

# Predição de Preços de Energia com modelos supervisionados de Machine Learning

Kaique Moraes da Silva <sup>1</sup>

<sup>1</sup>Faculdade de Computação e Informática (FCI)  
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

10410548@mackenzie.br

**Resumo.** Este projeto aplica técnicas de Machine Learning para predição de preços de energia elétrica utilizando dados históricos de consumo, geração, condições meteorológicas e preços. O objetivo é desenvolver um modelo preditivo capaz de estimar os preços futuros da energia com base em variáveis correlacionadas, auxiliando na tomada de decisões estratégicas no setor energético. A metodologia emprega algoritmos de regressão disponíveis na biblioteca scikit-learn, incluindo análise exploratória dos dados, pré-processamento, treinamento de modelos e avaliação de desempenho. Os resultados esperados incluem a identificação dos principais fatores que influenciam os preços de energia e a construção de um modelo com capacidade preditiva satisfatória, medida através de métricas como RMSE e  $R^2$ .

**Palavras-chave:** Machine Learning, Predição de Preços, Energia Elétrica, Regressão, scikit-learn.

**Endereço do repositório do GitHub:** <https://github.com/kaique-moraes/energy-predict>

**Endereço do vídeo do Youtube:** <https://youtu.be/r64regjGPdE>

## 1. Introdução

O mercado de energia elétrica é caracterizado por alta volatilidade de preços, influenciado por diversos fatores como demanda, capacidade de geração, condições climáticas, sazonalidade e políticas regulatórias. A previsão precisa dos preços de energia é fundamental para diversos agentes do setor, incluindo distribuidoras, geradores, consumidores industriais e traders de energia.

Nos últimos anos, técnicas de Inteligência Artificial, especialmente Machine Learning, têm demonstrado grande eficácia na análise de séries temporais e predição de variáveis complexas no setor energético. A capacidade de processar grandes volumes de dados históricos e identificar padrões não lineares torna essas técnicas particularmente adequadas para o problema de predição de preços.

A predição precisa de preços de energia proporciona benefícios significativos para o planejamento operacional e estratégico das empresas do setor. Para distribuidoras e geradores, permite otimização da alocação de recursos e redução de custos operacionais.

Para consumidores industriais, possibilita melhor planejamento de demanda e negociação de contratos mais vantajosos.

Além disso, a volatilidade dos preços de energia pode impactar diretamente a economia, afetando desde o custo de produção industrial até as tarifas pagas pelos consumidores finais. Ferramentas preditivas baseadas em Machine Learning podem contribuir para maior estabilidade e eficiência do mercado energético.

O objetivo geral deste projeto é desenvolver um modelo de Machine Learning capaz de prever os preços de energia elétrica com base em dados históricos de consumo, geração, condições meteorológicas e preços anteriores.

- Realizar análise exploratória do dataset para identificar padrões, correlações e características relevantes;
- Preparar e transformar os dados para aplicação de algoritmos de Machine Learning;
- Implementar e comparar diferentes algoritmos de regressão para predição de preços;
- Avaliar o desempenho dos modelos utilizando métricas apropriadas;
- Identificar as variáveis mais relevantes para a predição de preços de energia.

Este projeto empregando o uso dos conceitos de Machine Learning, especificamente com a biblioteca scikit-learn para implementar algoritmos voltados à solução de um problema supervisionado de regressão no contexto do mercado de energia.

