# *Análise Preditiva e Estratégica Aplicada ao Comércio Internacional*

*Autores:* Letícia Araújo Costa, Ándre Carvalho, Ítalo Timbó Santos

*Disciplina:* Análise Preditiva

*Data:* 23 de junho de 2025

## *Resumo*

Este trabalho explora a aplicação de técnicas de machine learning para analisar e prever dinâmicas do comércio internacional. Utilizando um dataset público de transações comerciais, foi desenvolvido um modelo de regressão baseado no algoritmo Random Forest para prever o valor financeiro (Valor) das transações. A metodologia abrange uma abordagem analítica em quatro níveis: descritivo, diagnóstico, preditivo e prescritivo. O modelo de regressão alcançou um Coeficiente de Determinação ($R^2$) de 0,991, demonstrando alta capacidade de ajuste aos dados. No entanto, a análise crítica revela que, apesar da precisão percentual, as métricas de erro absoluto (MAE e RMSE) são elevadas, uma consequência direta da escala e da variância extrema dos dados. O estudo conclui que o modelo é uma ferramenta robusta para o planejamento estratégico de alto nível, permitindo a derivação de recomendações (prescrições) para otimização de políticas comerciais e estratégias corporativas, ao mesmo tempo que destaca as limitações e os caminhos para trabalhos futuros.

*Palavras-chave:* Machine Learning, Análise Preditiva, Comércio Internacional, Random Forest, Análise Prescritiva.

## *1. Introdução*

O comércio internacional é um pilar da economia global, caracterizado por sua alta complexidade, dinamismo e pelo vasto volume de dados gerado. A capacidade de analisar esses dados para identificar padrões, diagnosticar

causas, prever tendências futuras e prescrever ações estratégicas confere uma vantagem competitiva decisiva para governos e corporações.

Neste contexto, a ciência de dados e o machine learning emergem como ferramentas fundamentais. A análise tradicional, muitas vezes limitada a estatísticas descritivas, pode ser aprimorada por modelos preditivos que aprendem a partir de dados históricos para realizar previsões acuradas.

O objetivo principal deste trabalho é desenvolver um modelo de regressão para prever o valor financeiro (Valor em USD) de transações comerciais. Para além da previsão, o projeto estrutura-se em uma análise completa que abrange desde a descrição do cenário atual até a recomendação de ações futuras, demonstrando o ciclo completo de uma solução de data science aplicada a um problema de negócio real.

## *2. Metodologia*

A metodologia foi estruturada em três pilares: a descrição e preparação do dataset, a definição do framework analítico e o processo de modelagem preditiva.

### *2.1. Descrição e Preparação do Dataset*

O estudo utilizou o dataset df_exportacao_importacao_classificado.csv, que consiste em um DataFrame da biblioteca Pandas com *27.384 registros* e *7 colunas*. A estrutura do dataset é a seguinte:

- Fluxo_Comercial: Categórica (2 valores: 'Exportação', 'Importação').
- Economia_Relatora: Categórica (264 economias distintas, incluindo países e blocos).
- Produto_Setor: Categórica (18 setores econômicos).
- Economia_Parceira: Categórica (3 valores agregados: 'Mundo', 'União Europeia', 'Comércio Extra UE').
- Ano: Categórica (4 anos distintos).
- Tipo: Categórica (3 tipos distintos).
- Valor: Numérica (float64), variável alvo da nossa previsão.

Uma análise inicial da qualidade dos dados revelou *136 entradas nulas* na coluna Valor. Como esta é a variável alvo, os registros correspondentes foram removidos do dataset para garantir a integridade do treinamento do modelo.

A estatística descritiva da variável Valor (após a limpeza) revelou características cruciais:

- *Contagem:* 27.248
- *Média:* $8,81 \times 10^{10}$
- *Desvio Padrão (std):* $8,04 \times 10^{11}$
- *Mínimo:* 0.0
- *Mediana (50%):* $7,79 \times 10^{8}$
- *Máximo:* $2,56 \times 10^{13}$

A notável disparidade entre a média e a mediana, aliada a um desvio padrão quase dez vezes superior à média, aponta para uma distribuição de dados com forte assimetria positiva, indicando a presença de outliers de altíssimo valor que influenciam significativamente as métricas estatísticas.

## *2.2. Framework Analítico*

O trabalho foi guiado por um framework de quatro estágios:

1. *Análise Descritiva:* O que aconteceu? Sumarização dos dados para entender o panorama.
2. *Análise Diagnóstica:* Por que aconteceu? Investigação das causas por trás dos padrões observados.
3. *Análise Preditiva:* O que vai acontecer? Construção do modelo para prever resultados futuros.
4. *Análise Prescritiva:* O que devemos fazer? Geração de recomendações acionáveis a partir dos insights.

## *2.3. Modelagem Preditiva*

O processo de construção do modelo de regressão seguiu os seguintes passos:

1. *Pré-processamento:* As variáveis categóricas (Fluxo_Comercial, Economia_Relatora, etc.) foram transformadas em representações numéricas através de técnicas de encoding, tornando-as adequadas para o algoritmo de machine learning.
2. *Seleção do Algoritmo:* Optou-se pelo *Random Forest Regressor. Esta escolha se justifica por sua alta performance em problemas de regressão,

sua robustez a *outliers (particularmente relevante para este dataset) e sua capacidade de capturar relações complexas e não lineares entre as variáveis sem a necessidade de escalar os dados.

