A machine learning-based data matching pipeline for linking electronic invoices to public procurement processes

Lucas de Souza Vieira[1], **Jônata Tyska Carvalho**[1]

[1] Informatics and Statistics Department, Federal University of Santa Catarina (UFSC), Florianópolis, Brazil

lucas.souza.vieira@grad.ufsc.br, jonata.tyska@ufsc.br

Abstract. *The digitization of public procurement has created new opportunities for computational auditing to detect fraud and irregularities. However, the data is spread across multiple sources, often without a key that would allow direct integration. This work proposes a machine learning pipeline to establish correspondences between procurement and electronic invoice records. The methodology combines semantic similarity and syntactic distance metrics applied to item and document descriptions. Experiments conducted on real-world data, comprising 1,000 procurement processes and 9.8 million electronic invoices, demonstrated high effectiveness. The optimized Random Forest model achieved an F1-score of 0.98. The results indicate that the proposed approach is a promising solution for linking fragmented records of public procurement processes.*

1. Introduction

The digitalization of public administration has promoted the modernization of governmental processes through the adoption of digital technologies [Bomfim 2025]. As a consequence, public procurement processes have come to be conducted in digital environments, enabling the broad availability of information on procurement procedures on transparency portals, including the tender notice, the signed contract, the budget commitment, and the amounts effectively paid at the end of the process.

In the final stage of this workflow, the electronic invoice, issued by the creditor, stands out, validating the goods or services delivered and serving as fiscal proof of the transaction. The acronym NF-e, originally derived from the Portuguese term *Nota Fiscal Eletrônica*, is used throughout this paper to refer to electronic invoices, following the official terminology adopted in Brazil. In this context, the availability of digital data has significantly expanded the potential for extracting and analyzing information derived from procurement procedures, enabling the use of computational techniques for the investigation and detection of frauds and irregularities [Silva and Rocha 2010].

Public procurement represents a significant share of government expenditure and plays an important economic role. Government spending on public procurement in Brazil corresponds to approximately 16% of the country's Gross Domestic Product (GDP) [OECD 2022]. Given the magnitude of these resources, frauds and irregularities in public contracting generate substantial economic impacts. Studies indicate that between 10% and 25% of the value of public contracts may be lost due to corruption and fraudulent practices [UNODC 2013]. Based on this estimate, potential losses in Brazil could range from approximately US$ 36.9 billion to US$ 92.3 billion annually.

Therefore, one of the most common types of fraud in public procurement is overpricing [TCU 2011], characterized by the contracted company charging prices above market value. This type of practice can be identified by constructing price databases based on NF-es, enabling comparison of prices charged across different contexts for the same item or service. However, despite the potential of these approaches and the challenges inherent in constructing a reliable price database, public data sources often do not provide the invoices for completed purchases. Furthermore, the lack of standardization in data entry across public procurement records and NF-es poses additional challenges [Brasil 2024]. Problems such as incomplete or inconsistent descriptions, incoherent values, incorrect units of measurement, and incompatible quantities are common, largely due to the use of free-text fields.

These limitations hinder the identification and linking of entities across different datasets, compromising the effectiveness of public transparency [Soylu et al. 2022]. Therefore, the development of methods to establish links between procurement data and their corresponding invoices is necessary.

To fill this gap, this work proposes a machine learning-based approach to automate the linkage between public procurements and electronic invoices, leveraging the textual descriptions of items in both datasets. For this purpose, a labeled dataset containing real public procurement records was built, on which four supervised classification algorithms were evaluated: Random Forest, XGBoost, LightGBM, and Logistic Regression.

The training dataset was produced using a deterministic linkage heuristic, resulting in a distribution of 76% negative-matching records and 24% positive-matching records. Among the evaluated models, Random Forest presented the best performance, achieving a Precision of 98.3% and a Recall of 97.7%. Additionally, class balancing techniques, including Random Undersampling and SMOTE, as well as hyperparameter optimization procedures, were investigated. The results indicated that hyperparameter optimization yielded additional performance gains, whereas balancing techniques did not yield benefits due to the predominantly textual nature of the attributes used.

The results demonstrate the feasibility of the proposed approach for automatically linking public procurements to electronic invoices, achieving high reliability and substantially reducing incorrect associations.

