Detecting Fraud in Public Procurement: A GMM-Based Approach to Analyzing Tender Data

Fernando Augusto Schmitz[1 2], **Lívia Ferrão**[1], **Matheus Machado dos Santos**[1], **Márcio Castro**[1], **Jônata Tyska Carvalho**[1]

[1]Federal University of Santa Catarina (UFSC), Brazil
[2]Santa Catarina Prosecutor's Office (MPSC), Brazil

faschmitz@mpsc.mp.br, livia.ferrao16@hotmail.com,
matheus.m.santos@posgrad.ufsc.brmarcio.castro,jonata.tyska@ufsc.br

Abstract. *Corruption and bid rigging in public procurement distort competition and increase the costs of products and services for public institutions, causing problems in different societal domains. The current availability of public data in digital format brings opportunities for applying machine learning to build solutions that help to deal with corruption. However, there are many challenges, like data sparsity and modeling complexity. Furthermore, confirmed cases of fraudulent tenders are limited, making applying traditional supervised learning techniques unfeasible. This work proposes a novel methodology for analyzing patterns using Gaussian Mixture Models (GMM) to identify suspicious bidding patterns when only a few fraudulent cases are known. Our methodology tests the similarity of unlabeled tenders, which can be fraudulent or not, with the fraudulent cases in different subspaces for defining a risk indicator. We run experiments in a dataset with tender data for acquiring heavy equipment purchases in which only a few cases are known as fraudulent. Results showed that our GMM-based methodology effectively provides a risk indicator ranking, highlighting risky tenders, making it a valuable tool for public agencies to enhance transparency and accountability in procurement.*

1. Introduction

Corruption is defined as the abuse of power for personal or third-party gain, encompassing the misappropriation of resources and the manipulation of government policies and processes [Rothstein 2011]. The Corruption Perceptions Index (CPI) by Transparency International [International 2022] ranks Brazil 94th among 180 countries assessed, with over half of the countries scoring below 40 on a scale from 0 (highly corrupt) to 100 (very clean). This highlights a significant corruption problem within the country, particularly in public processes such as government tenders.

Fraud in public tenders, especially through bid rigging and restricting competition, often called favoritism, is a common method of corruption. This practice involves manipulating the tender process to select a particular contractor. Typical methods include tailoring the specifications to the unique capabilities of the preferred vendor, unjustified requirements that exclude potential bidders, or otherwise skewing the bidding process in favor of a pre-selected party. Such activities undermine the principles of fairness, competitiveness, and transparency that govern public procurements.

One of the significant challenges in combating tender fraud is the lack of labeled data on fraudulent activities[Rabuzin and Modrušan 2019]. Without clear examples of fraud, developing models to detect such activities can be particularly challenging. The scarcity of labeled data limits the ability of investigative bodies to train predictive models that can identify irregular patterns and potential fraud in tender processes.

Operation Patrola, initiated by the Prosecutor's Office of Santa Catarina (MPSC), brought numerous instances of fraudulent public tenders to light. Data from this operation and information reported by municipalities to the Court of Accounts of Santa Catarina (TCE/SC) provide a unique dataset of tenders with potential fraudulent activities. This dataset is a foundation for developing a methodology to identify other potentially fraudulent tenders through feature modeling and clustering.