3. *Treinamento e Avaliação:* O dataset foi dividido em conjuntos de treino e teste para avaliar a capacidade de generalização do modelo. As métricas de avaliação selecionadas foram:
    - *Erro Absoluto Médio (MAE):* Mede a média dos erros absolutos entre os valores previstos e os reais.
    - *Raiz do Erro Quadrático Médio (RMSE):* Similar ao MAE, mas penaliza mais os erros maiores.
    - *Coeficiente de Determinação ($R^2$):* Indica a proporção da variância da variável alvo que é explicada pelo modelo.

## *3. Resultados e Discussão*

Esta seção apresenta os resultados obtidos em cada etapa do framework analítico.

### *3.1. Análise Descritiva*

A análise exploratória revelou que o dataset agrega dados de comércio de 264 economias distintas com três parceiros gerais: 'Mundo', 'União Europeia' e 'Comércio Extra UE'. Isso indica que o foco da análise não são as relações bilaterais, mas sim o desempenho comercial de uma nação perante grandes blocos. A distribuição da variável Valor é o achado mais significativo, com sua extrema assimetria e variância, sugerindo que o comércio global é dominado por um número relativamente pequeno de fluxos de valor maciço.

### *3.2. Análise Diagnóstica*

A variância extrema observada é diagnosticada como uma consequência da agregação de diferentes níveis de comércio no mesmo dataset. Por exemplo, o registro de "Mercadorias totais" da China com o "Mundo" coexiste com o de "Têxteis" de um país pequeno, gerando uma escala de valores que abrange mais de 13 ordens de magnitude. A dominância de certos setores em economias específicas (ex: 'Produtos agrícolas' no Brasil) foi diagnosticada como um reflexo da especialização econômica e das vantagens comparativas daquela nação.

## *3.3. Análise Preditiva*

O modelo Random Forest treinado apresentou os seguintes resultados no conjunto de teste:

| Métrica | Valor |
|---|---|
| *MAE* | ≈ 10,2 bilhões de dólares |
| *RMSE* | ≈ 74,57 bilhões de dólares |
| *$R^2$* | *0,991* |

A discussão crítica desses resultados é fundamental. O *$R^2$ de 0,991* é excepcionalmente alto, indicando que o modelo se ajusta quase perfeitamente aos dados. Isso é plausível porque os dados agregados (comércio de um país com o 'Mundo' em um setor) são inerentemente mais estáveis e previsíveis do que fluxos bilaterais voláteis.

Em contrapartida, o *MAE e o RMSE* são na casa dos bilhões de dólares. Este resultado, que à primeira vista parece pobre, é uma consequência direta da escala e do desvio padrão de 804 bilhões de dólares da variável Valor. Quando os valores podem chegar a trilhões, um erro médio de 10 bilhões é proporcionalmente pequeno, embora financeiramente significativo. Isso define o escopo de uso do modelo: excelente para direção estratégica, mas impreciso para projeções financeiras granulares.

## *3.4. Análise Prescritiva*

A capacidade preditiva do modelo habilita a geração de recomendações estratégicas:

- *Para Políticas Públicas:* Um governo pode utilizar o modelo para simular o impacto de focar em determinados Produto_Setor para exportação para a União Europeia. Se o modelo prevê um crescimento de Valor para o setor de 'Equipamentos de telecomunicações', a prescrição seria criar políticas de incentivo e remover barreiras para empresas dessa área, maximizando o potencial de superávit comercial.
- *Para Estratégia Corporativa:* Uma empresa pode usar as previsões de importação de uma Economia_Relatora para estimar o tamanho do mercado endereçável. Com base nisso, pode prescrever metas de vendas, planejar investimentos em logística e otimizar o gerenciamento de inventário para as rotas e produtos mais promissores.

## *4. Conclusão*

Este trabalho demonstrou com sucesso a aplicação de um ciclo completo de ciência de dados para a análise do comércio internacional. Foi construído um modelo preditivo de alta acurácia ($R^2=0,991$) que, quando interpretado criticamente, se revela uma poderosa ferramenta para o planejamento estratégico. A análise crítica das métricas de erro em função da distribuição dos dados foi essencial para compreender as capacidades e limitações do modelo.

### *4.1. Limitações e Trabalhos Futuros*

Apesar do sucesso, o estudo possui limitações que abrem caminhos para pesquisas futuras:

1. *Granularidade dos Dados:* A variável Economia_Parceira é muito agregada. Um trabalho futuro poderia utilizar dados de comércio bilateral para permitir diagnósticos e prescrições mais específicos.
2. *Enriquecimento de Features:* O modelo poderia ser aprimorado com a inclusão de variáveis macroeconômicas, como PIB, taxas de câmbio, tarifas e índices de produção industrial.
3. *Análise de Feature Importance:* Uma análise detalhada da importância das features do Random Forest poderia quantificar o impacto de cada variável, fortalecendo a análise diagnóstica.
4. *Modelos de Séries Temporais:* Para previsões que capturem melhor as dependências temporais, poderiam ser explorados algoritmos como ARIMA ou redes neurais recorrentes (LSTM).