Journal of Production, Operations and Systems Management

e-ISSN: 1984-2430

Open Access

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

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

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

How to reference this paper:

Santos, J. M. R., & Santos, B. S. Machine learning algorithms applied to the estimation of discounts in basic food basket items. *Revista GEPROS*, *20*, e025006. DOI: 10.15675/gepros.3037

| **Submitted**: 24/10/2024
| **Approved:** 18/09/2025
| **Published:** 13/11/2025

Editor: Prof. Dr. Paula de Camargo Fiorini

[1]Federal Technological University of Paraná (UFTPR), Londrina – Paraná (PR) – Brazil. Undergraduate student in Production Engineering.
[2]Federal Technological University of Paraná (UFTPR), Londrina – Paraná (PR) – Brazil. Adjunct Professor in the Department of Production Engineering. Researcher in the Optimization and Data Mining Research Group (GPOMD).

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.

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.

Introduction

The public procurement market in Brazil accounted, on average, for 12.5% of the country's Gross Domestic Product (GDP) between 2006 and 2017 (Ribeiro & Inácio Júnior, 2019). In general, all purchasing processes carried out by the public administration, autonomous agencies, or government foundations must comply with bidding and contracting rules currently established by Law No. 14,133/21.

The bidding process aims to select the most advantageous proposal for the government, ensuring that all interested parties can participate on equal terms, provided they meet the criteria predefined by the contracting authority (Mello, 2015). The current legislation defines several bidding modalities; however, for the acquisition of common goods and services, the *pregão* (reverse auction) modality is used, preferably in its electronic form (Freitas et al., 2021).

To set the maximum admissible price for the purchase, the prices obtained in market research must be critically assessed, disregarding unfeasible or excessively high values. The final reference amount is defined by a simple mean, a median, or the lowest value obtained in the research, which may be conducted directly with suppliers and specialized websites or by consulting records of similar contracts in the public sector. During the bidding procedure, bidders submit proposals and compete by placing successive bids over a specified period. Once the bidding phase is concluded, the best proposal that meets the requirements is declared the winner (Amorim, 2017). Therefore, the final contracted price may differ from the initially defined reference value, depending on the bids submitted by suppliers.

Given their nature as government data, information on the items tendered is available to all citizens. The Court of Accounts of the State of Paraná (TCE-PR) maintains in its database all bidding information from public administration entities in the municipalities of Paraná, making it publicly accessible through the "Portal de Informações para Todos." Data mining and machine learning techniques can be applied to these databases through the Knowledge Discovery in Databases (KDD) process, with the aim of finding patterns and extracting useful information (Ghazal & Hammad, 2022).

Machine learning is a subfield of artificial intelligence characterized by the development of self-learning algorithms that derive knowledge from data to make predictions, identify rules and relationships among variables, and gradually improve the performance of the estimating model (Géron, 2019).

According to Oliveira, Rêgo, and Diniz (2019), the reference price in a bidding process is subject to biases or even collusion between public agents and potential bidders, which can cause it to deviate from the real market value. Thus, by estimating the discount offered for an item, one can assess the quality of the reference price since—assuming it is compatible with the market—the discount percentage should be similar to that estimated from previous public contracts.

França (2021) emphasizes that the greater the quantity purchased of a given item, the lower the acquisition cost; this also applies to public procurement. By analyzing the percentage of discounts offered for similar items purchased in different quantities, it may be possible to observe economies of scale, especially in procurement processes carried out by smaller entities.

In this context, this article evaluates machine learning models that estimate the discount offered by the winning proposal in tenders for items in the national basic food basket, conducted by municipal public administration entities in the state of Paraná. To this end, three ML techniques were employed: Random Forest, XGBoost, and Artificial Neural Networks (ANNs), along with analyses of feature contributions in the estimation of 18 basic food-basket items.

In addition to this introduction, the study comprises a theoretical framework that briefly addresses concepts such as data mining, regression techniques, and bidding processes. Next, the applied methodology is presented, along with the dataset and research steps. This is followed by the results and discussion and, finally, the concluding remarks.

Theoretical Framework

This chapter addresses the concepts and stages of Knowledge Discovery in Databases (KDD), data mining, supervised learning, estimation (regression) techniques, evaluation metrics, and the bidding process, which is the focus of this study.

Knowledge Discovery in Databases

The KDD process aims to identify patterns in data and present them in a way that facilitates their assimilation as knowledge (Lara et al., 2014). Ghazal and Hammad (2022) define KDD as “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.”

The KDD process is carried out in a sequence of steps. First, data from different sources must be integrated and selected according to the needs of the analysis. The next step is cleaning, in which errors, noise, and missing values are treated before moving to transformation (third

step), where data are prepared for mining through summarization or aggregation techniques. Once transformed, the data are ready for mining, using methods to identify patterns. The patterns found must then be evaluated to determine whether they indeed represent useful knowledge. Finally, visualization and knowledge representation techniques are applied to present the extracted knowledge to stakeholders.

Each of these stages is important and contributes to the final outcome, which is the assimilation of knowledge from data. Data mining techniques may include machine learning algorithms, which can be of different types, as described in the following subsection.

