UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

THIAGO BIGOTTE GULLO

APLICAÇÃO DE MACHINE LEARNING NA RESOLUÇÃO DO PROBLEMA DE CORTE UNIDIMENSIONAL COM SOBRAS APROVEITÁVEIS E PREDIÇÃO DE DEMANDA

BAURU
Novembro/2025

THIAGO BIGOTTE GULLO

APLICAÇÃO DE MACHINE LEARNING NA RESOLUÇÃO DO PROBLEMA DE CORTE UNIDIMENSIONAL COM SOBRAS APROVEITÁVEIS E PREDIÇÃO DE DEMANDA

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.
Orientador: Profa. Dra. Adriana Cristina Cherri Nicola
Coorientador: Dr. Douglas Nogueira do Nascimento

BAURU
Novembro/2025

G973a

Gullo, Thiago

Aplicação de Machine Learning na resolução do problema de cortes unidimensional com sobras aproveitáveis e predição de demanda / Thiago Gullo. -- Bauru, 2025

54 p.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências, Bauru

Orientadora: Adriana Cristina Cherri Nicola

Coorientadora: Douglas Nogueira do Nascimento

1. Previsão de Demanda. 2. Redes Neurais LSTM. 3. Problema de Corte com Sobras Aproveitáveis. 4. Minimização de Perdas. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Thiago Bigotte Gullo

Aplicação de Machine Learning na resolução do problema de corte unidimensional com sobras aproveitáveis e predição de demanda

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.

Banca Examinadora

Profa. Dra. Adriana Cristina Cherri Nicola
Departamento de Matemática
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Profa. Dra. Simone das Graças Domingues Prado
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Profa. Dra. Andrea Carla Goncalves Vianna
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Bauru, 12 de Novembro de 2025.

Agradecimentos

Gostaria de agradecer primeiramente aos meus pais, que tornaram possível a oportunidade de estudar em uma das melhores universidades do Brasil. Mesmo longe de casa, pude contar com todo o suporte, dedicação e amor deles, sem os quais essa caminhada não seria possível.

Agradeço também aos meus amigos de Araraquara e de Bauru, em especial ao Enzo e ao Lucca, pelo apoio constante e pelos momentos de leveza que tornaram os últimos anos muito mais especiais.

Registro minha gratidão à minha orientadora e amiga, Profa. Adriana, pelo acolhimento e pela orientação ao longo de quase toda a graduação, sempre compartilhando conhecimento e contribuindo de forma decisiva para minha formação. Agradeço também ao meu coorientador, Douglas, pelo suporte constante e pela dedicação essencial durante todo esse período.

Por fim, agradeço à minha namorada, Isadora, por estar comigo em todos os momentos da graduação, sempre me ajudando, me ouvindo e me apoiando em tudo.

Resumo

Este trabalho propõe uma abordagem que visa a otimização da produção industrial, combinando técnicas avançadas de inteligência artificial para previsão de demanda em problemas de corte com sobras aproveitáveis. A proposta visa aprimorar o planejamento produtivo e minimizar perdas ao longo do processo, tornando-o mais sustentável e reduzindo custos operacionais. Dentre as abordagens adotadas, destaca-se o uso de redes neurais recorrentes do tipo LSTM (*Long Short-Term Memory*), capazes de identificar padrões temporais complexos em séries históricas de vendas, sazonalidades e variáveis externas, como o preço da matéria-prima. Esse modelo preditivo fornece estimativas mais precisas de demanda futura, auxiliando a tomada de decisões estratégicas voltadas à minimização de perdas. Essas técnicas computacionais foram implementadas para a obtenção de soluções para o Problema de Corte de Estoque com Sobras Aproveitáveis (PCESA), que é uma extensão do problema clássico de corte cujo foco está no aproveitamento das sobras geradas durante o processo produtivo. Para a condução dos experimentos, foi desenvolvida uma base de dados artificial com o objetivo de simular condições realistas de mercado. Essa base considera múltiplas regras para refletir o comportamento de clientes, suas demandas específicas e as oscilações diárias no preço de commodities, representando com maior fidelidade os desafios enfrentados na indústria. A avaliação dos resultados concentrou-se na minimização das perdas totais ao longo de um horizonte de planejamento, permitindo uma análise aplicada ao contexto empresarial, com foco na tomada de decisão por equipes de planejamento e negócios. Adicionalmente, foram conduzidas análises quantitativas com base nos resultados de treinamento e generalização, demonstrando a efetividade da integração entre a previsão de demanda e a otimização de processos de corte, na redução de desperdícios e no aprimoramento da produção industrial.

Palavras-chave: Previsão de Demanda, Redes Neurais LSTM, Problema de Corte com Sobras Aproveitáveis, Minimização de Perdas.

Abstract

This work proposes an approach aimed at optimizing industrial production by combining advanced artificial intelligence techniques for demand forecasting in cutting problems with usable leftovers. The goal is to improve production planning and minimize losses throughout the process, making it more sustainable and reducing operational costs. Among the approaches adopted, the use of recurrent neural networks of the LSTM (Long Short-Term Memory) type stands out, capable of identifying complex temporal patterns in historical sales data, seasonality, and external variables such as raw material prices. This predictive model provides more accurate estimates of future demand, aiding strategic decision-making focused on loss minimization. These computational techniques were implemented to find solutions for the Stock Cutting Problem with Usable Leftovers (PCESA), which is an extension of the classic cutting problem centered on utilizing leftovers generated during the production process. For conducting the experiments, an artificial database was developed to simulate realistic market conditions. This database considers multiple rules to reflect customer behavior, specific demands, and daily fluctuations in commodity prices, more faithfully representing the challenges faced by the industry. The evaluation of results focused on minimizing total losses over a planning horizon, enabling an analysis applicable to the business context, with an emphasis on decision-making by planning and business teams. Additionally, quantitative analyses were conducted based on training and generalization results, demonstrating the effectiveness of integrating demand forecasting with cutting process optimization in reducing waste and improving efficiency.

Keywords: Demand Forecasting, LSTM Neural Networks, Cutting Stock Problem with Usable Leftovers, Loss Minimization.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

PCE	Problemas de Corte de Estoque
PCESA	Problema de Corte de Estoque com Sobras Aproveitáveis
LSTM	*Long Short-Term Memory*
RNNs	Redes Neurais Recorrentes
ReLU	*Rectified Linear Unit*
MSE	*Mean Squared Error*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
OHE	*One Hot Encoder*
SHFE	*Shanghai Futures Exchange*
IPCA	Índice Nacional de Preços ao Consumidor Amplo
IBGE	Instituto Brasileiro de Geografia e Estatística
MDIC	Ministério do Desenvolvimento, Indústria, Comércio e Serviços
SCND	Projeto de Rede de Cadeia de Suprimentos

Sumário

1 Introdução

A produção industrial enfrenta desafios críticos no planejamento de seus processos, nos quais decisões inadequadas, como a gestão ineficiente de estoques ou estratégias subótimas de uso de matérias primas, podem resultar em perdas substanciais. Em problemas de corte de estoque, o uso sem planejamento da matéria-prima pode gerar desperdícios que comprometem não apenas a rentabilidade e a competitividade das empresas, mas também representam um custo de oportunidade significativo. Estudos empíricos indicam que a redução dessas perdas, por meio de sistemas integrados de gestão da produção e dos estoques, está diretamente associada a margens de lucro mais elevadas (HOFER; EROGLU; HOFER, 2012).

Além dos impactos econômicos, há consequências ambientais igualmente relevantes. O descarte excessivo de recursos e o consumo desnecessário de matérias-primas, intensificados por falhas no planejamento produtivo, contribuem para a degradação ambiental, como evidenciado pelo crescente volume de resíduos industriais em setores de alta intensidade de recursos, como o têxtil e o metalúrgico.

Diante desse cenário, este trabalho propõe uma abordagem para otimizar a produção industrial, combinando técnicas de previsão de demanda com modelagem matemática. Na primeira etapa da abordagem, foram desenvolvidos modelos preditivos de aprendizado de máquina, baseados especificamente em redes neurais do tipo *Long Short-Term Memory* (LSTM). Esses modelos de previsão foram utilizados para estimar a demanda no contexto do Problema de Corte de Estoque com Sobras Aproveitáveis (PCESA).

O PCESA trata do corte de objetos em estoque para atender à produção de itens com diferentes quantidades e tamanhos, buscando a otimização de uma função objetivo. Seu diferencial está na consideração explícita de sobras aproveitáveis (*retalhos*), que podem ser utilizadas em cortes futuros, aumentando a eficiência do processo. Nesse contexto, tanto objetos padronizados (comprados pela empresa) quanto retalhos armazenados são utilizados no atendimento da demanda. O objetivo central é minimizar as perdas de material e, consequentemente, reduzir custos produtivos e impactos ambientais, promovendo um uso mais sustentável e eficiente dos recursos. Diversos métodos de solução foram propostos na literatura para tratar o PCESA. Neste trabalho, utiliza-se o modelo matemático desenvolvido por Arenales et al. (2015), que permite lidar explicitamente com as sobras geradas, incorporando-as ao processo de decisão.

Para a previsão de demanda foi utilizada redes neurais recorrentes LSTM, arquitetura proposta por Hochreiter e Schmidhuber (1997), que se destaca pela capacidade de capturar padrões temporais complexos em séries históricas de vendas, como sazonalidades e tendências,

além de integrar variáveis exógenas como por exemplo, flutuações no preço de aço bruto. Essa arquitetura emprega mecanismos de controle (*input*, *forget* e *output gates*) e memória celular, permitindo preservar dependências de longo prazo e mitigar o problema do desaparecimento do gradiente. Tais características tornam o modelo particularmente adequado para lidar com não-linearidades em dados industriais, gerando previsões consistentes que servem de base para decisões estratégicas de produção e alocação de recursos.

No contexto dos problemas de corte, a previsão de demanda desempenha papel central no planejamento produtivo e constitui um dos principais desafios enfrentados pelas indústrias. A complexidade em prever com precisão os itens a serem produzidos decorre da sazonalidade dos pedidos, das particularidades dos clientes e da variabilidade temporal da demanda. Tais fatores tornam o planejamento mais incerto e exigem análises dinâmicas e atualizadas. A ausência de métodos eficazes para lidar com essas incertezas pode acarretar escassez de matéria-prima, excesso ou falta de produtos em estoque e alocação ineficiente de recursos. Como consequência, surgem desequilíbrios entre oferta e demanda, desperdícios de material e ineficiências nos processos de corte, comprometendo tanto os custos operacionais quanto a sustentabilidade da operação industrial.

O objetivo geral deste trabalho é viabilizar um planejamento industrial mais eficiente, integrando a previsão da demanda por meio de técnicas de aprendizado de máquina, como LSTM, com a otimização do uso de matéria-prima na produção de itens por meio da aplicação de modelos matemáticos para o problema de corte de estoque unidimensional com sobras aproveitáveis, otimizando a utilização dos recursos disponíveis. Tal objetivo foi atingido a partir do cumprimento dos seguintes objetivos específicos:

- Explorar e avaliar as melhores técnicas de aplicação da LSTM para previsão de demanda no contexto industrial;
- Desenvolver um modelo preditivo de demanda utilizando redes neurais LSTM;
- Revisar a literatura sobre o PCESA unidimensional e o uso de matheurísticas;
- Implementar o modelo proposto por Arenales et al. (2015);
- Implementar a matheurística *Kernel Search* para a obtenção de soluções inteiras para o PCESA;
- Integrar a previsão de demanda com a estratégia de corte de estoque;
- Comparar os resultados obtidos com e sem a predição de demanda no problema de corte de estoque com sobras aproveitáveis.

Os próximos capítulos deste documento estão organizados da seguinte forma: o Capítulo 2 revisa os conceitos essenciais referentes ao PCESA e à técnicas de previsão de demanda; o Capítulo 3 detalha a base de dados usada nos experimentos e como ela foi construída a partir de variáveis operacionais e macroeconômicas; na parte metodológica, o Capítulo 4 descreve os procedimentos de preparação dos dados e hiperparâmetros para o treinamento do modelo de predição de demanda, bem como os métodos de solução aplicados ao PCESA; o Capítulo 5 apresenta e discute os resultados obtidos pelos experimentos computacionais realizados, tanto em relação à previsão de demanda quanto à qualidade das soluções encontradas para o PCESA; por fim, o Capítulo 6 sintetiza as principais conclusões do presente trabalho.

2 Fundamentação Teórica

Neste capítulo, serão apresentadas a definição e a revisão bibliográfica acerca dos estudos relacionados ao PCESA. Serão também explorados os conceitos fundamentais sobre técnicas de predição de demanda, com ênfase no modelo preditivo baseado em redes neurais LSTM, bem como as principais métricas utilizadas em aprendizado de máquina para avaliação e validação de modelos.

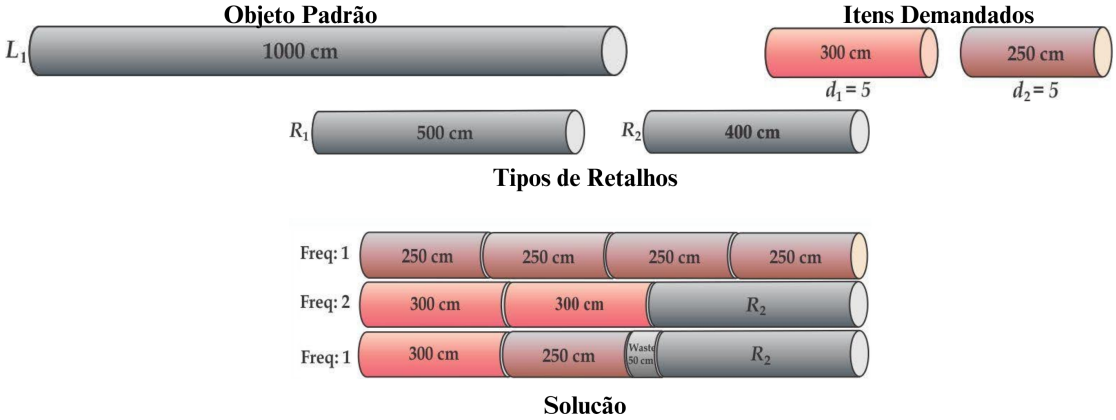
2.1 Problema de Corte e Estoque com Sobras Aproveitáveis (PCESA)

Os Problemas de Corte de Estoque (PCE) são desafios clássicos da otimização combinatória, amplamente encontrados na prática industrial e alvo frequente de estudos e publicações científicas. Esses problemas ocorrem quando um conjunto de objetos disponíveis em estoque deve ser cortado para atender à demanda por itens com tamanhos e quantidades específicas, buscando-se a otimização de uma função objetivo, a qual, geralmente, foca na minimização do desperdício ou no uso eficiente de recursos. Exemplos típicos incluem o corte de barras de aço, bobinas de papel e alumínio, placas metálicas e de madeira, bem como placas de circuito impresso.