2. Related Work

Recent literature has demonstrated a growing interest in the application of Artificial Intelligence (AI) and Natural Language Processing (NLP) to handle unstructured data in public procurement, particularly for fraud detection, description standardization, and the integration of heterogeneous databases. Such challenges are directly related to a lack of patterns and noisy textual descriptions in invoices and procurement records.

To understand the state of the art, Gomes and Colaço [Gomes and Colaço 2022] conducted a systematic literature mapping focused on the classification of incongruent descriptions in invoices. The authors analyzed 15 primary studies and found that supervised machine learning is the predominant approach, present in 60% of the studies, with an emphasis on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The study shows that the field is relatively new, with significant growth since

2021, and highlights an important gap: the scarcity of experiments with rigorous protocols that enable validation in real-world government auditing scenarios.

In the context of specific applications, Fontes et al. (2024) [Fontes et al. 2024] proposed DescMiner, a hierarchical classifier for fraud detection in electronic pharmaceutical invoices. The work addresses problems such as incorrect associations of Mercosur Common Nomenclature (NCM) codes and a lack of standardization in supplier descriptions. Using classical text-mining techniques, such as TF-IDF, the authors extracted structured attributes (e.g., active ingredient, dosage, and pharmaceutical form) and evaluated various similarity metrics. The results indicated the statistical superiority of Cosine Similarity over Euclidean and Manhattan distances, achieving an F1-Score of 99.82%.

Advancing to more general scenarios with limited data, Alves et al. (2025) [Alves et al. 2025] presented STEPMatch, an approach to record linkage for short, noisy descriptions, typical of NF-es. To address the scarcity of annotated data in Portuguese, the authors employed Cross-Lingual Learning (CLL) techniques, transferring knowledge from English corpora to the Brazilian context. The solution combines information retrieval methods, such as BM25, with semantic re-ranking using Cross-Encoder models built on the BERT architecture. This strategy achieved an F1-Score of 98.6%, demonstrating high effectiveness even in scenarios with low standardization.

In summary, the studies analyzed demonstrate a methodological evolution in the processing of fiscal data, ranging from approaches based on vector similarity within specific domains to the use of advanced language models for record linkage. However, challenges related to integrating diverse data sources and validating approaches in real-world auditing scenarios persist. In this context, this work seeks to advance the field by proposing an approach that links procurement data and invoices, explicitly addressing the problems of a lack of standardization and inconsistencies in textual descriptions.

3. Methodology

Figure 1 illustrates the methodological workflow adopted throughout this work. The process encompasses everything from integrating heterogeneous databases to evaluating machine learning models. The workflow is divided into four macro stages: (i) extraction and integration of public procurement and invoice data; (ii) generation of a stratified sample and mitigation of granularity disparities; (iii) feature engineering; and (iv) training and validation of the proposed classifiers.

3.1. Data Acquisition and Integration

The dataset was constructed based on an analysis of distinct open data sources from the state of Santa Catarina, which were divided into two categories. First, public procurement data originates from several sources: some contain the procurement notice and purchase items, while others are more complete, including commitments, liquidations, and payments. The second source is from the Santa Catarina State Department of Finance (SEF/SC) and contains detailed records of electronic invoices, including header metadata and invoice items.

The raw data were subjected to a data processing procedure aimed at standardization and inconsistency cleaning. Thus, the first step of the analysis consisted of searching for columns that would enable the relationship between the two sources, since there are