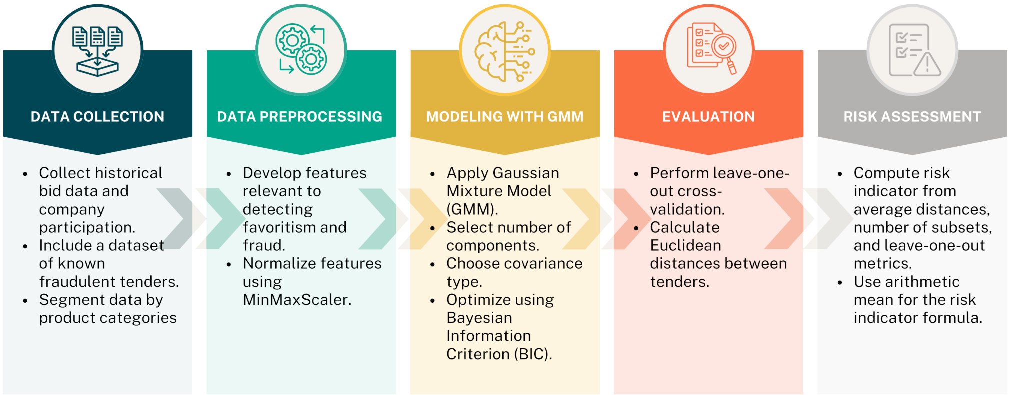
This study aims to propose a generic methodology for detecting potential favoritism in public tenders. The idea behind the methodology is to start from a known base of a few fraudulent cases and identify other processes similar to these frauds across a wide range of feature subspaces. This approach aims to provide a comprehensive framework that can be applied to various datasets from different tenders, enhancing the detection of fraudulent activities and improving transparency and accountability in public procurement. The experiments demonstrate that the methodology was able to identify similarities in one or more feature subspaces in 3,016 out of 5,652 tenders, across four different product categories in the context of "Operação Patrola". Additionally, the methodology was effective in providing a risk ranking that highlights the tenders with the highest potential for irregularities. The main contributions of this work are two-fold: first, a machine learning methodology that produces a ranking of tenders that are more similar to fraudulent cases in different subspaces, indicating high-risk tenders that require investigation. And secondly, a dataset with real cases for studying favoritism detection in public procurement processes.

2. Related Work

In public procurement, detecting corruption and favoritism remains a significant challenge that undermines the integrity of administrative processes and public trust. Recent technological advancements have enabled researchers and practitioners to employ sophisticated data mining techniques to tackle these issues effectively. This paper reviews various studies that have leveraged such methodologies across different jurisdictions, highlighting the effectiveness and challenges in identifying corrupt activities within public procurement systems.

For example, Torres-Berru and Lopez-Batista applied an array of clustering algorithms including K-Means, Self-Organizing Maps (SOM), Support Vector Machines (SVM), and Principal Component Analysis (PCA) in Ecuador. Their study achieved an impressive 95% accuracy in identifying anomalies in procurement contract qualifications, demonstrating the robust potential of these techniques in uncovering favoritism and corruption.[Goryunova et al. 2021]

Similarly, in Russia, Goryunova and Baklan tackled the challenge of limited labeled data by utilizing Positive-Unlabelled (PU) learning to analyze single-bidder auctions, often seen as corruption indicators. Their findings revealed that 53.86% of these auctions were suspicious, with corruption markers such as matching



DATA COLLECTION

- Collect historical bid data and company participation.
- Include a dataset of known fraudulent tenders.
- Segment data by product categories

DATA PREPROCESSING

- Develop features relevant to detecting favoritism and fraud.
- Normalize features using MinMaxScaler.

MODELING WITH GMM

- Apply Gaussian Mixture Model (GMM).
- Select number of components.
- Choose covariance type.
- Optimize using Bayesian Information Criterion (BIC).

EVALUATION

- Perform leave-one-out cross-validation.
- Calculate Euclidean distances between tenders.

RISK ASSESSMENT

- Compute risk indicator from average distances, number of subsets, and leave-one-out metrics.
- Use arithmetic mean for the risk indicator formula.

Figure 1. Methodology Diagram

bids to the reserve price and no previous interactions between bidders and procurers [Goryunova et al. 2021]. This underscores the complexity of detecting corruption, which often involves interpreting subtle indicators that may not be overtly illegal but suggest irregularities.

In Croatia, the work of Rabuzin and Modrušan emphasized the importance of analyzing single-bid tenders due to similar issues with labeled data. They used the number of bids as a proxy for suspicious activities, advocating for a meticulous review of tender documents that provide essential procurement details. Their research utilized text-mining techniques and machine-learning methods, such as Naive Bayes, logistic regression, and support vector machines, to develop a model that detects suspicious one-bid tenders. Their findings demonstrated that knowledge extracted from tender documentation can effectively identify suspicious tenders, highlighting the significance of detailed text analysis in corruption detection [Rabuzin and Modrušan 2019].

In Brazil, implementing a Decision Support System using data mining algorithms, graph theory, clustering, and regression analysis marked a significant advancement in fraud detection within public procurement. This system has substantially enhanced the identification of corruption risk patterns, aiding major law enforcement operations and demonstrating the effectiveness of systematic risk detection approaches [Velasco et al. 2021].