Data mining and machine learning

The data mining stage was defined by Tan et al. (2019) as the exploration of datasets to identify important correlations among variables contained in often extensive datasets and to summarize them to facilitate the understanding of these relationships and access to useful data. This process includes the application of algorithms to data to extract new and useful knowledge (Aggarwal, 2015).

Building models and understanding the rules of a large dataset manually can be very costly. For this problem, machine learning (ML) offers an alternative. According to Raschka (2015), the development of algorithms that "learn" by themselves has contributed to improving data processing performance and capacity. One of the main objectives of research in ML is the recognition of complex patterns to make intelligent decisions based on analyzed data.

ML is organized into several groups according to the purpose of the algorithm, but the main categories are supervised and unsupervised algorithms. Supervised algorithms acquire knowledge from training datasets that contain paired input–output combinations. Based on this training, the algorithm seeks to predict outputs for each input in a test set, and its accuracy is then evaluated (Radhoush et al., 2023).

If the expected outputs are continuous numerical values, the task is called regression; however, if there is a discrete number of possibilities or categories, the task is called classification. Examples of supervised learning applications include spam detection, pattern recognition, natural language processing, and predictive analyses based on regression or classification (Bonaccorso, 2017). This study employs supervised learning techniques, specifically regression algorithms, to estimate discount values for basic food basket items.

Regression techniques

To estimate numerical values based on an input dataset, several models exist within the domain of machine learning. This section presents the algorithms used in this study to estimate discounts on basic food basket items, as well as the performance metrics of the regressors.

Random Forest

This technique belongs to the class of ensemble algorithms—a class of models that combine the results of multiple individual algorithms into a joint solution (Xu & Yin, 2021). Random Forest is an ensemble of decision trees in which each tree depends on values from a random and independent sampling vector (Breiman, 2001).

Decision trees are a supervised approach where prediction is structured like a tree, composed of a root, nodes, branches, and leaves. Random Forest uses multiple decision trees generated through bootstrapping, where some inputs may be used more than once while others may not be selected, maintaining the original sample size. The trees also receive a much smaller subset of variables chosen randomly from the original set (Géron, 2019).

These characteristics enable the method to perform efficiently on large, high-dimensional datasets while also allowing for variable importance measurement and outlier detection. Compared to a single traditional decision tree, Random Forest generally achieves higher accuracy with lower computational cost (Santos et al., 2024).

XGBoost

XGBoost, short for *eXtreme Gradient Boosting*, is another machine learning algorithm used for value prediction that has gained considerable prominence in recent years. Models such as XGBoost are regarded as state-of-the-art for classification tasks and are widely used in competitions (Sagi & Rokach, 2021). Based on the concept of Gradient Boosting Decision Trees (GBDT), this technique also uses multiple decision trees to achieve higher accuracy but employs a different process.

In contrast to Random Forest, each new tree in XGBoost receives weighted data, prioritizing the correction of errors made in the previous step (Dhaliwal et al., 2018).

Much of the success of this technique is due to its execution being more than ten times faster than other methods available at the time of its development, in addition to being scalable, resource-efficient, capable of handling sparse data effectively, and able to process datasets

exceeding traditional computational limits through out-of-core methods (Chen & Guestrin, 2016).

Artificial Neural Networks

The ANNs were developed based on the biological functioning of neurons (Halužan Vasle & Moškon, 2024), which learn through the strength of synaptic connections in response to external stimuli.

Artificial neurons receive inputs, perform computational operations, and transmit outputs to other neurons, like biological networks. Learning occurs through the adjustment of connection weights (synapses), which are modified in response to incorrect predictions, meaning that the network's architecture directly influences its performance (Aggarwal, 2018). One of the most widely used ANN architectures in ML is the multilayer perceptron, composed of three main layers: input, hidden, and output.

Some advantages of neural networks compared with other learning approaches include robustness and tolerance to faulty neurons; the ability to generalize from prior experiences to new situations; the capacity to handle inconsistent or noisy data; and the distributed nature of computational work, enabling parallel or distributed processing (Mijwil, 2021).

Evaluation Metrics

To assess supervised ML algorithms, the predicted results of a test dataset are compared with actual values to determine prediction quality. For regression tasks, specific error metrics are commonly used. The Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) are among the most widely adopted. These metrics always yield non-negative values, and the goal is to achieve error values close to zero (Sampaio et al., 2019).

The MAE is computed by summing the absolute errors and dividing by the number of predictions. RMSE, in contrast, is based on the squared values of the errors, making it more sensitive to large errors and outliers (Willmott & Matsuura, 2005). The formulas are shown in Equations (1) and (2):

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

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

Where n is the number of evaluated instances, y_i is the actual value of instance i, and $\hat{y}_i$ is the predicted value of instance i.

Public procurement: Bidding processes

The acquisition of goods and services by the public sector has peculiarities when compared with private sector purchases, with procurement generally conducted through bidding processes. Bidding can be defined as the administrative procedure required to permit the sale, acquisition, leasing, concession, or contracting of any goods or services under conditions defined by the government entity, in which all interested parties are invited to submit proposals to select the most advantageous offer for the administration according to pre-established and publicly disclosed parameters (Mello, 2015).