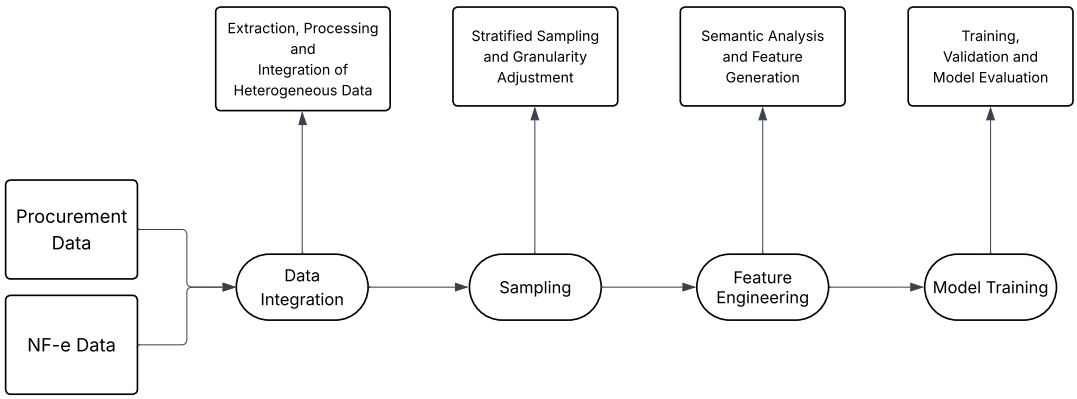


Figure 1. General workflow of the process

no direct or explicit linkage keys between the tables from the two origins. The tables used, and their original columns, are shown in Table 1.

Based on this exploratory analysis, columns with possible correspondences between attributes were identified, serving as the links for integration. These identified columns are presented in Table 2.

Based on the manually filtered columns, a linkage process was initiated using a deterministic heuristic, first to validate the linkage process and the relationship between the columns, and subsequently to serve as the dataset for training the models. The heuristic process is based on a structured workflow of filtering and exact matching between columns, described below:

- **Context Mapping:** Initially, the procurement context is mapped by joining the Commitment, Liquidation, and Procurement Process tables, extracting the operation dates and the identification document of the winning creditor (CNPJ or CPF). These credentials are then matched against the Electronic Invoice (NF-e) database to find linkage candidates.
- **Exclusion Filters:** Once the candidate set is defined, a sequence of filters is applied. The first is the *Temporal Filter*, which requires the NF-e issuance date to be equal to or later than the commitment date and, if available, earlier than the liquidation date. The second is the *Location Filter*, which validates whether the tax identification number (CNPJ) of the invoice recipient is the same as the CNPJ of the Procurement Management Unit.
- **Financial Validation:** Finally, records that survive the filters undergo value verification. The total net value of the NF-e is calculated and compared with the commitment value and with each liquidation installment individually. To absorb common tax rounding errors, a mathematical tolerance of R$ 0.50 was established in the matching process.

The NF-e that presents financial correspondence with the commitment or with one of its liquidations is then classified as a true link. This approach achieved high precision when public procurement processes contained complete information on commitments, liquidation, and payments.

Table 1. Original columns of the tables from public procurement and NF-e sources.

Table	Attributes
procurement_process	id_procurement_process, notice_number, procurement_object_description, estimated_total_value, bid_opening_date, management_unit_id, budgetary_unit_id, status, procurement_process_number.
procurement_item	id_procurement_item, procurement_item_description, approval_date, procurement_item_quantity, unit_of_measure, id_procurement_process.
management_unit	id_management_unit, management_unit_name, management_unit_zip_code, cnpj_management_unit, entity_id.
commitment	id_commitment, commitment_number, commitment_value, commitment_description, commitment_date, id_procurement_process, cnpj_cpf_creditor.
liquidation	id_liquidation, liquidation_date, liquidation_value, id_commitment.
commitment_payment	id_commitment_payment, payment_date, payment_value, id_liquidation, id_commitment.
nfe	id_nfe, issue_date, cnpj_issuer, cpf_issuer, issuer_name, issuer_trade_name, issuer_zip_code, cnpj_recipient, recipient_name, recipient_zip_code, cpf_recipient, access_key.
nfe_item	id_nfe_item, id_nfe, id_item, ncm_product, gtin_product, product_description, commercial_quantity, commercial_unit, commercial_unit_value, discount_value, freight_value, insurance_value, other_expenses_value, commercial_total_value, net_total_value, net_unit_value.

Table 2. Mapping of correspondences and linkage rules between public procurement and NF-e attributes.