## **2. Descrição do Problema**

O problema abordado neste projeto consiste na predição de preços de energia elétrica em mercados spot ou de curto prazo. Os preços de energia são determinados pela interação entre oferta e demanda, sendo influenciados por múltiplos fatores:

### **Fatores de Demanda:**

- Consumo histórico de energia
- Sazonalidade (dia da semana, hora do dia, mês do ano)
- Temperatura e condições climáticas
- Atividade econômica e industrial

### **Fatores de Oferta:**

- Capacidade de geração disponível
- Mix de fontes energéticas (renováveis, fósseis, nuclear)
- Condições meteorológicas que afetam geração renovável (vento, radiação solar)
- Manutenções programadas e falhas não programadas

### **Características do Problema:**

O problema pode ser formulado como uma tarefa de regressão supervisionada, onde:

- **Variável alvo (y):** Preço da energia elétrica em um determinado período (janela temporal)

- **Variáveis preditoras (X):** Consumo, geração por fonte, temperatura, velocidade do vento, radiação solar, precipitação, indicadores temporais, entre outros

#### **Desafios:**

- Não linearidade das relações entre variáveis
- Presença de sazonalidade em múltiplas escalas temporais
- Volatilidade e possíveis valores extremos (*spikes de preço*)
- Necessidade de tratamento de dados faltantes
- Correlações complexas entre variáveis meteorológicas e geração/consumo

Portanto, pode-se perceber que a questão da pesquisa é: Como algoritmos de aprendizado podem efetivamente manter a precisão preditiva em ambientes de alta frequência e volatilidade?

Este problema demanda soluções que combinem eficiência computacional, capacidade adaptativa para lidar com mudanças nos padrões e robustez estatística para manter qualidade preditiva mesmo com dados limitados.

### **3. Referencial Teórico**

#### **3.1. Machine Learning e Aprendizado Supervisionado**

Machine Learning é um subcampo da Inteligência Artificial que permite que sistemas computacionais aprendam e melhorem a partir da experiência sem serem explicitamente programados. O aprendizado supervisionado é uma categoria de Machine Learning onde o algoritmo aprende a partir de dados rotulados, ou seja, dados de entrada associados a respostas conhecidas.

No contexto de predição de preços, utilizamos algoritmos de regressão, que buscam aprender uma função  $f(X)$  que mapeia variáveis de entrada  $X$  para uma variável contínua de saída  $y$  (preço).

#### **3.2. Algoritmos de Regressão**

##### **3.2.1. Regressão Linear**

A Regressão Linear é o algoritmo mais simples e interpretável, que assume uma relação linear entre as variáveis preditoras e a variável alvo. O modelo pode ser representado como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$$

Apesar de sua simplicidade, a Regressão Linear pode servir como baseline para comparação com modelos mais complexos.

##### **3.2.2. Random Forest**

Random Forest é um algoritmo de ensemble que constrói múltiplas árvores de decisão durante o treinamento e produz a predição através da média das predições individuais. Características importantes:

- Reduz overfitting através da aleatoriedade na construção das árvores
- Robusto a outliers e valores faltantes

- Capaz de capturar relações não lineares complexas
- Fornece medidas de importância das variáveis

### 3.2.3. Gradient Boosting

Gradient Boosting é outra técnica de ensemble que constrói árvores de forma sequencial, onde cada nova árvore tenta corrigir os erros das árvores anteriores. É conhecido por:

- Alta capacidade preditiva em diversos domínios
- Flexibilidade no tratamento de diferentes tipos de dados
- Possibilidade de ajuste fino através de hiperparâmetros
- Requer cuidado para evitar overfitting

## 3.3. Pré-processamento de Dados

O pré-processamento é etapa crucial que pode determinar o sucesso do modelo. Técnicas comuns incluem:

### Tratamento de Valores Faltantes:

- Remoção de registros incompletos
- Imputação por média, mediana ou métodos mais sofisticados
- Interpolação temporal para dados de séries temporais

### Normalização e Padronização:

- Min-Max Scaling: transforma dados para intervalo  $[0,1]$
- Standardization: transforma dados para média 0 e desvio padrão 1
- Importante para algoritmos sensíveis à escala das variáveis

### Engenharia de Features:

- Criação de variáveis derivadas (hora do dia, dia da semana, mês)
- Transformações não lineares
- Interações entre variáveis
- Features baseadas em janelas temporais (médias móveis, lags)