Building on existing data mining approaches for detecting public procurement fraud, our research introduces a novel methodology designed to identify favoritism. This methodology uses a few known fraud cases to detect similar patterns in various subspaces. This methodology extends previous works by focusing on a broader range of feature subspaces, enhancing adaptability across different tenders and datasets. While earlier studies primarily targeted specific corruption indicators within limited contexts, our approach provides a comprehensive framework that is both flexible and scalable.

3. Methodology

This study introduces a methodology that utilizes Gaussian Mixture Models (GMM) to examine tenders to bolster transparency and curb malpractices in public procurement. The proposed model ranks unlabeled tender cases based on their similarity to known fraudulent cases. This ranking is calculated by considering the similarity of each tender to fraudulent cases across multiple feature subspaces. Furthermore, the proposed methodology can be applied in any scenario where few fraudulent cases are known for a given product, allowing for its application across multiple datasets and procurement contexts, as illustrated in the methodology diagram (Figure 1).

The analysis begins with the assumption of a pre-segmented dataset, either by product or as a single product dataset. This study utilizes a unified dataset combining historical data on bids and company participation with another dataset of known fraudulent tenders. It is important to note that the construction and segmentation of the dataset are beyond the scope of this paper, focusing instead on the application of the methodology. From this dataset, it is essential to enrich it with features that possibly correlate with the main objective of detecting favoritism in tenders. In this work, we propose and test a new methodology with a given set of features, such as the proportion of winning bids from a particular company, prior meetings with the managing unit, and the number of tenders a company has participated in, as described in Table 3. However, the more possibly relevant features are created, the better the risk estimate provided by our methodology will be. In other words, using more features can provide deeper insights and enhance the robustness of the fraud detection process.

Our methodology employs an unsupervised learning model, the Gaussian Mixture Model (GMM), which is executed separately for each product category to tailor the analysis to specific procurement contexts. This model tests several applications using all possible combinations of feature subspaces, ensuring that each product's unique characteristics are appropriately considered. The GMM is a probabilistic statistical model that assumes all data instances are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Due to the GMM model's sensitivity to variable scales, a normalization technique, specifically MinMaxScaler, adjusts all features to the same range, generally from 0 to 1. This normalization is crucial as differences in scales can disproportionately influence the learning process, affecting the model's distribution of weights and the convergence of the Expectation-Maximization (EM) algorithm. By equalizing the scales, each feature contributes equally to forming Gaussian components, preventing features with greater amplitude from dominating the learning process and ensuring a more balanced and effective analysis.

For each subset of evaluated features, the number of components and the covariance matrix settings used in the model must be selected, which can be Spherical, Tied, Diag, and Full. These types of covariance in GMM define the variance and correlation structure among the variables of each component, accommodating different data complexities. The methodology uses the Bayesian Information Criterion (BIC) to determine the optimal number of components and the type of covariance in a process known as hyperparameter optimization. The BIC criterion seeks to balance how well the model fits the data and its complexity, favoring simpler models that still adequately capture the data's structure.

After selecting the best GMM configuration, the model's ability to group the known frauds into the same cluster is evaluated. A criterion for selecting a feature subspace as one of the subspaces that will be considered for calculating the similarity of unlabeled cases to fraudulent cases is defined based on a threshold where only feature subsets that manage to group a defined percentage of the data into the same cluster are utilized. To operationalize this, the predict method is employed to assign each data point to the most probable cluster based on the learned parameters of the Gaussian components.

An adapted technique of leave-one-out cross-validation is employed to evaluate the performance of the best model selected according to the BIC criterion. For each feature subspace, a known fraudulent tender is left out, and after running the model with the test set, it is tested whether the tender left out is grouped with at least another known fraudulent tender. This results in a relative recall, given that we only know some positively labeled cases, assuming that some unlabeled instances are also positive and should be flagged as suspect for further investigation.