Uma variação relevante do PCE é o Problema de Corte de Estoque com Sobras Aproveitáveis (PCESA). Nessa variação, as sobras geradas durante o processo de corte podem ser aproveitadas em cortes futuros, desde que contribuam para a redução de perdas. Essa característica introduz uma complexidade adicional ao problema, exigindo estratégias de gestão de estoque mais elaboradas. Diversos estudos têm se dedicado ao PCESA, propondo modelos matemáticos e métodos de solução que buscam equilibrar eficiência operacional e sustentabilidade.

A figura 1 ilustra um exemplo fictício de como o PCESA pode ser aplicado a um problema prático. Suponha que esteja disponível em estoque um objeto padrão de 1000 cm e que seja necessário cortar 5 itens de 300 cm e 5 itens de 250 cm. Consideram-se como sobras aproveitáveis (ou retalhos) aquelas peças cujo comprimento seja igual a 500 cm ou 400 cm. Na solução apresentada, uma combinação de corte eficiente para atender a essa demanda, com a menor perda possível de apenas 50 cm, é obtida a partir da geração de retalhos. Esse exemplo demonstra como o aproveitamento de materiais pode ser uma estratégia eficaz para reduzir desperdícios e otimizar o uso do estoque disponível no futuro, desde que os retalhos retornam ao estoque.

Figura 1 – Exemplo PCESA



Fonte: (ARENALES et al., 2015)

A resolução do PCESA tem grande impacto econômico por contribuir com a redução do uso de matérias-primas, além de um impacto ambiental significativo, uma vez que reflete a uma exploração menos intensiva dos recursos naturais de onde são extraídas as matérias-primas e reduz a quantidade de resíduos gerados.

Na literatura científica relacionada ao PCESA, Scheithauer (1991) propôs uma abordagem restrita voltada a problemas de corte unidimensionais, nos quais apenas um tipo de objeto é considerado. Nessa abordagem, há o controle da produção de retalhos, sendo necessário definir previamente os tamanhos das sobras que poderão ser aproveitadas. Para lidar com essa característica, o autor modificou a estratégia clássica de geração de colunas, originalmente proposta por Gilmore e Gomory (1963), inserindo itens adicionais que não possuem demanda direta, permitindo que sejam considerados como sobras futuras, disponíveis para cortes.

Com o objetivo de minimizar as perdas ou concentrá-las em um único objeto, Gradišar, Jesenko e Resinovič (1997) desenvolveram um modelo matemático inteiro, juntamente com uma heurística denominada COLA, aplicada ao problema de corte de rolos em uma indústria têxtil. Posteriormente, Abuabara e Morabito (2009) reestruturaram esse modelo como um modelo de programação inteira mista, adaptando-o ao contexto do aproveitamento de sobras em processos produtivos da indústria aeronáutica.

Cherri, Arenales e Yanasse (2009) modificaram heurísticas construtivas e residuais existentes na literatura com o objetivo de minimizar as sobras geradas em cada padrão de corte, assegurando que fossem suficientemente pequenas para serem descartadas como perdas ou grandes o bastante para retornarem ao estoque de maneira viável. Cui e Yang (2010), por sua vez, adaptaram o modelo original proposto por Scheithauer (1991) para considerar múltiplos tipos de objetos disponíveis em estoque, além de introduzirem restrições quanto à quantidade de retalhos permitidos por padrão de corte, ampliando a aplicabilidade do modelo em contextos mais

complexos.

Posteriormente, Cherri et al. (2014) realizaram um levantamento abrangente da literatura sobre o PCESA. Neste estudo, os autores apresentaram as aplicações do problema, os modelos matemáticos e heurísticas desenvolvidas, uma análise crítica dos resultados obtidos nos trabalhos revisados e sugestões para direcionamentos futuros de pesquisa nessa área.

Embora diversos autores tenham proposto modelos matemáticos para representar o PCESA, a abordagem para o tratamento das sobras geradas sempre foi realizada de forma implícita, utilizando estratégias previamente definidas para a geração e manipulação de retalhos. Diferentemente dessas abordagens, Arenales et al. (2015) desenvolveram um modelo matemático que incorpora explicitamente as sobras na modelagem do PCESA. O modelo proposto permite a pré-definição dos comprimentos das sobras e a imposição de limites à quantidade de retalhos gerados, além de possibilitar a priorização do uso das sobras armazenadas. Essa característica é particularmente importante em aplicações práticas, como no corte de bobinas de aço, nas quais, após o desembalamento, o material está sujeito à oxidação em um curto período de tempo, exigindo seu uso preferencial. Para resolver o problema, os autores relaxaram a condição de integralidade do modelo proposto, permitindo que as variáveis de decisão assumissem valores reais, e utilizaram o método simplex com geração de colunas de Gilmore e Gomory (1963). Soluções ótimas contínuas foram geradas, porém, nenhum procedimento heurístico foi proposto para obtenção de soluções inteiras.

Coelho et al. (2017) apresentaram um modelo matemático e duas heurísticas para resolver o PCESA, considerando a possibilidade de que as sobras geradas durante o processo de corte pudessem retornar ao estoque ou serem comercializadas com outras empresas. Os autores também discutiram as implicações sustentáveis da abordagem proposta, destacando benefícios como a redução do impacto ambiental, o aumento da lucratividade, a melhoria da imagem institucional da empresa, entre outros.

Khan et al. (2020) estudaram o PCESA no contexto do corte de bobinas em uma fábrica de papel, incorporando a restrição relacionada ao número limitado de facas disponíveis no processo de corte. Já Ravelo, Meneses e Santos (2020) propuseram uma abordagem heurística e duas *matheurísticas* para resolver instâncias do PCESA, tanto com base em dados reais quanto em problemas presentes na literatura. Segundo os autores, os métodos apresentaram desempenho satisfatório em todos os testes realizados.

Cherri et al. (2023) abordaram o problema de corte de estoque com a possibilidade de gerar sobras como um problema estocástico de duas fases com recurso. Um modelo matemático para representar este problema foi proposto, e uma abordagem de geração de colunas foi desenvolvida para o resolver o problema. Senergues et al. (2025) apresentaram uma revisão da literatura envolvendo o aproveitamento de sobras para problemas uni, bi e tridimensionais. Características,

particularidades, métodos de solução e aplicações foram destacadas pelos autores.

Devido à necessidade de se obter soluções inteiras para os problemas de corte, neste projeto foi desenvolvido uma matheurística baseada na abordagem *Kernel Search*, com o objetivo de gerar soluções inteiras a partir dos resultados contínuos obtidos por Arenales et al. (2015). A técnica *Kernel Search* tem sido amplamente aplicada em diferentes contextos de problemas combinatórios, como no Problema da Mochila Binária (ANGELELLI; MANSINI; SPERANZA, 2012), no problema de localização de hubs em múltiplos períodos (WEMANS, 2016), e em problemas de localização de instalações sem restrições de capacidade (FILIPPI et al., 2021).

2.2 Modelo Matemático para resolução do PCESA

Como mencionado anteriormente, o modelo matemático proposto por Arenales et al. (2015) trata explicitamente as sobras geradas durante o processo de corte. Dessa forma, a produção dos itens demandados pode ocorrer tanto a partir de padrões de corte aplicados sobre objetos padronizados adquiridos pela empresa quanto a partir das sobras resultantes de cortes anteriores, armazenadas em estoque.

O objetivo do modelo consiste em atender às demandas minimizando as perdas geradas no processo. As sobras a serem aproveitadas possuem comprimentos pré-definidos, podem ser geradas em quantidades limitadas de acordo com restrições de capacidade de armazenamento e não são contabilizadas como perdas.

Os parâmetros do modelo proposto por Arenales et al. (2015) são definidos por:

- S: número de tipos de objetos padronizados. São denotados objetos do tipo s, $s \in \{1, ..., S\}$;
- R: número de tipos de sobras em estoque. São denotadas sobras do tipo $S+s$, $s \in \{1, ..., R\}$;
- e_s: número de objetos/sobras tipo s disponíveis em estoque, $s = 1, ..., S + R$;
- m: número de tipos de itens demandados;
- d_i: demanda do item do tipo i, $i = 1, ..., m$;
- J_s: conjunto de padrões de corte para o objeto do tipo s, $s = 1, ..., S + R$;
- $J_s(k)$: conjunto de padrões de corte para objetos padronizados do tipo s com sobra do tipo $S + k$, $k = 1, ..., R$, $s = 1, ..., S$;
- a_{ijs}: número de itens do tipo i no padrão de corte j para objeto do tipo s, $i = 1, ..., m$, $s = 1, ..., S + R$, $j \in J_s$;

- a_{ijsk}: número de itens do tipo i no padrão de corte j para o objeto s com sobra do tipo $S+k$, $i = 1, ..., m$, $k = 1, ..., R$, $s = 1, ..., S$, $j \in J_s(k)$;
- c_{js}: perda por cortar o objeto/sobra tipo s no padrão de corte j, $s = 1, ..., S + R$, $j \in J_s$;
- c_{jsk}: perda por cortar o objeto s no padrão de corte j gerando uma sobra do tipo $S + k$, $s = 1, ..., S$, $k = 1, ..., R$, $j \in J_s(k)$;
- U: número máximo de sobras.

As variáveis de decisão que compõem o modelo podem ser descritas como:

- x_{js}: número de objetos do tipo s cortados de acordo com o padrão de corte j, $s = 1, ..., S+R$, $j \in J_s$;
- x_{jsk}: número de objetos do tipo s cortados de acordo com o padrão de corte j e gerando uma sobra do tipo $S + k$, $s = 1, ..., S$, $k = 1, ..., R$, $j \in J_s(k)$.

O modelo matemático proposto pelos autores é dado pela equação (2.1) a (2.6):

$$\text{Min } f(x) = \sum_{j \in J_s} \sum_{s=1}^{S} c_{js} x_{js} + \alpha' \sum_{j \in J_s(k)} \sum_{k=1}^{R} \sum_{s=S+1}^{S+R} c_{jsk} x_{jsk} + \alpha'' \sum_{j \in J_s} \sum_{s=S+1}^{S+R} c_{js} x_{js} \tag{2.1}$$

Sujeito a:

$$\sum_{j \in J_s} a_{ijs} x_{js} + \sum_{j \in J_s(k)} \sum_{k=1}^{R} a_{ijsk} x_{jsk} + \sum_{s=1}^{S+1} a_{ijs} x_{js} = d_i, \quad i = 1, ..., m \tag{2.2}$$

$$\sum_{j \in J_s} x_{js} + \sum_{j \in J_s(k)} \sum_{k=1}^{R} x_{jsk} \leq e_s, \quad j \in J_s(k), \quad s = 1, ..., S \tag{2.3}$$

$$\sum_{j \in J_s} x_{js} \leq e_s, \quad s = S + 1, ..., S + R \tag{2.4}$$

$$\sum_{j \in J_s(k)} \sum_{k=1}^{R} x_{jsk} - \sum_{s=1}^{S} \sum_{j \in J_s} x_{js} \leq U - \sum_{s=S+1}^{S+R} e_s, \quad j \in J_s, \quad s = S + 1, ..., S + R \tag{2.5}$$

$$x_{js}, x_{jsk} \geq 0, \quad s = 1, ..., S + R, \quad j \in J_s \text{ e inteiro}$$
$$x_{jsk} \geq 0, \quad k = 1, ..., R, \quad s = 1, ..., S, \quad j \in J_s(k) \text{ e inteiro} \tag{2.6}$$

No modelo (2.1)-(2.6), a função objetivo (2.1) visa minimizar a perda total resultante do corte de objetos padronizados, com ou sem geração de sobras, bem como do corte das sobras

disponíveis em estoque. Os parâmetros α' e α'' são utilizados para incentivar, respectivamente, a geração controlada de sobras e o reaproveitamento daquelas já armazenadas.

As restrições (2.2) asseguram o atendimento total da demanda dos itens. As restrições (2.3) e (2.4) limitam a quantidade de objetos padronizados e de sobras que podem ser cortadas, respeitando a disponibilidade em estoque. As restrições (2.5) impõem um limite à quantidade de sobras que podem ser geradas, com base na capacidade de armazenamento. Por fim, as condições de não negatividade e integralidade das variáveis de decisão são definidas por (2.6).

Devido ao grande número de variáveis e às condições de integralidade, é computacionalmente inviável resolver o modelo (2.1)-(2.6) de forma ótima definindo previamente todos os possíveis padrões de corte. Diante disso, a seguir sera descrito um método de solução para a relaxação linear desse modelo, obtida ao relaxar as condições de integralidade (2.6). Para a obtenção de soluções inteiras, foi utilizada a heurística Kernel Search, que também sera apresentada nas próximas seções.

2.3 Geração de Padrões de Corte

Para a resolução da relaxação linear do modelo (2.1)-(2.6), foram implementados os subproblemas propostos por Arenales et al. (2015) com o objetivo de gerar padrões de corte, derivados da técnica de geração de colunas idealizada por Gilmore e Gomory (1963), conhecida como *column generation*. Essa abordagem requer o cálculo do custo reduzido de cada coluna candidata à inserção, utilizando os multiplicadores simplex (ou variáveis duais) associados às restrições do modelo.

