



ANAIS

PREVISÃO DE DEMANDA DE MATERIAIS DE MANUTENÇÃO NO SETOR SUCROENERGÉTICO: APLICAÇÃO DE MODELOS PREDITIVOS E MACHINE LEARNING

JOSE HENRIQUE GARRIDO DE LIMA

garridolima@hotmail.com

FATEC - FACULDADE DE TECNOLOGIA ESTADO DE SÃO PAULO

RESUMO: A gestão de materiais de manutenção, reparo e operação (MRO) no setor sucroenergético é desafiada pela marcante sazonalidade do ciclo de safra e por padrões de demanda intermitentes em milhares de SKUs. Este estudo desenvolve e avalia um framework de previsão de demanda que combina a classificação automática de perfis de consumo, pelo esquema de categorização de Syntetos, Boylan e Croston (2005), com limiares revisados por Kostenko e Hyndman (2006), à avaliação comparativa de sete modelos preditivos: um modelo de referência (Naive Sazonal), modelos estatísticos clássicos (ARIMA, SARIMA, Holt-Winters, CrostonTSB, SBA) e o LightGBM, algoritmo de machine learning avaliado em três versões sucessivas, correspondentes a diferentes configurações de hiperparâmetros e variáveis de entrada (features). A validação seguiu backtesting com holdout de 12 meses, precedido por 48 meses de treinamento, sobre 2.296 SKUs classificados em quatro quadrantes de demanda, medindo acurácia por WMAPE e direção do erro por Bias. Os resultados mostram que nenhum modelo domina uniformemente todos os quadrantes: o LightGBM lidera no quadrante de demanda regular (Smooth), com WMAPE de 49,6%, enquanto no quadrante Erratic (5,6% dos SKUs) o próprio modelo de referência obtém o menor erro (78,8%). O refinamento sucessivo do LightGBM reduziu o WMAPE e corrigiu parte do viés sistemático nos quadrantes mais intermitentes, com todas as versões treinadas em menos de oito segundos, contra 1.375,5 segundos do SARIMA. O quadrante Intermittent, que concentra 54,3% dos SKUs, revelou-se o mais desafiador — todos os nove modelos registraram WMAPE superior a 121%, já que a alta proporção de meses sem consumo torna o sinal ruidoso para qualquer método, estatístico ou de máquina.

PALAVRAS CHAVE: Previsão de Demanda; Machine Learning; LightGBM; Demanda Intermitente; Gestão de Estoques MRO; Setor Sucroenergético

ABSTRACT: The management of maintenance, repair, and operations (MRO) materials in the sugar-energy sector is challenged by the pronounced seasonality of the harvest cycle and by intermittent demand patterns across thousands of SKUs. This study develops and evaluates a demand forecasting framework that combines automatic classification of consumption profiles, using the categorization scheme of Syntetos, Boylan, and Croston (2005), with thresholds revised by Kostenko and Hyndman (2006), with a comparative assessment of seven predictive models: a benchmark model (Seasonal Naive), classical statistical models (ARIMA, SARIMA, Holt-Winters, Croston-TSB, SBA), and LightGBM, a machine learning algorithm evaluated across three successive versions corresponding to different hyperparameter configurations and input features. Validation followed a backtesting approach with a 12-month holdout period, preceded by 48 months of training, applied to 2,296 SKUs classified into four demand quadrants, measuring accuracy via WMAPE and error direction via Bias. Results show that no single model uniformly dominates across all quadrants: LightGBM leads in the regular demand quadrant (Smooth) with a WMAPE of 49,6% whereas in the Erratic quadrant (5.6% of SKUs) the benchmark model itself achieves the lowest error (78.8%). Successive refinement of LightGBM reduced WMAPE and corrected part of the systematic bias in the more intermittent quadrants, with all versions trained in under eight seconds, compared to 1,375.5 seconds for SARIMA. The Intermittent quadrant, which accounts for 54.3% of SKUs, proved to be the most challenging: all nine models recorded WMAPE above 121%, as the high proportion of zero-demand months renders the signal noisy for any method, whether statistical or machine learning-based.

KEY WORDS: Demand Forecasting; Machine Learning; LightGBM; Intermittent Demand; MRO Inventory Management; Sugarcane-energy Sector

1. INTRODUÇÃO

A gestão de materiais de manutenção, reparo e operação (MRO) em empresas sucroenergéticas é particularmente desafiadora em função da marcante sazonalidade do ciclo produtivo. Na região Centro-Sul do Brasil, responsável pela maior parte da produção nacional de cana-de-açúcar, a safra se estende de abril a novembro, impondo um ritmo de consumo de peças e materiais para manutenção automotiva e implementos, classificados como MRO Automotivo, profundamente distinto do período de entressafra. Essa dinâmica bifásica, combinada a milhares de SKUs (stock keeping unit) com muitos períodos sem consumo, cria padrões de demanda heterogêneos que tornam os modelos tradicionais de planejamento, baseados apenas em médias históricas fixas, insuficientes para uma gestão eficiente.

A adoção de sistemas ERP tornou possível o registro sistemático de todas as movimentações de materiais ao longo dos anos. Contudo, conforme apontado por Lima e Campos (2013), ainda há lacunas significativas na utilização analítica desses dados para o planejamento da demanda. A disponibilidade de séries históricas longas abre a oportunidade para a aplicação de modelos quantitativos avançados de previsão, combinando técnicas de inteligência artificial e modelos estatísticos especializados para forecasting, sendo o forecasting estatístico, segundo Makridakis, Spiliotis e Assimakopoulos (2018), uma disciplina em transformação diante da crescente disponibilidade de métodos de aprendizado de máquina, cuja capacidade de generalização a partir de múltiplas séries simultâneas representa uma ruptura metodológica frente à abordagem tradicional de ajuste individual por série.

