Journal of Production, Operations and Systems Management

e-ISSN: 1984-2430

Open Access

ALGORITMOS DE APRENDIZADO DE MÁQUINA APLICADOS NA ESTIMAÇÃO DE DESCONTOS EM ITENS DA CESTA BÁSICA

MACHINE LEARNING ALGORITHMS APPLIED TO THE ESTIMATION OF DISCOUNTS IN BASIC FOOD BASKET ITEMS

Jose Mateus Rodrigues dos SANTOS[1]
e-mail: jose.2000@alunos.utfpr.edu.br

Bruno Samways dos SANTOS[2]
e-mail: brunosantos@utfpr.edu.br

Como referenciar este artigo:

Santos, J. M. R., & Santos, B. S. Algoritmos de aprendizado de máquina aplicados na estimação de descontos em itens da cesta básica. *Revista GEPROS*, *20*, e025006. DOI: 10.15675/gepros.3037

| **Submetido em**: 24/10/2024
| **Aprovado em**: 18/09/2025
| **Publicado em**: 13/11/2025

Editora: Profa. Dra. Paula de Camargo Fiorini

turnitin
ARTIGO SUBMETIDO AO SISTEMA DE SIMILARIDADE

[1]Universidade Tecnológica Federal do Paraná (UFTPR), Londrina – Paraná (PR) – Brasil. Graduando em Engenharia de Produção.
[2]Universidade Tecnológica Federal do Paraná (UFTPR), Londrina – Paraná (PR) – Brasil. Professor Adjunto do Departamento de Engenharia de Produção. Pesquisador do Grupo de Pesquisa em Otimização e Mineração de Dados (GPOMD).

(cc) BY-NC-SA

RESUMO

Objetivo: O objetivo deste trabalho foi avaliar modelos de aprendizado de máquina (ou *machine learning*, ML) para estimar descontos ofertados em licitações de itens da cesta básica nacional, realizando posteriormente uma análise sobre os atributos mais relevantes dentro destes estimadores. **Referencial Teórico**: Os modelos de ML podem ser aplicados em vários contextos, incluindo áreas da engenharia, medicina, saúde pública, economia, entre outros. Estes algoritmos podem encontrar padrões ocultos que técnicas tradicionais da estatística podem ser incapazes de detectar, tornando-os propício para análises preditivas e interpretativas quanto às características dos descontos ofertados em licitação de itens da cesta básica. **Metodologia/Abordagem**: Foram utilizados os algoritmos Random Forest, XGBoost e Redes Neurais Artificiais, e avaliadas pelo erro absoluto médio e pela raiz do erro quadrático médio. Os dados foram obtidos junto ao Tribunal de Contas do Estado do Paraná e incluíram 18 itens da cesta básica. **Resultados**: Em geral, o modelo XGBoost apresentou os melhores resultados quanto aos erros para os itens avaliados na cesta básica. Já para os atributos mais relevantes, entendeu-se que a variável "quantidade" dos itens "pão" e "manteiga" foram as mais importantes para se estimar o desconto, enquanto o "ano de homologação" do pedido foi importante para estimar o desconto do "óleo de soja", "pão francês", "carne bovina", "arroz" e "feijão". **Contribuições, implicações práticas e sociais**: Estes modelos podem fornecer informações relevantes para gestores e órgãos fiscalizadores, auxiliando no planejamento orçamentário, detecção de fraudes e negociação de preços em licitações, contribuindo para uma gestão pública mais econômica e transparente. Porém, verificou-se que os modelos foram limitados para estimar o valor do desconto em situações de grandes picos de variação. **Originalidade/Valor**: A partir da etapa do pré-processamento, sugere-se uma análise profunda quanto ao sistema que disponibiliza os dados, pois os arquivos estão subdivididos e não padronizados, fazendo com que pessoas fora da academia ou do mundo corporativo tenham dificuldades em extraí-los. Os modelos de ML podem auxiliar gestores públicos, no atendimento aos princípios e tendências trazidos pela Nova Lei de Licitações, em um mercado que representa 12% do PIB nacional.

PALAVRAS-CHAVE*:* Aprendizado de máquina. Administração Pública. Licitação. Cesta Básica. Compras Públicas.

(cc) BY-NC-SA

ABSTRACT

Purpose: This study evaluates machine learning (ML) models for estimating discounts in public procurement of national basic food basket items and analyzes the most relevant features influencing these estimations. **Theoretical framework**: ML models can be applied in various fields, including engineering, medicine, public health, and economics. These algorithms are capable of uncovering hidden patterns that traditional statistical techniques might miss, making them suitable for predictive and interpretative analyses regarding the characteristics of discounts offered in the procurement of basic food basket items. **Methodology/Approach**: The algorithms Random Forest, XGBoost, and Artificial Neural Networks were employed and evaluated using mean absolute error and root mean square error. The data were obtained from the Court of Accounts of the State of Paraná and included 18 items from the basic food basket. **Findings**: Overall, the XGBoost model exhibited the best performance based on error metrics. Regarding feature importance, the 'quantity' variable was most significant for estimating discounts on bread and butter, whereas the 'year of approval' was a key factor for soybean oil, French bread, beef, rice, and beans. **Research, practical & social implications**: These models can provide valuable information to managers and oversight bodies, supporting budget planning, fraud detection, and price negotiation in public procurement, thereby contributing to more cost-effective and transparent public administration. However, the models showed limitations in estimating discount values in situations of large variation peaks. **Originality/Value**: From the preprocessing stage, we suggest a deeper analysis of the data management system, as the fragmented and non-standardized files make data extraction difficult for individuals outside of academia or the corporate sector. ML models can assist public managers in aligning with the principles and trends introduced by the New Public Procurement Law, in a market that accounts for 12% of the national GDP.

KEYWORDS: Machine Learning. Public Administration. Public Procurement. Basic Food Basket. Public Purchasing.

(cc) BY-NC-SA

Introdução

O mercado de compras públicas no Brasil representou, em média, 12,5% do Produto Interno Bruto (PIB) do país entre os anos de 2006 e 2017 (Ribeiro & Inácio Júnior, 2019). Em regra, todos os processos de compra realizados pela administração pública, autarquias ou fundações governamentais devem seguir as normas de licitação e contratos, estabelecidas atualmente pela Lei n.º 14.133/21.

O processo licitatório tem por objetivo a seleção da proposta mais vantajosa ao governo, garantindo que todos os interessados possam participar de forma isonômica, desde que atendam a determinados critérios predefinidos pelo órgão (Mello, 2015). Existem diversas modalidades de licitação definidas na legislação atual; porém, para a aquisição de bens e serviços comuns, faz-se uso da modalidade pregão, preferencialmente em sua forma eletrônica (Freitas et al., 2021).

Para a definição do valor máximo admissível para a compra, os valores obtidos na pesquisa de mercado devem ser analisados de forma crítica, desconsiderando valores inexequíveis ou excessivamente elevados. O valor final é definido via simples média, mediana ou pelo menor dos valores obtidos na pesquisa, que podem ser realizados diretamente com fornecedores e sites especializados ou por meio do histórico de contratações similares na esfera pública. Durante o trâmite licitatório, os licitantes apresentam propostas, competindo entre si em lances consecutivos por um determinado período. Concluída a fase de lances, a melhor proposta que atenda aos requisitos é declarada vencedora (Amorim, 2017). Portanto, o valor final para a contratação pode ser diferente do valor inicialmente definido, variando de acordo com os lances dados pelos fornecedores.

Por se tratar de dados governamentais, as informações dos itens licitados estão disponíveis a todos os cidadãos. O Tribunal de Contas do Estado do Paraná (TCE-PR) mantém em seu banco de dados todas as informações das licitações de entidades da administração pública dos municípios do Paraná, disponibilizando-as ao público por meio do “Portal de Informações para Todos”. Técnicas de mineração de dados e aprendizado de máquina podem ser aplicadas a essas bases de dados por meio do processo de descoberta de conhecimento em bancos de dados, com o objetivo de encontrar padrões e extrair informações úteis (Ghazal & Hammad, 2022).