Os multiplicadores duais são representados por $\pi = (\pi^1, \pi^2, \pi^3, \pi^4)$, sendo:

- π^1: vetor de dimensão m, associado à restrição de atendimento da demanda (restrição 2.2);
- π^2: vetor de dimensão S, associado à limitação da quantidade de objetos padronizados disponíveis (restrição 2.3);
- π^3: vetor de dimensão R, associado à limitação das sobras disponíveis em estoque (restrição 2.4);
- π^4: escalar correspondente à restrição de geração de sobras (restrição 2.5).

As colunas, que representam diferentes padrões de corte, são classificadas em três tipos: colunas associadas a objetos padronizados, colunas associadas a objetos padronizados com geração de sobras e colunas associadas a sobras em estoque. Em cada iteração do algoritmo de geração de colunas, é escolhida a coluna com menor custo reduzido, determinado pela solução

dos subproblemas descritos a seguir. O custo reduzido mede o potencial de uma nova coluna em melhorar a solução atual, refletindo o ganho efetivo obtido ao incorporá-la ao conjunto de padrões de corte do modelo (2.1)-(2.6).

2.3.0.1 Colunas associadas a objetos padronizados

Essas colunas representam padrões que cortam objetos de tamanho padronizado, e possuem a seguinte estrutura:

$$\mathbf{a}_s = \left(\underbrace{a_{1s} \cdots a_{ms}}_{m\text{-vetor (padrão de corte)}} \quad | \quad \underbrace{0 \ldots 1 \ldots 0}_{S\text{-vetor (coluna identidade)}} \quad | \quad \underbrace{0 \ldots 0}_{R\text{-vetor de zeros}} \quad | \, 0 \right)^T$$

O melhor padrão de corte para esse tipo de coluna pode ser obtido por meio das equações 2.7 e 2.8 de custo reduzido:

$$\text{Max } \hat{c}(a_s) = -\sum_{i=1}^{m} (l_i + \pi_i^1) \cdot a_{is} + (L_s - \pi_s^2) \tag{2.7}$$

Sujeito a:

$$\sum_{i=1}^{m} l_i a_{is} <= L_s, \quad s = 1, ..., S \tag{2.8}$$

onde l_i representa o comprimento do item i, com $i = 1, \ldots, m$, e $(a_{1s}, \ldots, a_{ms})^T$ indica a quantidade de itens cortados a partir de um objeto de comprimento L_s, para $s \in \{1, \ldots, S\}$.

2.3.0.2 Colunas associadas a objetos padronizados com geração de sobras

Essas colunas representam padrões de corte aplicados sobre objetos padronizados que geram uma sobra de comprimento L_{S+k} e seguem a estrutura:

$$\mathbf{a}_s(k) = \left(\underbrace{a_{1s} \cdots a_{ms}}_{m\text{-vetor (padrão de corte)}} \quad | \quad \underbrace{0 \ldots 1 \ldots 0}_{S\text{-vetor (coluna identidade)}} \quad | \quad \underbrace{0 \ldots 0}_{R\text{-vetor de zeros}} \quad | \, 1 \right)^T$$

O padrão de corte mais eficiente é identificado com base nas equações 2.9 e 2.10 de custo reduzido:

$$\text{Max } \hat{c}(a_s(k)) = -\sum_{i=1}^{m} (\alpha' l_i + \pi_i^1) \cdot a_{is} + \left[\alpha'(L_s - L_{S+k}) - \pi_s^2 - \pi_4 \right] \tag{2.9}$$

Sujeito a:

$$\sum_{i=1}^{m} l_i a_{is} <= (L_s - L_{S+k}), \quad s = 1, ..., S, k = 1..., R \tag{2.10}$$

2.3.0.3 Colunas associadas a sobras em estoque

Essas colunas correspondem a padrões de corte aplicados diretamente às sobras armazenadas em estoque e apresentam a estrutura:

$$\mathbf{a}_s = \left(\underbrace{a_{1s} \cdots a_{ms}}_{m\text{-vetor (padrão de corte)}} \mid \underbrace{0 \ldots 0}_{S\text{-vetor de zeros}} \mid \underbrace{0 \ldots 1 \ldots 0}_{R\text{-vetor (coluna identidade)}} \mid -1 \right)^T$$

A equação 2.11 e 2.12 de custo reduzido utilizada para determinar o melhor padrão é:

$$\text{Max } \hat{c}(a_s) = -\sum_{i=1}^{m} (\alpha'' l_i + \pi_i^1) \cdot a_{is} + \left[\alpha'' L_s - \pi_s^3 + \pi_4 \right] \tag{2.11}$$

Sujeito a:

$$\sum_{i=1}^{m} l_i a_{is} <= L_s, \quad s = S+1, ..., S+R \tag{2.12}$$

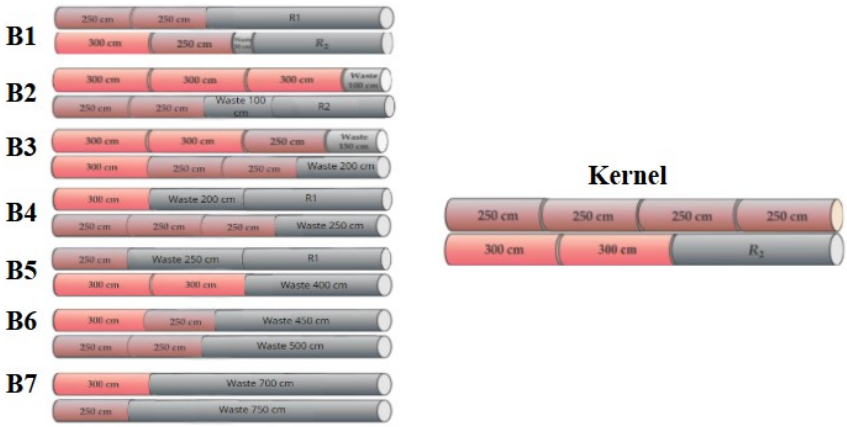
2.4 Heurística Kernel Search

O método de geração de colunas aplicado ao modelo (2.1) a (2.6) fornece de modo eficiente soluções continuas para o PCESA. Entretanto, na prática, tais soluções não são aplicáveis pois o valores das frequências de cada padrão de corte (numero de objetos/retalhos cortados de acordo com determinado padrão) são decimais. Para a obtenção de soluções inteiras apropriadas para situações práticas reais, foi utilizada a heurística *Kernel Search*. Esse procedimento consiste em particionar os padrões gerados após a obtenção da solução contínua do modelo PCESA em dois grupos: um núcleo (*kernel*) e subconjuntos auxiliares denominados *buckets*. O algoritmo *Kernel Search* é estruturado em duas fases principais: Construção e Extensão.

Na fase de Construção, após a resolução do modelo matemático, todos os padrões de cortes obtidos são ordenadas de acordo com um critério pré-definido, como por exemplo, a menor perda gerada em cada padrão. O *kernel* inicial é então formado pela seleção de um conjunto limitado de padrões. Os padrões remanescentes são distribuídos em *buckets*, também seguindo um critério de ordenação. Em seguida, o problema é resolvido considerando-se apenas as variáveis presentes no *kernel* inicial, agora com as variáveis de decisão restritas a valores inteiros, o que resulta em uma solução inteira factível.

A figura 2 ilustra essa etapa do processo. No exemplo, baseado no mesmo problema apresentado na figura 1, os padrões resultantes da solução ótima são separados para compor o *kernel* inicial, enquanto os padrões restantes são ordenados de acordo com a perda crescente e agrupados em *buckets* (B1 a B7), cada um contendo dois padrões.

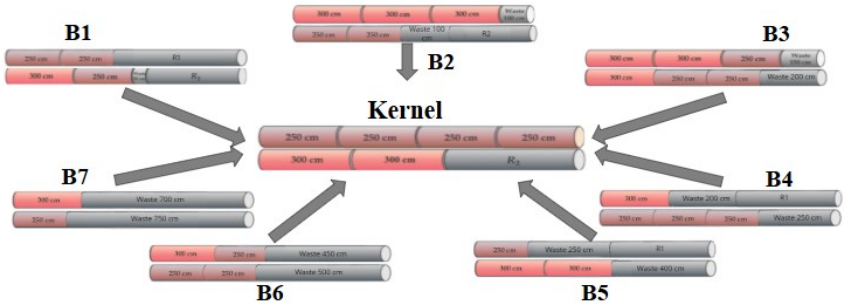
Figura 2 – *Kernel Search* - Etapa Construção



Fonte: Elaborada pelo Autor

Na fase de Extensão, busca-se aprimorar a solução obtida na etapa anterior. Conforme ilustrado na figura 3, para cada *bucket*, resolve-se novamente o problema, agora considerando as variáveis do *kernel* em conjunto com as do *bucket* em análise. Impõe-se que pelo menos uma variável do *bucket* seja incluída na solução. Caso essa nova solução seja melhor que a atual, as variáveis do *bucket* utilizadas são incorporadas permanentemente ao *kernel*. Esse processo iterativo permite explorar sistematicamente combinações adicionais de variáveis, refinando progressivamente a qualidade da solução.

Figura 3 – *Kernel Search* - Etapa Extensão



Fonte: Elaborada pelo Autor

2.5 Previsão de Demanda

A previsão precisa da demanda proporciona benefícios estratégicos significativos na gestão da cadeia de suprimentos, incluindo a redução de custos de estoque em até 30%, a melhoria no nível de serviço ao cliente, e a otimização do planejamento de produção (CHOPRA; MEINDL, 2019; SILVER; PYKE; PETERSON, 1998). Segundo Chopra e Meindl (2019), uma previsão acurada permite às empresas equilibrar adequadamente a oferta e a demanda, minimizando tanto excessos quanto faltas de estoque, enquanto Fisher (1997) destaca que a capacidade de prever demandas futuras é fundamental para alinhar as estratégias da cadeia de suprimentos às características dos produtos. Em um ambiente global marcado por rápidas transformações econômicas e elevada incerteza (AAMER, 2018), a dificuldade em prever com precisão a demanda dos clientes pode gerar o conhecido Efeito Chicote (*Bullwhip Effect*), no qual pequenas variações na previsão da demanda real são amplificadas ao longo da cadeia de suprimentos, provocando distorções informacionais, ineficiências operacionais, acúmulo excessivo de estoques ou rupturas de produtos (NORRMAN; NASLUND, 2019).

Diante desse cenário, a crescente disponibilidade de grandes volumes de dados, caracterizados pelos **"5Vs"** (Volume, Velocidade, Variedade, Veracidade e Valor) (JEBLE; KUMARI; PATIL, 2017), impulsionou a adoção de métodos analíticos avançados capazes de extrair padrões relevantes e apoiar decisões mais assertivas. Neste contexto, técnicas de aprendizado de máquina ganham destaque pela capacidade de modelar relações complexas e não lineares, superando limitações de abordagens tradicionais. No âmbito industrial, essa característica é particularmente relevante, uma vez que a acurácia das previsões de demanda afeta diretamente a eficiência produtiva, a alocação de recursos e a qualidade do nível de serviço oferecido ao cliente (YANI; PRIYATNA; AAMER, 2019).

Este projeto adota uma arquitetura baseada em redes neurais recorrentes do tipo LSTM, amplamente empregada em domínios diversos, como previsão de consumo de gás (ZHA et al., 2022), tradução automática (SUTSKEVER; VINYALS; LE, 2014), reconhecimento de voz (ZAZO et al., 2018), aplicações financeiras (MUNCHARAZ, 2020), e síntese de fala (ZEN; SAK, 2015). No presente trabalho, a LSTM é aplicada ao problema de previsão de demanda de barras de aço para corte, constituindo etapa essencial para a integração com o modelo de otimização de cortes com aproveitamento de sobras, visando maximizar a eficiência operacional e minimizar perdas no processo produtivo.

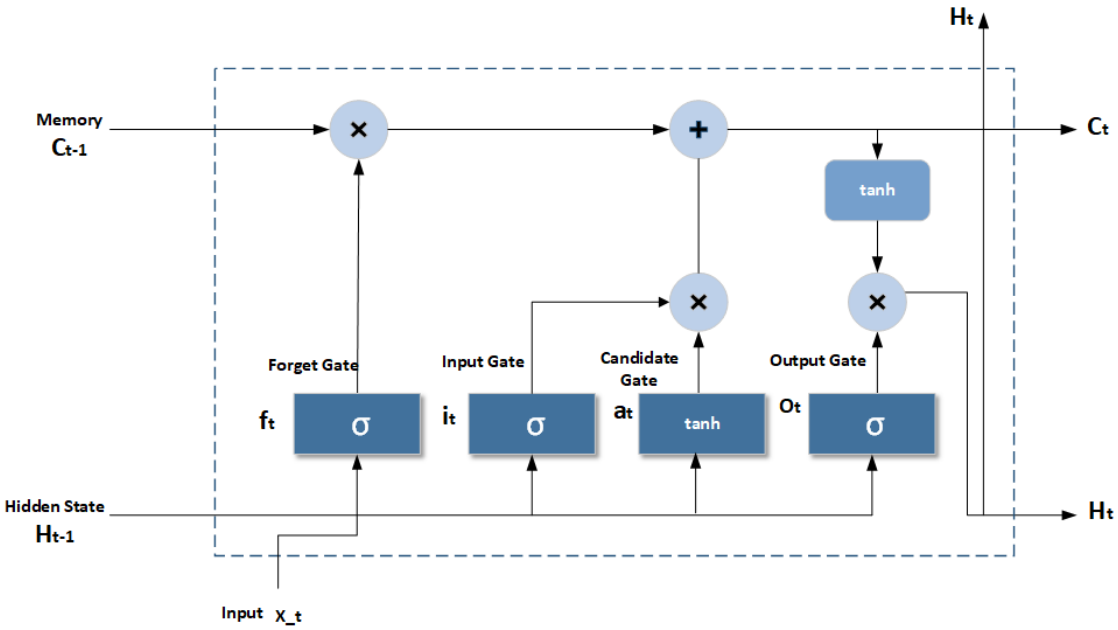
2.5.1 *Long Short-Term Memory* (LSTM)

As redes LSTM, introduzidas por Hochreiter e Schmidhuber (1997), são uma evolução das redes neurais recorrentes (RNNs), projetadas especificamente para superar o problema do desapa-

recimento do gradiente, que limita a capacidade das RNNs tradicionais de capturar dependências de longo prazo. A principal vantagem das LSTMs está na sua habilidade de armazenar, atualizar e esquecer informações de forma controlada ao longo do tempo, o que as torna eficazes para a modelagem de séries temporais e dados sequenciais.