O presente estudo descreve o desenvolvimento e a avaliação de um framework de machine learning para previsão de demanda de 2.296 SKUs de uma empresa sucroenergética com estrutura industrial e agrícola integrada, abrangendo o processamento da cana-de-açúcar e as atividades de manejo dos canaviais. A metodologia combina classificação automática de perfis de consumo via esquema de categorização de Syntetos, Boylan e Croston (2005), com limiares revisados por Kostenko e Hyndman (2006), com a avaliação de múltiplos modelos preditivos estatísticos e o LightGBM, algoritmo de gradient boosting cuja família de modelos dominou as soluções de melhor desempenho na competição M5 (Makridakis, Spiliotis e Assimakopoulos, 2022). Os modelos foram validados por backtesting — técnica que consiste em simular a previsão sobre um período passado já conhecido, permitindo comparar diretamente o valor previsto ao que de fato ocorreu, com período de teste (holdout) de 12 meses, correspondente a uma safra completa já realizada — precedido por 48 meses utilizados para treinamento. O objetivo central é identificar quais modelos apresentam maior acurácia na previsão de demanda no consumo com medição por WMAPE (*Weighted Mean Absolute Percentage Error*, ou Erro Percentual Absoluto Médio Ponderado) e *Bias* (viés) com especial atenção às dificuldades inerentes à demanda intermitente em ambientes sazonais agroindustriais.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Classificação de demanda e o esquema de categorização

A literatura de gestão de estoques classifica os padrões de demanda em função de dois eixos principais: a frequência de ocorrência e a variabilidade do tamanho das requisições. Syntetos, Boylan e Croston (2005), em estudo validado em 3.000 séries reais da indústria de produção automotiva, propuseram um esquema de categorização baseado em dois parâmetros: o intervalo médio entre demandas (ADI, do inglês *Average Demand Interval*), que mede a frequência de ocorrência, onde valores maiores indicam maior intermitência — e o coeficiente de variação ao quadrado dos tamanhos de demanda (CV^2), que mede a variabilidade do volume

consumido em cada evento. Os autores criticam explicitamente a prática arbitrária de categorizar padrões de demanda sem fundamentação:

"The common practice in the inventory control software industry is to arbitrarily categorize those demand patterns and then proceed to select an estimation procedure and optimize the forecast parameters." (SYNTETOS; BOYLAN; CROSTON, 2005, p. 495)

Como alternativa, os autores propõem que os próprios modelos de previsão definam onde uma categoria termina e outra começa: calcularam, tanto analiticamente quanto testando em grandes bases de dados reais, a partir de qual ponto o método Croston deixa de ser o estimador mais preciso e sua correção (SBA) passa a superá-lo, medindo o erro pelo Erro Quadrático Médio. A fronteira entre as categorias é fixada exatamente nesse ponto de virada. Os limiares resultantes definem quatro quadrantes: suave (*smooth*, $ADI \leq 1,32$ e $CV^2 \leq 0,49$), errática (*erratic*, $ADI \leq 1,32$ e $CV^2 > 0,49$), intermitente (*intermittent*, $ADI > 1,32$ e $CV^2 \leq 0,49$) e irregular (*lumpy*, $ADI > 1,32$ e $CV^2 > 0,49$). Kostenko e Hyndman (2006) revisitaram essa derivação e demonstraram que a fronteira original era uma aproximação: a partir de um resultado analítico exato, os autores mostraram que os limiares corretos são $ADI = 4/3$ ($\approx 1,33$) e $CV^2 = 0,5$. A diferença é pequena e não altera a lógica de categorização, mas o refinamento é relevante do ponto de vista teórico e comprovam ainda mais o método — por essa razão, os limiares de Kostenko e Hyndman (2006) foram os adotados no presente estudo.

A vantagem dessa abordagem é que a categoria de um item deixa de ser um rótulo abstrato e passa a indicar diretamente qual modelo tende a prever melhor aquele padrão — a classificação decorre diretamente do comportamento histórico do próprio SKU, calculado a partir do ADI e do CV^2 observados na janela de treinamento. No presente estudo, os quatro quadrantes teóricos do esquema — *Smooth*, *Erratic*, *Intermittent* e *Lumpy* — foram mantidos separados, em vez de consolidados em categorias operacionais mais amplas, permitindo testar empiricamente, para cada quadrante puro, qual entre os modelos avaliados apresenta melhor desempenho preditivo. A distribuição dos SKUs da base deste estudo nos quadrantes é apresentada na Seção 3.3.

2.2 Modelos de previsão de demanda

Segundo Hyndman e Athanasopoulos (2021), um método de previsão é qualquer procedimento reproduzível que gera previsões a partir de dados passados, variando desde regras simples de repetição até modelos estatísticos com estrutura probabilística explícita. A escolha do modelo adequado depende diretamente do padrão de demanda do item: séries regulares, intermitentes e esporádicas respondem melhor a famílias de modelos distintas. Esse mesmo referencial tem sido adotado em pesquisas brasileiras recentes, como a de Saraiva (2025), que comparou métodos estatísticos e de aprendizado de máquina na previsão de demanda logística para comércio eletrônico B2C. Esta seção organiza os sete modelos avaliados neste estudo segundo as principais famílias reconhecidas na literatura de previsão de séries temporais — todos aplicados uniformemente aos quatro quadrantes do esquema de categorização, sem atribuição prévia de um modelo "correto" por categoria, de modo que a correspondência entre teoria e desempenho empírico pudesse ser testada, e não assumida.

Modelo de referência (baseline): o Naive Sazonal (sNaive) projeta cada período futuro como o último valor observado na mesma posição do ciclo sazonal anterior — no caso de séries mensais, o mesmo mês do ano anterior. Trata-se do mais simples dos métodos aplicáveis a séries sazonais: não estima nenhum parâmetro, limitando-se a replicar o padrão do ciclo imediatamente anterior. Como a demanda de materiais MRO no setor sucroenergético é estruturada pelo ciclo anual de safra e entressafra (Seção 1), esse é o benchmark teoricamente apropriado ao presente estudo. Hewamalage, Ackermann e Bergmeir (2023) advertem que a

adoção de um benchmark inadequadamente fraco é uma das falhas metodológicas mais recorrentes na literatura de previsão, produzindo ganhos aparentes que não se sustentam frente a uma referência corretamente especificada.

Modelos autorregressivos (Box-Jenkins): família baseada na autocorrelação — o modelo estima a relação entre o valor atual e os valores passados da própria série (BOX et al., 2015). O ARIMA (*AutoRegressive Integrated Moving Average*) modela a série a partir de três componentes — autorregressivo, de diferenciação e de médias móveis —, sem incorporar sazonalidade explícita. O SARIMA (*Seasonal ARIMA*) estende o modelo com um componente sazonal adicional, adequado a séries com ciclo anual bem definido, como o calendário de safra; a comparação entre os dois permite isolar empiricamente o ganho atribuível à modelagem da sazonalidade.

Suavização exponencial: família que pondera as observações recentes com peso maior que as antigas, via parâmetros de suavização ajustados aos dados. O Holt-Winters modela simultaneamente três componentes: nível, tendência e sazonalidade, sendo por isso adequado a séries com padrão sazonal recorrente (HOLT, 1957; WINTERS, 1960).

Modelos para demanda intermitente: família dedicada a séries com períodos sem consumo. O SBA (*Syntetos-Boylan Approximation*) corrige o viés sistemático identificado no método original de Croston, aplicando um fator de ajuste multiplicativo (SYNTETOS; BOYLAN, 2005). O TSB (*Teunter-Syntetos-Babai*) substitui a atualização do intervalo por uma probabilidade de ocorrência atualizada a cada período, tornando o modelo mais responsivo a mudanças recentes no padrão de consumo, inclusive a possível obsolescência do item (TEUNTER; SYNTETOS; BABAI, 2011).