O aprendizado de máquina é um subcampo da inteligência artificial que tem por característica o desenvolvimento de algoritmos de autoaprendizado que obtêm conhecimento a

(cc) BY-NC-SA

partir de dados para realizar predições, identificando as regras e relações entre variáveis e melhorando, de forma gradativa, o desempenho do modelo estimador (Géron, 2019).

De acordo com Oliveira, Rêgo e Diniz (2019), o valor de referência de uma licitação está sujeito a vieses ou até mesmo a conluio entre os agentes públicos e possíveis interessados, o que pode fazer com que o valor se afaste do real. Assim, a partir da estimação do desconto oferecido para o item, pode-se avaliar a qualidade do valor de referência, já que, considerando que o valor seja compatível com o mercado, o percentual de desconto deve ser similar ao estimado com base nos dados de contratações anteriores.

França (2021) destaca que, quanto maior a quantidade comprada de determinado item, menor será o custo de aquisição, o que se aplica também ao campo das licitações. Por meio da análise do percentual de descontos oferecidos para itens similares em compras de quantidades diferentes, pode ser possível observar essa diferença do ganho de escala, principalmente em processos de compra de entidades de menor porte.

Nesse contexto, o presente artigo buscou avaliar modelos de aprendizado de máquina que estimem o valor de desconto ofertado pela proposta vencedora em licitações de itens da cesta básica nacional, realizadas por entidades da administração pública municipal do estado no Paraná. Para tal, foram utilizadas três técnicas de *machine learning* (ML): Random Forest, XGBoost e Redes Neurais Artificiais, além de análises de contribuição de atributos na estimação de 18 itens da cesta básica.

Compõem este estudo, além desta introdução, o referencial teórico, que aborda de forma resumida conceitos como mineração de dados, técnicas de regressão e processos licitatórios. Na sequência, é apresentada a metodologia aplicada, bem como o conjunto de dados e as etapas da pesquisa. Em seguida, são apresentados os resultados obtidos e as discussões e, por fim, as considerações finais.

Fundamentação Teórica

Este capítulo aborda os conceitos e fases da Descoberta de Conhecimento em Bases de Dados, Mineração de Dados, a aprendizagem supervisionada, as técnicas para estimação (ou regressão), métricas de avaliação e como funciona o processo licitatório, foco da aplicação da pesquisa.

(cc) BY-NC-SA

Descoberta de conhecimento em banco de dados

O processo de descoberta de conhecimento em bancos de dados, do inglês *Knowledge Discovery in Databases* (KDD), tem como objetivo encontrar padrões nos dados e apresentá-los de tal modo que sua absorção como conhecimento seja facilitada (Lara et al., 2014). Ghazal e Hammad (2022) definem o KDD como "o processo não trivial de identificar padrões válidos, novos, potencialmente úteis e, em última instância, compreensíveis em dados".

O processo de KDD é executado em uma sequência de sete passos. Na primeira etapa, realiza-se a limpeza dos dados, visando remover ruídos e inconsistências. Em seguida, os dados, que podem ser provenientes de diferentes fontes, devem ser integrados e selecionados de acordo com a necessidade da análise. O próximo passo é a transformação, na qual erros, ruídos e dados faltantes são tratados, tornando-os adequados à mineração por meio de técnicas de resumo ou agregação. Posteriormente, os dados estão prontos para serem minerados, utilizando-se métodos para identificar padrões. Os padrões encontrados na etapa de mineração devem ser avaliados, a fim de verificar se representam, de fato, conhecimento útil. Por fim, devem ser utilizadas técnicas de visualização e representação do conhecimento para apresentar o conhecimento extraído aos interessados.

Cada uma dessas etapas tem sua importância e contribuição para o resultado, que é a assimilação do conhecimento a partir dos dados. As técnicas de mineração de dados podem incluir algoritmos de aprendizagem de máquina, e estes podem ter diferentes tipos, que serão descritos na próxima subseção.

Mineração de dados e aprendizado de máquina

A etapa da mineração dos dados foi definida por Tan et al. (2019) como a exploração de conjuntos de dados, visando encontrar correlações importantes entre variáveis contidas no conjunto, que é frequentemente extenso, e fazer um resumo dos dados de forma a facilitar a compreensão dessas relações e o acesso aos dados úteis. Esse processo inclui a aplicação de algoritmos aos dados, objetivando a extração de conhecimentos novos e úteis (Aggarwal, 2015).

Criar modelos e entender as regras de uma base de dados extensa de forma manual pode ser muito custoso. Para tal problema, o ML, ou aprendizado de máquina, pode ser uma alternativa. Segundo Raschka (2015), o desenvolvimento de algoritmos que "aprendem" por si mesmos tem contribuído para aumentar a performance e a capacidade de processamento dos dados. Um dos principais objetivos das pesquisas em aprendizado de máquina é o reconhecimento de padrões complexos com o intuito de tomar decisões inteligentes com base

(cc) BY-NC-SA

nos dados analisados.

O ML é organizado em vários grupos, de acordo com o objetivo do algoritmo; porém, os principais grupos são os algoritmos supervisionados e os não supervisionados. Os algoritmos supervisionados obtêm conhecimento a partir de conjuntos de dados de treinamento que possuem combinações de dados de entrada e saída em pares ordenados. A partir do aprendizado pelo conjunto de treinamento, o algoritmo procura prever as saídas para cada uma das entradas de um conjunto de teste, onde sua acurácia é avaliada (Radhoush et al., 2023).

Caso os valores que se espera prever sejam valores numéricos contínuos, a tarefa é chamada de regressão; porém, se existe apenas um número discreto de possibilidades ou categorias, então a tarefa é chamada de classificação. Alguns exemplos da aplicação dessas tarefas de aprendizagem supervisionada são: detecção de *spam*; detecção de padrões; processamento de linguagem natural; e análises preditivas baseadas em regressão ou classificação (Bonaccorso, 2017). Este estudo fará uso de técnicas de aprendizagem supervisionada, especificamente os algoritmos de regressão, para estimar os valores dos descontos dos itens da cesta básica.

Técnicas de regressão

Para estimar valores numéricos com base em um conjunto de dados de entrada, existem vários modelos no âmbito do aprendizado de máquina. Nesta seção, estão apresentados os algoritmos utilizados neste estudo para a estimação do desconto do item da cesta básica, bem como as métricas de desempenho dos regressores.

Random Forest

Essa técnica pertence à classe de algoritmos *ensemble*. Esse tipo de modelo tem por característica o treinamento e a execução de vários algoritmos, cujos resultados são utilizados para formar uma solução conjunta (Xu & Yin, 2021). Este algoritmo é uma combinação de árvores de decisão, em que cada uma delas depende de valores de um vetor de amostragem aleatório e independente (Breiman, 2001).

Árvores de decisão são uma abordagem supervisionada em que a previsão é definida de acordo com uma estrutura de decisão similar a uma árvore, sendo composta de uma raiz, nós, galhos e folhas. A técnica Random Forest utiliza várias árvores de decisão; porém, as árvores geradas passam por um processo de amostragem por substituição (*bootstrapping*), em que algumas entradas podem ser utilizadas mais de uma vez, enquanto outras podem não ser

(cc) BY-NC-SA

escolhidas, mantendo o tamanho da amostra original. As árvores também recebem uma quantidade de variáveis muito menor do que o conjunto original, escolhidas aleatoriamente (Géron, 2019).

Essas características permitem que o método atue de forma eficiente em bases de dados extensas e com grande dimensionalidade, oferecendo ainda a possibilidade de medir a importância das variáveis para a predição e detectar *outliers*, além de ser um algoritmo com maior acurácia e menor custo computacional quando comparado à implementação de apenas uma árvore de decisão tradicional (Santos et al., 2024).

XGBoost

Outro algoritmo utilizado para a predição de valores por meio do aprendizado de máquina é o XGBoost, sigla para *eXtreme Gradient Boosting*, que possui grande notoriedade no cenário recente. Modelos como o XGBoost são considerados a mais alta tecnologia em tarefas de classificação, amplamente utilizados em competições (Sagi & Rokach, 2021). Com base no conceito de *Gradient Boosting Decision Trees*, essa técnica também utiliza várias árvores de decisão para atingir um resultado com maior acurácia, porém realiza esse processo por um método diferente.

Em contraste com a metodologia utilizada no Random Forest, cada uma das novas árvores recebe os dados com pesos, visando priorizar a acurácia na predição dos erros do passo anterior (Dhaliwal et al., 2018).