O funcionamento das redes LSTM é viabilizado por uma arquitetura que permite o processamento e a propagação eficiente de informações ao longo do tempo, por meio de conexões internas entre as células. Como ilustrado na figura 4, a informação percorre dois caminhos distintos: o primeiro está associado ao fluxo de memória de longo prazo (*"Memory"*), enquanto o segundo está relacionado ao processamento do estado oculto atual (*Hidden State*), responsável pela saída imediata da célula e pela transmissão do contexto para os próximos passos temporais. Essa estrutura garante que informações relevantes sejam preservadas, atualizadas e descartadas de forma seletiva e eficaz.

Figura 4 – Arquitetura Célula LSTM



Fonte: Amin (2018)

O primeiro caminho, representado pelo *"Memory"* é responsável por armazenar e preservar informações relevantes ao longo da sequência. Ele começa a partir do estado da célula anterior (C_{t-1}), que passa por diferentes portas (*gates*) para definir quais informações serão retidas ou descartadas.

- *Forget Gate* (Porta do Esquecimento): Decide quais informações do estado da célula anterior (C_{t-1}) devem ser mantidas ou descartadas. A decisão é tomada a partir da combinação do

estado oculto anterior (H_{t-1}) e da entrada atual (X_t) processada por uma função sigmoide que gera valores entre 0 e 1. Valores próximos de 0 indicam que a informação será descartada, enquanto valores próximos de 1 indicam que será preservada.

- *Input Gate* (Porta de Entrada) controla quais novas informações devem ser adicionadas à célula de memória. Assim como o *Forget Gate*, ele recebe o estado oculto anterior (H_{t-1}) e a entrada atual (X_t) como entradas, e o processamento ocorre em duas etapas. Primeiramente, uma função sigmoide regula quais partes da nova informação devem ser atualizadas, gerando um coeficiente entre 0 e 1 para filtrar os dados mais relevantes. Em seguida, uma função *tanh* cria um vetor candidato, com valores processados para a atualização da célula, variando entre -1 e 1, permitindo representar tanto a magnitude quanto a polaridade das informações. A multiplicação entre esse vetor e o filtro sigmoide determina quais informações serão incorporadas à célula, garantindo que apenas os dados mais úteis sejam armazenados e utilizados no aprendizado de longo prazo. O resultado final dessa operação é somado ao que foi mantido após a etapa do *Forget Gate*, formando o novo estado da célula (C_t).

O segundo caminho corresponde ao *Hidden State* e é responsável por determinar a saída da célula no instante atual, além de transmitir o contexto para as próximas unidades da rede:

- *Output Gate* (Porta de Saída): Controla quais partes do estado da célula atual (C_t) serão usadas para gerar a saída (H_t). Para isso, recebe (H_{t-1}) e (X_t), aplicando uma função sigmoide que regula as informações a serem expostas, enquanto (C_t) é transformado por uma função *tanh*, que ajusta seus valores para garantir uma transição das informações para a próxima célula. O produto dessas duas operações gera a saída H_t, que desempenha dois papéis fundamentais: fornecer a saída imediata da LSTM no tempo t, e atuar como entrada para a próxima célula. Esse mecanismo permite que a rede combine informações passadas e presentes, ajustando dinamicamente suas previsões de acordo com o contexto sequencial.

2.6 Métricas

As métricas e funções empregadas neste trabalho desempenham papéis complementares e fundamentais. Durante a fase de treinamento, a função de ativação *Rectified Linear Unit* (*ReLU*), a função de perda *Mean Squared Error* (MSE) e a técnica de regularização *Early Stopping* são cruciais para guiar o aprendizado da rede LSTM, monitorar seu progresso e evitar sobreajuste. Posteriormente, métricas como o *Mean Absolute Error* (MAE) e o *Mean Absolute Percentage Error* (MAPE) serão utilizadas para uma avaliação quantitativa e qualitativa robusta dos resultados obtidos, permitindo a comparação com outros modelos ou abordagens.

Em conjunto, essas métricas e funções fornecem uma base sólida para o treinamento e a validação do modelo, permitindo que a LSTM aprenda de forma mais eficiente e produza previsões mais acuradas. Essa combinação garante tanto a qualidade do ajuste durante o aprendizado quanto a precisão na análise dos resultados obtidos.

2.6.1 Função de Ativação *ReLU*

A função de ativação é um componente vital em redes neurais, sendo responsável por introduzir não-linearidades no modelo. Fundamentalmente, sua função é determinar se um neurônio deve ser ativado ou não, com base na soma ponderada de suas entradas, permitindo assim que a rede capture relações complexas e não lineares presentes nos dados (GOODFELLOW et al., 2016). Neste trabalho, foi adotada a função de ativação *Rectified Linear Unit* (ReLU), definida pela equação 2.13:

$$\mathit{ReLU}(x) = \max(0, x) \tag{2.13}$$

Dentre suas principais vantagens, destacam-se a simplicidade computacional e a eficiência no treinamento de redes. A *ReLU* mitiga o problema do gradiente desaparecente, comum em funções de ativação saturantes (como a sigmoide ou tangente hiperbólica), pois sua derivada é 1 para entradas positivas (GLOROT; BORDES; BENGIO, 2011). Isso resulta em um processo de retropropagação mais estável e acelera significativamente a convergência do modelo.

A escolha da taxa de aprendizado está intrinsecamente ligada à eficácia da *ReLU*. A taxa de aprendizado determina a magnitude dos ajustes aplicados aos pesos da rede a cada iteração do algoritmo de otimização. Um valor excessivamente alto pode impedir a convergência, fazendo com que o modelo oscile em torno do mínimo da função de perda, enquanto um valor demasiadamente baixo pode tornar o treinamento lento e propenso a ficar preso em mínimos locais. Portanto, a sinergia entre a função de ativação *ReLU* e uma taxa de aprendizado adequadamente calibrada é essencial para um aprendizado estável e eficiente da rede LSTM.

2.6.2 *Mean Squared Error* (MSE)

Para otimizar os pesos da rede LSTM durante o treinamento, foi utilizada como função de perda a métrica MSE. Esta é uma medida amplamente adotada em problemas de regressão, calculada pela média dos quadrados das diferenças entre os valores previstos e os observados (HASTIE, 2009). Sua formulação é dada pela equação 2.14:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.14}$$

onde y_i representa o valor real (alvo) e $\hat{y}_i$ o valor previsto pelo modelo para a i-ésima observação em um conjunto de n amostras.

Uma característica importante do MSE é a sua propriedade de penalizar erros grandes de forma mais significativa do que erros pequenos, devido ao termo quadrático. Isso incentiva o modelo a produzir previsões não apenas corretas, mas também consistentes, com menor variância em relação aos valores reais (GOODFELLOW et al., 2016).

2.6.3 *Early Stopping*

O *Early Stopping* é uma técnica de regularização utilizada para prevenir o sobreajuste. O sobreajuste ocorre quando o modelo se adapta excessivamente aos dados de treinamento, incluindo seu ruído e flutuações aleatórias, perdendo assim a capacidade de generalizar para dados não vistos (GOODFELLOW et al., 2016).

A técnica de *Early Stopping* monitora continuamente o desempenho do modelo durante o treinamento e interrompe o processo de forma antecipada quando o erro de validação não apresenta melhoria significativa após um número pré-definido de épocas, denominado "paciência". Dessa maneira, os pesos do modelo no momento da interrupção correspondem àqueles que proporcionaram a melhor capacidade de generalização, e não necessariamente aos obtidos na última época de treinamento (PRECHELT, 2002). Além de controlar a duração do treinamento, o *Early Stopping* é computacionalmente eficiente e frequentemente determinante para o sucesso prático de modelos de redes neurais.

2.6.4 *Mean Absolute Error* (MAE)

O *Mean Absolute Error* (MAE) é uma métrica fundamental para avaliação de modelos de regressão que mede a magnitude média dos erros absolutos entre as previsões do modelo e os valores reais. Diferente do MSE, o MAE não eleva os erros ao quadrado, resultando em penalização linear para todos os resíduos e tratando desvios grandes e pequenos de forma proporcional (WILLMOTT; MATSUURA, 2005).

Sua formulação é dada pela equação 2.15:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.15}$$

onde y_i representa o valor real, $\hat{y}_i$ o valor previsto pelo modelo e n o número total de observações.

Por apresentar a mesma unidade de medida da variável-alvo, o MAE possui uma interpretação intuitiva e direta, representando o erro médio esperado para uma previsão (HYNDMAN; KOEHLER, 2006).

2.6.5 *Mean Absolute Percentage Error* (MAPE)

O *Mean Absolute Percentage Error* (MAPE) é uma métrica de erro relativo frequentemente utilizada para expressar a acurácia de modelos de previsão em termos percentuais. Ela mede o erro absoluto médio em relação à magnitude dos valores reais, oferecendo uma perspectiva relativa do desempenho do modelo.

Sua formulação é definida pela equação 2.16:

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{2.16}$$

A principal vantagem do MAPE é permitir comparações diretas entre modelos aplicados a diferentes conjuntos de dados ou escalas de valores, tornando-o especialmente útil em análises de séries temporais e problemas de previsão (HYNDMAN; KOEHLER, 2006).

3 Base de Dados

Para viabilizar a implementação e o treinamento do modelo preditivo baseado em aprendizado de máquina, foi desenvolvida uma base de dados artificial. O objetivo central dessa construção foi incorporar características empíricas reportadas na literatura e em artigos científicos internacionais, de modo a criar cenários realistas de comportamento da demanda. Com isso, possibilita-se a realização de análises comparativas mais consistentes e a obtenção de resultados robustos nas etapas subsequentes do estudo.

A base de dados contempla as seguintes variáveis: *Data*, *Código do Item*, *Depósito*, *Variáveis Macroeconômicas (Preço, Dólar, Disponibilidade, Inflação e Importação)* e *Demanda*. Foram considerados 30 itens distintos, distribuídos em três categorias de porte: Pequenos (P), Médios (M) e Grandes (G), cada uma composta por 10 itens. A identificação de cada item foi realizada de acordo com sua categoria, seguida de um número sequencial, resultando nos códigos P1 a P10, M1 a M10 e G1 a G10.

Além disso, para cada cliente, foi construída uma base de dados individual, incorporando suas respectivas especificidades de consumo. O horizonte temporal simulado abrange o período de 1º de janeiro de 2017 a 30 de abril de 2025, permitindo capturar a dinâmica da demanda ao longo do tempo e refletir tanto as influências de fatores macroeconômicos quanto as características particulares de cada cliente.

A tabela 1 ilustra a estrutura da base de dados gerada, apresentando um exemplo de registro consolidado. A seguir, é fornecida uma explicação detalhada do significado e funcionamento de cada variável.

Tabela 1 – Exemplo da formação da base de dados

Dia	Código	Depósito	Preço	Dólar	Disponibilidade	Inflação	Importação	Demanda
03/01/2017	P1	D2	442,0	3,264	2,404	0,38	201	368
04/01/2017	G4	D1	440,0	3,216	2,404	0,38	201	194
05/01/2017	M2	D3	435,0	3,198	2,404	0,38	201	260

Fonte: Elaborada pelo autor

3.1 Variáveis Macroeconômicas

As variáveis macroeconômicas são utilizadas como indicadores temporais para simular a variação da demanda ao longo do tempo, refletindo os impactos de fatores externos à operação

logística. Tais variáveis influenciam diretamente o comportamento de consumo, produção e custo dos materiais, permitindo uma simulação mais realista e próxima do comportamento de mercado.

3.1.1 Preço do Aço Bruto

Para representar o custo da matéria-prima na variável *Preço*, utilizou-se a série histórica diária do *Steel Rebar Futures*, um contrato de vergalhão de aço amplamente negociado em dólares norte-americanos na Bolsa de Futuros de Xangai (*Shanghai Futures Exchange - SHFE*) (Investing.com, 2025a). Esse indicador reflete as expectativas do mercado sobre o preço do aço, servindo como referência global para os custos de produção em diversos setores da indústria metalúrgica e da construção civil.

O *Steel Rebar Futures* é amplamente utilizado como proxy para o preço do aço bruto, pois suas flutuações são influenciadas por fatores como demanda internacional, políticas comerciais, disponibilidade de insumos e variações cambiais. Neste estudo, a variação temporal dessa série foi utilizada para ajustar dinamicamente a demanda simulada dos itens, refletindo o impacto de alterações nos custos produtivos ao longo do tempo.

O gráfico da figura 5 apresenta a evolução do preço do aço durante o período analisado. Observam-se picos significativos entre 2021 e 2022, indicando momentos de elevada volatilidade do mercado, enquanto o período mais recente demonstra uma tendência de estabilização, com os preços variando em torno de uma faixa intermediária.

Figura 5 – Evolução do Preço do Aço

Fonte: Elaborada pelo Autor

3.1.2 Taxa de Câmbio (Dólar)

A cotação do dólar foi considerada como uma variável macroeconômica relevante, obtida a partir do conjunto de dados históricos diários disponível em Investing.com (2025b). A taxa de câmbio do dólar norte-americano exerce influência significativa sobre o mercado interno brasileiro, especialmente nos setores industriais que dependem de insumos importados ou mantêm relações comerciais internacionais.

Como principal moeda de referência no comércio global, as variações na cotação do dólar, ilustradas na figura 6, afetam diretamente os custos de produção, transporte e aquisição de matérias-primas. Dessa forma, essa variável foi incorporada à base de dados como um fator externo que influencia a demanda dos itens simulados.