Despite the mandatory nature of bidding, there are exceptions to this rule, such as direct contracting, which may be waived or deemed unnecessary in accordance with the legislation (Freitas et al., 2021). Currently, public procurement in Brazil is governed by Law No. 14,133, enacted on April 1, 2021, which replaced Law No. 8,666/93. In the legal field, Law No. 14,133/21 is commonly referred to as the New Procurement Law.

According to Law No. 14,133/21, the objectives of the bidding process are to ensure the selection of the most advantageous proposal, safeguard the right to equal treatment among bidders, prevent contracts at prices incompatible with the market, and foster innovation and sustainable national development.

To achieve these objectives, some principles established by Brazilian law must also be observed, providing interpretive criteria for dealing with unforeseen situations (Amorim, 2017). Article 5 of the New Procurement Law lists principles such as legality, impartiality, morality, publicity, efficiency, public interest, administrative probity, equality, planning, transparency, effectiveness, segregation of duties, justification, adherence to the notice, objective judgment, legal certainty, reasonableness, competitiveness, proportionality, promptness, cost-effectiveness, and sustainable national development.

The legislation requires that the public procurement process begin with a detailed description of the need for contracting a specific item, including the definition of technical characteristics, execution conditions, payment, and guarantees. Subsequently, to establish the maximum price to be paid for the item, the administration must conduct market surveys in accordance with legal criteria.

Price estimation must be conducted critically by consulting sources that accurately reflect the market, such as similar public contracts, specialized e-commerce platforms, and suppliers in the sector. To determine the reference price, unfeasible or excessively high values must be disregarded, and the final amount must be set using the mean, median, or lowest market price (Amorim, 2017).

After defining the object to be tendered, the procurement rules are set and later disclosed in the public notice, ensuring that all potential participants are fully aware of the conditions and are treated equally. Before the publication of the notice, the law requires a risk analysis to identify potential threats to the success and proper execution of the contract. Once these steps are completed and legal approval is obtained, the process is published and carried out in accordance with the rules of the bidding document.

Methodology

This section describes the dataset used for the analysis, the workflow applied during the data preparation process, and the tools employed. The study is characterized as applied research, with a quantitative approach, experimental procedures, and the use of analytical research methods.

Dataset

Data from the Court of Accounts of the State of Paraná (TCE-PR), available through the "Portal de Informações para Todos," were used to achieve the objective of this study. These data are updated weekly and contain information on fuel consumption, contracts and agreements, revenues and expenses of the agency, travel allowances, procurement processes, and public works from all municipalities in the state.

The data are provided in compressed files, with one for each municipality, year, and table. For this analysis, the "LicitacaoVencedor" table was used, covering all municipalities

between 2019 and 2023. Originally in XML format, the data were converted into CSV using the ElementTree and Pandas libraries in Python.

After conversion, the dataset was complemented with additional information on company size and municipality size. Company size was extracted from the National Registry of Legal Entities (CNPJ), managed by the Federal Revenue Service of Brazil (RFB). These data, updated monthly, were retrieved on September 9, 2023, in CSV format. To incorporate municipality size, 2022 demographic census data published by the Brazilian Institute of Geography and Statistics (IBGE) were used. The dataset, provided in XLS format, includes the population of all 5,570 municipalities in the country and was updated on June 22, 2023.

To categorize municipalities according to population, the following categories were defined based on the characteristics of Paraná's municipalities, which are generally small: Category 1 – Up to 10,000 inhabitants; Category 2 – Between 10,000 and 30,000 inhabitants; Category 3 – Between 30,000 and 60,000 inhabitants; Category 4 – Between 60,000 and 120,000 inhabitants; Category 5 – Above 120,000 inhabitants.

An additional column named *desconto* (discount) was created to represent the percentage discount offered in the proposals, calculated as the ratio between the maximum item value and the proposed value, subtracted by one.

After these additions, the dataset contained 14,510,460 rows and 38 columns, of which 10 were considered essential for analysis. Due to the large data volume, the analysis was conducted in smaller parts using the *chunksize* parameter of the Pandas library. The selected attributes are listed in Table 1.

Table 1

List of features selected for the analysis

Feature	Description
nmMunicipio	Name of the municipality where the entity is located
dsModalidadeLicitacao	Bidding modality
nrQuantidade	Quantity of the item procured
vlMaximoUnitarioitem	Maximum unit value of the item
dsItem	Item description
idTipoEntregaProduto	Type of product delivery
desconto	Percentage discount offered in the proposal
catMunicipio	Municipality category, according to population

porteEmpresa	The company size of the legal entity that submitted the proposal
anoHomologacao	Year in which the process was approved

Workflow

According to Morabito and Pureza (2018), research methodology in the field of operations research involves abstracting situations using analytical (mathematical and statistical) and experimental (simulation) techniques in the case of quantitative models. This article defines variables of interest and their relationships to describe system behavior, thus characterizing it as model-oriented empirical research.

To obtain data related to the procurement of national basic food basket items, several filters were applied. Only instances of the *pregão* modality were considered, as this is the mandatory modality for acquiring common goods and services, in which price competition occurs. Only winning bids were included. Instances with a null or zero maximum value and rows with a discount equal to or less than zero were also excluded.