Finally, a possible risk indicator for evaluating each tender could be the inverse of the average Euclidean distance to the known fraudulent tenders in the selected feature subspaces (since lower distances indicate a higher risk), the number of clusters using the selected feature subsets where the tender appeared, and the average relative recall value of that GMM setup. These variables are combined by weighting each normalized metric (α for the Inverse Normalized Distance, β for the Normalized Subsets, and γ for the Normalized Relative Recall) to reflect better their relative importance in indicating risk. The adjusted formula is shown in Equation 1:

$$\text{Risk} = \frac{\alpha \times (1 - \text{Norm. Distance}) + \beta \times \text{Norm. Subsets} + \gamma \times \text{Norm. Recall}}{\alpha + \beta + \gamma} \tag{1}$$

This weighting scheme allows for the calibration of the metric such that labeled fraudulent cases achieve higher risk indicators. Adjusting the weights based on empirical evidence or expert consultation can help ensure that the most indicative metrics of fraudulent activity are emphasized.

To complement the detailed description of the methodology, Figure 2 illustrates the application of the Gaussian Mixture Model (GMM) in different feature subspaces. Each quadrant of the figure represents a distinct subset of features used in the analysis. The red points indicate known fraudulent cases, the gray points represent other cases, the yellow points are weak suspects, and the green points are strong suspects. The yellow points vary among the different subspaces, while the green points remain consistent across all subspaces. This indicates that cases frequently grouped together in different subspaces are stronger suspects. The proximity of the points within each feature subspace showcases how different combinations of features can affect the distance, highlighting the importance of appropriately selecting feature subspaces for effective fraud detection. The visual representation underscores the methodology's ability to discern patterns and similarities between known fraudulent cases and potential suspects based on their feature characteristics.

4. Experiments

The application of this methodology is exemplified through a case study involving "Operação Patrola" an investigation initiated by the Prosecutor's Office of Santa Catarina (MPSC) to address widespread corruption and fraud in public procurement of heavy machinery. This case study provides a concrete example of how the methodology can be used to provide a ranked list of tenders based on their fraud risk, determined by their similarity to known fraudulent tenders in different feature subspaces. By leveraging data

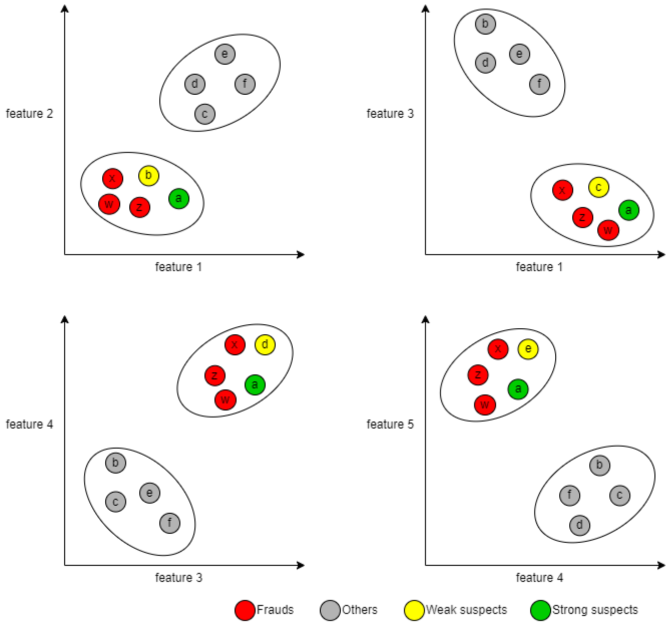


Figure 2. GMM application in different feature subspaces. Red points are known frauds, gray are other cases, yellow are weak suspects (vary across subspaces), and green are strong suspects (consistent across subspaces). Frequent identification across subspaces indicates stronger suspicion.

from the operation, including information on identified fraudulent tenders, enrichment with historical bid data, and detailed analysis using defined features and advanced modeling techniques, the methodology demonstrates its effectiveness in uncovering patterns of fraud and aiding in the efforts to enhance transparency and accountability in public procurement.

4.1. Dataset

“Operação Patrola” initiated by the Prosecutor’s Office of Santa Catarina (MPSC), addresses widespread corruption and fraud in public procurement of heavy machinery. This operation began with significant legal actions against officials, including a former mayor of Tangará, who were implicated in manipulating tenders to siphon public funds, with misappropriated amounts exceeding R$ 569,000 ([Ministério Público de Santa Catarina 2022b]).