Grande parte do sucesso dessa técnica se deve ao fato de sua execução ser mais de dez vezes mais rápida que outros métodos à época de sua criação, além de ser escalável, utilizar menos recursos, lidar de forma eficiente com dados esparsos e ter a capacidade de processar dados que excedam a capacidade computacional tradicional por meio de métodos de computação *out-of-core* (Chen & Guestrin, 2016).

Redes Neurais Artificiais

As Redes Neurais Artificiais foram desenvolvidas inspirando-se no funcionamento biológico dos neurônios (Halužan Vasle & Moškon, 2024), que aprendem por meio da intensidade das conexões entre eles, chamadas conexões sinápticas, em resposta a estímulos externos.

Os neurônios artificiais recebem entradas, realizam o trabalho computacional nos dados e os enviam para outros neurônios, de forma semelhante às redes biológicas. O aprendizado

(cc) BY-NC-SA

ocorre por meio da variação dos pesos transmitidos nessas ligações (sinapses), que são modificados em caso de predição incorreta pelo conjunto anterior de pesos. Logo, a arquitetura dessas redes influencia diretamente os resultados por elas obtidos (Aggarwal, 2018). Uma das arquiteturas mais utilizadas nessa técnica de ML é a *perceptron* multicamadas, composta por três principais camadas: a camada de entrada, a camada oculta e a camada de saída.

Algumas vantagens das redes neurais em relação a outros tipos de abordagens de aprendizado são a sua robustez e tolerância a neurônios deteriorados; a habilidade de aprender com eventos para tomar decisões a partir de eventos similares; a capacidade de lidar com dados inconsistentes e com alta incidência de ruído; e a forma coletiva de execução do trabalho computacional, podendo operar de forma paralela ou distribuída (Mijwil, 2021).

Métricas de avaliação

A fim de avaliar os algoritmos de aprendizado de máquina supervisionado, os resultados previstos de um conjunto de dados de teste são comparados com os dados reais, buscando-se determinar a qualidade da previsão realizada. Para quantificar os erros, algumas métricas específicas podem ser utilizadas para a tarefa de regressão. O *Mean Absolute Error* (MAE), ou erro médio absoluto, e o *Root Mean Squared Error* (RMSE), ou raiz do erro quadrático médio, são algumas das medidas de erro mais conhecidas. Essas métricas apresentam sempre valores não negativos como resultado, tendo como objetivo alcançar valores de erro próximos de zero (Sampaio et al., 2019).

Para o cálculo do MAE, somam-se os valores absolutos de cada um dos erros, a fim de obter o erro total e, com esses dados, calcular o valor médio a partir da quantidade de previsões realizadas. Já no RMSE, a soma é realizada a partir do quadrado de cada valor de erro absoluto, o que torna essa métrica mais sensível a erros maiores e *outliers* (Willmott & Matsuura, 2005). As equações 1 e 2 mostram o cálculo com os parâmetros necessários:

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

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

Onde n é o número de instâncias avaliadas y_i corresponde ao valor real da instância i, e $\hat{y}_i$ é o valor estimado da instância i.

(cc) BY-NC-SA

Compras públicas: Licitações

A aquisição de bens e serviços pelo funcionalismo público possui peculiaridades quando comparada às compras do meio privado, tendo como regra geral a compra por meio de processos licitatórios. A licitação pode ser definida como o procedimento administrativo necessário para permitir a alienação, aquisição, locação, concessão ou contratação de quaisquer bens ou serviços, de acordo com condições definidas pela pessoa governamental, sendo convocados todos os interessados na apresentação de propostas, visando selecionar a proposta mais vantajosa à administração em função de parâmetros preestabelecidos e divulgados (Mello, 2015).

Apesar da obrigatoriedade da licitação, existem duas exceções a essa regra — as chamadas contratações diretas —, casos em que o processo pode ser dispensado ou considerado inexigível, observando-se as regras definidas na legislação (Freitas et al., 2021). Atualmente, as licitações são regidas pela Lei n.º 14.133, de 1º de abril de 2021, que substituiu a Lei n.º 8.666/93. É comum, no meio jurídico, referir-se à Lei 14.133/21 como Nova Lei de Licitações.

De acordo com a Lei n.º 14.133/21, os objetivos do processo licitatório são: garantir a seleção da proposta mais vantajosa; resguardar o direito de tratamento isonômico entre os licitantes; evitar contratações com preço incompatível com o mercado; e incentivar a inovação e o desenvolvimento nacional sustentável.

Para que esses objetivos sejam atingidos, devem ser seguidos também alguns princípios definidos pela legislação brasileira, que estabelecem critérios interpretativos que possibilitam lidar com situações não previstas nas normas (Amorim, 2017). Os princípios citados no artigo 5º da Nova Lei de Licitações são os princípios da legalidade, da impessoalidade, da moralidade, da publicidade, da eficiência, do interesse público, da probidade administrativa, da igualdade, do planejamento, da transparência, da eficácia, da segregação de funções, da motivação, da vinculação ao edital, do julgamento objetivo, da segurança jurídica, da razoabilidade, da competitividade, da proporcionalidade, da celeridade, da economicidade e do desenvolvimento nacional sustentável.

A legislação define que o processo de compra nos órgãos públicos deve iniciar-se com a fundamentada descrição da necessidade de contratação de determinado item, sendo necessário definir as características técnicas e as condições de execução, pagamento e garantias exigidas. Em seguida, para definir o valor máximo a ser pago por determinado item, a administração pública deve voltar-se ao mercado para realizar orçamentos que obedeçam a critérios legais.

A estimativa de preços deve ser realizada de forma crítica, consultando fontes que

(cc) BY-NC-SA

representem de fato o mercado, como: contratações públicas similares, comércio eletrônico especializado e fornecedores do ramo. Para a definição do preço de referência, devem ser desconsiderados os valores inexequíveis e excessivamente elevados, obtendo-se, por meio da média, da mediana ou do menor preço obtido, o valor parâmetro do processo licitatório (Amorim, 2017).

Após a definição do objeto a ser licitado, são estabelecidas as regras da contratação, que serão divulgadas posteriormente por meio de edital, garantindo, assim, que todos os eventuais participantes tenham total ciência das regras e possam ter tratamento isonômico. Ainda antes da divulgação do certame, a legislação determina a realização de uma análise de riscos que possam comprometer o sucesso e a boa execução contratual. Concluídas essas etapas e após a aprovação da assessoria jurídica, o processo é divulgado e realizado de acordo com as regras definidas no instrumento convocatório.

Metodologia

Nesta seção, são descritos o conjunto de dados utilizado para a análise, o fluxo de trabalho aplicado durante o processo de preparação dos dados e as ferramentas utilizadas. Quanto à classificação deste estudo, tem-se por propósito a pesquisa aplicada, de abordagem quantitativa, procedimentos experimentais e utilização da pesquisa analítica como método.

O conjunto de dados

Para atingir o objetivo deste estudo, foram utilizados dados disponibilizados pelo Tribunal de Contas do Estado do Paraná (TCE-PR), por meio do “Portal de Informações para Todos”. Os dados são atualizados semanalmente e contêm informações sobre consumo de combustíveis, contratos e convênios firmados, receitas e despesas do órgão, diárias, licitações e obras de todos os municípios do Estado.

Ao realizar o download, os dados são apresentados em arquivo compactado, que contém vários outros arquivos compactados, sendo um para cada município, ano e tabela. Para a análise, foram utilizados os dados da tabela “LicitacaoVencedor” de todos os municípios entre os anos de 2019 e 2023. Originalmente em formato XML, os dados foram convertidos para “CSV” usando as bibliotecas ElementTree e Pandas, em linguagem Python.

Após a conversão, optou-se por realizar uma complementação nos dados com informações sobre o porte das empresas e dos municípios. O porte das empresas foi extraído do Cadastro Nacional da Pessoa Jurídica (CNPJ), gerenciado pela Secretaria Especial da Receita

(cc) BY-NC-SA

Federal do Brasil (RFB). As informações são atualizadas mensalmente e foram obtidas em 9 de setembro de 2023. Os dados são disponibilizados em formato "CSV". Para incluir o porte do município, foram utilizados dados do Censo Demográfico de 2022, divulgados pelo Instituto Brasileiro de Geografia e Estatística (IBGE). As informações são disponibilizadas em formato "XLS" e contêm dados sobre a população de todos os 5.570 municípios do país, atualizados em 22 de junho de 2023.