## 3.4. Validação e Avaliação de Modelos

Para dados de séries temporais, é fundamental respeitar a ordem temporal na validação:

### Time Series Split:

- Divide os dados mantendo a sequência temporal
- Evita vazamento de informação do futuro para o passado (*data leakage*)
- Simula cenário real de predição

### Métricas de Avaliação:

- **MAE (Mean Absolute Error):** média dos erros absolutos, na mesma unidade da variável alvo
- **RMSE (Root Mean Squared Error):** penaliza erros maiores de forma quadrática
- **R<sup>2</sup> (Coeficiente de Determinação):** proporção da variância explicada pelo modelo (0 a 1)
- **MAPE (Mean Absolute Percentage Error):** erro percentual médio, útil para comparações

### 3.5. Aplicações de ML em Mercados de Energia

Diversos estudos têm demonstrado a eficácia de técnicas de Machine Learning para predição de preços de energia:

- Modelos baseados em árvores (Random Forest, Gradient Boosting) frequentemente apresentam bom desempenho devido à capacidade de capturar não linearidades
- Features temporais e sazonais são frequentemente identificadas como importantes preditoras
- A inclusão de variáveis meteorológicas melhora significativamente as predições em sistemas com alta penetração de energias renováveis
- Ensemble methods tendem a superar modelos individuais

## 4. Metodologia

### 4.1. Dataset

O dataset foi obtido do Kaggle (Nicholas Jhana – *Energy Consumption, Generation, Prices and Weather*) por meio da API kagglehub. O conjunto de dados contém registros temporais com informações sobre:

- **Variável alvo:** preços de energia elétrica.
- **Variáveis preditoras:** consumo total e por setor, geração por fonte (fóssil, nuclear, eólica, solar, hidrelétrica), variáveis meteorológicas (temperatura, velocidade do vento, radiação solar, precipitação) e indicadores temporais (data, hora).

O dataset é adequado ao problema proposto por integrar múltiplas dimensões influenciadoras dos preços de energia, possibilitando análise das relações entre oferta, demanda e condições ambientais.

### 4.2. Análise Exploratória dos Dados

A análise exploratória visa compreender estrutura, qualidade e características dos dados antes da modelagem.

#### Análise Descritiva:

- Estatísticas descritivas das variáveis numéricas (média, mediana, desvio padrão, quartis).
- Identificação de valores faltantes e sua proporção.

- Análise das distribuições por meio de histogramas.
- Detecção de outliers com boxplots e quartis.

#### **Análise Temporal:**

- Visualização da série temporal de preços para identificar tendências.
- Verificação de padrões sazonais em escalas diária, semanal, mensal e anual.
- Análise de autocorrelação.
- Decomposição em tendência, sazonalidade e resíduo.

#### **Análise de Correlações:**

- Construção da matriz de correlação.
- Identificação de variáveis mais correlacionadas ao preço.
- Avaliação de multicolinearidade entre preditores.
- Heatmaps e scatter plots para relações relevantes.

#### **Análise de Relações:**

- Relação entre consumo e preços.
- Impacto das fontes de geração no preço.
- Influência de variáveis meteorológicas.
- Estudo de eventos extremos e seus efeitos.

### **4.3. Preparação dos Dados**

#### **Limpeza e Tratamento:**

- Imputação de valores faltantes por interpolação temporal ou mediana.
- Tratamento de outliers conforme contexto.
- Verificação de consistência temporal.
- Correção de erros de formatação ou valores inválidos.

#### **Engenharia de Features:**

- Extração de componentes temporais (hora, dia da semana, mês etc.).
- Criação de variáveis cíclicas (seno/cosseno).
- Criação de *lags* ( $t-1$ ,  $t-24$ ,  $t-168$ ).
- Cálculo de médias móveis (24h, 7 dias, 30 dias).
- Features de interação (ex.: consumo/geração).
- Transformações logarítmicas quando apropriado.

#### **Seleção de Features:**

- Análise de importância e correlação.
- Remoção de variáveis altamente correlacionadas ( $r > 0,9$ ).