Machine learning e modelos globais: machine learning é uma subárea da inteligência artificial dedicada ao desenvolvimento de algoritmos que melhoram seu desempenho em uma tarefa por meio da experiência (dados), reduzindo a necessidade de regras explicitamente programadas (JORDAN; MITCHELL, 2015). Diferentemente dos modelos anteriores, ajustados individualmente para cada série a partir apenas do seu próprio histórico, os modelos de machine learning aplicados a séries temporais costumam adotar uma abordagem global: um único modelo é treinado simultaneamente sobre múltiplas séries, aprendendo padrões compartilhados entre elas — estratégia conhecida como cross-learning. O LightGBM (*Light Gradient Boosting Machine*), desenvolvido pela Microsoft Research e publicado em 2017, é um algoritmo de gradient boosting baseado em árvores de decisão, distinguindo-se por técnicas de amostragem e agrupamento de atributos que aumentam significativamente a velocidade de processamento sem perda de acurácia frente a implementações anteriores de boosting (KE et al., 2017).

2.3 Configuração do LightGBM: função de perda Tweedie e hiperparâmetros

O LightGBM (KE et al., 2017) constrói, de forma sequencial, um conjunto de árvores de decisão simples, no qual cada nova árvore corrige os erros deixados pelas anteriores. Para isso, o algoritmo precisa de um critério que defina o que conta como "erro" durante esse ajuste — esse critério é chamado de função de perda.

A opção padrão de função de perda do LightGBM é mais adequada a dados que variam de forma suave em torno de uma média. Uma alternativa é a função de perda Tweedie, discutida por Januschowski et al. (2022) como particularmente adequada a dados de alvo esparsos, ou seja, com excesso de valores iguais a zero, por lidar melhor com séries que têm muitos valores nulos intercalados com valores positivos. Esse padrão corresponde justamente à demandas intermitentes e irregulares caracterizada na Seção 2.1, o que motiva a avaliação comparativa dessa função de perda frente à opção padrão neste estudo (Seção 3.5). Quando utilizada, essa

função de perda possui um parâmetro interno, chamado potência de variância, que controla o quanto o modelo tolera picos ocasionais de consumo.

Adicionalmente, este estudo avalia o efeito de hiperparâmetros que limitam a complexidade do LightGBM — como o número de árvores criadas e a profundidade de cada uma — frente à configuração padrão da biblioteca. Hiperparâmetros mais restritivos reduzem o risco de o modelo memorizar padrões específicos apenas dos dados de treinamento (fenômeno chamado sobreajuste, ou *overfitting*), o que pode ajudar o modelo a generalizar melhor para os meses futuros que ele nunca viu. As configurações efetivamente testadas, e os resultados de cada uma, são apresentados nas Seções 3.5 e 4.5.

2.4 Engenharia de features

A engenharia de features consiste na transformação de dados brutos em variáveis que representem explicitamente padrões relevantes ao problema, aumentando a capacidade de um modelo de aprendizado extrair sinal útil dos dados disponíveis. Diferentemente dos modelos estatísticos clássicos, ajustados a cada série individualmente e que incorporam implicitamente sua própria estrutura de autocorrelação, o LightGBM — por operar sobre dados tabulares sem noção intrínseca de tempo — depende inteiramente de que essa estrutura temporal seja tornada explícita por meio de variáveis derivadas.

As features construídas para esse fim podem ser classificadas em duas categorias. **Features internas** são derivadas exclusivamente do histórico de consumo do próprio SKU — é o caso das features adotadas neste estudo, descritas a seguir, incluindo a variável de período do ciclo agrícola (safra ou entressafra), construída a partir do próprio calendário de consumo observado. **Features externas**, por sua vez, incorporam informação alheia ao histórico de consumo do item — por exemplo, a cotação do dólar, dados de planejamento das ordens de manutenção, ou variáveis climáticas. Este estudo não incorporou features externas, restringindo-se a variáveis internas; a incorporação de fontes externas de informação estruturada constitui uma direção natural para trabalhos futuros.

Neste estudo, adotaram-se duas famílias de features internas amplamente empregadas na literatura de previsão de séries temporais (HYNDMAN; ATHANASOPOULOS, 2021): variáveis de defasagem (lags), que informam ao modelo o valor de consumo observado em períodos anteriores específicos — capturando autocorrelação de curto prazo e sazonalidade anual quando a defasagem corresponde a doze períodos —; e variáveis de janela móvel (rolling), que resumem o comportamento recente da série por meio de estatísticas como média e máximo calculados sobre uma janela de meses anteriores, suavizando ruído e fornecendo ao modelo uma noção de nível e magnitude recente. A ambas as famílias se aplicou um deslocamento de um período antes do cálculo da janela, garantindo que nenhuma feature utilizasse informação do próprio mês a ser previsto — prevenção necessária contra vazamento de dados (data leakage), que inflaria artificialmente a acurácia aparente do modelo sem corresponder a um ganho real de capacidade preditiva. Complementarmente, incluiu-se a variável de ciclo agrícola citada acima e uma variável de recência, que mede o número de meses transcorridos desde o último consumo registrado — sinal particularmente relevante para os quadrantes de demanda intermitente e irregular.

2.5 Métricas de acurácia

A métrica principal de acurácia dos modelos adotada neste estudo é o WMAPE (*Weighted Mean Absolute Percentage Error*), um erro percentual que pondera cada período pelo seu volume de consumo — de modo que erros em meses de alto consumo pesam mais que erros em meses de baixo consumo. Quanto maior o resultado, maior o erro do modelo:

$$WMAPE = \sum |real - previsto| / \sum real \times 100$$

O WMAPE evita as distorções do MAPE tradicional, que se torna matematicamente indefinido ou artificialmente inflado em séries com muitos períodos de consumo igual a zero — situação frequente em demanda intermitente (HYNDMAN; ATHANASOPOULOS, 2021).

Ainda assim, um problema estrutural permanece mesmo com o WMAPE: modelos estatísticos projetam valores de consumo esperado matematicamente acima de zero sempre que exista algum sinal histórico na série, por menor que seja — mesmo em meses nos quais o consumo real observado é nulo. Esse comportamento é inerente à natureza probabilística dos modelos de séries temporais e está documentado na literatura especializada (TEUNTER; SYNTETOS; BABAI, 2011).