A fim de realizar a categorização dos municípios de acordo com a população, foram definidas as seguintes categorias, com base nas características próprias dos municípios do estado do Paraná, que possuem baixa população: Categoria 1 – até 10.000 habitantes; Categoria 2 – entre 10.000 e 30.000 habitantes; Categoria 3 – entre 30.000 e 60.000 habitantes; Categoria 4 – entre 60.000 e 120.000 habitantes; e Categoria 5 – acima de 120.000 habitantes.

Uma coluna adicional, denominada "desconto", foi criada para representar o percentual de desconto oferecido nas propostas, calculado pela razão entre o valor máximo do item e o valor proposto, subtraindo-se 1.

Após essas agregações, o conjunto de dados finalizou com 14.510.460 linhas e 38 colunas, das quais 10 foram consideradas essenciais para a análise. Devido ao grande volume de dados, a análise foi realizada em partes menores, utilizando o parâmetro *chunksize* da biblioteca Pandas. Os atributos selecionados estão dispostos na Tabela 1.

Tabela 1
Lista de atributos selecionados para a análise

Atributo	Descrição
nmMunicipio	Nome do município sede da entidade
dsModalidadeLicitacao	Modalidade da licitação
nrQuantidade	Quantidade licitada do item
vlMaximoUnitarioitem	Valor máximo unitário do item
dsItem	Descrição do item
idTipoEntregaProduto	Tipo de entrega do produto
desconto	Percentual de desconto oferecido em proposta
catMunicipio	Categoria do município, conforme população
porteEmpresa	Porte empresarial da Pessoal Jurídica que apresentou proposta para o item
anoHomologacao	Ano da homologação do processo

Fluxo de trabalho

Para Morabito e Pureza (2018), a metodologia de pesquisa na área de pesquisa operacional envolve a abstração de situações utilizando técnicas analíticas (matemáticas e estatísticas) e experimentais (simulação), no caso dos modelos quantitativos. O presente artigo

(cc) BY-NC-SA

define variáveis de interesse e as relações que descrevem o comportamento de um sistema, caracterizando-se, assim, como uma pesquisa empírica orientada a modelos.

Para obter os dados referentes às contratações de itens da cesta básica nacional, foram aplicados alguns filtros nos dados. Foram utilizadas apenas instâncias da modalidade licitatória Pregão, que é a modalidade obrigatória para aquisição de bens e serviços comuns, na qual existe disputa de preços. Foram também consideradas apenas as propostas vencedoras da licitação. Instâncias com valor máximo nulo ou igual a zero e linhas com desconto menor ou igual a zero foram desconsideradas.

Visando obter dados relativos aos itens da cesta básica nacional, foram selecionados apenas os dados cuja descrição do item se inicia com uma das palavras-chave listadas na sequência, que foram definidas a partir dos itens considerados pelo Departamento Intersindical de Estatística e Estudos Socioeconômicos (DIEESE) como parte da cesta básica nacional. A lista original foi modificada, visando separar alguns itens que são descritos de forma muito abrangente pelo DIEESE. Os itens considerados foram: carne bovina, carne suína, carne de aves/frango, leite, feijão, arroz, farinha de trigo, farinha de milho, farinha de mandioca, batata, tomate, pão francês, café, banana, açúcar, óleo de soja, manteiga e margarina. Após a filtragem dos resultados, restaram 97.226 instâncias válidas.

Para permitir a análise e o uso de ferramentas de ML em cada item, foi necessária a categorização dos dados. Para maior eficiência, técnicas de processamento de linguagem natural e ML foram aplicadas na clusterização dos itens, utilizando as bibliotecas *Natural Language Toolkit* e *scikit-learn*, juntamente com a técnica de agrupamento *k-means*. Como resultado, cada item foi classificado em um dos dezoito grupos definidos pelo autor como parte da cesta básica. Após a clusterização e o processamento dos itens, restaram 79.667 registros. Foi então adicionada a coluna “Item”, identificando a qual item da lista cada linha se refere.

O próximo passo no pré-processamento dos dados foi a análise e o tratamento de *outliers* nos atributos “vlMaximoUnitarioItem” e “desconto”. A análise foi segmentada por item e realizada utilizando o método *Interquartile Range* (IQR), em que valores além do intervalo 1,5IQR, abaixo do 25º ou acima do 75º percentil, foram considerados *outliers* e, então, excluídos. Essa exclusão dos *outliers* foi realizada para permitir aos modelos de ML um aprendizado de acordo com os padrões dos descontos, e não a detecção de valores atípicos, que pode ser considerada outra tarefa de mineração de dados. Por fim, 73.861 instâncias foram mantidas.

Para possibilitar melhor desempenho no uso dos dados para o treinamento dos modelos

(cc) BY-NC-SA

de ML, realizou-se o processo de binarização e normalização de alguns atributos. As variáveis "idTipoEntregaProduto", "catMunicipio", "porteEmpresa" e "anoHomologacao" foram binarizadas por meio da função *get_dummies*, da biblioteca Pandas e, utilizando o método *MinMax*, disponível na biblioteca *scikit-learn*, foi executada a normalização dos dados da coluna "nrQuantidade".

Após o pré-processamento dos dados, iniciou-se a análise de hiperparâmetros para as técnicas de ML. O conjunto de dados foi dividido em subconjuntos por item, com 20% destinado ao teste. Foram utilizadas as ferramentas GridSearchCV e PyCaret para a análise dos parâmetros, ambas com validação cruzada. O GridSearchCV requer a especificação de parâmetros e valores a serem testados, enquanto o PyCaret utiliza busca randômica. A Tabela 2 apresenta os parâmetros e valores testados pelo GridSearchCV para cada modelo. A parametrização final dos modelos foi baseada nos melhores resultados obtidos.

Tabela 2

Hiperparâmetros e valores testados com a busca Grid Search

Modelo	Parâmetro	Valores
Random Forest	*n_estimators*	[100, 200, 500, 1000]
	criterion	["squared_error", "absolute_error", "friedman_mse"]
	max_features	["sqrt", "log2", None]
Redes Neurais Artificiais	*hidden_layer_sizes*	[100, 500, 1000]
	activation	['identity', 'logistic', 'tanh', 'relu']
	solver	['lbfgs', 'adam']
	learning_rate	['constant', 'invscaling', 'adaptive']
	max_iter	[200, 500, 1000]
XGBoost	*max_depth*	[3, 5, 10]
	learning_rate	[0.01, 0.08, 0.15, 0.2]
	subsample	[0.3, 0.5, 0.8]
	colsample_bytree	[0.3, 0.5, 0.8]
	booster	['gbtree', 'gblinear', 'dart']

Resultados e Discussão

Nesta seção, serão apresentados os resultados obtidos a partir da execução das melhores combinações de parâmetros para cada um dos modelos e itens objeto deste estudo. Serão expostas também discussões a respeito da interpretação dos resultados, sua importância e formas de aplicação prática dos modelos no cotidiano dos possíveis beneficiados por este estudo.

(cc) BY-NC-SA

Análise das métricas

Com base nos resultados da análise de hiperparâmetros (Tabela 2), foram definidas as melhores combinações para otimizar o desempenho dos modelos. Devido à diversidade dos itens, foi necessária uma análise segmentada, resultando em múltiplos modelos, cada um otimizado para as características específicas de cada item. As Tabelas 3 e 4 ilustram a diferença percentual entre o melhor modelo e os demais resultados para as métricas MAE e RMSE, respectivamente.