Figura 6 – Evolução da Cotação do Dólar

Fonte: Elaborada pelo autor

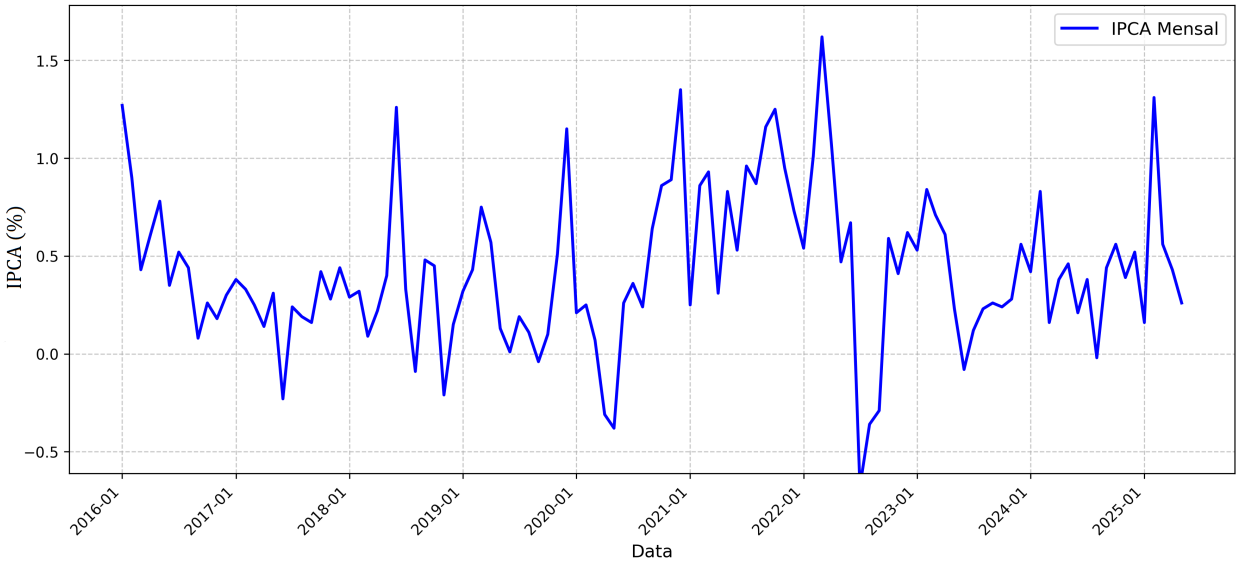
3.1.3 Inflação

A inflação é um fenômeno econômico que representa o aumento contínuo e generalizado dos preços de bens e serviços, resultando na desvalorização da moeda e na redução do poder de compra. Considerando sua influência sobre o comportamento do mercado e o consumo, a inflação foi incorporada à simulação como uma variável macroeconômica exógena, impactando diretamente a demanda dos itens ao longo do tempo.

Para representar esse fenômeno de forma fidedigna à realidade econômica brasileira, adotou-se a série histórica mensal do Índice Nacional de Preços ao Consumidor Amplo (IPCA),

disponível em (Instituto Brasileiro de Geografia e Estatística (IBGE), 2025). Reconhecido como o principal indicador oficial de inflação no país, o IPCA é meticulosamente apurado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), instituição governamental responsável pela produção e disseminação de estatísticas oficiais. Este índice desempenha um papel crucial na economia nacional, servindo como referência central para o sistema de metas inflacionárias do Banco Central, além de fundamentar reajustes contratuais, políticas salariais e análises macroeconômicas estratégicas. Sua variação ao longo do tempo pode ser observada na figura 7.

Figura 7 – Evolução da Inflação (IPCA)



Fonte: Elaborada pelo autor

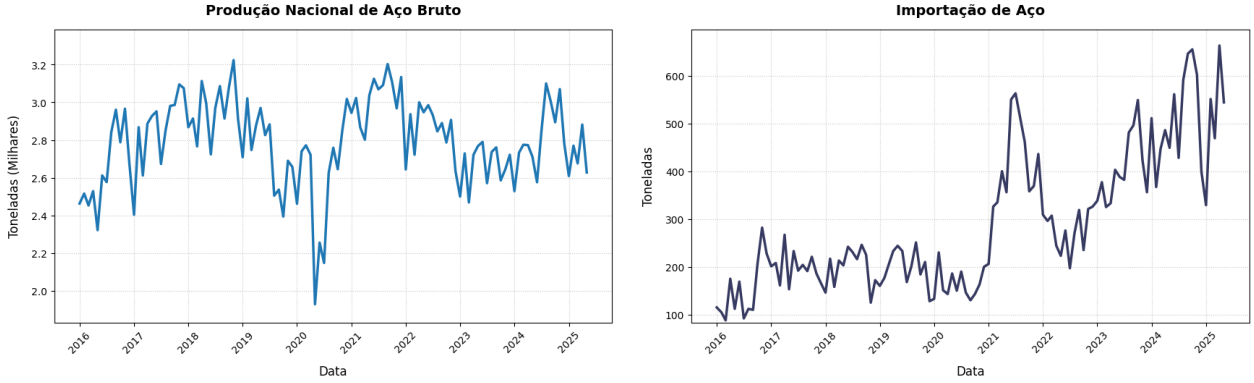
3.1.4 Importação e Disponibilidade Nacional de Aço

A disponibilidade interna de aço, mensurada pela produção das siderúrgicas nacionais, e os volumes de importação, ambas expressas em toneladas métricas, constituem variáveis macroeconômicas fundamentais para a análise da dinâmica de oferta e demanda do setor siderúrgico brasileiro. Esses indicadores exercem influência direta na formação de preços dos produtos acabados, atuando como determinantes críticos tanto nos custos de produção quanto nos padrões de consumo do mercado doméstico. A produção nacional reflete a capacidade instalada e a utilização da capacidade produtiva do parque siderúrgico, enquanto as importações funcionam como variável de ajuste para suprir déficits pontuais de oferta ou aproveitar vantagens competitivas internacionais. Juntas, essas métricas permitem calcular o *market share* da produção doméstica e o grau de dependência de suprimentos externos, aspectos cruciais para a compreensão da competitividade setorial.

Os dados utilizados abrangendo todo o período de análise foram obtidos junto ao Instituto Aço Brasil (Instituto Aço Brasil, 2025) (figura 8), entidade representativa da cadeia produtiva do aço no país. O Instituto reúne as principais empresas produtoras de aço do Brasil e disponibiliza dados estatísticos mensais, relatórios setoriais e informações de mercado. Muitos desses dados contam com o apoio do Ministério do Desenvolvimento, Indústria, Comércio e Serviços (MDIC), conferindo legitimidade e abrangência à base utilizada.

A inclusão dessas informações na base de dados permite capturar oscilações estruturais relevantes, especialmente em um contexto em que a cadeia produtiva está sujeita a fatores globais, como variações no comércio internacional, custos de insumos e políticas industriais.

Figura 8 – Evolução da Produção e Importação do Aço



Fonte: Elaborada pelo autor

3.2 Demanda

A modelagem da demanda em cadeias de suprimentos é um tema de grande relevância acadêmica e prática, sobretudo em indústrias de base, como a siderúrgica, cujo desempenho depende diretamente de variáveis econômicas e dinâmicas de mercado. Para este projeto, a fundamentação teórica da modelagem da demanda baseia-se em duas abordagens complementares da literatura.

A primeira abordagem considera a sensibilidade da demanda ao preço dentro da otimização da rede de suprimentos. Hu et al. (2024) propõem um modelo multiproduto e multiperíodo de projeto de rede de cadeia de suprimentos (SCND), em que a demanda de cada produto é tratada como função inversa de seu preço de venda, refletindo o princípio econômico clássico de que aumentos de preço tendem a reduzir o consumo. O objetivo central do estudo é maximizar o lucro global da cadeia de suprimentos, determinando simultaneamente as melhores configurações de rede, estoques e decisões de precificação ao longo de múltiplos períodos.

Em paralelo, a influência de fatores macroeconômicos externos sobre a demanda de commodities, como o aço, é igualmente relevante. Mehmanpazir, Khalili-Damghani e Hafezalkotob (2019) investigaram essa relação utilizando dados reais da indústria do Irã, adotando um modelo de regressão logarítmica múltipla (Log-Log Linear) para estimar funções de oferta e demanda. Essa abordagem permite capturar relações não-lineares de forma linearizada, em que os coeficientes de regressão (β_i) representam elasticidades constantes, ou seja, a variação percentual da demanda em resposta a uma variação percentual de cada variável econômica. O estudo evidenciou que a taxa de câmbio (dólar), a inflação (IPCA), os níveis de importação, a disponibilidade interna de aço e o preço do aço bruto exercem impacto direto e mensurável sobre a demanda.

Inspirado por essas contribuições, este projeto adota uma formulação linear simplificada, adequada ao tratamento de dados simulados. A modelagem assume que a variação percentual da demanda de cada cliente depende linearmente das variações percentuais das principais variáveis macroeconômicas, ponderadas por coeficientes de elasticidade obtidos em (MEHMANPAZIR; KHALILI-DAMGHANI; HAFEZALKOTOB, 2019). A cada novo pedido registrado, a demanda de todos os itens do cliente é atualizada em função dessas alterações, de modo a refletir consistentemente o impacto das condições econômicas sobre o consumo individual e preservar a coerência temporal entre os itens.

A equação 3.1 aplicada para calcular a variação percentual da demanda (ΔD) é:

$$\begin{aligned} \Delta D = {} & 0.867 \cdot \Delta Disponibilidade + 0.162 \cdot \Delta Importação \\ & - 0.067 \cdot \Delta Inflação + 0.119 \cdot \Delta Preço \\ & + 0.068 \cdot \Delta Dólar \end{aligned} \tag{3.1}$$

Essa variação é então aplicada à demanda de cada item por cliente conforme a equação 3.2:

$$Demanda = Demanda \cdot (1 + \Delta D) \tag{3.2}$$

em que:

- ΔD: variação da demanda;
- $\Delta Disponibilidade$: variação na oferta de aço no mercado interno em relação ao último pedido do cliente;
- $\Delta Importação$: variação na importação de aço em relação ao último pedido do cliente;

- $\Delta Inflação$: variação da inflação (IPCA) em relação ao último pedido do cliente;
- $\Delta Preço$: variação no preço do aço bruto em relação ao último pedido do cliente;
- $\Delta Dólar$: variação na taxa de câmbio (dólar) em relação ao último pedido do cliente.

Na simulação, a atualização da demanda de todos os itens considera a variação percentual das variáveis macroeconômicas com base no último pedido do cliente. Esse mecanismo valoriza clientes com maior frequência de compras, cujas demandas se estabilizam devido à menor oscilação, ao mesmo tempo em que aplica correções proporcionais para clientes com intervalos maiores entre pedidos. Dessa forma, a base de dados consegue refletir de maneira equilibrada tanto os efeitos de curto quanto de longo prazo, capturando variações estruturais e pontuais no comportamento da demanda.

3.3 Cliente

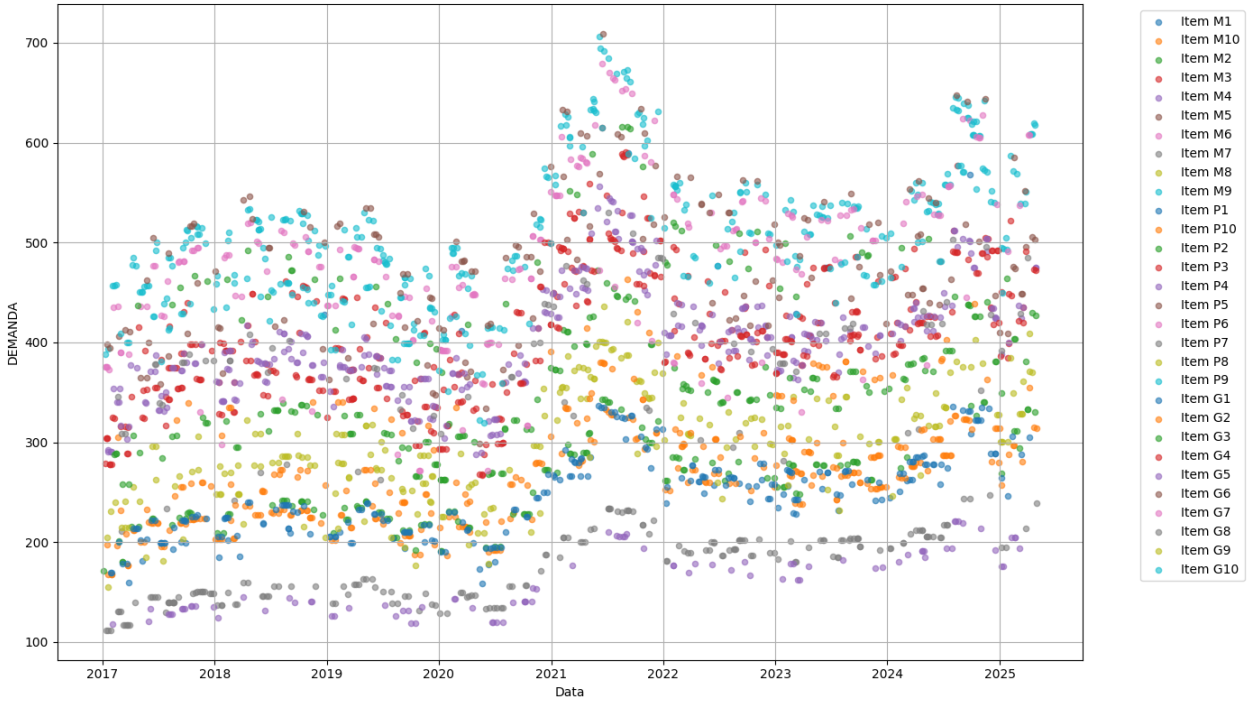
A feature *Cliente* representa o agente responsável por iniciar o processo de aquisição dos itens, sendo, portanto, a entidade que demanda os produtos ao longo do horizonte de simulação. Foram considerados 20 clientes distintos, identificados pelo código C seguido de uma sequência de 1 a 20, resultando em C1 a C20.

Cada cliente apresenta comportamentos específicos de consumo, refletindo particularidades em suas demandas para diferentes itens. A associação entre um cliente e um item é definida por uma probabilidade aleatória no intervalo $[0, 100]$, simulando possíveis relações comerciais observadas no mercado real.

Após estabelecida a associação, a demanda inicial de cada item vinculado a um cliente é gerada aleatoriamente no intervalo de 100 a 400 unidades. A partir desse ponto, a demanda evolui dinamicamente ao longo de cada dia da simulação, sendo influenciada por variáveis exógenas presentes na base de dados, como o preço da matéria-prima e indicadores macroeconômicos, de modo a reproduzir oscilações de mercado e padrões de compra mais realistas conforme descritas na Seção 3.1.