To ensure the analysis focused on basic food basket items, only data in which the item description began with one of the keywords listed below were selected. These keywords were defined based on items considered by the Inter-Union Department of Statistics and Socioeconomic Studies (DIEESE) as part of the national basic food basket. The original list was modified to separate some items broadly defined by DIEESE. The items considered were: beef, pork, poultry/chicken, milk, beans, rice, wheat flour, corn flour, cassava flour, potato, tomato, French bread, coffee, banana, sugar, soybean oil, butter, and margarine. After filtering, 97,226 valid instances remained.

For the analysis and use of ML tools on each item, data categorization was required. To improve efficiency, natural language processing and ML techniques were applied for item clustering, using the Natural Language Toolkit and scikit-learn libraries, along with the k-means clustering technique. As a result, each item was classified into one of the 18 groups defined by the author as part of the basic food basket. After clustering and processing, 79,667 records remained. A new column, “Item,” was then added to identify which list item each row referred to.

The next step in preprocessing was the analysis and treatment of outliers in the attributes “vlMaximoUnitarioItem” and “desconto.” The analysis was conducted by item, using the

Interquartile Range (IQR) method, with values beyond 1.5 IQR below the 25th percentile or above the 75th percentile considered outliers and excluded. This exclusion was intended to allow ML models to learn according to discount patterns rather than detect atypical values, which can be treated as a separate data mining task. In total, 73,861 instances were retained.

To enhance performance in training ML models, binarization and normalization of some attributes were performed. The variables "idTipoEntregaProduto," "catMunicipio," "porteEmpresa," and "anoHomologacao" were binarized using the *get_dummies* function from the Pandas library. Additionally, the "nrQuantidade" column was normalized using the MinMax method from scikit-learn.

After preprocessing, hyperparameter analysis was conducted for the ML techniques. The dataset was divided into item-specific subsets, with 20% allocated to testing. The tools GridSearchCV and PyCaret were used for parameter analysis, both employing cross-validation. GridSearchCV requires explicit specification of parameters and values to be tested, while PyCaret performs a random search. Table 2 presents the parameters and values tested with GridSearchCV for each model. Final parameterization of the models was based on the best results obtained.

Table 2

Hyperparameters and values tested using Grid Search

Model	Parameter	Values
Random Forest	*n_estimators*	[100, 200, 500, 1000]
	criterion	["squared_error," "absolute_error," "friedman_mse"]
	max_features	["sqrt," "log2," None]
Artificial Neural Networks	*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”]

Results and Discussion

This section presents the results obtained from the optimal parameter combinations for each of the models and items under study. Furthermore, it discusses the interpretation and significance of these results, along with the potential practical applications of the models for their intended beneficiaries.

Metrics Analysis

Based on the results of the hyperparameter analysis (Table 2), the best parameter combinations were defined to optimize the performance of the models. Due to the diversity of items, a segmented analysis was required, resulting in multiple models, each optimized for the specific characteristics of each item. Tables 3 and 4 illustrate the percentage difference between the best-performing model and the other results for the MAE and RMSE metrics, respectively.

Table 3

Percentage difference in MAE between models by item

Item	Random Forest		Artificial Neural Networks		XGBoost	
	GS	PC	GS	PC	GS	PC
Beef	3.24%	0.076	4.43%	13.22%	9.4%	1.37%
Pork	13.07%	0.062	17.09%	32.13%	14.01%	14.05%
Poultry/Chicken	4.57%	0.09	1.32%	3.42%	0.27%	2.57%
Milk	10.78%	5.99%	12.24%	2.93%	9.17%	0.076
Beans	12.46%	0.1	6.21%	2.65%	6.15%	3.91%
Rice	9.48%	0.092	16.39%	11.47%	4.5%	2.1%
Wheat flour	7.69%	1.92%	4.02%	2.29%	0.094	8.71%
Corn flour	8.5%	2.57%	2.57%	6.04%	0.14	3.41%
Cassava flour	5.10%	1.19%	7.22%	2.52%	0.07%	0.14
Potato	12.83%	5.26%	17.01%	37.26%	11.5%	0.12
Tomato	23.55%	0.12	29.43%	17.29%	13.61%	10.6%
French bread	27.91%	14.49%	19.4%	28.56%	18.47%	0.033
Coffee	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
Sugar	8.59%	1.6%	0.067	4.64%	5.59%	3.51%
Soybean oil	15.73%	0.069	7.85%	10.67%	3.41%	2.24%
Butter	23.98%	0.07	19.42%	32.02%	21.21%	25.09%
Margarine	14.22%	0.12	8.77%	11.55%	4.28%	5.72%

Note. PyCaret (PC); *GridSearch* (GS).

