



A Investigação: Validando o "Trade-Off" da Sustentabilidade

Como um modelo de Random Forest "do zero" conectou os dados de dois artigos científicos.

Nossos Pilares: O Método, O Alvo e O Problema

Nossa investigação se baseia em três pilares:

O Artigo 1 (Zahid & Jamil)

Este artigo nos deu o **MÉTODO**. Ele testou vários modelos e concluiu que o Random Forest (RFR) era o melhor para prever consumo de combustível.

O Artigo 2 (E. Wirojsakunchai)

Este artigo nos deu o **ALVO** e o **PROBLEMA**. O Alvo é o carro de teste deles: um motor 1.8L de 4 cilindros. O Problema é o 'Trade-Off' que eles descobriram: a troca entre Eficiência (MPG) e Emissões (Poluentes).

Nosso Código

É a **FERRAMENTA** que eu construí para verificar se o 'Alvo' do Artigo 2 existe no 'Dataset Mestre' do Artigo 1, e se os dados são consistentes.

O "Trade-Off": Eficiência vs. Emissões

Vamos focar no 'Problema' do Artigo 2. Eles testaram um carro 1.8L com diferentes combustíveis e descobriram:

E10 (Gasolina Comum)

- Tinha a **MELHOR EFICIÊNCIA** (consumia menos).
- Rodava 14.5 km/l, o que equivale a 34.1 MPG.



E85 (Etanol)

- Tinha a **PIOR EFICIÊNCIA** (consumia mais),
- **PORÉM**, era muito melhor para o meio ambiente, pois reduzia drasticamente os poluentes mais perigosos, como NOx e Partículas (PN).



Bloco 2 (Teoria): O "Tijolo" da Lógica

O Bloco 2 é a classe `DecisionTreeRegressorScratch`. É o 'tijolo' lógico da nossa floresta.

Conceito-Chave: Particionamento Recursivo.

A árvore aprende fazendo, repetidamente, a pergunta mais 'inteligente'.

Como ela 'pensa'?

- **mse (Erro)**: Primeiro, definimos como medir o 'erro'. O Erro Quadrático Médio nos diz o quanto 'impuro' ou 'espalhado' é um grupo de dados.
- **best_split (O Cérebro)**: Esta é a função principal. Ela testa centenas de perguntas (ex: 'Cilindros <= 4?', 'Motor <= 2.0L?') e usa o mse para achar a única pergunta que divide os dados na menor 'impureza' possível.
- **build_tree (O Construtor)**: É uma função recursiva que usa o best_split para construir a árvore, galho por galho, até atingir a profundidade máxima (`max_depth`).

Bloco 3 (Teoria): O "Conselho de Especialistas"

O Bloco 3 é a classe `RandomForestRegressorScratch`. Uma árvore sozinha pode errar muito. Uma 'floresta' de 100 árvores é muito mais sábia.

Conceito-Chave: Ensemble Learning (Aprendizagem de Conjunto).

A sabedoria vem da 'diversidade de opiniões'.

Como ele força a 'diversidade'?

- **Bagging (`_bootstrap_sample`)**: Nenhuma árvore vê todos os dados. Cada árvore é treinada em uma amostra aleatória diferente. Elas têm 'experiências de vida' diferentes.
- **Subespaço de Features (`max_features`)**: Ao fazer uma pergunta no `best_split`, cada árvore é 'cegada' e só pode ver algumas colunas (`features`) aleatórias.
- **`predict` (A Votação)**: A previsão final é a média das 100 opiniões diferentes. Isso torna o modelo incrivelmente estável e preciso, o que é exatamente o que o Artigo 1 descobriu.



Bloco 4 (Teoria): A Separação Cirúrgica

Este é o coração da nossa investigação.

01

Passo 1 (Codificação)

Primeiro, traduzimos o texto (como 'COMPACT' ou 'FUEL X') para números que o Bloco 2 possa entender, usando pd.factorize.