A figura 9 apresenta um exemplo prático para o Cliente C6, demonstrando a trajetória temporal da demanda para cada item. Observa-se que, apesar de se tratar do mesmo cliente, cada item exibe uma curva de evolução distinta, resultado das particularidades de consumo individuais e da influência das variáveis econômicas ao longo do tempo.

Figura 9 – Variação da Demanda por Item do Cliente C6



Fonte: Elaborada pelo autor

3.4 Depósito

A feature *Depósito* representa os centros logísticos responsáveis pelo armazenamento e distribuição dos itens, os quais também funcionam como filiais produtivas no sistema simulado. Foram considerados três depósitos distintos, identificados como D1, D2 e D3.

A escolha do depósito que atenderá a demanda de um determinado cliente é feita com base em uma distribuição probabilística, que simula afinidades logísticas ou preferências operacionais. Para cada cliente, a probabilidade de atendimento por D1 é sorteada no intervalo $[20, 80]$. A partir desse valor, a probabilidade de D2 é determinada como um novo valor sorteado dentro do intervalo $[20, 80]$, respeitando o limite da soma total de 100. Por fim, a probabilidade restante é atribuída automaticamente a D3, garantindo que a soma das probabilidades para cada cliente seja sempre igual a 100.

A capacidade dos depósitos constitui um elemento essencial para a representação realista das restrições operacionais do sistema, estando diretamente associada tanto ao potencial produtivo quanto à capacidade de estocagem dos itens. Para sua definição, adotou-se uma abordagem baseada na simulação da maior requisição de demanda possível, considerando o comportamento das variáveis macroeconômicas relevantes. Essa estimativa da capacidade máxima foi obtida a partir da fórmula apresentada na Seção 3.2, projetando-se o cenário de maior criticidade e elevando

a demanda a fim de capturar a sua expansão máxima.

A simulação utilizou dados históricos referentes ao período de 01/01/2016 a 31/12/2016, com o propósito de identificar as condições macroeconômicas mais favoráveis ao crescimento da demanda. Para isso, foram selecionados o menor índice de disponibilidade interna de aço e, de forma coerente, o nível de importação correspondente ao mesmo mês, de modo a manter consistência entre os indicadores. Além disso, consideraram-se o menor preço do aço bruto e a menor taxa de câmbio do dólar, ambos favorecendo o aumento expressivo do consumo. Já para a inflação (IPCA), foi utilizado o maior valor registrado, uma vez que seu impacto sobre a demanda é inversamente proporcional, maximizando assim a variação negativa em contraste com os efeitos positivos das demais variáveis.

A partir dessa simulação, determinou-se a demanda máxima potencial, sobre a qual foi aplicada uma margem adicional de 50% com o objetivo de assegurar o adequado dimensionamento operacional ao longo de um horizonte de aproximadamente oito anos consecutivos. A demanda ajustada foi então distribuída entre os depósitos, atribuindo-se 30% a D1, 30% a D2 e 40% a D3, em consonância com a estratégia de priorização do armazenamento e do fluxo logístico. Os valores finais para cada depósito encontram-se sintetizados na tabela 2, a qual apresenta a capacidade máxima considerada para o planejamento operacional.

Tabela 2 – Capacidade máxima dos depósitos

Depósito	**Capacidade Máxima (Itens)**
D1	6.678
D2	6.678
D3	8.904

Fonte: Elaborada pelo autor

4 Metodologia

Neste capítulo, é descrita a metodologia adotada para o desenvolvimento do trabalho, contemplando as etapas do processo, as ferramentas empregadas, as bibliotecas utilizadas e a configuração aplicada ao aprendizado de máquina.

4.1 Treinamento do Modelo de Predição de Demanda

Com o propósito de predizer a demanda aplicada ao PCESA, foram realizadas diversas etapas de pré-processamento e ajustes na base de dados descrita anteriormente, com foco específico no treinamento de um modelo de aprendizado de máquina baseado na arquitetura LSTM (Subseção 2.5.1). A escolha desse modelo se justifica por sua capacidade de capturar dependências temporais de longo prazo, preservando informações de estados anteriores e incorporando o contexto histórico no processo de previsão. Esse mecanismo é particularmente relevante em séries temporais, em que a dinâmica de variação da demanda não pode ser compreendida apenas por valores pontuais. Dessa forma, as tratativas adotadas visaram otimizar a fase de treinamento, ampliando a ineficiência do modelo e aprimorando sua capacidade de generalização ao longo do tempo.

4.1.1 Scaler *MinMaxScaler*

Um dos desafios no uso de variáveis macroeconômicas e operacionais como entrada do modelo é a diferença de ordem de grandeza entre elas: enquanto algumas se encontram em escala de unidades, outras atingem valores na casa das centenas e milhares. Essa disparidade compromete o desempenho de algoritmos baseados em redes neurais, uma vez que pesos associados a variáveis de maior magnitude tendem a dominar o processo de otimização. Para mitigar esse efeito, empregou-se o *MinMaxScaler*, técnica de normalização que transforma cada variável para um intervalo comum, tipicamente $[0, 1]$, conforme a Equação 4.1, em que cada valor x da variável original é ajustado com base em seu valor mínimo (x_{min}) e máximo (x_{max}) observados na base de dados. No presente trabalho, o *MinMaxScaler* foi empregado tanto nas *lag features*, apresentadas na Seção 4.1.4 derivadas da demanda histórica, quanto nas variáveis macroeconômicas conforme descrito na Seção 4.1.5.

$$x' = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{4.1}$$

Essa estratégia garantiu que todas as variáveis fossem representadas em uma escala padronizada, promovendo equilíbrio em sua contribuição para o processo de treinamento do modelo e mitigando distorções decorrentes de diferenças de magnitude entre elas.

4.1.2 Feature *Item*

Como o modelo trabalha exclusivamente com variáveis numéricas, foi necessário transformar as variáveis categóricas antes do treinamento. Para isso, inicialmente aplicou-se a técnica de *Label Encoding*, que converte cada categoria em um valor numérico inteiro. Essa abordagem é simples e eficiente, mas pode levar o modelo a interpretar as categorias como se existisse uma ordem ou hierarquia entre elas. Por exemplo, se a variável categórica 'Item' for codificada como P1 = 0, P2 = 1, ..., M1 = 11, ..., G1 = 21, o modelo poderia inferir incorretamente que G1 é 'maior' que M1 e que M1 é 'maior' que P1, como se houvesse uma relação de magnitude entre os itens, quando na realidade eles representam apenas categorias distintas.

Para corrigir esse problema, aplicou-se em seguida a técnica de *One Hot Encoder* (OHE). Nessa abordagem, cada categoria é convertida em uma nova coluna binária, que assume o valor 1 quando o item está presente e 0 caso contrário. Assim, os itens P1 a P10, M1 a M10 e G1 a G10 são transformados em múltiplas colunas, uma para cada item, em que apenas a coluna correspondente ao item em questão terá valor 1 em cada registro. Dessa forma, o modelo não atribui nenhuma ordem ou hierarquia artificial entre os itens, tratando-os corretamente como categorias independentes.

4.1.3 Feature *Data*

A variável temporal (*Data*) desempenha um papel crucial na previsão da demanda, pois é o elemento que codifica a sequência cronológica dos eventos, permitindo ao modelo LSTM identificar padrões cíclicos, sazonalidades e tendências de longo prazo. Para que o modelo possa interpretar com precisão a natureza multifacetada do tempo, a variável de data foi submetida a um processo de transformação e decomposição, separando-a em componentes discretos e contínuos.

O Dia do Mês possui uma característica de periodicidade que não pode ser representada de forma adequada por uma codificação numérica linear (e.g., 1 a 31). Nessa codificação simples, a distância numérica entre o dia 31 e o dia 1 é máxima, o que contradiz a realidade temporal em que esses dias são adjacentes no ciclo mensal.

Para preservar a proximidade intrínseca entre o final e o início do mês, o Dia foi transformado em duas variáveis contínuas, derivadas das funções trigonométricas seno (Equação 4.2) e cosseno (Equação 4.3) resultando na Figura 10. Essa transformação mapeia cada dia para um

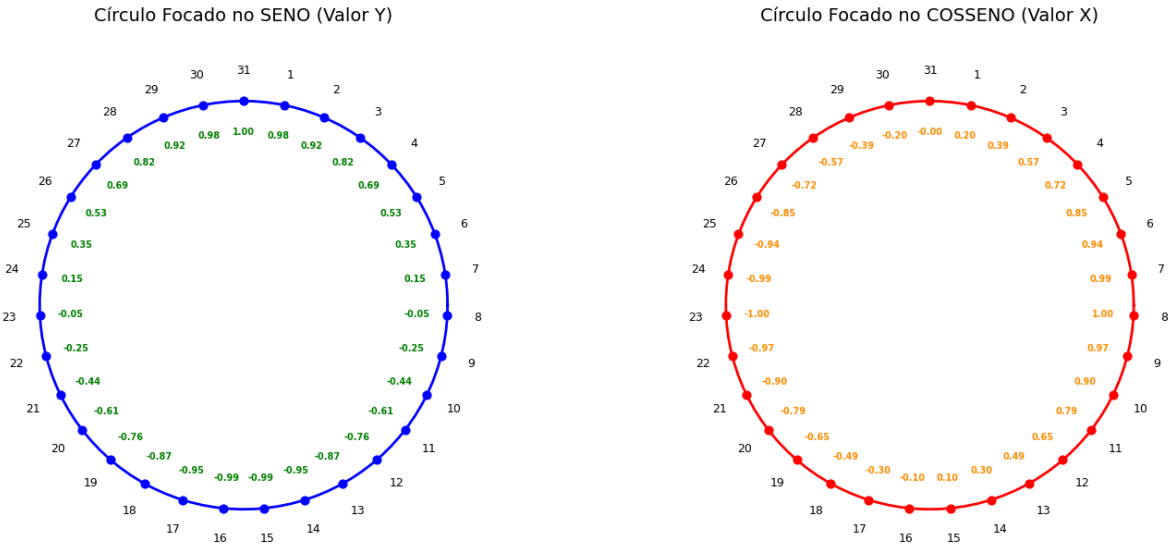
ponto único sobre o círculo unitário, preservando a continuidade entre o final e o início do mês e garantindo que a proximidade temporal seja representada de forma consistente.

$$\text{Dia}_{\text{Seno}} = \sin\left(\frac{2\pi \cdot (\text{Dia})}{31}\right) \tag{4.2}$$

$$\text{Dia}_{\text{Cosseno}} = \cos\left(\frac{2\pi \cdot (\text{Dia})}{31}\right) \tag{4.3}$$

A representação bidimensional resultante garante que o modelo reconheça a natureza cíclica da variável, tratando a transição do último para o primeiro dia do mês como contínua.

Figura 10 – Círculo Trigonométrico dos Dias do Mês



Fonte: Elaborada pelo autor

Além disso, os meses e anos foram tratados com OHE. Cada um dos 12 meses foi convertido em uma coluna binária, na qual apenas o mês correspondente a cada registro recebe valor 1. Isso permite que o modelo capture efeitos sazonais, como aumentos de demanda em determinados períodos, sem supor qualquer hierarquia entre os meses. De forma similar, os anos foram transformados em colunas binárias separadas, permitindo ao modelo identificar mudanças estruturais ou choques específicos em determinados períodos, como crises econômicas ou variações macroeconômicas de longo prazo. Essas transformações garantem uma representação completa e adequada da dimensão temporal, preservando a independência das categorias e a periodicidade dos ciclos, o que melhora a capacidade do modelo de aprender padrões sazonais e temporais relevantes para a previsão da demanda.

4.1.4 *Features* para Dependência Temporal

Para capturar a relação sequencial entre a demanda atual e o histórico de consumo de cada item, foram utilizadas *lag features*, que correspondem às últimas demandas registradas pelo mesmo cliente para o item em questão. Como a demanda é atualizada a cada novo pedido realizado, independentemente do item solicitado (conforme descrito na Seção 3.2), essas variáveis desempenham papel essencial na modelagem da dependência temporal.

A variável central considerada é a demanda anterior do mesmo item para o mesmo cliente, que atua como referência direta para a identificação de padrões de correlação temporal com a demanda futura. Dessa forma, o modelo pode explorar de maneira consistente a continuidade no comportamento de consumo individual.

Para assegurar a comparabilidade com as demais entradas do modelo, as *lag features* foram submetidas ao processo de normalização via *MinMaxScaler*, de modo a ajustar sua escala em relação às variáveis macroeconômicas. Esse procedimento evita que discrepâncias de magnitude entre os atributos distorçam o processo de aprendizado.

Com essa estratégia, o modelo é capaz de capturar tanto a dependência temporal intrínseca do consumo quanto os efeitos das variáveis externas, ampliando sua capacidade de generalização e resultando em previsões de demanda mais robustas e precisas.

4.1.5 Variáveis Macroeconômicas

A inclusão de variáveis macroeconômicas na modelagem da demanda é crucial para que o modelo LSTM capture a sensibilidade do consumo a fatores externos e, consequentemente, melhore sua capacidade de generalização. Para representar essa influência de forma eficaz, foi realizada uma transformação direta dos dados. O método consistiu na criação de cinco novas features, cada uma associada a uma variável macroeconômica descritas na Seção 3.1. O valor de entrada para cada uma dessas features é dado pela variação percentual ($\Delta\%$) da variável em questão, calculada em comparação entre o valor registrado no momento do último pedido do cliente para este código de item e o valor atual do pedido. Essa abordagem concentra a informação na dinâmica de mudança e sua magnitude real, permitindo que o modelo identifique com clareza a taxa e a direção da variação econômica que precede a nova demanda, resultando em um aprendizado mais robusto sobre a relação temporal entre choques macroeconômicos e o padrão de consumo individual.