NF-e Attributes	Public Procurement Attributes	Relationship Description
cnpj_issuer	cnpj_cpf_creditor	Unique identification of the supplier (creditor).
cnpj_recipient	cnpj_management_unit	Location validation: the invoice recipient must be the Management Unit associated with the procurement process.
issue_date	commitment_date, liquidation_date, payment_date	Temporal filter: the invoice issuance date must be $\geq$ the commitment date and $\leq$ the liquidation/payment date.
net_total_value	liquidation_value, payment_value	Financial consistency: value equivalence with a tolerance of R$ 0.50 to account for rounding differences.
product_description	procurement_item_description, commitment_description, procurement_object_description	Semantic and lexical analysis: textual correspondence between the invoiced item, the specific procured item, and the macro procurement object.
commercial_unit	unit_of_measure	Lexical analysis: correspondence between the units of measure in the procurement item and the NF-e item.
commercial_quantity	procurement_item_quantity	Volume validation: compatibility between requested items and delivered items.

However, its use as a definitive solution is limited by data availability. In practice, many procurement processes in public databases do not provide all the financial information required at the time of consultation. Consequently, the deterministic heuristic cannot establish links in these cases, reducing its coverage in large-scale applications.

Thus, the heuristic was used only to construct the labeled dataset. It is important to highlight that, because the deterministic heuristic requires strict, 100% agreement across all administrative, temporal, and financial constraints, its primary limitation is not a high rate of false positives, but rather low coverage, as it may fail to link valid pairs due to missing data or slight variations. To mitigate this limitation and increase coverage, the dataset generated for training incorporated relaxed candidate selection rules, using root CNPJ and spatial pre-filters, while ensuring that the final labels were based on strict financial and temporal agreement.

After identifying these true links, the financial attributes used exclusively for validation were removed from the training process. The machine learning model was trained exclusively using information available in the primary sources, such as item descriptions and product-related characteristics. The objective is to enable the model to learn matching patterns that identify links even in the absence of the financial information required by the original heuristic, effectively scaling up the matching process beyond the rigid boundaries of deterministic rules.

3.2. Sampling and Granularity Processing

With the baseline heuristic established, the dataset for training the supervised model was constructed. The data selection was based on procurement processes opened between 2020 and 2025 that already had confirmed links with NF-es through the baseline heuristic. To ensure representativeness of the sample and avoid bias from selecting only simple purchases, a stratified sampling technique was applied. The dataset was divided into four complexity classes based on the volume of commitments in the process: 1 to 10, 11 to 100, 101 to 500, and more than 500. From each stratum, 250 processes were randomly selected, totaling a sample of 1,000 procurement processes.

After defining the sample, it was matched with the procurement process items table. The objective of this stage was to extract the essential data for linkage, retrieving textual and numerical attributes such as the detailed description, unit of measure, and requested quantity.

Subsequently, to resolve the granularity difference and enable matching at the Procurement-NF-e level, the multiple items of each procurement process were aggregated into a single record. For the textual variables, quantities, units of measure, and descriptions were first grouped into lists. Then, a textualization algorithm was applied that concatenated this information element by element, creating a continuous text representation (e.g., quantity + unit + description). An example is shown in Table 3.

In the next stage, the necessary tables were joined to add structural information to the model. The management unit table was joined to add the unit name, ZIP code, and CNPJ; the commitment table was joined to add the date and creditor document (CNPJ or CPF); and the liquidation table was joined to add the liquidation date. The use of commitment and liquidation information was based on applying relaxed heuristic rules to obtain candidate NF-es.

Table 3. Consolidation of procurement items for granularity reduction.

Before: Individual Items with Repeated Header

Proc. ID	Procurement Object	Item ID	Item Description	Qty.	Unit
6888	Aquisições parceladas de medicamentos da revista ABC-FARMA...	175330	MEDICAMENTO DE REFERENCIA	1.0	UN
6888	Aquisições parceladas de medicamentos da revista ABC-FARMA...	175331	MEDICAMENTO GENERICO E SIMILIAR	1.0	UN

After: Single Record with Textual Concatenation

Proc. ID	Procurement Object	ID List	Complete Item Description
6888	Aquisições parceladas de medicamentos da revista ABC-FARMA...	[175330, 175331]	1.0 UN de MEDICAMENTO DE REFERENCIA. 1.0 UN de MEDICAMENTO GENERICO E SIMILIAR

Note: Original procurement object and item descriptions are presented in Portuguese, as they correspond to real procurement records. Example translations: "Aquisições parceladas de medicamentos da revista ABCFARMA" = "Installment-based procurement of medicines listed in the ABCFARMA catalog"; "MEDICAMENTO DE REFERENCIA" = "Brand-name medicine"; "MEDICAMENTO GENERICO E SIMILIAR" = "Generic and equivalent medicine".