Tabela 3
Diferença percentual do MAE entre os modelos por item

Item	Random Forest		Redes Neurais Artificiais		XGBoost	
	GS	PC	GS	PC	GS	PC
Carne bovina	3,24%	0,076	4,43%	13,22%	9,4%	1,37%
Carne suína	13,07%	0,062	17,09%	32,13%	14,01%	14,05%
Carne de aves/frango	4,57%	0,09	1,32%	3,42%	0,27%	2,57%
Leite	10,78%	5,99%	12,24%	2,93%	9,17%	0,076
Feijão	12,46%	0,1	6,21%	2,65%	6,15%	3,91%
Arroz	9,48%	0,092	16,39%	11,47%	4,5%	2,1%
Farinha de trigo	7,69%	1,92%	4,02%	2,29%	0,094	8,71%
Farinha de milho	8,5%	2,57%	2,57%	6,04%	0,14	3,41%
Farinha de mandioca	5,10%	1,19%	7,22%	2,52%	0,07%	0,14
Batata	12,83%	5,26%	17,01%	37,26%	11,5%	0,12
Tomate	23,55%	0,12	29,43%	17,29%	13,61%	10,6%
Pão francês	27,91%	14,49%	19,4%	28,56%	18,47%	0,033
Café	5,45%	3,85%	9,42%	7,47%	1,81%	0,12
Banana	20,95%	5,93%	11,07%	15,78%	15,28%	0,097
Açúcar	8,59%	1,6%	0,067	4,64%	5,59%	3,51%
Óleo de soja	15,73%	0,069	7,85%	10,67%	3,41%	2,24%
Manteiga	23,98%	0,07	19,42%	32,02%	21,21%	25,09%
Margarina	14,22%	0,12	8,77%	11,55%	4,28%	5,72%

Nota. *PyCaret* (PC); *GridSeach* (GS).

(cc) BY-NC-SA

Tabela 4
Diferença percentual do RMSE entre os modelos por item

Item	Random Forest		Redes Neurais Artificiais		XGBoost	
	GS	PC	GS	PC	GS	PC
Carne bovina	4,35%	7,97%	9,63%	3,48%	0,1	9,41%
Carne suína	10,98%	6,3%	12,33%	9,13%	0,088	16,34%
Carne de aves/frango	16,6%	12,19%	0,11	8,6%	2,83%	26,78%
Leite	11,57%	15,51%	1,11%	0,1	1,51%	6,71%
Feijão	16,74%	8,08%	2,67%	0,13	4,32%	5,65%
Arroz	9,86%	2,02%	7,53%	2,68%	0,12	7,43%
Farinha de trigo	11,81%	12,73%	3,02%	2,15%	0,12	10,67%
Farinha de milho	14,84%	10,95%	5,64%	7,4%	0,16	10,14%
Farinha de mandioca	8,39%	6,8%	4,87%	3,32%	0,17	6,25%
Batata	9,31%	17,04%	5,08%	10,06%	0,17	10,87%
Tomate	28,72%	10,57%	7,69%	6,53%	0,17	14,8%
Pão francês	28,72%	42,76%	20,58%	22,08%	13,24%	0,049
Café	8,91%	8,21%	4,94%	3,4%	0,15	0,56%
Banana	15,4%	10,88%	0,14	2,51%	3,6%	0,45%
Açúcar	13,08%	14,18%	4,41%	0,084	3,53%	2,42%
Óleo de soja	13,92%	9,75%	0,49%	3,75%	0,091	2,52%
Manteiga	10,96%	1,69%	0,1	8,56%	3,82%	13,23%
Margarina	15,88%	9,07%	1,41%	3,49%	0,16	16,95%

Nota. *PyCaret* (PC); *GridSeach* (GS).

Para a métrica MAE (Tabela 3), a técnica *Random Forest* apresentou os melhores resultados em nove dos dezoito itens, enquanto o XGBoost foi superior em oito itens, e as Redes Neurais Artificiais obtiveram o menor MAE em um item. Já para a métrica RMSE (Tabela 4), as Redes Neurais Artificiais tiveram o melhor desempenho em seis itens, com os demais mostrando melhor desempenho com o XGBoost. A variação percentual entre os resultados das técnicas foi pequena, indicando estabilidade no modelo.

Os dados mostram que houve, em média, aproximadamente 10% de divergência entre os métodos e técnicas para o mesmo conjunto de dados dividido por item. É possível verificar também alguns casos em que técnicas específicas se mostraram mais eficientes, como ocorre no desempenho da técnica XGBoost para o conjunto do item 12, que apresentou RMSE distante em mais de 20% dos resultados concorrentes.

Neste estudo, houve considerável variação na quantidade de dados disponíveis para cada um dos itens. A Tabela 5 exibe as métricas dos modelos com melhor RMSE (usado como referência) para cada item, ordenadas pela quantidade de instâncias do conjunto de dados.

(cc) BY-NC-SA

Tabela 5
Métricas e quantidade de instâncias por item

Item	RMSE	MAE	# Instâncias
Carne bovina	0,104412	0,082633	11.597
Leite	0,101889	0,078598	6.800
Arroz	0,116956	0,096484	6.360
Feijão	0,127605	0,105741	5.402
Café	0,148016	0,122982	4.702
Açúcar	0,083791	0,070306	4.584
Batata	0,165434	0,132006	4.341
Margarina	0,155359	0,126859	3.908
Banana	0,137533	0,108011	3.491
Farinha de trigo	0,115433	0,094102	3.376
Carne suína	0,088143	0,070393	3.309
Óleo de soja	0,090947	0,071826	3.070
Farinha de milho	0,164186	0,136869	3.035
Farinha de mandioca	0,172336	0,143207	2.966
Carne de aves/frango	0,107111	0,091439	2.787
Tomate	0,168482	0,137294	2.206
Manteiga	0,102312	0,083512	998
Pão francês	0,048789	0,033271	929

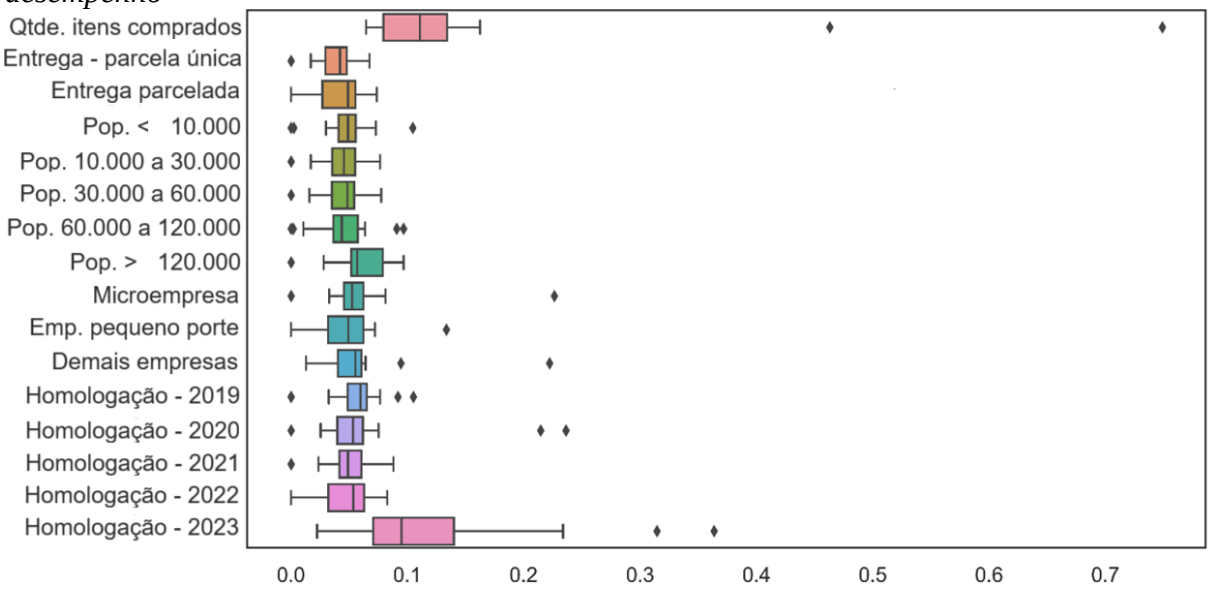
Observa-se, na Tabela 5, que o açúcar, a carne suína, o óleo de soja e o pão francês obtiveram erros baixos em relação ao RMSE e ao MAE, mesmo tendo sido treinados com menos instâncias do que a carne bovina, o leite e outros itens com mais registros disponíveis.

Importância dos atributos para os modelos

As predições por meio das técnicas Random Forest e XGBoost permitem observar, de forma simples, quais foram os atributos que tiveram maior importância para os resultados. A Figura 1 exibe a importância dos atributos com boxplots para os modelos que obtiveram os melhores valores de RMSE.