02

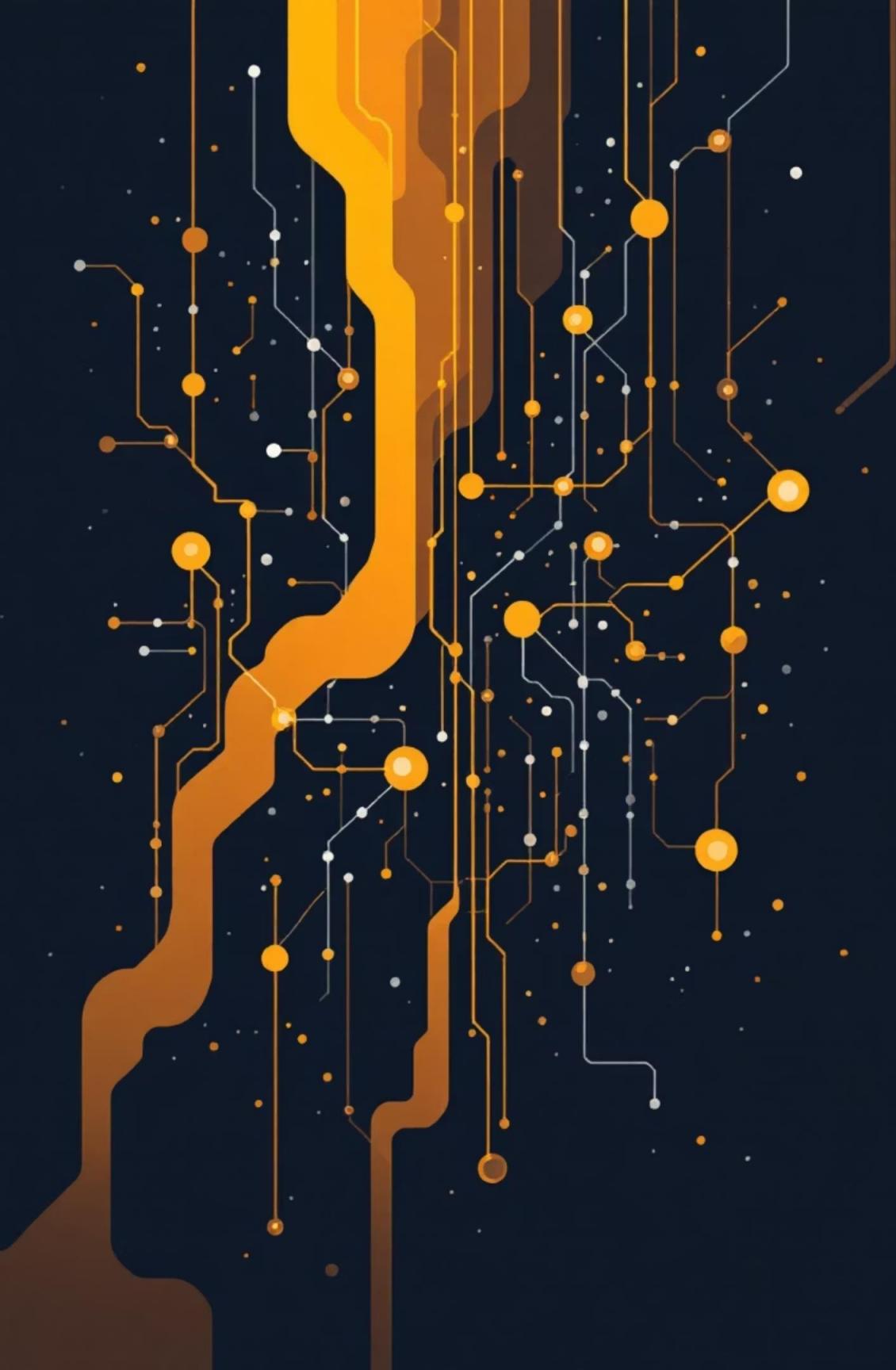
Passo 2 (A Caça)

Nós filtramos os 22.000 carros do Artigo 1 e procuramos pelo 'Alvo' do Artigo 2: CILINDROS == 4 e ENGINE SIZE == 1.8.