Table 4
Percentage difference in RMSE between models by item

Item	Random Forest		Artificial Neural Networks		XGBoost	
	GS	PC	GS	PC	GS	PC
Beef	4.35%	7.97%	9.63%	3.48%	0.1	9.41%
Pork	10.98%	6.3%	12.33%	9.13%	0.088	16.34%
Poultry/Chicken	16.6%	12.19%	0.11	8.6%	2.83%	26.78%
Milk	11.57%	15.51%	1.11%	0.1	1.51%	6.71%
Beans	16.74%	8.08%	2.67%	0.13	4.32%	5.65%
Rice	9.86%	2.02%	7.53%	2.68%	0.12	7.43%
Wheat flour	11.81%	12.73%	3.02%	2.15%	0.12	10.67%
Corn flour	14.84%	10.95%	5.64%	7.4%	0.16	10.14%
Cassava flour	8.39%	6.8%	4.87%	3.32%	0.17	6.25%
Potato	9.31%	17.04%	5.08%	10.06%	0.17	10.87%
Tomato	28.72%	10.57%	7.69%	6.53%	0.17	14.8%
French bread	28.72%	42.76%	20.58%	22.08%	13.24%	0.049
Coffee	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%
Sugar	13.08%	14.18%	4.41%	0.084	3.53%	2.42%
Soybean oil	13.92%	9.75%	0.49%	3.75%	0.091	2.52%
Butter	10.96%	1.69%	0.1	8.56%	3.82%	13.23%
Margarine	15.88%	9.07%	1.41%	3.49%	0.16	16.95%

Note. PyCaret (PC); *GridSearch* (GS).

For the MAE metric (Table 3), the Random Forest technique yielded the best results for nine out of eighteen items, while XGBoost outperformed the others in eight items, and ANNs achieved the lowest MAE in one item. Regarding the RMSE metric (Table 4), ANNs showed the best performance in six items, while the remaining items exhibited superior performance with XGBoost. The percentage variation among the results of the techniques was small, indicating stability in the models.

On average, the data revealed approximately a 10% divergence between methods and techniques applied to the same dataset segmented by item. It is also possible to observe specific cases in which certain techniques proved more efficient, such as the performance of XGBoost for item 12, whose RMSE was more than 20% lower compared to competing results.

In this study, we observed a considerable variation in the amount of data available for each item. Table 5 presents the metrics of the models with the best RMSE (used as a reference) for each item, ordered by the number of instances in the dataset.

Table 5
Metrics and number of instances

Item	RMSE	MAE	# Instances
Beef	0.104412	0.082633	11,597
Milk	0.101889	0.078598	6,800
Rice	0.116956	0.096484	6,360
Beans	0.127605	0.105741	5,402
Coffee	0.148016	0.122982	4,702
Sugar	0.083791	0.070306	4,584
Potato	0.165434	0.132006	4,341
Margarine	0.155359	0.126859	3,908
Banana	0.137533	0.108011	3,491
Wheat flour	0.115433	0.094102	3,376
Pork	0.088143	0.070393	3,309
Soybean oil	0.090947	0.071826	3,070
Corn flour	0.164186	0.136869	3,035
Cassava flour	0.172336	0.143207	2,966
Poultry/Chicken	0.107111	0.091439	2,787
Tomato	0.168482	0.137294	2,206
Butter	0.102312	0.083512	998
French bread	0.048789	0.033271	929

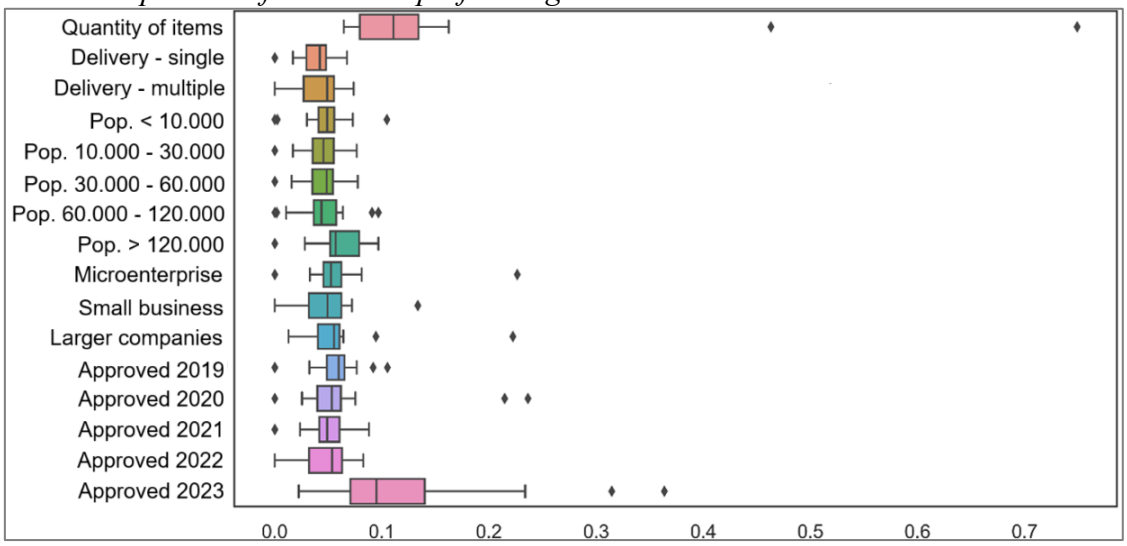
As shown in Table 5, sugar, pork, soybean oil, and French bread achieved low errors in terms of RMSE and MAE, even when trained with fewer instances compared to beef, milk, and other items with larger numbers of records available.