Complementarmente, o Bias mede a direção sistemática do erro:

$$Bias (\%) = (\sum previsto - \sum real) / \sum real \times 100$$

Um Bias positivo indica superestimação sistemática — o modelo tende a prever mais do que o efetivamente consumido —, enquanto um Bias negativo indica subestimação sistemática. Diferentemente do WMAPE, que penaliza igualmente erros para mais e para menos, o Bias revela a direção predominante do erro, informação relevante do ponto de vista gerencial de estoques por exemplo: a superestimação sistemática tende a gerar excesso de estoque e capital imobilizado, enquanto a subestimação sistemática tende a gerar rupturas de abastecimento.

3. METODOLOGIA

3.1 Base de dados e tratamento preliminar

Os dados de consumo foram extraídos do módulo de gestão de materiais do sistema ERP, restritos aos registros correspondentes a saídas de consumo. Após a consolidação dos registros transacionais em uma base mensal por SKU, a série de cada SKU foi completada com zeros nos meses sem registro de consumo (*zero-fill*), a partir do primeiro mês em que o item apresentou consumo real, evitando penalizar artificialmente itens introduzidos recentemente no catálogo com meses de ausência anteriores à sua própria existência no registro mestre de materiais. Os SKUs com um mês ou menos de consumo registrado no histórico completo foram excluídos da análise, por não permitirem o cálculo do coeficiente de variação necessário à classificação SBC (Seção 2.1). A base final resultante contempla 2.296 SKUs, ao longo de 60 meses de histórico mensal.

3.2 Modelos de previsão e implementação computacional

A implementação computacional dos sete modelos foi realizada em linguagem Python, com o uso de bibliotecas de código aberto especializadas em previsão de séries temporais para grandes volumes de dados, capazes de ajustar centenas ou milhares de modelos individuais de forma automatizada. Os seis modelos estatísticos foram implementados por meio da biblioteca `'statsforecast'`, que automatiza a seleção de parâmetros. No caso do ARIMA e do SARIMA, a ordem do modelo (parâmetros autorregressivo, de diferenciação e de médias móveis, e seus equivalentes sazonais) é selecionada automaticamente por meio de busca via critério de informação (AIC), conforme o algoritmo proposto por Hyndman e Khandakar (2008) — automação fundamental para a aplicação desses modelos em escala, sem a qual seria necessário especificar manualmente a ordem de cada um dos milhares de modelos ajustados.

Complementarmente, o LightGBM foi implementado por meio da biblioteca `'mlforecast'`, integrada à biblioteca `'lightgbm'`. Para cada um dos 12 meses do período de teste,

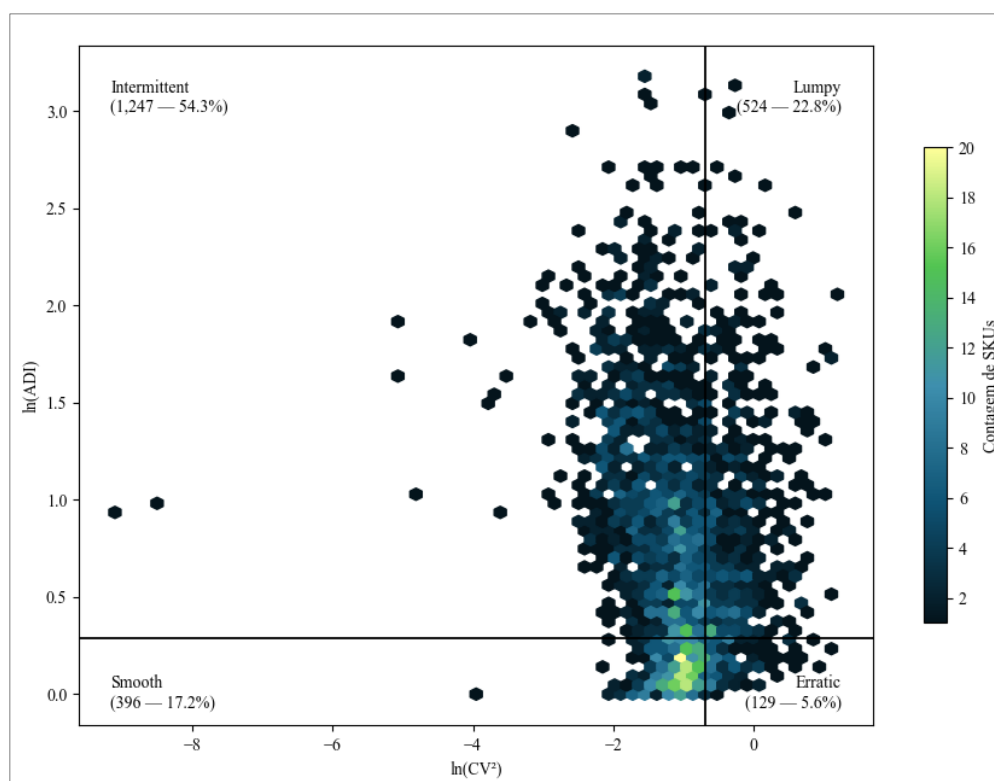
o modelo gera primeiro a previsão do mês corrente e, em seguida, utiliza esse valor previsto — nunca o valor real, que permanece desconhecido durante a validação — como se fosse o dado mais recente disponível, para então calcular as features e gerar a previsão do mês seguinte; esse encadeamento se repete até completar os 12 meses do período de teste. Esse procedimento garante que a avaliação do LightGBM seja submetida à mesma restrição de informação enfrentada pelos demais modelos: todos preveem múltiplos meses consecutivos à frente (previsão multi-passos, ou seja, vários meses seguidos numa mesma rodada, sem atualização com dados de teste entre eles) usando apenas o histórico de treinamento, sem nunca ter acesso, em nenhuma etapa, aos 12 meses reais do período de teste.

Como apoio às pesquisas bibliográficas conduzidas neste estudo e à construção do código Python utilizado na implementação computacional dos modelos, empregou-se o modelo Claude Sonnet 4.6 (Anthropic), sob supervisão e validação integral do autor em cada etapa do desenvolvimento.

3.3 Classificação de demanda

A classificação de cada SKU nos quatro quadrantes do esquema de categorização (Seção 2.1) foi realizada a partir do ADI e do CV^2 calculados exclusivamente sobre a janela de treinamento (48 meses), excluindo deliberadamente os 12 meses reservados para o período de teste. Essa restrição evita que a classificação de um SKU seja influenciada por informação do período que os modelos deveriam prever como se fosse desconhecido — um vazamento indireto de dados que comprometeria a validade do desenho de backtesting. A Figura 1 apresenta a distribuição resultante dos SKUs da base deste estudo nos quatro quadrantes, com os eixos em escala logarítmica, sendo 74 SKUs com $CV^2 = 0$ – consumo sempre igual quando ocorre (não representados no gráfico).