(cc) BY-NC-SA

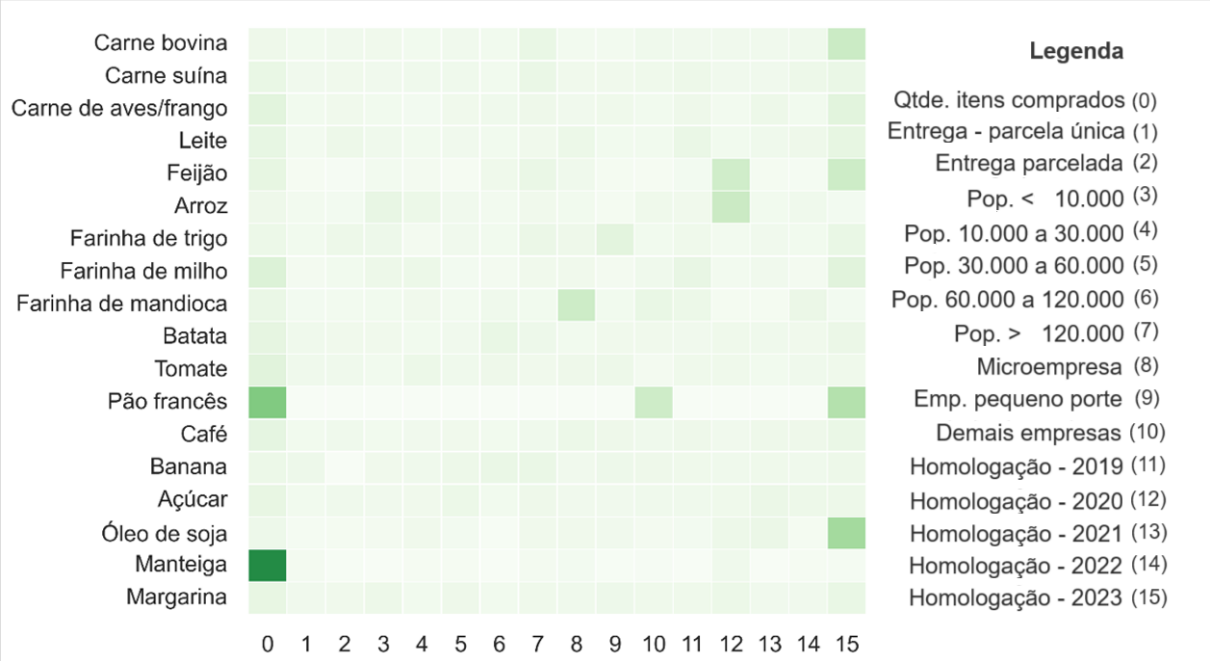
Figura 1
Importância de atributos para os modelos Random Forest e XGboost com o melhor desempenho



Nota-se que existe considerável variação e ocorrência de *outliers* em alguns dos atributos, principalmente quanto ao ano em que o processo licitatório foi homologado (que, em geral, também é o mesmo ano em que a disputa ocorreu) e à quantidade licitada do item. Devido à característica deste estudo, obtiveram-se modelos diferentes, que naturalmente apresentam divergências. A Figura 2 mostra, por meio de um mapa de calor, quais são os atributos mais importantes para cada um dos itens.

(cc) BY-NC-SA

Figura 2
Importância de atributos por item



Nota-se que a quantidade (em unidades) do item e o ano de homologação de 2023 se sobressaem com maior importância em grande parte dos itens, destacadas por meio das cores mais escuras da Figura 2. Observa-se, por exemplo, que os modelos dos itens “Feijão” e “Arroz” possuem “anoHomologação2020” com importância acima da média, o que pode ser comprovado com dados reais do mercado. Segundo o IBGE, na região de Curitiba/PR, em 2020, o arroz e o feijão tiveram um aumento de 74,13% e 51,68%, respectivamente (IBGE, 2024). Esse aumento no valor de mercado influencia diretamente o percentual de descontos que os fornecedores oferecem aos órgãos públicos, devido à incerteza do mercado e à morosidade dos processos de reajuste econômico-financeiro.

Essa convergência entre a importância dos atributos nas predições e os dados reais do mercado corrobora a confiabilidade e o comportamento correto do modelo. Os modelos treinados com GridSearchCV e PyCaret mostram resultados similares na importância dos atributos, apesar das diferenças de parametrização. No entanto, o PyCaret pode destacar mais alguns atributos específicos, como nos itens “Tomate” e “Pão francês”.

Análise dos resultados

As estimações geradas pelos modelos treinados podem ser analisadas de forma visual por meio de gráficos de linha, comparando o previsto com o real (conjunto de teste). Dada a natureza deste estudo, que contém 18 modelos diferentes, optou-se por expor aqui apenas a

(cc) BY-NC-SA

análise dos dois modelos da mediana do RMSE (Figuras 3 e 4). As figuras exibem uma amostragem aleatória de cem ocorrências, que representam 15% e 8% do total de instâncias de teste nos itens "Arroz" e "Farinha de trigo", respectivamente. Essa quantidade foi definida visando permitir uma visualização mais precisa das variações, porém representa de forma satisfatória a totalidade dos dados.

Figura 3

Comparativo entre o valor estimado e o valor real do modelo para o item "arroz"

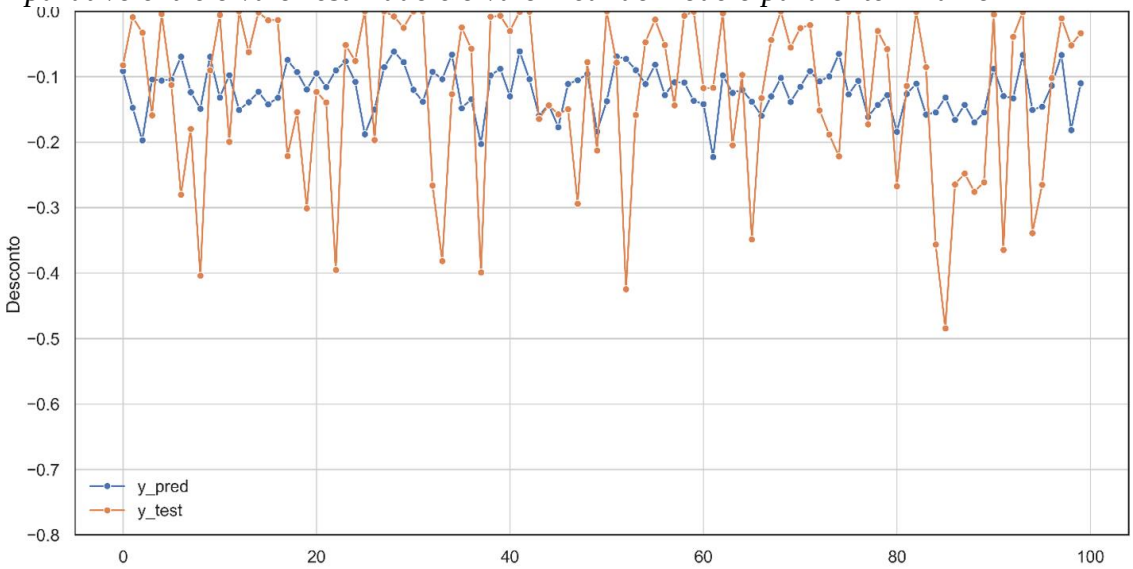
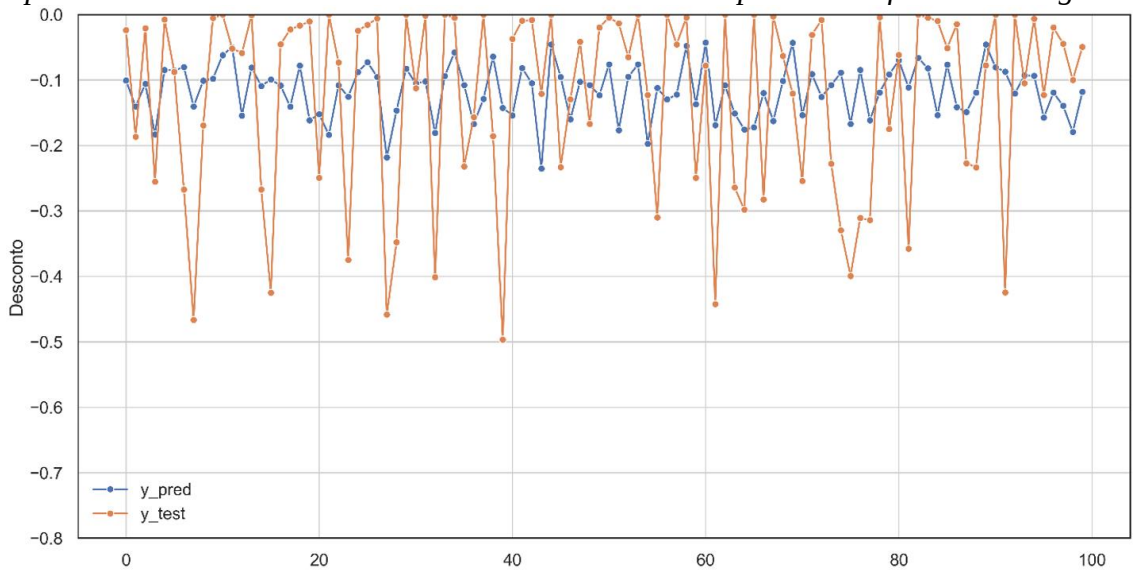