To extract candidate NF-es, a filtering strategy was adopted. During database extraction, the search was restricted to the creditor documents (CNPJ) in the sample, using a flexible spatial pre-filter based on the first 8 digits of the CNPJ and the first 3 digits of the ZIP code. The use of the CNPJ root is justified by the particularities of public administration: it is common for a procurement process to be conducted by a central entity, such as a city hall, while the NF-e is delivered and billed in the name of an associated municipal department or entity. Since they are public entities under the same administration, the first eight digits indicate the parent organization to which the CNPJ belongs. Following the same logic, the first three digits of the ZIP code correspond to the city and its region. Additionally, the same temporal filter used in the baseline heuristic was applied, confirming that the NF-e issuance date was perfectly contained between the commitment date and the liquidation deadline.

Next, the items belonging to the candidate NF-es were extracted and joined. The same granularity strategy used in procurement processes was applied, grouping all products on an invoice into a single record. An example of this processing is shown in Table 4. Finally, the target variable for the supervised model was constructed, assigning 1 (true link) when the procurement process associated with the NF-e matched the procurement process, and 0 (false) otherwise.

To finalize the dataset, identifying and redundant variables were removed. All attributes used exclusively for structural matching and deterministic filtering were discarded. This cleaning process aims to avoid model overfitting by preventing the algorithm from memorizing specific entities rather than learning linkage patterns.

3.3. Feature Engineering

With the sample in hand, the feature engineering process began with semantic analysis using the "paraphrase-multilingual-MiniLM-L12-v2" language model. After vectorizing

Table 4. Textualization and granularity adjustment process for NF-e items.

Before: Individual NF-e Items

NF-e ID	Item ID	Item Description	Qty.	Unit
1367386	927	ATADURA CREPE 12 CM PCT C/12	679.0	duzia
1367386	937	ATADURA CREPE 15 CM PCT C/12	592.0	pacote
1367386	948	BOBINA GRAU CIRURGICO 250MM X 100MM	285.0	rolo

After: Single Textualized Record

NF-e ID	ID List	Complete Item Description
1367386	[927, 937, 948]	679.0 duzia de ATADURA CREPE 12 CM PCT C/12. 592.0 pacote de ATADURA CREPE 15 CM PCT C/12. 285.0 rolo de BOBINA GRAU CIRURGICO 250MM X 100MM

Note: Original product descriptions and units are presented in Portuguese, as they correspond to real NF-e records. Translation of the examples: "ATADURA CREPE 12 CM PCT C/12" = "12 cm Crepe Bandage Pack (12 units)"; "ATADURA CREPE 15 CM PCT C/12" = "15 cm Crepe Bandage Pack (12 units)"; "BOBINA GRAU CIRURGICO 250MM X 100MM" = "Surgical Grade Roll 250 mm × 100 mm"; "duzia" = "dozen"; "pacote" = "package"; "rolo" = "roll".

the unique texts into embeddings, cosine similarity was used to assess the contextual similarity between the datasets. Three semantic score columns were generated by comparing NF-e data with the source procurement process: Procurement Object vs. NF-e Item Description; Procurement Item Description vs. NF-e Item Description; and Complete Procurement Item Description vs. Complete NF-e Item Description.

The next step in feature engineering involved lexical analysis of the data using Fuzzy Similarity. To capture spelling variations, abbreviations, and typographical errors, the Weighted Ratio metric was used. This technique was applied in two validation fronts: first, the structured quantity and unit of measure strings (e.g., "100 units", "100 g") from the procurement items were compared against those from the NF-e; subsequently, the spelling correspondence between the name of the purchasing institution and the name of the recipient in the NF-e was measured. To mitigate false positives, a preprocessing stage was implemented to remove institutional noise and generic terms (e.g., "City Hall", "Municipality", "Fund", "Department", and connectors). Finally, all original fields underwent structural cleaning, retaining only the processed features.