Figura 1. Classificação dos SKUs no esquema de categorização Syntetos, Boylan e Croston, 2005 ($ADI \times CV^2$, escala logarítmica natural – limiares Kostenko e Hyndman, 2006).



Fonte: Spiliotis et al. (2021), adaptado pelo autor.

3.4 Construção de features

As features descritas na Seção 2.4 foram calculadas para cada SKU ao longo de toda a sua série histórica. Em todos os casos, aplicou-se um cuidado técnico importante: cada feature de um determinado mês foi construída utilizando apenas informação de meses anteriores, nunca do próprio mês que o modelo deveria prever. Essa regra é chamada de defasagem (ou *shift*, em inglês) e é fundamental porque, na prática, o modelo nunca terá acesso ao valor real do mês futuro no momento em que precisa gerar sua previsão — apenas ao histórico até o mês anterior. Um modelo treinado dessa forma incorreta pareceria extremamente preciso nos testes, mas essa precisão seria artificial e não se repetiria quando o modelo fosse efetivamente utilizado para prever o futuro.

3.5 Configuração e versões do LightGBM avaliadas

A configuração do LightGBM foi refinada ao longo de três versões experimentais sucessivas, permitindo isolar o efeito de cada alteração sobre a acurácia. Os resultados de cada versão são apresentados e discutidos na Seção de Resultados:

Tabela 1. Versões avaliadas do LightGBM.

Versão	Configuração
V0	Configuração padrão da biblioteca (função de perda de regressão / erro quadrático médio; demais hiperparâmetros nos valores default).
V1	Função de perda Tweedie (potência de variância = 1,2 — Seção 2.3) e hiperparâmetros de regularização mais conservadores que o padrão da biblioteca: número de árvores (<i>n_estimators</i>) reduzido de 200 para 100, profundidade máxima (<i>max_depth</i>) limitada de ilimitada para 4, número de folhas (<i>num_leaves</i>) reduzido de 31 para 15, e taxa de aprendizado (<i>learning_rate</i>) reduzida de 0,1 para 0,05.
V2	Mantém a configuração de V1, acrescida de quatro features adicionais: defasagem de 6 meses (<i>lag_6</i>), máximo móvel de 6 meses (<i>rolling_max_6</i>), média móvel de 12 meses (<i>rolling_mean_12</i>) e máximo histórico (<i>max_historico</i>) — Seção 2.4.

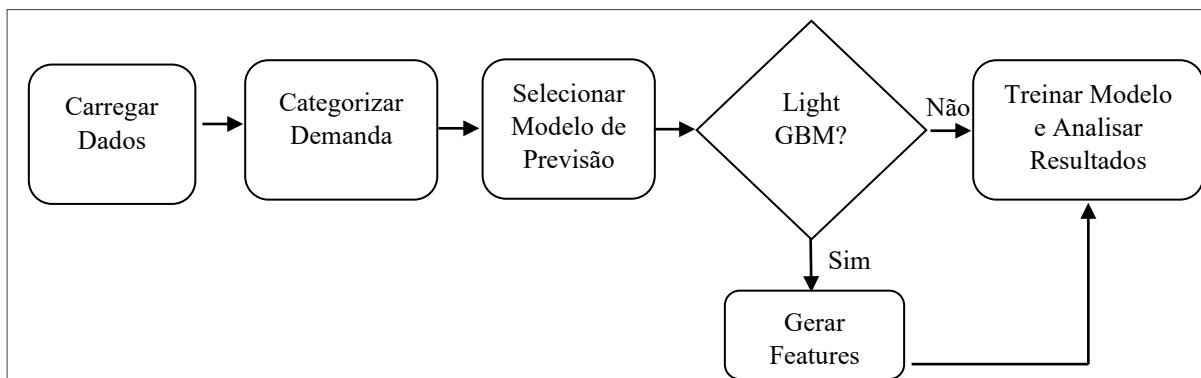
Fonte: Elaborado pelo autor.

3.6 Desenho do experimento: split treino/teste

A validação de todos os modelos seguiu um desenho de backtesting com holdout fixo, ou seja, uma única divisão entre período de treinamento e período de teste, aplicada igualmente a todos os modelos, em vez de múltiplas divisões alternativas testadas repetidamente: os primeiros 48 meses do histórico foram utilizados exclusivamente para treinamento, e os 12 meses subsequentes — correspondentes a uma safra completa já realizada — foram reservados como período de teste, ocultos de todos os modelos durante o ajuste. A previsão gerada para os 12 meses de teste foi comparada ao consumo real observado no mesmo período, utilizando as métricas descritas na Seção 2.5 (WMAPE e Bias).

A Figura 2 apresenta uma visão geral do fluxo de execução do framework proposto, detalhado nas subseções seguintes.

Figura 2. Fluxo de execução do framework de previsão de demanda.



Fonte: Elaborado pelo autor.

4. RESULTADOS E DISCUSSÃO

4.1 Tempo de treinamento

A Tabela 2 apresenta o tempo de treinamento de cada modelo por rodada, consolidado ao longo dos quatro quadrantes, medido em segundos numa única execução completa da etapa de seleção e execução dos modelos.

Tabela 2. Tempo de treinamento consolidado dos quadrantes, por modelo.

Modelo	Tempo_treino (segundos)
SARIMA	1.375,5
ARIMA	610,5
Holt-Winters	39,6
LightGBM_v2	7,1
LightGBM_v0	5,1
LightGBM_v1	4,6
CrostonTSB	1,8
SBA	0,6
sNaive	0,2

Fonte: Elaborado pelo autor.