Feature importance for the models

The Random Forest and XGBoost techniques allow for a straightforward identification of the features that had the greatest impact on the results. Figure 1 presents the feature importance through boxplots for the models that achieved the best RMSE values.

Figure 1

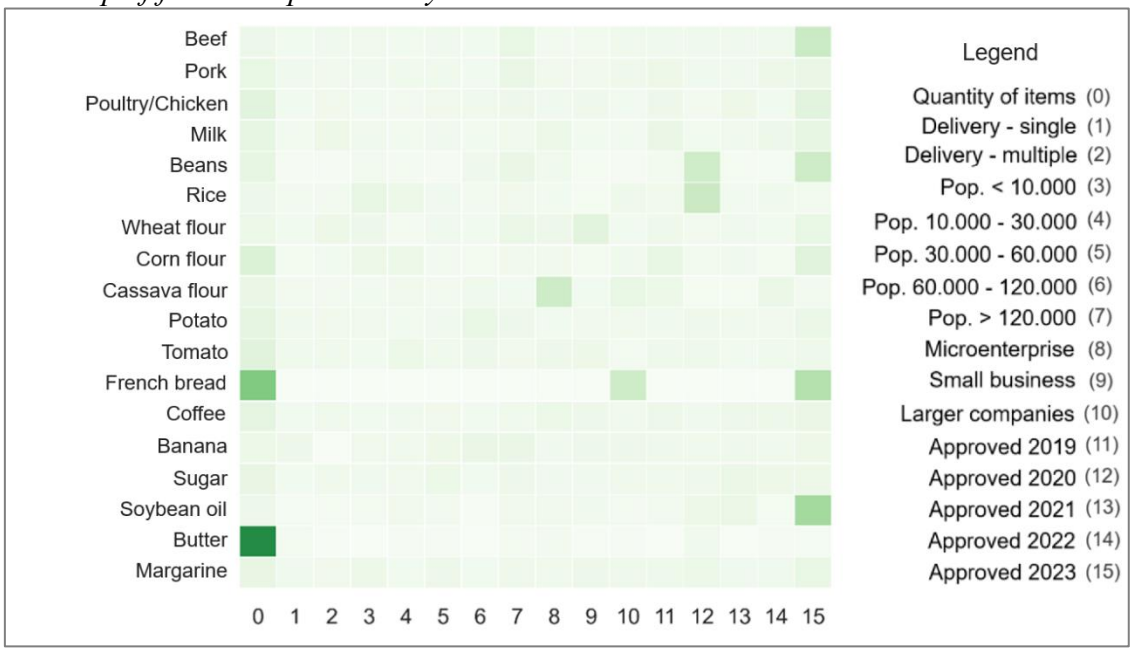
Feature importance for the best-performing Random Forest and XGBoost models



We noted that there was considerable variation and the occurrence of outliers in some of the features, particularly regarding the year in which the bidding process was approved (which, in general, also corresponds to the year in which the bidding took place) and the quantity of the item tendered. Due to the nature of this study, the results consist of different models that naturally present divergences. Figure 2 presents a heatmap showing the most important features for each item.

Figure 2

Heatmap of feature importance by item



It is evident that the quantity (in units) of the items and the homologation year of 2023 stand out as having greater importance in most items, highlighted by the darker colors in Figure 2. For instance, the models for the items beans and rice show that "yearHomologation2020" has above-average importance; this finding can be corroborated by real market data. According to IBGE, in the Curitiba/PR region in 2020, rice and beans experienced price increases of 74.13% and 51.68%, respectively (IBGE, 2024). This rise in market value directly influences the discount percentages that suppliers offer to public agencies, due to market uncertainty and the slowness of economic–financial adjustment processes.

This convergence between feature importance in the predictions and actual market data supports the reliability and proper behavior of the model. Models trained with GridSearchCV and PyCaret show similar results regarding feature importance, despite differences in parameterization. However, PyCaret may highlight additional specific features, as observed in the cases of "Tomato" and "French bread."

Results analysis

The estimations generated by the trained models can be visually analyzed using line charts, comparing the predicted values with the actual ones (test set). Given that this study involves 18 different models, we present the analysis for only two models, corresponding to the median RMSE (Figures 3 and 4). The figures display a random sample of one hundred occurrences, representing 15% and 8% of the total test instances for rice and wheat flour, respectively. This sampling size was chosen to provide a more precise visualization of the variations while still satisfactorily representing the entirety of the data.