3.4. Model Training

Model training began by preparing the training and test sets using the holdout technique. The dataset was partitioned into 80% for training and 20% for testing. This division was performed in a stratified manner to preserve the original class proportions of the target variable in both sets.

The models used in the experiment were Random Forest, XGBoost, and LightGBM, chosen for their efficiency in binary classification of tabular data. Logistic Regression was used as the baseline. After training, the models were evaluated using Precision, Recall, F1-score, and AUC-ROC.

Finally, the importance of each attribute was extracted from the best-performing model to determine the variables' weights in the classification process. By analyzing

feature importance coefficients, it was possible to rank the mathematical and natural language processing features that contributed most to link prediction.

4. Experiments and Results

The application of the feature engineering pipeline resulted in a robust dataset, totaling 9,810,171 candidate combinations processed from the sample of 1,000 procurement processes. The dataset presented a slight class imbalance, with the target variable distributed at approximately 76% for the majority class (absence of linkage) and 24% for the minority class (confirmed linkage).

The obtained results demonstrate that decision tree-based algorithms significantly outperformed the Logistic Regression baseline. As presented in Table 5, the Random Forest model stood out as the most efficient classifier for the proposed problem, achieving an F1-score of 98%, a Recall of 97.7%, and a Precision of 98.3%, indicating a high capacity to correctly identify true links without compromising the occurrence of false positives.

Given the nature of the problem, false links, that is, false positives, represent a major issue, as they can generate auditing errors and false accusations; therefore, the Precision metric was prioritized in the analysis. Random Forest achieved high precision, demonstrating superior reliability in identifying true links. In contrast, the Logistic Regression model, although computationally simpler, presented the lowest performance, with a Precision of 74.7%, evidencing that the relationships among the extracted textual and mathematical variables exhibit complex non-linear patterns that require models with greater capacity.

Table 5. Performance comparison of the models on the test set (N = 1,962,035)

Model	Precision	Recall	F1-Score	ROC-AUC	Accuracy
Random Forest	**0.9826**	**0.9768**	**0.9797**	**0.9990**	**0.9852**
XGBoost	0.8417	0.8512	0.8463	0.9449	0.8862
LightGBM	0.8238	0.8381	0.8305	0.9350	0.8735
Logistic Regression	0.7467	0.7207	0.7317	0.8713	0.8141

The superior performance of Random Forest can be attributed to its robustness to the class imbalance in the dataset, as well as its intrinsic ability to capture interactions among the developed NLP features (Embeddings and Fuzzy).

4.1. Decision Analysis and Confusion Matrix

To further understand the behavior of the Random Forest classifier, an analysis of its confusion matrix and feature importance ranking was conducted. Figure 2 presents the distribution of predictions on the test set. The model exhibited an extremely low false positive rate, which translates into the high precision previously reported.

Additionally, the explainability of the model was evaluated through feature importance, as illustrated in Figure 3. The attribute with the highest decision weight was `simil_fuzzy_mu_recipient`, indicating that the syntactic divergence between the management unit and the invoice recipient is the main indicator of inconsistency.

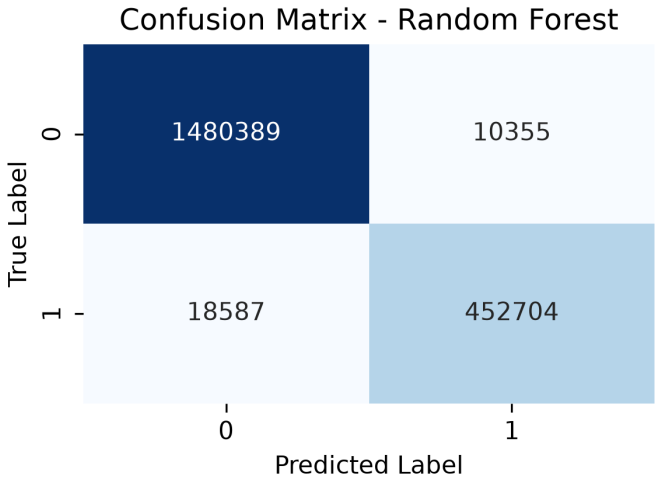


Figure 2. Confusion Matrix of the Random Forest model (Test Set)

Subsequently, the NLP variables (`simil_semantic_complete_text`, `simil_semantic_item_product` and `simil_semantic_object_product`) demonstrated a similar role, being responsible for identifying entities that refer to the same products. Finally, `simil_fuzzy_quantity_unit` was the attribute with the lowest influence, due to its low information content.