- Consideração de conhecimento de domínio.
- Testes de diferentes combinações.

#### **Divisão dos Dados:**

- *Split* temporal para evitar *data leakage*.
- Treino: 80%
- Teste: 20%
  - O conjunto de teste simula previsões futuras reais.

#### **Normalização:**

- Uso de `StandardScaler` com *fit* apenas no treino.
- Aplicação separada em treino e teste.
- Variáveis cíclicas não normalizadas.

### **4.4. Modelagem**

#### **Algoritmos Selecionados:**

##### **Regressão Linear:**

- Modelo *baseline* para referência.
- Assume relações lineares.
- Alta interpretabilidade.
- Regularização Ridge para mitigar *overfitting*.

##### **Random Forest Regressor:**

- Ensemble de árvores com média.
- Captura relações não lineares.
- Resistente a outliers.
- Oferece importância das features.

##### **Gradient Boosting Regressor:**

- Ensemble sequencial corrigindo erros anteriores.
- Alto desempenho preditivo.
- Flexibilidade de hiperparâmetros.
- Maior risco de *overfitting* se não ajustado.

#### **Processo de Treinamento:**

- Implementação com `scikit-learn`.
- Validação com **`TimeSeriesSplit (5 folds)`**.
- **Grid Search** para otimização.
- Monitoramento contínuo para evitar *overfitting*.

### **4.5. Avaliação e Validação**

## **Métricas**

- **MAE**: erro médio absoluto.
- **RMSE**: penaliza grandes erros.
- **R<sup>2</sup>**: variância explicada.
- **MAPE**: erro percentual.

## **Análise Comparativa:**

- Comparação tabular de métricas.
- Identificação do melhor modelo.
- Tempos de treino e predição.
- Balanceamento entre complexidade e desempenho.

## **Interpretabilidade:**

- Importância de features para RF e GBoost.
- Coeficientes da Regressão Linear.
- Identificação de variáveis mais impactantes.

## **Validação Robusta:**

- Comparação visual entre valores reais e preditos.
- Análise de resíduos.
- Testes por subperíodos.
- Avaliação em condições extremas.

## **4.6. Ferramentas e Tecnologias.**

- Python 3.12
- pandas, numpy, matplotlib, seaborn, scikit-learn, kagglehub
- Jupyter Notebook / Google Colab
- Git/GitHub

## **5. Resultados e discussão**

A análise concentrou-se em sete variáveis representativas do sistema energético – fontes fósseis de geração (gás e carvão), carga real, preços (dia seguinte e real) e condições climáticas (temperatura e vento). As distribuições destas variáveis revelaram comportamentos heterogêneos: geração térmica com alta variabilidade operacional, carga com padrões multimodais, clima com forte sazonalidade e preços marcados por volatilidade, caudas pesadas e múltiplos regimes. A matriz de correlação confirmou relações coerentes com a teoria econômica: preços influenciados por oferta, demanda e clima, além de forte autocorrelação entre preços atuais e futuros. As correlações cruzadas entre variáveis reforçaram a necessidade de modelos capazes de capturar interações e não-linearidades.

A análise temporal revelou mudanças estruturais nos preços ao longo dos anos, com fases alternadas de estabilidade e alta volatilidade. Estes “desvios de conceito” indicam violações na suposição de estacionaridade temporal, tornando modelos propensos a perda de desempenho ao serem aplicados em períodos com regimes diferentes dos de treinamento. O histograma dos preços confirmou uma distribuição assimétrica com eventos extremos genuínos e múltiplos modos, indicando que abordagens lineares simples podem ser insuficientes. A divisão temporal entre treino e teste respeitou a sequência histórica, essencial para evitar validação irrealisticamente otimista e para capturar corretamente o impacto do concept drift.