Figure 3

Comparison between estimated and actual values for the model of the item "rice"

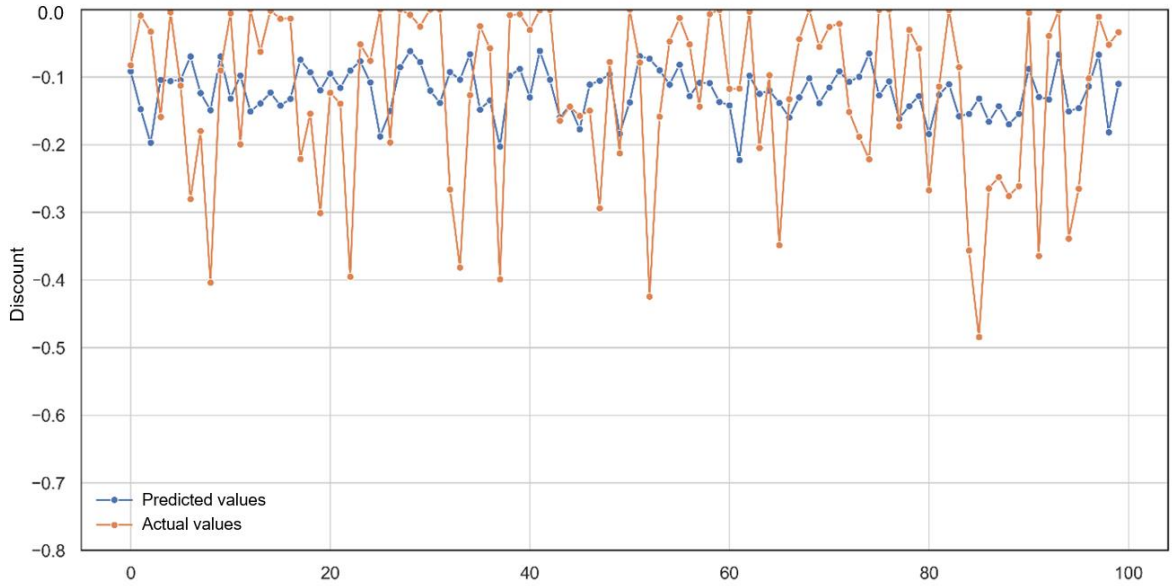
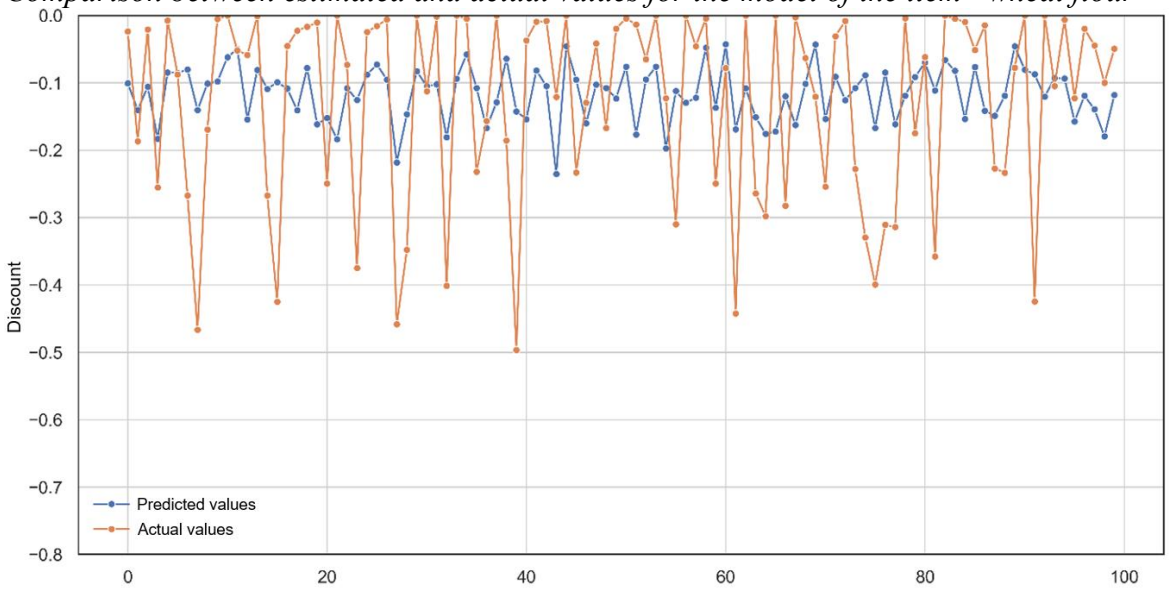


Figure 4

Comparison between estimated and actual values for the model of the item "wheat flour"



The predictions generated by the models for the items rice and wheat flour demonstrated a limited ability to follow the negative or positive peaks of the actual data. It is worth noting that the predictive models for the items cassava flour and French bread produced quite limited results. The former was unable to track larger variations of the item (remaining close to a simple mean), while the latter, which does not typically present significant discounts, produced estimates around an average discount between 0.1 and 0.2, failing to predict discounts above these values.

The models for the remaining items analyzed exhibited behavior similar to those previously presented (Figures 3 and 4), facing the same limitations in tracking more relevant variations. The performance of these models could be improved by incorporating a greater number of features, such as the time of year in which purchases were made, the microregion of the purchasing entities, and the number of companies participating in the bidding process.

Discussion and model validation

For public agencies, obtaining estimates of expected discounts in upcoming bidding processes is of great importance for planning and administrative purposes. According to Xerez (2013), public agencies' budgets estimate revenue and define expenditures, serving as an important source of information for decision-making regarding goals and priorities for a given period. Furthermore, the 1988 Federal Constitution established the requirement for the Multi-Year Plan, the Budget Guidelines Law, and the Annual Budget Law, aiming to link the public budget to institutional planning.