Figura 4

Comparativo entre o valor estimado e valor real do modelo para item "farinha de trigo"



As predições realizadas a partir dos modelos dos itens "Arroz" e "Farinha de trigo" apresentam limitada capacidade de acompanhar os picos negativos ou positivos dos dados reais. Vale destacar que os modelos de predição do desconto para os itens "Farinha de mandioca" e

BY-NC-SA

"Pão francês" tiveram resultados bastante limitados, sendo que o primeiro não conseguiu acompanhar variações maiores do item (ficando em torno de uma média simples), enquanto o segundo, por si só, já não possui muitos descontos relevantes, mas o modelo manteve uma média de desconto entre 0,1 e 0,2, não conseguindo estimar descontos acima desses valores.

Os modelos dos demais itens analisados possuem comportamento similar aos expostos anteriormente (Figuras 3 e 4), sofrendo com as mesmas limitações em acompanhar variações mais relevantes. O desempenho desses modelos pode ser aprimorado por meio da inserção de maior quantidade de atributos, como, por exemplo, a época do ano em que foram realizadas as compras, a microrregião em que estão inseridas as entidades compradoras e a quantidade de empresas participantes no processo licitatório.

Discussão e validação do modelo

Para os órgãos públicos, obter estimações a respeito dos descontos esperados em processos licitatórios a serem realizados tem grande importância para fins de planejamento e administração das entidades. Segundo Xerez (2013), o orçamento dos órgãos públicos estima a receita e determina as despesas, sendo uma importante fonte de informações para a tomada de decisões em relação às metas e prioridades de determinado período. Além disso, a Constituição Federal de 1988 estabeleceu a criação do Plano Plurianual, da Lei de Diretrizes Orçamentárias e da Lei Orçamentária Anual, com a finalidade de vincular o orçamento público ao planejamento das entidades.

Em posse de valores estimados para os descontos a serem ofertados em relação aos valores de mercado dos itens, torna-se viável à equipe de planejamento orçamentário estimar corretamente os valores a serem destinados para tais aquisições, possibilitando alocar da melhor forma a receita do órgão. De fato, as estimações devem ser analisadas à luz de seus erros e consideradas apenas após minuciosa análise de riscos e fatores externos. Contudo, seu caráter acessório pode auxiliar na tomada de decisões baseadas em dados.

Silva et al. (2023) defendem que um dos principais indicadores de fraudes em licitações públicas é a ocorrência de sobrepreços, definidos pela Lei n.º 14.133/2021 como "preço orçado para licitação ou contratado em valor expressivamente superior aos preços referenciais de mercado" (Brasil, 2021). A ocorrência de sobrepreço pode gerar comportamentos anormais nos descontos oferecidos pelos licitantes quando há competitividade no certame. Dessa forma, previsões obtidas por modelos de ML, como os apresentados neste estudo, podem auxiliar na detecção dessas variações quando os descontos oferecidos ficam muito abaixo do estimado,

(cc) BY-NC-SA

filtrando possíveis casos de sobrepreço para posterior análise do valor máximo, a fim de verificar sua compatibilidade com o mercado.

De forma similar, quando o valor limite fica aquém do valor real praticado pelo mercado, a tendência é que ocorram descontos abaixo do esperado. Assim, é possível identificar essas imprecisões na estimativa por meio da comparação dos descontos ofertados com os estimados, possibilitando também detectar possíveis falhas da equipe de contratação em sua responsabilidade de negociar os valores das propostas.

Além das situações citadas, outro âmbito em que o uso de modelos preditores pode ter valor é na análise de possíveis casos de inexequibilidade das propostas. Fatores como a inexperiência do vencedor, o número de competidores, deficiência de julgamento e até mesmo comportamentos irracionais podem levar licitantes a apresentar propostas irreais, gerando inexecução, atrasos e abandono de contratos (Signor et al., 2022). Com isso, descontos excessivamente altos sobre o valor máximo definido pelo órgão — desde que esse valor reflita o mercado — podem servir como parâmetro para que a administração pública busque maneiras de prevenir prejuízos e verificar a real capacidade do licitante em atender às exigências estabelecidas no certame.

Sobre a validação dos modelos utilizados, entende-se que eles conseguiram executar o treinamento e obter resultados relevantes, exceto para os itens da cesta básica “Farinha de mandioca” e “Pão francês”, que não se ajustaram de acordo com as variações naturais no conjunto de teste, mostrando basicamente uma média simples. Entretanto, a estimação foi realizada de forma satisfatória para o restante dos itens, com erros de 0,07% nos melhores experimentos.

Uma vez que os dados foram obtidos diretamente de um portal oficial do governo, os dados de entrada para análise são considerados de boa qualidade, sendo isso crucial para uma análise consistente após os testes e a avaliação do modelo (Morabito & Pureza, 2018). Situações atípicas, como grandes descontos em determinados registros, precisam ser exploradas com maior profundidade, caracterizando uma possibilidade de estudo futuro. Assim, o uso de ML na previsão de descontos pode auxiliar órgãos reguladores, controladorias internas, equipes de contratação e gestores das entidades na tomada de decisão com base em dados, bem como na prevenção e detecção de fraudes e contratos inexequíveis.

É importante destacar que as etapas de obtenção e pré-processamento dos dados para este estudo tiveram maior complexidade do que o esperado, especialmente quanto a dois pontos: (i) o formato de disponibilização dos arquivos e (ii) a falta de padronização dos itens.

(cc) BY-NC-SA

O TCE-PR disponibiliza em seu portal de informações opções para busca de licitações por meio de filtros, em uma plataforma amigável e de fácil compreensão. Porém, para obter detalhes sobre os itens licitados em cada processo, é necessário acessá-los individualmente. A obtenção de dados com maior quantidade de informações só é possível a partir de arquivos anuais compactados, que contêm vários outros arquivos compactados, cada um deles contendo diversos arquivos em formato XML.

Dessa forma, foram necessárias várias etapas para conseguir extrair as informações desejadas. O acesso seria facilitado caso as informações fossem disponibilizadas em formato CSV, por exemplo, eliminando várias etapas de pré-processamento que dificultam o uso dos dados para pesquisas e análises que podem trazer benefícios ao órgão regulador e aos seus jurisdicionados.

A falta de padronização dos itens exigiu que fossem aplicadas técnicas de processamento de linguagem natural para que os itens, objeto deste estudo, pudessem ser categorizados. Nota-se que os diversos órgãos descrevem os itens de formas distintas, sem padrão aparente.

A Lei n.º 14.133/2021 recomenda a utilização de catálogos eletrônicos de padronização das especificações de produtos(Brasil, 2021). A definição de descrições padronizadas, ou indexadas a um cadastro-base, pode auxiliar na análise dos itens pelo órgão. O Tribunal de Contas do Estado de Mato Grosso (TCE-MT), por exemplo, mantém o Catálogo de Materiais e Serviços, que é um banco de especificações de itens licitáveis de uso obrigatório para todos os órgãos sob sua jurisdição (https://servicos.tce.mt.gov.br/consulta-item). O governo federal também oferece ferramenta para tal padronização, por meio do “Catálogo Compras.gov.br” (https://catalogo.compras.gov.br/).