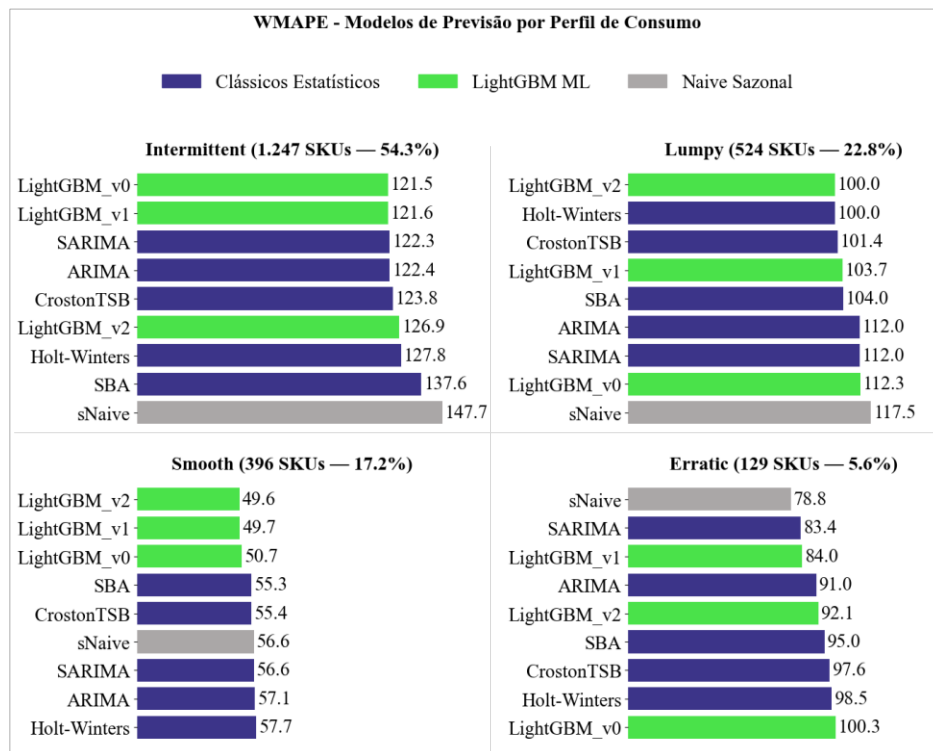
Os modelos da família Box-Jenkins (SARIMA e ARIMA) apresentam o maior tempo de processamento, superando em mais de duas ordens de magnitude o tempo de treinamento do LightGBM e dos modelos de demanda intermitente (CrostonTSB, SBA). Essa diferença é atribuível à busca automática da ordem do modelo via critério de informação (Seção 3.2), que testa múltiplas combinações de parâmetros para cada um dos 2.296 SKUs da base. O LightGBM, apesar de operar sobre toda esta categoria de materiais de um quadrante em uma única chamada, apresenta tempo de treinamento inferior a 8 segundos em todas as versões avaliadas — evidência do ganho de velocidade de processamento típico de modelos globais frente a modelos ajustados individualmente por série.

4.2 Acurácia por quadrante

A Figura 3 apresenta o WMAPE obtido por cada modelo em cada um dos quatro quadrantes do esquema de categorização, calculado sobre o período de teste (12 meses), ordenado por WMAPE dentro de cada quadrante, e a Figura 4 o Bias — com o LightGBM

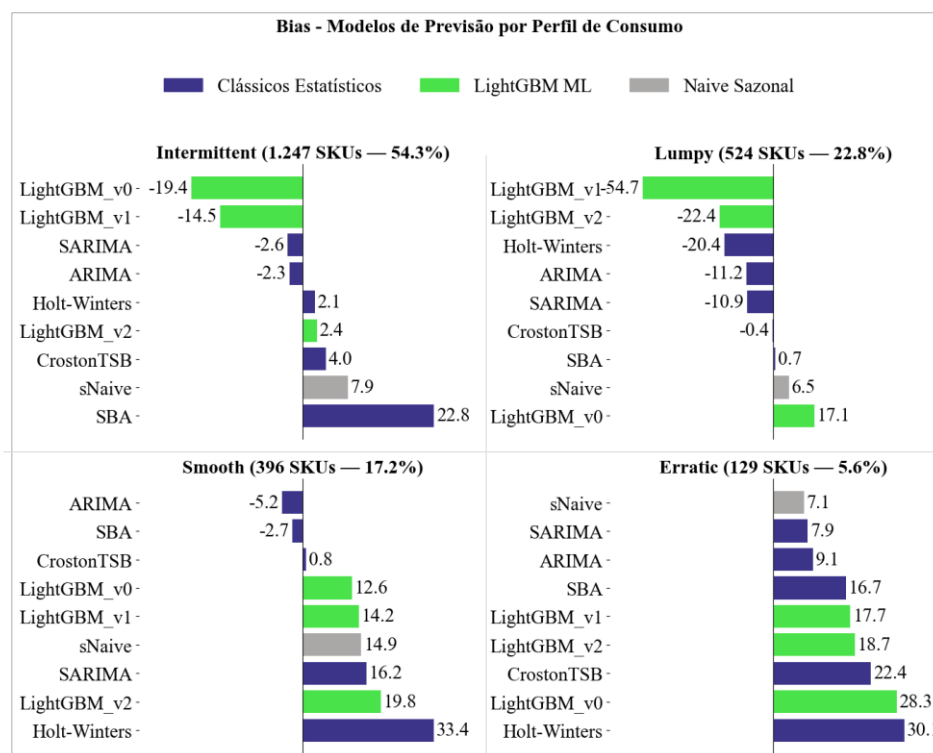
destacado em verde, o modelo de referência (Naive Sazonal) em cinza e os demais modelos estatísticos clássicos em azul.

Figura 3. WMAPE (%) por modelo, dentro de cada quadrante do esquema de categorização.



Fonte: Elaborado pelo autor.

Figura 4. Bias (%) por modelo, dentro de cada quadrante do esquema de categorização.



Fonte: Elaborado pelo autor.

A Figura 4 apresenta o Bias obtido por cada um dos nove modelos nos quatro quadrantes, ampliando para a totalidade dos modelos a análise de direção do erro anteriormente restrita às três versões do LightGBM (Seção 4.5). No Intermittent, a direção do erro é mista, mas assimétrica em magnitude: cinco dos nove modelos superestimam o consumo — Holt-Winters, LightGBM_v2, CrostonTSB, Naive Sazonal e, com a maior magnitude do quadrante, o SBA (22,8%) —, enquanto os quatro restantes subestimam, com destaque para as duas primeiras versões do LightGBM, responsáveis pela subestimação mais acentuada (−19,4% e −14,5%). No Lumpy, a subestimação passa a predominar de fato, presente em seis dos nove modelos, e se intensifica: o LightGBM_v1 registra ali a maior subestimação de todo o estudo (−54,7%). Nos quadrantes Smooth e Erratic, a direção se inverte com maior nitidez: no Smooth, sete dos nove modelos superestimam o consumo; no Erratic, a totalidade dos nove modelos apresenta Bias positivo — em ambos os casos, com o Holt-Winters à frente (33,4% no Smooth e 30,1% no Erratic).

4.3 Box-Jenkins: ARIMA versus SARIMA

A comparação entre ARIMA e SARIMA permite isolar empiricamente o ganho atribuível à modelagem explícita da sazonalidade (Seção 2.2). A Tabela 3 apresenta o WMAPE de ambos os modelos nos quatro quadrantes, e a diferença entre eles.

Tabela 3. WMAPE (%) — comparação entre ARIMA e SARIMA.

Quadrante	ARIMA	SARIMA	Diferença (SARIMA x ARIMA)
Smooth	57,1	56,6	−0,5
Erratic	91,0	83,4	−7,6
Intermittent	122,4	122,3	−0,1
Lumpy	112,0	112,0	0,0

Fonte: Elaborado pelo autor.