Using the estimated discount values for market prices, budget planning teams are able to correctly forecast the amounts to be allocated for such acquisitions, thereby enabling better allocation of public resources. However, estimates must be interpreted in light of their errors and considered only after careful risk assessment and analysis of external factors. Nevertheless, their complementary nature may support more data-driven decision-making.

Silva et al. (2023) argue that one of the main indicators of fraud in public procurement is the occurrence of overpricing, defined by Law 14.133/2021 as "a bid budget or contract price significantly higher than reference market prices" (Brazil, 2021). Overpricing may result in abnormal discount patterns in competitive bidding. In this regard, predictions generated by ML models, such as those presented in this study, may assist in detecting such variations when offered discounts are far below estimates, filtering possible cases of overpricing for subsequent verification of their compatibility with market values.

Similarly, when the maximum reference price is set below the actual market value, the tendency is for discounts to be lower than expected. In this case, discrepancies can be identified by comparing offered and estimated discounts, also revealing possible shortcomings in the contracting team's responsibility to negotiate proposal values.

Beyond these situations, predictive models can also add value in assessing potential cases of unfeasible proposals. Factors such as the inexperience of the winner, the number of competitors, deficient evaluation, and even irrational behaviors may lead bidders to submit

unrealistic offers, resulting in contract nonperformance, delays, or abandonment (Signor et al., 2022). Thus, excessively high discounts relative to the maximum reference price—provided this price reflects market conditions—may serve as a parameter for public administrations to anticipate risks and assess the real capacity of the bidder to fulfill the contract requirements.

Regarding the validation of the models used, we found that they were able to successfully execute training and produce relevant results, except for the items “cassava flour” and “French bread,” which failed to adjust to the natural variations of the test set, essentially yielding a simple average. However, estimation was satisfactorily achieved for the remaining items, with errors as low as 0.07% in the best experiments.

Because the data were obtained directly from an official government portal, the input data are considered of good quality, which is crucial for consistent analysis following testing and model evaluation (Morabito & Pureza, 2018). Atypical cases, such as unusually large discounts in specific records, require deeper exploration, representing a potential avenue for future research. Thus, the application of ML in discount forecasting may support regulatory agencies, internal audit offices, procurement teams, and entity managers in making data-driven decisions, as well as in preventing and detecting fraud and unfeasible contracts.

It is important to emphasize that the data collection and preprocessing stages in this study were more complex than initially expected, particularly with respect to two aspects: (i) the format in which the files were made available, and (ii) the lack of standardization of items.

The Court of Accounts of Paraná (TCE-PR) provides procurement search options through filters on a user-friendly platform. However, to obtain details about the items in each bidding process, it is necessary to access them individually. More comprehensive datasets are only available in annual compressed files, which contain multiple other compressed files, each with several XML documents.

Consequently, several steps were required to extract the desired information. Access would be facilitated if the data were provided in CSV format, for example, eliminating many preprocessing stages that currently hinder research and analyses that could benefit both the regulator and the audited entities.

The lack of item standardization required the application of natural language processing techniques so that the items studied could be categorized. Different agencies describe items in various ways, without an apparent standard.

Law No. 14.133/2021 recommends the use of electronic catalogs to standardize product specifications (Brazil, 2021). The adoption of standardized descriptions or indexing to a base

registry could assist in item analysis by the agencies. For example, the Court of Accounts of Mato Grosso (TCE-MT) maintains a Catalog of Materials and Services — a mandatory database of item specifications for all entities under its jurisdiction (https://servicos.tce.mt.gov.br/consulta-item). The federal government also offers such a tool through the "Compras.gov.br Catalog" (https://catalogo.compras.gov.br/).

These tools can support the application of machine learning models by simplifying item classification, thereby enabling more reliable analyses with reduced preprocessing effort.

Conclusions

The objective of this study was to evaluate ML models for estimating discount values offered by winning bidders in procurement processes for supplying items in the Brazilian basic food basket. To this end, Random Forest, ANN, and XGBoost techniques were applied to datasets of the 18 selected items.

As a result, an average RMSE of 0.122152 and an average MAE of 0.099196 were obtained among the selected models. The analysis of feature importance identified that the quantity tendered and the year of homologation were the attributes that most influenced estimation in most models. However, some variation exists, as these are models for distinct items, each subject to different external factors.

The ML models trained yielded satisfactory results for nearly all items, demonstrating that they can serve as a foundation for discount estimation in certain situations, as shown by the RMSE and MAE metrics. This study suggests that future research focus on a deeper analysis of outliers and specific variables that may also influence discount estimation but lie beyond the scope of this work, such as economic indicators (SELIC interest rate, inflation, among others).

The field of public procurement has undergone significant changes, with the New Public Procurement Law bringing both improvements and challenges for public agencies. The use of technology to increase the effectiveness and efficiency of processes can support the provision of high-quality public services to the population.

REFERENCES

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*. Brasília: Senado Federao, Coordenação de Edições Técnicas. 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.

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