O modelo de Regressão Linear apresentou desempenho forte no treino ( $R^2 \approx 0,92$ ), com erros médios baixos e viés mínimo. No teste, mostrou degradação moderada e controlada:  $R^2$  caiu para  $\approx 0,84$ , e os erros aumentaram levemente, mas mantiveram-se estáveis. A interpretabilidade dos coeficientes e a robustez temporal refletiram a vantagem estrutural da linearidade sob regimes mutáveis. O Random Forest, configurado com 100 árvores e profundidade 15, superou amplamente a Regressão Linear no treino ( $R^2 \approx 0,94$ ; erros cerca de 20% menores). Entretanto, no teste sofreu forte deterioração:  $R^2$  caiu para  $\approx 0,80$ , MAE e RMSE superaram os da Regressão Linear, e o modelo apresentou grande sensibilidade ao concept drift. A flexibilidade que permitiu capturar não-linearidades e interações profundas no treino tornou-se vulnerabilidade quando aplicado a novos regimes de mercado.

A comparação integrada mostrou que a Regressão Linear, embora menos expressiva, generalizou melhor temporalmente. O Random Forest apresentou overfitting às particularidades históricas do período de treino, perdendo desempenho em cenários futuros. Os gráficos comparativos (resíduos, previsões vs. valores reais, distribuição de erros) apontam que o modelo linear produz erros mais uniformes e estáveis, enquanto o Random Forest apresenta maior dispersão e instabilidade em períodos de mudança estrutural.

Em síntese, o estudo demonstrou que a complexidade do mercado energético – não-linear, volátil e sujeito a concept drift – exige equilíbrio entre capacidade preditiva e robustez. Embora modelos avançados capturem melhor os padrões históricos, modelos mais simples podem oferecer previsões mais confiáveis diante de mudanças nos regimes de mercado.

## 6. Conclusão

A avaliação comparativa entre Regressão Linear e Random Forest para previsão de preços de energia revelou que ambos os modelos conseguem capturar parte substancial da estrutura preditiva dos dados – atingindo MAPE entre 4,94% e 5,51% no teste, valores adequados para aplicações reais de suporte à decisão e gestão de risco. A Regressão Linear estabeleceu um baseline robusto ( $R^2 = 0,8382$ , MAE  $\approx \text{€}2,87/\text{MWh}$ ), enquanto o Random Forest, embora mais vulnerável a mudanças temporais, ainda explicou aproximadamente 80% da variância futura ( $R^2 = 0,7972$ ). Esses resultados confirmam que as sete variáveis escolhidas capturam, de forma sintética, a essência da formação de preços no sistema elétrico — combinando oferta, demanda, persistência histórica e clima.

Contudo, a principal conclusão do estudo não reside na comparação direta dos algoritmos, mas na evidência inequívoca de concept drift, que surge como a limitação

estrutural dominante para modelos preditivos em mercados energéticos. Os gráficos temporais e os “pontos de inflexão” identificados mostram que os preços não constituem um processo estacionário; regimes mudam abrupta ou gradualmente, invalidando premissas de estabilidade que modelos tradicionais pressupõem.

Três evidências convergentes fundamentam esse diagnóstico: (1) degradações substanciais de  $R^2$  do treino para o teste — 8,25pp para Regressão Linear e 14,48pp para Random Forest; (2) inversões sistemáticas de viés (MPE mudando de negativo para positivo em ambos os modelos), revelando deslocamentos do nível médio dos preços; e (3) deterioração proporcionalmente maior do modelo mais flexível, cuja vantagem no treino torna-se desvantagem no teste, evidenciando sobreajuste a regimes específicos.

Essas evidências indicam que o problema fundamental não é falta de capacidade dos modelos, mas sim incompatibilidade entre modelos estáticos e dados não-estacionários. O Random Forest captura não-linearidades reais, mas também padrões temporários que perdem validade quando o regime muda; já a Regressão Linear, ao impor restrições estruturais, generaliza melhor temporalmente — mesmo sendo menos expressiva.