4.1.6 Hiperparâmetros

No desenvolvimento do modelo LSTM, os hiperparâmetros foram definidos de forma criteriosa com o intuito de equilibrar a capacidade de aprendizado e a generalização. A arquitetura foi composta por duas camadas LSTM em sequência: a primeira, com 64 neurônios, voltada à extração de padrões complexos presentes nos dados, e a segunda, com 25 neurônios, responsável por realizar uma compressão representacional mais focada na tarefa de predição. Em ambas as camadas, adotou-se a função de ativação ReLU (Rectified Linear Unit), que introduz não linearidades de forma eficiente e contribui para mitigar o problema do desaparecimento do gradiente. Para reduzir o risco de sobreajuste, aplicou-se um *Dropout* de 20% entre as camadas, desativando aleatoriamente parte dos neurônios durante o treinamento e, assim, ampliando a capacidade de generalização do modelo. O treinamento foi conduzido com a função de perda MSE, tradicional em problemas de regressão por penalizar desvios quadráticos e favorecer previsões consistentes. Para otimização, adotou-se o algoritmo Adam em conjunto com uma estratégia de decaimento manual da taxa de aprendizado em três estágios progressivos: inicial de 0.01 por 15 épocas, permitindo rápida convergência; em seguida 0.001 por 35 épocas, para estabilização dos parâmetros; e, por fim, 0.0001 por 10 épocas, assegurando um refinamento da solução. Em sincronia, os tamanhos de lote foram igualmente ajustados de maneira gradual, passando de 32 para 16 e, por fim, para 8, a fim de explorar diferentes dinâmicas de atualização do gradiente ao longo do processo de treinamento. Complementarmente, implementou-se o *callback Early Stopping* com paciência de 10 épocas, monitorando a função de perda e interrompendo o treinamento em caso de estagnação, além de restaurar automaticamente os melhores pesos obtidos.

4.2 Resolução do PCESA

A resolução do modelo matemático descrito na Seção 2.1 começa com a definição de um conjunto inicial de padrões de corte. Essa etapa é essencial, uma vez que o método de geração de colunas requer que o problema mestre seja iniciado a partir de uma base viável, a qual permite a construção incremental de soluções melhores ao longo das iterações.

Três tipos de padrões são utilizados no início do processo: *i)* padrões unitários que geram apenas um item (cada padrão é associado a um tipo de item); *ii)* padrões unitários que geram o maior numero possível de itens do mesmo tipo dentro do comprimento do objeto cortado (cada padrão é associado a um tipo de item); e *iii)* padrões que produzem exatamente dois itens distintos (cada padrão é associado a um par de tipos de itens), obtidos a partir de heurísticas construtivas que buscam combinações viáveis de comprimentos. Embora sejam pouco eficientes em termos de aproveitamento do material, os padrões unitários asseguram que sempre haverá uma solução factível, já que qualquer demanda individual pode ser atendida.

A relaxação linear do modelo matemático é então resolvida pelo método de geração de colunas descrito na seção 2.3. Em cada iteração do método, o subproblema busca identificar novos padrões de corte com custo reduzido negativo. Enquanto tais padrões existirem, eles são incorporados ao problema mestre, melhorando progressivamente a solução. O processo se encerra quando não há mais colunas atrativas a serem adicionadas, isto é, quando não se encontram padrões com custo reduzido negativo. Nesse ponto, considera-se que o algoritmo atingiu a solução ótima para a relaxação do modelo.

Após a obtenção da solução ótima contínua, é aplicada a heurística *Kernel Search* considerando todos os padrões de corte produzidos pelo método de geração de colunas. O *Kernel* Inicial é formado pela combinação dos padrões de corte presentes na solução obtida na etapa anterior, acrescidos de padrões identidade, cuja inclusão é necessária para garantir a factibilidade da solução. Os demais padrões de corte são ordenados de acordo com a perda associada a cada um, em ordem crescente, e organizados em subconjuntos (*buckets*) de, no máximo, 15 padrões. Essa separação é feita para todos os tipos de padrões de corte, associados a objetos padronizados (com e sem geração de retalhos) e retalhos.

Na sequência, inicia-se a fase de Extensão, cujo objetivo é identificar a melhor solução inteira possível a partir da exploração dos padrões de corte disponíveis. Para cada *bucket*, o modelo 2.1-2.6 é resolvido considerando variáveis de decisão inteiras e utilizando, simultaneamente, os padrões do *Kernel* e do *bucket* em análise. Se a solução obtida com essa combinação for melhor de que a solução inteira atual, os padrões do *bucket* que compuserem a nova solução passam a ser incorporados ao *Kernel*, e o valor da solução é atualizado. Esse processo é repetido iterativamente até que todos os *buckets* tenham sido processados. Ao final, obtém-se a combinação de padrões de corte que resulta na melhor solução inteira possível dentro da abordagem proposta.

A implementação e resolução do modelo 2.1 a 2.6, do método de geração de colunas e da heurística *Kernel Search* foi realizada com a linguagem de programação Python e o solver Gurobi, versão 11.0.1. O Gurobi é um dos softwares de programação matemática mais avançados disponíveis atualmente, sendo amplamente utilizado em problemas de otimização linear, inteira e mista-inteira. Entre suas principais vantagens destacam-se a alta eficiência na busca por soluções ótimas, a escalabilidade para lidar com problemas de grande porte (com elevado número de variáveis e restrições) e o suporte a recursos avançados, como análise de sensibilidade e exploração de múltiplas soluções viáveis.

5 Resultados

Neste capítulo são apresentados e analisados os resultados obtidos a partir dos experimentos computacionais, realizados com a criação de 10 bases de dados para validação, que simulam diferentes configurações possíveis para o mês de maio de 2025. Essas bases, denominadas V1 a V10, representam variações das *features* como extensão do conjunto de dados descrito no Capítulo 3.

A análise comparativa concentra-se no desempenho do modelo LSTM, treinado conforme a metodologia descrita na Seção 4.1, em relação a duas abordagens tradicionais: a previsão baseada no último pedido do cliente, que pressupõe estabilidade da demanda sem variação entre pedidos consecutivos, e o cenário sem previsão de demanda, no qual a produção ocorre exclusivamente sob encomenda.

5.1 Previsão de Demanda

A avaliação comparativa entre o modelo LSTM e o método do último pedido foi conduzida mediante 10 cenários de validação distintos. As métricas de MAE e MAPE foram empregadas para quantificar o desempenho preditivo de ambas as abordagens ao longo do horizonte de planejamento.

Os resultados consolidados das simulações encontram-se sumarizados na tabela 3.

Tabela 3 – Comparação de Desempenho de Predição de Demanda

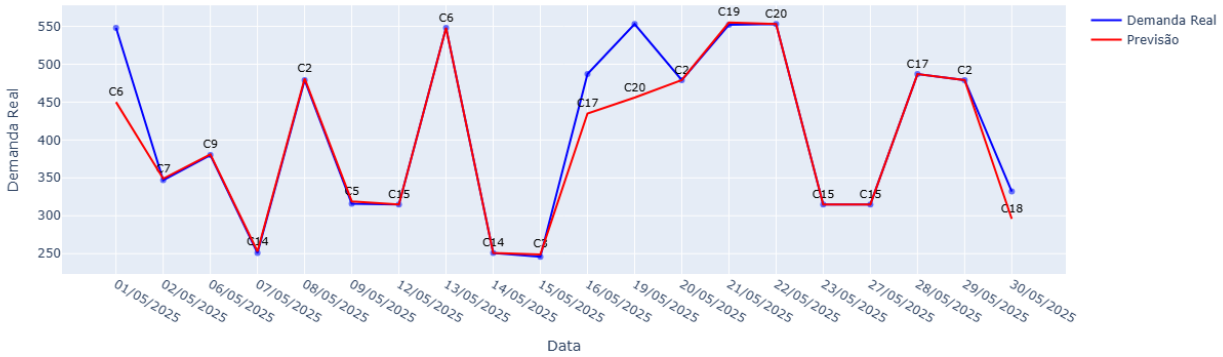
	LSTM		**Último Pedido**	
Cenário	**MAE**	**MAPE (%)**	**MAE**	**MAPE (%)**
V1	6,76	1,79	8,20	2,01
V2	6,88	1,80	9,29	2,24
V3	6,87	1,83	8,54	2,02
V4	6,96	1,83	8,51	2,04
V5	6,81	1,82	9,10	2,20
V6	6,74	1,72	9,35	2,27
V7	6,91	1,85	8,96	2,20
V8	6,79	1,79	8,83	2,13
V9	6,81	1,77	8,61	2,06
V10	6,82	1,80	9,05	2,13
Média	**6,83**	**1,80**	**8,84**	**2,13**

Fonte: Elaborada pelo autor

Os resultados demonstram superioridade consistente do modelo LSTM em todos os cenários de validação. A abordagem baseada em aprendizado de máquina registrou MAE médio de 6,83 e MAPE médio de 1,80%, representando reduções de aproximadamente 22,7% e 15,5%, respectivamente, em comparação ao método do último pedido (MAE: 8,84; MAPE: 2,13%).

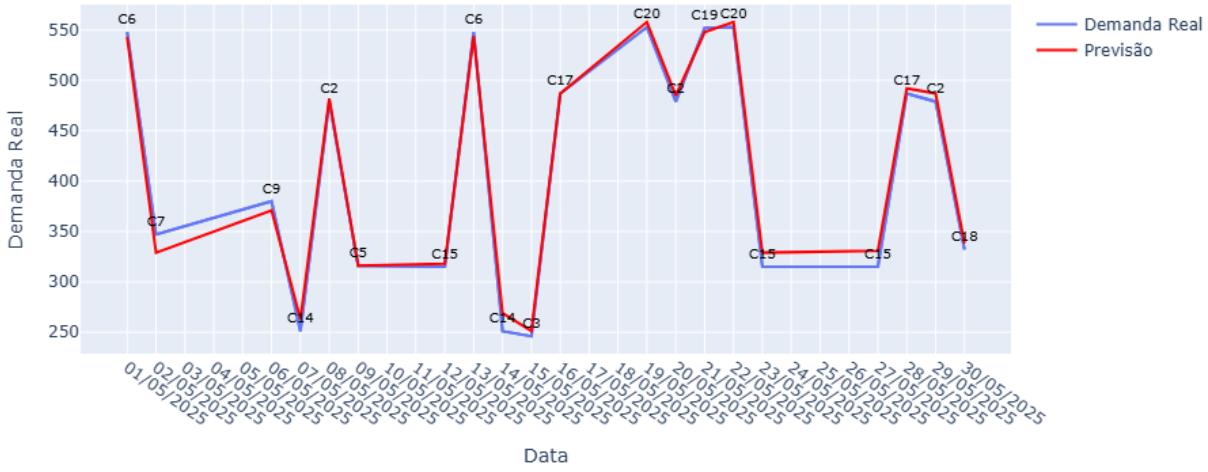
Esta diferença de desempenho pode ser visualizada nas figuras 11 e 12, que ilustram o comportamento preditivo de ambas as metodologias para o item G10 ao longo do horizonte de planejamento.

Figura 11 – Previsão Útilmo Pedido - Item G10



Fonte: Elaborada pelo autor

Figura 12 – Previsão Modelo LSTM - Item G10



Fonte: Elaborada pelo autor

A superioridade do modelo LSTM torna-se particularmente evidente em cenários de elevada volatilidade. Um exemplo ilustrativo é o caso do item G10 para o cliente C6 na data de 01/05/2025, no qual se destacam de forma clara as limitações da abordagem baseada no último pedido em capturar os efeitos de variações macroeconômicas significativas.

O último pedido do cliente C6 referente ao item G10 foi registrado em 09/01/2025, estabelecendo um intervalo de aproximadamente três meses e meio até a data em análise. Nesse período, as variáveis macroeconômicas apresentaram trajetória consistente de crescimento, como evidenciado nas representações gráficas da Seção 3 para cada *feature*, refletindo uma dinâmica expansionista que exerceu influência direta sobre a demanda e gerou condições propícias ao aumento do consumo.

A intensidade da atividade comercial nesse intervalo é evidenciada pelo registro de aproximadamente 90 pedidos de outros itens realizados pelo cliente C6, de modo que, mesmo na ausência de solicitações específicas para o item G10, a demanda foi continuamente atualizada em segundo plano, assegurando a consistência do perfil de consumo e a coerência de seu comportamento de compra.

Entre janeiro e abril, as condições macroeconômicas favoráveis impulsionaram incrementos sucessivos na variação da demanda, em conformidade com o modelo expresso pela Equação 3.1, resultando em crescimento acumulado substancial. Em maio, embora tenha sido observada uma retração em relação ao pico anterior, o nível de consumo permaneceu consideravelmente superior ao registrado no último pedido de janeiro, o que evidencia o efeito cumulativo das condições macroeconômicas positivas dos meses anteriores sobre a dinâmica da demanda.

Essa evolução revela as fragilidades inerentes ao método do último pedido, que desconsidera flutuações intermediárias e ignora a natureza não-linear das variáveis explicativas, conduzindo a previsões sistematicamente inferiores à demanda real. Em contrapartida, o modelo LSTM é capaz de identificar padrões temporais e capturar interações complexas entre variáveis macroeconômicas e demanda, ajustando suas previsões de acordo com as transformações do ambiente econômico.

5.2 Soluções para o PCESA

Nesta etapa, integrou-se a previsão de demanda ao solucionador PCESA, previamente abordado neste projeto. Ressalta-se que a demanda considerada pelo PCESA corresponde exatamente à demanda prevista para cada um dos três depósitos. Dessa forma, cada depósito produz antecipadamente os itens sob sua responsabilidade, caracterizando uma simulação de produção proativa com o objetivo é antecipar o pedido dos clientes.

Tanto para a abordagem baseada no último pedido quanto para o modelo LSTM, foram estabelecidas premissas para lidar com os erros inerentes aos processos preditivos, uma vez que as previsões não alcançam 100% de acurácia. Essas premissas resultam em dois cenários distintos:

- Demanda Prevista $>$ Demanda Real: Quando a previsão supera a demanda efetiva, ocorre a produção de itens excedentes. Esses itens não são descartados, mas armazenados como

estoque disponível. Assim, ao surgir um novo pedido no depósito, o consumo é atendido prioritariamente por esse estoque, com a devida baixa do item armazenado.