4.2. Data Balancing and Hyperparameter Optimization

Building on the previously best-performing model, the *Random Forest*, an empirical study was conducted to evaluate the impact of class-balancing techniques on system precision. The methodology employed Stratified Cross-Validation ($K = 5$) to preserve the original class proportions (*baseline* of 76/24) during the evaluations. Three sampling approaches were structured: *Undersampling*, SMOTE, and a hybrid approach, SMOTETomek. These techniques were tested by enforcing three new progressive target distributions: 70/30, 60/40, and 50/50.

The results demonstrated the Precision-Recall trade-off. The best performance occurred in the *Baseline* scenario (without balancing intervention), achieving a Precision of 98.1%, a Recall of 97.6%, and an *F1-Score* of 97.9%. The degradation of the model under balancing is justified by the vector nature of the predictive variables (semantic and linguistic distances). *Undersampling* removed complex negative examples, reducing the model's ability to reject false positives with high textual similarity. Simultaneously, SMOTE introduced noise by interpolating mathematical representations of texts that do not exist in the reality of public procurement, confusing the decision tree partitions.

As a statistical consequence of this manipulation, the search for extreme symmetry (50/50) through *SMOTE* maximized Recall to 98.3% (compared to 97.6% in the *baseline*), but at the expense of Precision, which decreased to 96% , reducing the *F1-Score* to 97.1%. Thus, it was decided to proceed to the hyperparameter optimization stage while maintaining the original imbalanced dataset, delegating classification rigor to Cost-Sensitive Learning to algorithmically penalize false positives.

Thus, a hyperparameter optimization stage was conducted to improve perfor-

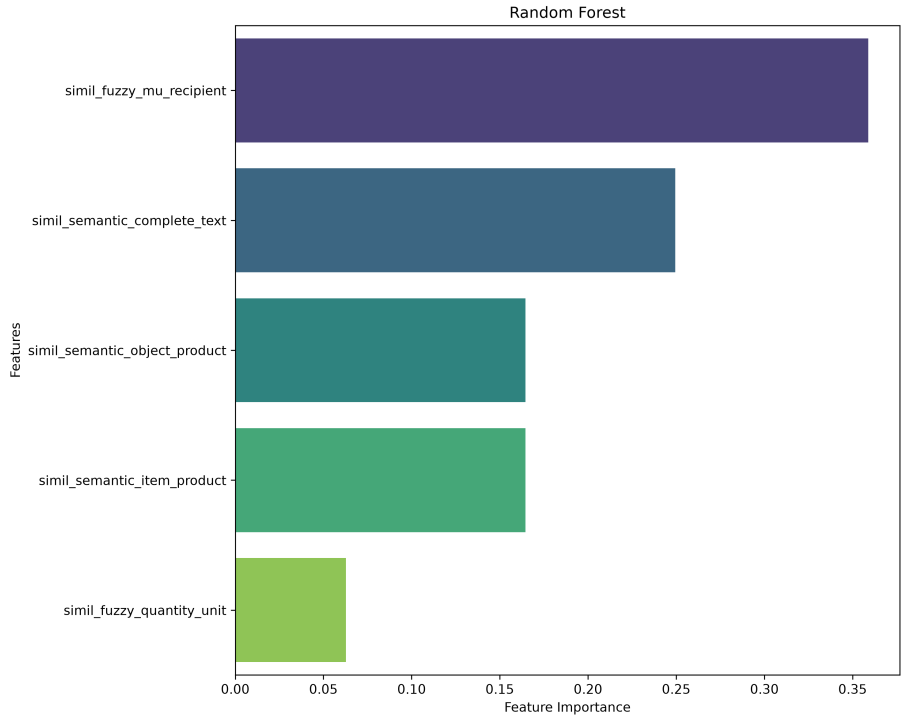


Figure 3. Feature Importance for the Classification Process

mance. For this purpose, the random search method was used, defining a search space composed of 5,760 possible hyperparameter combinations. The algorithm was configured to perform 50 iterations, evaluating a random sample from this search space.