O SARIMA supera o ARIMA de forma expressiva apenas no quadrante Erratic (redução de 7,6 pontos percentuais no WMAPE), sugerindo que o componente sazonal captura efetivamente o ciclo safra-entressafra quando a demanda ocorre com regularidade (ADI baixo), mesmo com alta variabilidade de tamanho (CV^2 elevado). Nos quadrantes Smooth e Lumpy, a diferença entre os dois modelos é praticamente nula, e no Intermittent é marginal (0,1 ponto percentual).

4.4 LightGBM versus modelos estatísticos clássicos

Considerando, em cada quadrante, o menor WMAPE entre as três versões do LightGBM e o menor WMAPE entre os seis modelos estatísticos clássicos (incluindo o Naive Sazonal), observa-se que nenhuma das duas abordagens domina uniformemente os quatro quadrantes. No quadrante Smooth, o LightGBM (V2, WMAPE = 49,6) supera com margem considerável o melhor modelo clássico (SBA, WMAPE = 55,3). No quadrante Erratic, o menor WMAPE cabe ao Naive Sazonal (78,8), seguido pelo SARIMA (83,4) e pela melhor versão do LightGBM (V1, com 84,0). No quadrante Lumpy, o LightGBM (V2, WMAPE = 100,0) e o Holt-Winters (WMAPE = 100,0) obtêm resultado idêntico, configurando um empate técnico, enquanto o Naive Sazonal registra o pior desempenho do quadrante (117,5). No quadrante Intermittent, a melhor marca é do LightGBM (V0, WMAPE = 121,5), seguido de perto pelo V1 (121,6) e pelo SARIMA (122,3); o Naive Sazonal, novamente, apresenta o maior erro entre todos os modelos

avaliados (147,7). O desempenho do Naive Sazonal no Erratic merece registro, ainda que o quadrante reúna apenas 129 dos 2.296 SKUs da base (5,6% do total). O resultado ilustra a observação de Hyndman e Athanasopoulos (2021) de que *"sometimes one of these simple methods will be the best forecasting method available"* — em tradução livre, "às vezes um destes métodos simples será o melhor método de previsão disponível", sem que isso configure uma regra geral.

Esse padrão indica que a superioridade do LightGBM sobre os modelos estatísticos clássicos não é universal, mas condicionada ao quadrante de demanda: o ganho de flexibilidade do aprendizado de máquina se traduz em acurácia superior no quadrante com maior regularidade e menor intermitência (Smooth), e em desempenho equivalente ao melhor modelo clássico nos quadrantes Lumpy e Intermittent, nos quais lidera ou empata dentro de margens inferiores a um ponto percentual. Esse padrão é consistente com o observado por Saraiva (2025) em outro contexto logístico: ao comparar modelos que incorporam mais dimensões de informação simultaneamente a modelos restritos a uma única dimensão, a autora constata que "os resultados obtidos mostraram que modelos de previsão com abordagem espaço-temporal tiveram desempenho superior aos modelos com abordagem apenas temporal.

4.5 Evolução das versões do LightGBM

A Tabela 4 apresenta o WMAPE e o Bias das três versões do LightGBM avaliadas (Seção 3.5), permitindo isolar o efeito de cada alteração de configuração sobre a acurácia e sobre a direção do erro.

Tabela 4. WMAPE (%) e Bias (%) das versões do LightGBM, por quadrante.

Quadrante	V0 WMAPE	V0 Bias	V1 WMAPE	V1 Bias	V2 WMAPE	V2 Bias
Smooth	50,7	12,6	49,7	14,2	49,6	19,8
Erratic	100,3	28,3	84,0	17,7	92,1	18,7
Intermittent	121,5	-19,4	121,6	-14,5	126,9	2,4
Lumpy	112,3	17,1	103,7	-54,7	100,0	-22,4

Fonte: Elaborado pelo autor.

A introdução da função de perda Tweedie (V1) produziu o ganho mais expressivo entre as três versões: o WMAPE melhorou em três quadrantes frente à V0, com destaque para o quadrante Erratic (redução de 16,3 pontos percentuais). O Bias também se aproximou de zero em três dos quatro quadrantes — exceto no Lumpy, no qual a V1 passou a subestimar fortemente (Bias de -54,7%), um efeito colateral não previsto pela mudança da função de perda isoladamente.

A adição das quatro features complementares (V2 — Seção 2.4) apresentou resultado heterogêneo: o WMAPE melhorou nos quadrantes Smooth e Lumpy frente à V1, mas piorou nos quadrantes Erratic e Intermittent. O Bias, por outro lado, melhorou de forma expressiva no Intermittent (de -14,5% para 2,4%, praticamente neutro) e no Lumpy (de -54,7% para -22,4%, reduzindo a subestimação a menos da metade), ainda que ao custo de um Bias um pouco mais positivo no Smooth e no Erratic. Esse padrão é consistente com um trade-off entre erro absoluto (WMAPE) e direção sistemática do erro (Bias): as novas features parecem ter corrigido parte da subestimação estrutural nos quadrantes de maior intermitência, sem necessariamente reduzir o erro absoluto nesses mesmos quadrantes.

4.6 O desafio do quadrante Intermittent

O quadrante Intermittent concentra 1.247 SKUs (54,3%), mais da metade de toda a categoria avaliada. Essa proporção confere ao desempenho neste quadrante um peso desproporcional sobre a qualidade geral do planejamento de materiais: mesmo que um modelo apresente excelente acurácia nos demais quadrantes, seu impacto agregado sobre a operação será limitado se seu desempenho no Intermittent for insatisfatório.

Os resultados da Seção 4.2 confirmam que o Intermittent é, de fato, o quadrante mais desafiador da base: todos os nove modelos avaliados registraram WMAPE superior a 121%, o maior piso entre os quatro quadrantes. O melhor desempenho coube ao LightGBM (V0, WMAPE = 121,5), com vantagem marginal sobre o V1 (121,6) e o SARIMA (122,3) — diferenças de magnitude inferior a um ponto percentual, que não permitem afirmar superioridade prática de qualquer abordagem sobre as demais. A maioria dos modelos avaliados se concentram numa faixa estreita acima de 121%, à exceção do benchmark, sugere que a dificuldade de prever demanda intermitente não decorre primariamente da sofisticação do modelo, mas de uma limitação estrutural mais profunda: com um histórico de treinamento de 48 meses e uma alta proporção de meses sem consumo, o sinal disponível para o aprendizado — seja estatístico, seja de máquina — é inerentemente escasso. A evolução das versões do LightGBM (Seção 4.5) reforça esse diagnóstico: mesmo as intervenções direcionadas especificamente à intermitência (função de perda Tweedie, features de recência e de máximo histórico) produziram ganhos de Bias, mas não foram suficientes para reduzir o WMAPE a um nível competitivo com os demais quadrantes. Dado o peso deste quadrante na composição desta categoria de materiais, a busca por abordagens que melhor capturem esse padrão específico de demanda permanece como a principal fronteira de melhoria para trabalhos futuros.

