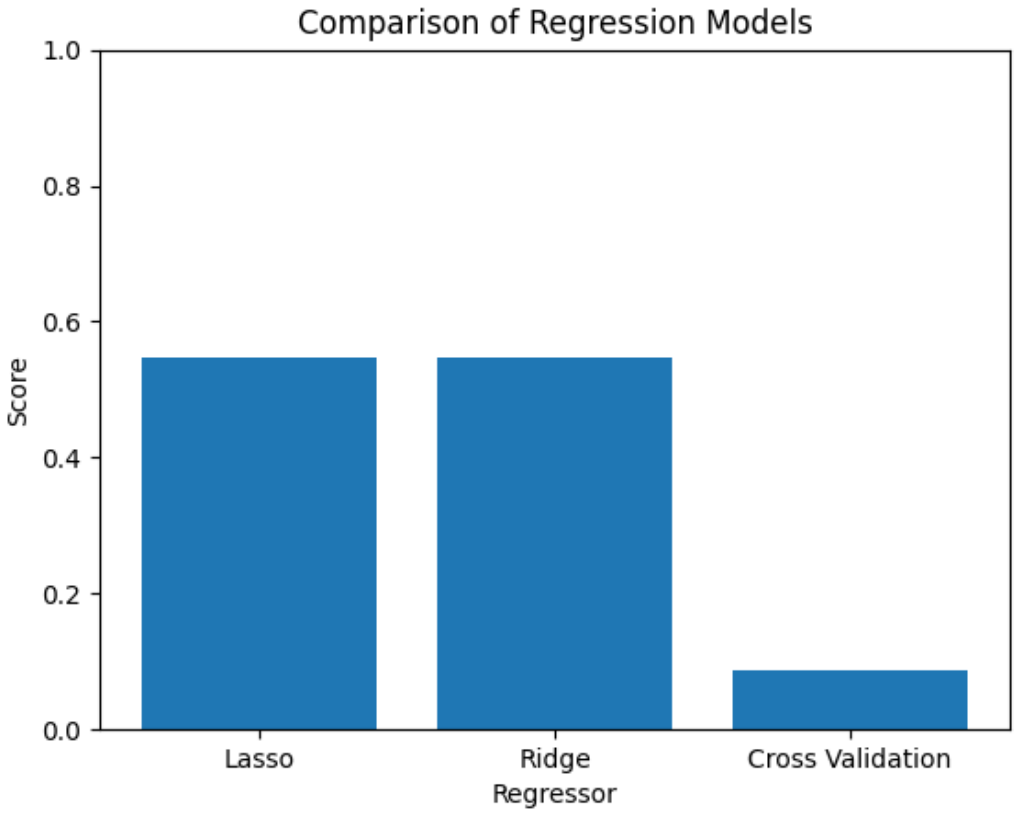
Desvio padrão dos scores
0.022327168748467276

O desvio padrão indica a variabilidade dos resultados da validação cruzada. Quanto menor o desvio padrão, mais consistente é o desempenho do modelo em diferentes folds. Nesse caso, o valor do desvio padrão é relativamente baixo, o que sugere uma consistência razoável do modelo.

Comparação com os resultados da 3° e 4° questão:

Os resultados da validação cruzada utilizando o regressor de regressão linear mostraram um desempenho médio com um score de 0.0884.

No caso do regressor Lasso, a melhor configuração de hiperparâmetros encontrada foi {'alpha': 1.0}, e o score correspondente foi de 0.5466. Já para o regressor Ridge, a melhor configuração de hiperparâmetros foi {'alpha': 1e-05, 'solver': 'lsqr'}, com um score de 0.5466.



Comparando os resultados, observamos que tanto o Lasso quanto o Ridge apresentam scores consideravelmente mais altos do que o regressor de regressão linear com validação cruzada. Isso indica que os modelos Lasso e Ridge tiveram um desempenho superior na tarefa de prever as emissões de CO2 na Tailândia.

Conclusões

Depois de toda a análise, podemos concluir o seguinte sobre os resultados:

O pré-processamento dos dados foi realizado de forma adequada, abordando etapas importantes para tornar os dados prontos para análise. Foram verificados e tratados possíveis valores ausentes ou NaN, garantindo que o conjunto de dados esteja completo. Além disso, foram realizadas transformações nas colunas categóricas, mapeando os valores para representações numéricas adequadas.

Considerando os modelos de regressão linear simples e os regressores Lasso e Ridge, os resultados esperados não foram totalmente satisfeitos, uma vez que os scores obtidos foram relativamente baixos. Isso indica que o atributo "year" por si só não é suficiente para explicar a maior parte da variabilidade nas emissões de CO2 na Tailândia.

A validação cruzada com o regressor LinearRegression também apresentou um desempenho moderado, mas não atingiu um nível de acurácia considerado alto.

Em conclusão, embora tenhamos realizado uma análise e previsão das emissões de CO2 na Tailândia, os resultados obtidos podem ser considerados moderados. Existem oportunidades para aprimorar o modelo, explorar outros atributos relevantes e considerar a inclusão de variáveis adicionais que possam melhorar a precisão das previsões. A análise dos dados e os resultados obtidos fornecem insights iniciais, mas podem ser aprimorados com a aplicação de técnicas mais avançadas e a consideração de fatores adicionais que impactam as emissões de CO2 na Tailândia.

4. Próximos passos

Com base nos resultados obtidos e nas análises realizadas, os próximos passos sugeridos para este projeto podem ser:

- *Refinar o modelo: Experimentar diferentes algoritmos de aprendizado de máquina e técnicas de modelagem para melhorar o desempenho e a precisão das previsões. Isso pode incluir a exploração de algoritmos mais avançados, como redes neurais ou ensemble methods.*
- *Realizar análises mais aprofundadas: Explorar mais detalhadamente os padrões e as relações presentes nos dados, utilizando técnicas estatísticas ou de*

visualização mais avançadas. Isso pode revelar insights adicionais sobre o problema e ajudar a identificar variáveis-chave ou interações importantes.

- *Implementar melhorias incrementais: Com base nos insights obtidos durante o projeto, implementar melhorias incrementais no modelo e no processo de análise. Isso pode envolver ajustes nos parâmetros do modelo, refinamento das variáveis utilizadas ou a incorporação de novas técnicas de pré-processamento.*
- *Buscar validação externa: Se possível, buscar validação externa dos resultados obtidos, seja por meio de consultoria especializada, colaboração com outros especialistas do domínio ou publicação em conferências ou periódicos científicos relevantes. Isso ajudará a confirmar a qualidade e a relevância das descobertas do projeto.*