- Demanda Prevista $<$ Demanda Real: Se a demanda prevista é inferior à demanda real, a diferença é caracterizada como *Demanda Residual*. Neste caso, considerando o pressuposto de que a produção do pedido é antecipada e a demanda dos clientes deve ser totalmente atendida, a produção da demanda residual é realizada no dia seguinte.

Para validação do modelo, foram realizados testes com 10 versões distintas do mês de maio, onde para cada variação mensal foram sorteados aleatoriamente os tamanhos dos itens conforme as categorias estabelecidas: Pequenos [50, 200], Médios [200, 500] e Grandes [500, 750]. Utilizou-se um objeto padrão de 1000 cm, permitindo gerar e aproveitar retalhos de 400, 500 e 600 cm, com capacidade de Armazenamento de Retalhos de 20 unidades. A simulação abrangeu todo o horizonte de produção, e os resultados consolidados são apresentados na tabela 4.

Tabela 4 – Comparação de Desempenho no Problema de Corte de Estoques

Cenário	Sem Predição			Modelo LSTM			Último Pedido		
	Total Cortado	Perda Abs.	Perda Rel. (%)	Total Cortado	Perda Abs.	Perda Rel. (%)	Total Cortado	Perda Abs.	Perda Rel. (%)
V1	94531411	12340289	13,06	94702076	12361324	13,00	94472300	12241300	12,78
V2	98925915	11357085	11,44	99034786	11245914	11,25	98895891	11271309	11,27
V3	95779605	8235695	9,95	95881876	8125524	9,82	95796441	8089359	9,86
V4	92126765	8243035	10,43	92307782	8205518	10,35	91968811	8194189	10,29
V5	93893496	9344804	10,04	94056215	9264085	9,89	93947850	9194550	9,79
V6	95035689	12941711	13,68	95183778	12923322	13,61	95054390	12778510	13,44
V7	88539597	7610303	9,23	88646290	7513810	9,13	88464549	7562051	9,11
V8	87702067	9350833	9,86	87827575	9305025	9,77	87706115	9334585	9,87
V9	87951429	7199971	8,37	88098195	7206405	8,35	87982121	7086179	8,16
V10	91027402	7780098	8,44	91119802	7724998	8,35	91008803	7711497	8,39
Total	925513376	94403824	10,20	926858375	93875925	10,13	925297271	93463529	10,10

Fonte: Elaborada pelo autor

A tabela 4 apresenta os resultados referentes ao tamanho total cortado - incluindo tanto objetos padronizados quanto retalhos - ao longo de todo o horizonte de planejamento. Adicionalmente, são mostradas as perdas absolutas e relativas para cada cenário. Observa-se que as perdas relativas são bastante próximas entre as três abordagens avaliadas, com variações mínimas. A abordagem baseada no último pedido obteve a menor perda relativa média (10,10%), representando uma redução de 0,03% em comparação com o modelo LSTM (10,13%) e de 0,10% em relação à abordagem sem predição (10,20%).

A similaridade nos valores de perda pode ser atribuída ao fato de que tanto o modelo LSTM quanto a abordagem do último pedido geram demandas muito próximas da demanda real (cenário sem predição). Cabe destacar que as variações observadas decorrem das diferenças nos itens gerados e utilizados, bem como do tratamento dado à demanda residual e ao estoque disponível.

Essas diferenças geram problemas de corte distintos em cada cenário, ainda que similares ao problema original, resultando em perdas próximas, porém não idênticas.

6 Conclusão

Neste trabalho foram abordadas técnicas de inteligência artificial para previsão de demanda no problema de corte de estoque com sobras aproveitáveis. O estudo teve por objetivo aprimorar o planejamento produtivo minimizando perdas ao longo do processo, tornando-o mais sustentável. De modo geral, os objetivos centrais do trabalho foram alcançados ao estruturar uma base de dados, desenvolver e treinar algoritmos de aprendizado de máquina baseados no modelo *Long Short-Term Memory* (LSTM), e implementar com sucesso o problema de corte unidimensional com sobras aproveitáveis (PCESA), visando a obtenção da solução ótima a partir do modelo proposto por Arenales et al. (2015). Adicionalmente, foi incorporada a metaheurística *Kernel Search*, que possibilitou a geração de soluções inteiras para as variáveis de decisão do PCESA, conferindo maior aderência à realidade prática e viabilizando uma validação consistente.

Os resultados obtidos a partir da comparação entre a abordagem baseada no Último Pedido e o modelo LSTM demonstraram a superioridade do LSTM em termos de proximidade com a demanda real. Sua capacidade de capturar padrões temporais mais complexos e adaptar-se a variações no comportamento da demanda evidencia vantagens estratégicas relevantes para o planejamento de médio e longo prazo, em consonância com as discussões apresentadas por Chopra e Meindl (2019).

No âmbito do problema de cortes com sobras aproveitáveis, embora o modelo LSTM tenha apresentado uma perda relativa marginalmente superior à obtida pela abordagem do Último Pedido, sua utilização justifica-se pelo fato de oferecer previsões mais alinhadas à demanda futura dos clientes. Essa característica potencializa ganhos estratégicos em planejamento e alocação de recursos, uma vez que pequenas diferenças na acurácia preditiva podem se traduzir em impactos significativos na eficiência operacional, sobretudo em problemas de corte, em que cada variação de demanda gera novos padrões de otimização.

Este estudo contribuiu de maneira relevante para a área de otimização, ao propor uma integração pouco explorada entre técnicas de aprendizado de máquina e a solução do PCESA na literatura. Ressalta-se, contudo, que algumas premissas foram necessárias para mitigar os efeitos dos erros de previsão, o que abre espaço para investigações futuras voltadas ao aprimoramento dessas estratégias. Nesse contexto, pesquisas futuras podem explorar arquiteturas alternativas de modelos de aprendizado de máquina, com o objetivo de aumentar a acurácia das previsões de demanda e fortalecer a integração entre modelos preditivos e métodos de otimização. Além disso, destaca-se o potencial de expandir esta abordagem para outras operações na área de otimização, como a incorporação do problema de cortes em múltiplos períodos, possibilitando uma solução

integrada e ainda mais aderente à dinâmica real dos processos produtivos.

Referências

AAMER, A. M. Outsourcing in non-developed supplier markets: a lean thinking approach. *International Journal of Production Research*, Taylor & Francis, v. 56, n. 18, p. 6048–6065, 2018.

ABUABARA, A.; MORABITO, R. Cutting optimization of structural tubes to build agricultural light aircrafts. *Annals of Operations Research*, Springer, v. 169, n. 1, p. 149–165, 2009.

AMIN, S. *Understanding LSTMs: LSTM Implementation from Scratch*. 2018. Acesso em: 19 dez. 2024. Disponível em: <https://medium.com/@samina.amin/understanding-lstms-lstm-implementation-from-scratch-18965a150eca>.

ANGELELLI, E.; MANSINI, R.; SPERANZA, M. Kernel search: a new heuristic framework for portfolio selection. *Computational Optimization and Applications*, v. 51, p. 345–361, 2012.

ARENALES, M. N.; CHERRI, A. C.; NASCIMENTO, D. N. d.; VIANNA, A. A new mathematical model for the cutting stock/leftover problem. *Pesquisa Operacional*, SciELO Brasil, v. 35, n. 3, p. 509–522, 2015.

CHERRI, A.; ARENALES, M.; YANASSE, H. The one-dimensional cutting stock problems with usable leftover: A heuristic approach. *European Journal of Operational Research*, v. 196, p. 897–908, 2009.

CHERRI, A.; ARENALES, M.; YANASSE, H.; POLDI, K.; VIANNA, A. The one-dimensional cutting stock problem with usable leftovers - a survey. *European Journal of Operational Research*, v. 236, p. 395–402, 2014.

CHERRI, A. C.; CHERRI, L. H.; OLIVEIRA, B. B.; OLIVEIRA, J. F.; CARRAVILLA, M. A. A stochastic programming approach to the cutting stock problem with usable leftovers. *European Journal of Operational Research*, Elsevier, v. 308, n. 1, p. 38–53, 2023.

CHOPRA, S.; MEINDL, P. Supply chain management. strategy, planning & operation. In: *Das Summa Summarum des Management: Die 25 wichtigsten Werke für Strategie, Führung und Veränderung*. [S.l.]: Springer, 2019. p. 265–275.

COELHO, K.; CHERRI, A.; BAPTISTA, E.; JABOOUR, C.; SOLER, E. Sustainable operations: The cutting stock problem with usable leftovers from a sustainable perspective. *Journal of Cleaner Production*, v. 167, p. 545–552, 2017.

CUI, Y.; YANG, Y. A heuristic for the one-dimensional cutting stock problem with usable leftover. *European Journal of Operational Research*, v. 204, p. 245–250, 2010.

FILIPPI, C.; GUASTAROBA, G.; HUERTA-MUñOZ, D.; SPERANZA, M. A kernel search heuristic for a fair facility location problem. *Computers and Operations Research*, v. 132, 2021.

FISHER, M. L. What is the right supply chain for your product? *Harvard business review*, v. 75, p. 105–117, 1997.

GILMORE, P.; GOMORY, R. A linear programming approach to the cutting stock problem - part ii. *Operations Research*, v. 11, p. 863–888, 1963.

GLOROT, X.; BORDES, A.; BENGIO, Y. Deep sparse rectifier neural networks. In: JMLR WORKSHOP AND CONFERENCE PROCEEDINGS. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. [S.l.], 2011. p. 315–323.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. v. 1.

GRADIŠAR, M.; JESENKO, J.; RESINOVIČ, G. Optimization of roll cutting in clothing industry. *Computers & Operations Research*, Elsevier, v. 24, n. 10, p. 945–953, 1997.

HASTIE, T. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer, 2009.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT press, v. 9, n. 8, p. 1735–1780, 1997.

HOFER, C.; EROGLU, C.; HOFER, A. R. The effect of lean production on financial performance: The mediating role of inventory leanness. *International Journal of Production Economics*, Elsevier, v. 138, n. 2, p. 242–253, 2012.

HU, H.; GUO, S.; ZHEN, L.; WANG, S.; BIAN, Y. A multi-product and multi-period supply chain network design problem with price-sensitive demand and incremental quantity discount. *Expert Systems with Applications*, Elsevier, v. 238, p. 122005, 2024.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International journal of forecasting*, Elsevier, v. 22, n. 4, p. 679–688, 2006.

Instituto Aço Brasil. *Estatísticas Mensais do Aço*. 2025. Acesso em: 28 set. 2025. Disponível em: <https://www.acobrasil.org.br/site/estatistica-mensal>.

Instituto Brasileiro de Geografia e Estatística (IBGE). *Índice Nacional de Preços ao Consumidor Amplo (IPCA) - Séries Históricas*. 2025. Acesso em: 28 set. 2025. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indice-nacional-de-precos-ao-consumidor-amplo.html>.

Investing.com. *Steel Rebar Futures Historical Data*. 2025. Acesso em: 28 set. 2025. Disponível em: <https://www.investing.com/commodities/steel-rebar-historical-data>.

Investing.com. *USD/BRL Historical Data*. 2025. Acesso em: 28 set. 2025. Disponível em: <https://www.investing.com/currencies/usd-brl>.

JEBLE, S.; KUMARI, S.; PATIL, Y. Role of big data in decision making. *Operations and Supply Chain Management: An International Journal*, v. 11, n. 1, p. 36–44, 2017.

KHAN, R.; PRUNCU, C.; NAEEM, K.; ABAS, M.; QAZI, S.; KHALID, Q. A mathematical model for reduction of trim loss in cutting reels at a make-to-order paper mill. *Applied Sciences*, v. 10, p. 1–18, 2020.

MEHMANPAZIR, F.; KHALILI-DAMGHANI, K.; HAFEZALKOTOB, A. Modeling steel supply and demand functions using logarithmic multiple regression analysis (case study: Steel industry in iran). *Resources Policy*, Elsevier, v. 63, p. 101409, 2019.

MUNCHARAZ, J. O. Comparing classic time series models and the lstm recurrent neural network: An application to s&p 500 stocks. *Finance, Markets and Valuation*, v. 6, n. 2, p. 137–148, 2020.

NORRMAN, A.; NASLUND, D. Supply chain incentive alignment: The gap between perceived importance and actual practice. *Operations and Supply Chain Management: An International Journal*, v. 12, n. 3, p. 129–142, 2019.

PRECHELT, L. Early stopping-but when? In: *Neural Networks: Tricks of the trade*. [S.l.]: Springer, 2002. p. 55–69.

RAVELO, S.; MENESES, C.; SANTOS, M. Meta-heuristics for the one-dimensional cutting stock problem with usable leftover. *Journal of Heuristics*, v. 26, n. 4, p. 585–618, 2020.

SCHEITHAUER, G. A note on handling residual length. *Optimization*, v. 22, p. 461–466, 1991.

SENERGUES, V.; BRAHIMI, N.; CHERRI, A.; KLEIN, F.; PÉTON, O. Cutting stock problem with usable leftovers: A review. *European Journal of Operational Research*, Elsevier, 2025.

SILVER, E. A.; PYKE, D. F.; PETERSON, R. *Inventory management and production planning and scheduling*. [S.l.]: Wiley New York, 1998. v. 3.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, v. 27, 2014.

WEMANS, A. M. C. d. S. *A heuristic approach for a multi-period capacitated single-allocation hub location problem*. Dissertação (Mestrado), 2016.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, v. 30, n. 1, p. 79–82, 2005.

YANI, L. P. E.; PRIYATNA, I. M. A.; AAMER, A. Exploring machine learning applications in supply chain management. In: *9th international conference on operations and supply chain management*. [S.l.: s.n.], 2019. p. 161–169.

ZAZO, R.; NIDADAVOLU, P. S.; CHEN, N.; GONZALEZ-RODRIGUEZ, J.; DEHAK, N. Age estimation in short speech utterances based on lstm recurrent neural networks. *IEEE Access*, IEEE, v. 6, p. 22524–22530, 2018.

ZEN, H.; SAK, H. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In: IEEE. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2015. p. 4470–4474.

ZHA, W.; LIU, Y.; WAN, Y.; LUO, R.; LI, D.; YANG, S.; XU, Y. Forecasting monthly gas field production based on the cnn-lstm model. *Energy*, Elsevier, v. 260, p. 124889, 2022.