5. CONCLUSÃO

Este estudo demonstrou a viabilidade de um framework de machine learning para a previsão de demanda de 2.296 SKUs em uma empresa sucroenergética com dinâmica sazonal, validando a eficácia de modelos globais frente a abordagens estatísticas tradicionais.

A análise segmentada por quadrantes revelou que o LightGBM (V2) apresentou o melhor desempenho no perfil Smooth (17,2% dos itens), atingindo um WMAPE de 49,6%, superando o melhor modelo estatístico (SBA, com 55,3%). No quadrante Erratic (5,6% dos itens), observou-se a importância de benchmarks robustos, uma vez que o modelo Naive Sazonal (78,8%) superou tanto o SARIMA (83,4%) quanto a melhor versão do LightGBM (84,0%) ao replicar o ciclo de safra.

A evolução técnica das versões do LightGBM trouxe ganhos. A introdução da função de perda Tweedie reduziu o WMAPE no quadrante Erratic de 100,3% para 84,0%. Complementarmente, a engenharia de features na versão V2 foi fundamental para a correção de viés sistemático: no quadrante Intermittent, o Bias passou de -14,5% para 2,4%, tornando-se praticamente neutro, enquanto no Lumpy a subestimação foi reduzida de -54,7% para -22,4%.

Do ponto de vista operacional, a eficiência computacional do framework é um diferencial. Enquanto o modelo SARIMA demandou 1.375 segundos para processar a base, as versões do LightGBM foram treinadas em menos de 8 segundos. Essa rapidez viabiliza o reprocessamento frequente necessário para o planejamento dinâmico de suprimentos.

Apesar dos avanços, o quadrante Intermittent permanece como o maior desafio, concentrando 54,3% dos SKUs e apresentando erros persistentemente elevados, com todos os modelos registrando WMAPE superior a 121%. Esta limitação estrutural, decorrente da escassez de sinal em séries com alta proporção de zeros, define novas fronteiras para pesquisas futuras.

6. REFERÊNCIAS BIBLIOGRÁFICAS

- BOX, George E. P.; JENKINS, Gwilym M.; REINSEL, Gregory C.; LJUNG, Greta M. **Time Series Analysis: Forecasting and Control**. 5. ed. Hoboken: John Wiley & Sons, 2015.
- CROSTON, John D. Forecasting and stock control for intermittent demands. **Operational Research Quarterly**, v. 23, n. 3, p. 289-303, 1972.
- HEWAMALAGE, Hansika; ACKERMANN, Klaus; BERGMEIR, Christoph. Forecast evaluation for data scientists: common pitfalls and best practices. **Data Mining and Knowledge Discovery**, v. 37, n. 2, p. 788-832, 2023.
- HOLT, Charles C. **Forecasting seasonals and trends by exponentially weighted moving averages**. Pittsburgh: Carnegie Institute of Technology, 1957.
- HYNDMAN, Rob J.; ATHANASOPOULOS, George. **Forecasting: Principles and Practice**. 3. ed. Melbourne: OTexts, 2021.
- HYNDMAN, Rob J.; KHANDAKAR, Yeasmin. Automatic time series forecasting: the forecast package for R. **Journal of Statistical Software**, v. 27, n. 3, p. 1-22, 2008.
- HYNDMAN, Rob J.; KOEHLER, Anne B. Another look at measures of forecast accuracy. **International Journal of Forecasting**, v. 22, n. 4, p. 679-688, 2006.
- JANUSCHOWSKI, Tim; WANG, Yuyang; TORKKOLA, Kari; ERKKILÄ, Timo; HASSON, Hilaf; GASTHAUS, Jan. Forecasting with trees. **International Journal of Forecasting**, v. 38, n. 4, p. 1473-1481, 2022.
- LIMA, José H. G.; CAMPOS, Ronaldo R. Aplicação de conceitos da gestão de estoques para melhoria em sistema ERP. **Interface Tecnológica**, v. 10, n. 1, p. 55-68, 2013.
- JORDAN, Michael I.; MITCHELL, Tom M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015.
- KE, Guolin et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: **Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)**. Red Hook: Curran Associates, 2017. p. 3149-3157.
- KOSTENKO, Andrey V.; HYNDMAN, Rob J. A note on the categorization of demand patterns. **Journal of the Operational Research Society**, v. 57, n. 10, p. 1256-1257, 2006.
- KRIGUER, Henrique. **Classificação multicritério de peças de reposição**. 2015. Dissertação (Mestrado em Engenharia Mecânica) – Faculdade de Engenharia e Ciências, Universidade Estadual Paulista (Unesp), Guaratinguetá, 2015.
- MAKRIDAKIS, Spyros; SPILIOTIS, Evangelos; ASSIMAKOPOULOS, Vassilios. Statistical and Machine Learning forecasting methods: Concerns and ways forward. **PLOS ONE**, v. 13, n. 3, e0194889, 2018.
- MAKRIDAKIS, Spyros; SPILIOTIS, Evangelos; ASSIMAKOPOULOS, Vassilios. The M5 competition: background, organization, and implementation. **International Journal of Forecasting**, v. 38, n. 4, p. 1325-1336, 2022.
- SARAIWA, Flaviane Azevedo. **Previsão de demanda logística para comércio eletrônico B2C utilizando métodos estatísticos e de aprendizado de máquina**. 2024. Dissertação (Mestrado em Engenharia de Produção) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2025.
- SPILIOTIS, Evangelos; MAKRIDAKIS, Spyros; KALTSOUNIS, Anastasios; ASSIMAKOPOULOS, Vassilios. Product sales probabilistic forecasting: An empirical evaluation using the M5 competition data. **International Journal of Production Economics**, v. 240, 108237, 2021.
- SYNTETOS, Aris A.; BOYLAN, John E. The accuracy of intermittent demand estimates. **International Journal of Forecasting**, v. 21, n. 2, p. 303-314, 2005.

ANAIS

SYNTETOS, Aris A.; BOYLAN, John E.; CROSTON, John D. On the categorization of demand patterns. **Journal of the Operational Research Society**, v. 56, n. 5, p. 495-503, 2005.

TEUNTER, Ruud H.; SYNTETOS, Aris A.; BABAI, Mohamed Zied. Intermittent demand: Linking forecasting to inventory obsolescence. **European Journal of Operational Research**, v. 214, n. 3, p. 606-615, 2011.

WINTERS, Peter R. Forecasting sales by exponentially weighted moving averages. **Management Science**, v. 6, n. 3, p. 324-342, 1960.