Essas ferramentas podem favorecer a utilização de modelos de aprendizado de máquina, dada a simplificação no processo de classificação dos itens, possibilitando análises mais confiáveis, com menor dispêndio de tempo nas etapas de pré-processamento.

Conclusões

O objetivo deste estudo foi avaliar modelos de ML para a estimação de valores de desconto ofertados pelo licitante vencedor de processos licitatórios para o fornecimento de itens da cesta básica nacional. Para atingir esse objetivo, foram aplicadas as técnicas Random Forest, Redes Neurais Artificiais e XGBoost a cada um dos conjuntos de dados dos 18 itens selecionados pelo autor.

(cc) BY-NC-SA

Como resultado, obteve-se um RMSE médio de 0,122152 e um MAE médio de 0,099196 entre os modelos selecionados. A partir da análise da importância das variáveis, identificou-se que a quantidade licitada e o ano em que o processo foi homologado foram os atributos que mais afetaram a estimação em grande parte dos modelos, embora exista certa variação, por se tratar de modelos para itens distintos, que estão sujeitos a diferentes fatores externos.

Os modelos de ML treinados apresentaram resultados satisfatórios para quase todos os itens, mostrando que podem servir de alicerce para a estimação de descontos em determinadas situações, como demonstrado nas métricas RMSE e MAE. Sugere-se, a partir deste estudo, uma análise aprofundada de valores atípicos e de variáveis específicas que também possam influenciar a estimação do desconto, as quais estão além do escopo deste trabalho, como índices econômicos (taxa de juros Selic, inflação, entre outros).

A área de compras públicas tem passado por mudanças, com a Nova Lei de Licitações trazendo melhorias e desafios aos órgãos públicos. O uso da tecnologia para o aumento da eficácia e da eficiência dos processos pode auxiliar na prestação de serviços públicos de qualidade à população.

(cc) BY-NC-SA

REFERÊNCIAS

Aggarwal, C. C. (2015). *Data Mining: The textbook*. Springer International Publishing. https://doi.org/10.1007/978-3-319-14142-8

Aggarwal, C. C. (2018). Neural Networks and Deep Learning. In *Neural Networks and Deep Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-319-94463-0

Amorim, V. A. J. de. (2017). *Licitações e contratos administrativos: teoria e jurisprudência*. https://www2.senado.leg.br/bdsf/handle/id/533714

Bonaccorso, G. (2017). *Machine Learning Algorithms: A reference guide to popular algorithms for data science and machine learning* (1st ed., Vol. 1). Packt Publishing.

Brasil. (2021). *Lei 14133 de Licitações e Contratos Administrativos*. https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1007/9781441993267_5

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. https://doi.org/10.1145/2939672.2939785

Dhaliwal, S. S., Nahid, A.-A., & Abbas, R. (2018). Effective Intrusion Detection System Using XGBoost. *Information*, *9*(7), 149. https://doi.org/10.3390/info9070149

França, R. M. M. (2021). *Compras públicas conjuntas e economias de escala: um estudo de caso do Sistema COFEN / Conselhos Regionais de Enfermagem* [Dissertação de Mestrado, Universidade de Brasília]. http://www.rlbea.unb.br/jspui/handle/10482/42165

Freitas, A. M. de, Prado, F. O., Alexandre, P. L. T., & Carmona, M. F. F. (2021). *Nova lei de licitações e contratos administrativos: comentários a lei no 14.133/2021* (2nd ed.). https://www2.senado.leg.br/bdsf/handle/id/588204

Géron, A. (2019). *Hands-on: Machine Learning with Scikit-Learn, Keras & Tensorflow* (2nd ed.). O'Reilly Media.

Ghazal, M. M., & Hammad, A. (2022). Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects. *International Journal of Construction Management*, *22*(9), 1632–1646. https://doi.org/10.1080/15623599.2020.1738205

Halužan Vasle, A., & Moškon, M. (2024). Synthetic biological neural networks: From current implementations to future perspectives. *BioSystems*, *237*, 105164. https://doi.org/10.1016/j.biosystems.2024.105164

Instituto Brasileiro de Geografia e Estatística. (2024). *Sistema Nacional de Índices de Preços ao Consumidor. Tabela 7063 - INPC*. https://sidra.ibge.gov.br/tabela/7063

Lara, J. A., Lizcano, D., Martínez, M. A., & Pazos, J. (2014). Data preparation for KDD through automatic reasoning based on description logic. *Information Systems*, *44*, 54–72. https://doi.org/10.1016/j.is.2014.03.002

M. Mijwil, M. (2021). Artificial Neural Networks Advantages and Disadvantages. *Mesopotamian Journal of Big Data*, *2021*, 29–31. https://doi.org/10.58496/MJBD/2021/006

Mello, C. A. B. de. (2015). *Curso de Direito Administrativo* (32nd ed.). Malheiros.

Morabito, R., & Pureza, V. (2018). Modelagem e Simulação. In Cauchick, P.A., (Coord.). Metodologia de Pesquisa em Engenharia de Produção e Gestão de Operações (3 ed., Chap. 8, pp. 165-195). Rio de Janeiro: Elsevier.

Oliveira, L. H. R. de, Rêgo, T. G. do, & Diniz, J. A. (2019). Previsão de Valores de Aquisições Governamentais: o Uso dos Conceitos de Data Science e Machine Learning. *XVI Congresso USP de Iniciação Científica Em Contabilidade*, 1–15.

(cc) BY-NC-SA

Radhoush, S., Whitaker, B. M., & Nehrir, H. (2023). An Overview of Supervised Machine Learning Approaches for Applications in Active Distribution Networks. *Energies*, *16*(16), 5972. https://doi.org/10.3390/en16165972

Raschka, S. (2015). *Python Machine Learning* (1st ed.). Packt Publishing Ltd.

Ribeiro, C. G., & Inácio Júnior, E. (2019). *O mercado de compras governamentais brasileiro (2006-2017): Mensuração e análise*.

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522–542. https://doi.org/10.1016/j.ins.2021.05.055

Sampaio, I. G., Bernardini, F., Paes, A., Andrade, E. de O., & Viterbo, J. (2019). Avaliação de Modelos de Predição e Previsão Construídos por Algoritmos de Aprendizado de Máquina em Problemas de Cidades Inteligentes. In *Tópicos em Sistemas de Informação: Minicursos SBSI 2019* (pp. 81–113). SBC. https://doi.org/10.5753/sbc.480.9.04

Santos, L. B., Gentry, D., Tryforos, A., Fultz, L., Beasley, J., & Gentimis, T. (2024). Soybean yield prediction using machine learning algorithms under a cover crop management system. *Smart Agricultural Technology*, *8*, 100442. https://doi.org/10.1016/j.atech.2024.100442

Signor, R., Marchiori, F. F., Raupp, A. B., Magro, R. R., & Lopes, A. de O. (2022). A nova lei de licitações como promotora da maldição do vencedor. *Revista de Administração Pública*, *56*(1), 176–190. https://doi.org/10.1590/0034-761220210133

Silva, M. O., Costa, L. L., Bezerra, G., Gomide, L. D., Hott, H. R., Oliveira, G. P., Brandão, M. A., Lacerda, A., & Pappa, G. (2023). Análise de Sobrepreço em Itens de Licitações Públicas. *Anais Do XI Workshop de Computação Aplicada Em Governo Eletrônico (WCGE 2023)*, 118–129. https://doi.org/10.5753/wcge.2023.230608

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson Prentice Hall.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. http://www.jstor.org/stable/24869236

Xerez, S. R. D. (2013). A evolução do orçamento público e seus instrumentos de planejamento. *Revista Científica Semana Acadêmica*, *01*(43), 1–19.

Xu, Q., & Yin, J. (2021). Application of Random Forest Algorithm in Physical Education. *Scientific Programming*, *2021*. https://doi.org/10.1155/2021/1996904

(cc) BY-NC-SA