As escolhas de modelagem foram metodologicamente defendidas: a Regressão Linear oferece interpretabilidade e robustez, com regularização implícita via estrutura; o Random Forest foi calibrado para equilibrar flexibilidade e controle de variância; e a divisão temporal foi essencial para evitar validação irrealista, expondo o impacto real do drift. Essa metodologia permitiu diagnosticar que o desafio não é aprimorar um algoritmo específico, mas sim criar sistemas de previsão adaptativos.

As implicações práticas desse diagnóstico são diretas e quantificáveis: diferenças aparentemente pequenas em MAE e RMSE traduzem-se em dezenas a centenas de milhares de euros anuais de exposição ao risco, especialmente porque o Random Forest comete erros extremos com maior frequência. Além disso, previsões com MAPE ~5% definem padrões de incerteza que precisam ser integrados a estratégias de hedge e operação.

As análises sugerem uma agenda clara para futuras pesquisas em modelagem adaptativa: modelos com janelas deslizantes, detectores automáticos de drift (CUSUM, Page-Hinkley, ADWIN), ensembles dinâmicos que balanceiem memória e adaptabilidade, redes neurais recorrentes e transformers capazes de “esquecer” padrões obsoletos, aprendizado online que atualiza modelos continuamente, incorporação de informação contextual externa (regulação, clima extremo, geopolítica), e frameworks híbridos que integrem conhecimento físico-econômico.

Esse conjunto de direções reflete a necessidade de transição para sistemas que tratam modelos como processos dinâmicos e não como artefatos estáticos. Para domínios com dados fortemente não-estacionários, como energia, a verdadeira métrica de sucesso não é maximizar  $R^2$  estático, mas manter desempenho estável e previsões confiáveis à medida que o mercado evolui.

Em síntese, as conclusões do estudo estabelecem que o conceito central não é “qual modelo performa melhor”, mas sim “como lidar com mudança”. Os resultados quantitativos e visuais revelam que adaptabilidade — e não complexidade — será a característica definidora da próxima geração de sistemas preditivos para mercados energéticos. O desafio deixa de ser puramente técnico e torna-se sistêmico: construir

infraestruturas capazes de monitorar drift, retrainar imediatamente, recalibrar incertezas e justificar decisões em ambientes cada vez mais dinâmicos e críticos para a transição energética.

Este trabalho é uma extensão do Trabalho de Conclusão de Curso que irei realizar, onde se propõe uma pesquisa dedicada ao uso de algoritmos adaptativos baseados em Árvores de Hoeffding — modelos projetados especificamente para aprendizado incremental e detecção natural de mudanças de distribuição. Essas árvores, ao processarem cada exemplo apenas uma vez e utilizarem limites de Hoeffding para decidir quando dividir nós, oferecem capacidade de adaptação contínua com baixo custo computacional, característica essencial em mercados energéticos de alta frequência. A investigação futura examinará não apenas a performance isolada dessas árvores, mas também sua integração em ensembles adaptativos — como Adaptive Random Forests — avaliando sua habilidade de responder a concept drift abruptos e graduais, manter estabilidade durante períodos estacionários e recuperar desempenho rapidamente após transições de regime. Essa linha de pesquisa busca consolidar um framework de previsão verdadeiramente online, capaz de aprender em tempo real, manter robustez e reduzir significativamente a deterioração observada em modelos estáticos, avançando um passo decisivo rumo a sistemas preditivos plenamente adaptativos para mercados de energia.

## 7. Referências bibliográficas

- BREIMAN, L. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5-32, 2001.
- FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, v. 29, n. 5, p. 1189-1232, 2001.
- GÉRON, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2ª ed. O'Reilly Media, 2019.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2ª ed. Springer, 2009.
- JAMES, G. et al. An Introduction to Statistical Learning with Applications in R. Springer, 2013.
- KAGGLE. Energy Consumption Generation Prices and Weather Dataset. Disponível em: <https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather>.
- PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825-2830, 2011.