Due to the large volume of data available for training, the optimization process was conducted on a stratified random sample corresponding to 10% of the training dataset, preserving the original class distribution. Each combination was evaluated using five-fold stratified cross-validation. As the selection criterion, the Precision metric was adopted, and various class weighting strategies were investigated to address class imbalance. At the end of the process, the hyperparameter values that achieved the best average performance during cross-validation were selected. The optimal configuration consisted of 1000 estimators, a maximum depth of 50, a minimum number of samples required for splitting equal to 2, a minimum number of samples per leaf equal to 1, a *log2* feature selection strategy, and class weighting $=\{0{:}1.2,\ 1{:}1\}$. When evaluated on the test set, the optimized model achieved a Precision of 98.4% , representing a slight improvement over the baseline configuration. In contrast, a small reduction was observed in the remaining metrics, with an F1-score of 98% and a Recall of 97.6% , while maintaining overall performance virtually equivalent to the original model.

5. Conclusion

This work introduced a machine learning pipeline to bridge the traceability gap between public procurement records and electronic invoices, overcoming structural barriers such as missing shared keys and noisy free-text descriptions. Evaluated on a massive real-world dataset, the optimized Random Forest model demonstrated high reliability with a precision of 98.4%, confirming its suitability for governmental auditing.

By enabling robust, large-scale data linkage, the proposed approach establishes a scalable technological foundation for continuous automated financial auditing. Future work will focus on incorporating additional financial and contractual variables—such as quotation and contract details—to further enhance the predictive capability and precision of public procurement oversight.

Declaration of AI Use.

During the preparation of this manuscript, the Gemini 3.1 Pro and ChatGPT systems were employed to assist with code and grammar revision. All content processed by this tool has been verified by the authors, who take full responsibility for the accuracy and integrity of the information presented.

References

Alves, A. L. F., Baptista, C. d. S., Diniz, J. I. M. S., Mendes, F. I. d. L., and Cunha, M. Q. (2025). An approach for product record linkage using cross-lingual learning and large language models. In *Proceedings of the International Conference on Enterprise Information Systems*.

Bomfim, P. B. (2025). Transformação digital na administração pública e privada: Impactos da digitalização na eficiência organizacional, transparência e inovação de serviços. In de Barros Ahrens, R., editor, *Administração Pública e Privada: Novas Tendências e Perspectivas*, volume 9, chapter 25.

Brasil, T. (2024). Recomendações e desafios técnicos do portal nacional de contratações públicas: Qualidade de dados. Technical report, Transparência Brasil.

Fontes, R. S., Colaço Júnior, M., de Oliveira, F. R., and de Medeiros Valentim, R. A. (2024). Avaliação experimental de um classificador para apoiar a detecção de fraudes em notas fiscais públicas de medicamentos. *Revista Gestão & Tecnologia*, 24(1).

Gomes, W. and Colaço, M. (2022). Applications of artificial intelligence for auditing and classification of incongruent descriptions in public procurement. In *Anais do XVIII Simpósio Brasileiro de Sistemas de Informação*, Porto Alegre, RS, Brasil. SBC.

OECD (2022). *OECD Responsible Business Conduct Policy Reviews: Brazil*. Paris.

Silva, C. V. S. and Rocha, H. A. d. (2010). Técnicas de mineração de dados como apoio às auditorias governamentais. *Revista da CGU*, 5(8):29–40.

Soylu, A., Corcho, Ó., Elvesæter, B., Badenes-Olmedo, C., Yedro-Martínez, F., Kovacic, M., Posinkovic, M., Medvešček, M., Makgill, I., Taggart, C., Simperl, E., Lech, T. C., and Roman, D. (2022). Data quality barriers for transparency in public procurement. *Information*, 13(2).

TCU (2011). Fiscobras 2011 - sobrepreço/superfaturamento lideram irregularidades em obras. Technical report, Tribunal de Contas da União.

UNODC (2013). *Guia de Combate à Corrupção em Compras e Contratações Públicas e na Gestão das Finanças Públicas*. UNODC.