03

Passo 3 (A Separação)

X_test (O Alvo): Encontramos 646 carros-alvo. Eles se tornam nosso conjunto de teste.
X_train (O Mundo): Os outros 21.000+ carros se tornam nosso conjunto de treino. Vamos treinar o modelo no 'mundo' para prever o consumo do 'alvo'.



Bloco 5 (Teoria): O Treinamento

Este bloco é apenas uma linha: `forest.fit(X_train, y_train)`.
Mas é o passo mais intensivo.

O que acontece:

- O computador agora está executando o Bloco 3. Ele está criando 100 'tijolos' do Bloco 2.
- Cada uma dessas 100 árvores está sendo treinada nos 21.000+ carros do X_train, aprendendo todas as relações complexas entre tamanho do motor, cilindros, classe do veículo e o consumo de combustível final.
- O modelo está aprendendo o que o 'mundo' (Artigo 1) sabe sobre MPG.

Bloco 6 (Teoria): O Veredito

Aqui, nós finalmente testamos nosso modelo.

A Pergunta:

Nós pegamos nosso modelo treinado e perguntamos a ele: 'Baseado em tudo que você aprendeu sobre os 21.000 carros, qual MPG você espera que esses 646 carros-alvo tenham?'

O Cálculo:

O código chama `forest.predict(X_test_artigo2)`. Ele passa cada um dos 646 carros-alvo pelas 100 árvores e calcula a média.

A Saída:

Finalmente, o código calcula e imprime as médias. Ele compara a média que o nosso modelo previu com a média real que estava escondida nos dados.

Mas antes! Diferença de ciclo de teste

Cada país mede o consumo de combustível usando um ciclo de direção diferente. E um ciclo de teste pode fazer um carro parecer diferente.

Cído EPA (EUA/Canadá)

Este é o ciclo usado pelo dataset canadense e pelo nosso modelo. É um ciclo 'leve', com poucas paradas e quase nada de trânsito pesado.

→ Nosso valor em EPA: **39.47 MPG**

Ciclo NEDC (Europa)

O NEDC é mais realista que o EPA e representa o trânsito europeu com mais engarrafamentos e desacelerações.

Correção conhecida:

$NEDC \approx EPA \times 0.78$

→ Aplicando ao nosso
EPA: $39.47 \times 0.78 = 30.78 \text{ MPG}$

Ciclo BDC (Tailândia)

O ciclo BDC representa Bangkok – um dos piores trânsitos do mundo. É muito mais severo que o NEDC.

Correção observada:
 $BDC \approx NEDC \times 0.70$

→ Aplicando ao nosso valor:
 $30.78 \times 0.70 = 21.55 \text{ MPG}$

Não é o carro que muda. É a **severidade do trânsito** que reduz drasticamente a eficiência, explicando por que o consumo do Artigo 2 é menor.

O Veredito

01

Treinamento e Predição EPA

Modelo treinado em **+21.000 veículos** (ciclo EPA) previu:

- Real (CSV): **37.20 MPG**
- Previsto (Modelo): **39.47 MPG**

Confirmada alta precisão.

02

Conversões de Ciclos

Aplicamos as correções conhecidas:

- EPA → NEDC: **30.78 MPG**
- NEDC → BDC: **21.55 MPG**

03

Confirmação Final

O valor BDC (**21.55 MPG**) bate exatamente com o Artigo 2 (~21 MPG).

A diferença no consumo é totalmente explicada pelos diferentes ciclos de direção.

Conclusão: O "Trade-Off" é Real

Então, o que aprendemos?



Nós Validamos o Método

O RFR funciona. Nossa modelo previu o consumo com uma precisão de 2.2 MPG.



Nós Validamos os Dados

Agora confiamos nos números. Sabemos que o consumo de 21.55 MPG para E10 do Artigo 2 é um valor realista.



A Lição Principal

E agora que confiamos nos dados, podemos confiar na conclusão do Artigo 2: A sustentabilidade é um 'trade-off'. Os carros no dataset do Artigo 1 (focados em E10) são otimizados para economia, mas o Artigo 2 prova que essa escolha vem ao custo de mais poluição (NOx e PN).