As the investigation unfolded, it was expanded with the support of the Special Anti-Corruption Group (GEAC) due to the complex network of corruption. The operation revealed a pattern where public officials and businessmen orchestrated bribes ranging from R$ 15,000 to R$ 45,000 per machine, leading to total illicit gains of over R$ 6 million. These findings have spurred ongoing efforts to recover the misappropriated funds, underscoring the operation’s significant impact on enhancing transparency and accountability in public dealings ([Ministério Público de Santa Catarina 2022a]).

Following revelations from ”Operação Patrola,” the need for stringent oversight of public procurement, particularly in municipalities, is highlighted. The e-Sfinge system, operated by the Tribunal de Contas do Estado de Santa Catarina (TCE/SC), is central to these efforts. This system gathers and consolidates public account data from municipalities, enhancing the transparency and oversight of fiscal management across the region.[Tribunal de Contas do Estado de Santa Catarina 2024]

Our examination of fraudulent tender proportions across various product categories, as detailed in Table 1, highlights distinct patterns of fraud susceptibility. It is crucial to note that the tenders not categorized as fraudulent are, in fact, unlabeled and their status remains unknown, potentially obscuring additional instances of fraud. To address this, our Gaussian Mixture Models (GMM) are applied separately to each product category ensuring tailored analysis that captures category-specific nuances. Subsequently, these analyses are collectively examined to identify overarching trends and anomalies, thus enhancing the robustness and specificity of our fraud detection methodology.

Product	Fraudulent Tenders	Total Tenders	Proportion (%)
Crawler Tractor	18	182	9.89
Excavator	33	4400	0.75
Motor Grader	22	251	8.76
Road Roller	19	819	2.32

Table 1. Proportion of Fraudulent Tenders by Product Category

In our analysis, a summarized table of statistical measures, as shown in Table 2, was utilized to comprehensively understand the key features within our dataset. This table, delineating count, mean, standard deviation, and other descriptive statistics for variables such as unit price, number of participants, and win rates, serves as an tool in our GMM-based methodology. It provides a quantitative baseline that aids in identifying abnormal patterns and assessing risk indicators.

	unit_price	num_partic	win	met	num	period	duration	unique
Mean	37,605.92	27.21	0.103	0.804	252.12	2,088.66	55.85	0.037
Std	125,676.31	13.04	0.187	0.397	278.78	1,229.28	27.36	0.027
Min	24.58	1.00	0.000	0.000	1.00	0.00	0.00	0.009
25%	119.00	33.00	0.049	1.000	46.00	950.00	64.00	0.033
50%	140.00	33.00	0.058	1.000	159.00	2,776.00	64.00	0.035
75%	200.00	36.00	0.075	1.000	351.50	3,080.00	73.00	0.035
Max	1,680,000.00	46.00	1.000	1.000	1,449.00	4,275.00	245.00	0.600

Table 2. Summary Statistics for Selected Variables

Reliable systems and databases are crucial data sources for analytical models that detect irregularities in procurement processes. By delivering robust data on municipal management, they facilitate the application of advanced techniques to identify and prevent fraudulent activities in public tenders. This synergy between comprehensive data collection and sophisticated data analysis assists public agencies in combating corruption, safeguarding public funds, and enhancing governance. This approach underscores the importance of having dependable and transparent data infrastructures to support effective public administration and accountability.

4.2. Data Preparation

The dataset for our study, derived from ”Operação Patrola,” provided a foundational basis for developing a robust analysis of procurement trends and potential fraud. To enhance

Table 3. Features for Assessing Procurement Processes

Feature	Description
Win	The proportion of winning bids from a particular company within the same Managing Unit raises concerns. Companies that consistently secure contracts from a single procurer yet fail to succeed in auctions elsewhere prompt critical inquiries regarding the fairness of the procurement process. Such patterns may indicate preferential treatment or suggest potential irregularities.
Met	Assess whether a company has had prior meetings with the Managing Unit in previous tenders. It serves as an indicator for identifying interactions between entities. A MET value of 1 indicates that a meeting occurred, while a MET value of 0 indicates no prior meetings.
Num	Tracks the number of tenders a company has participated in. Companies that have participated in only a few tenders are often labeled as 'one-day' companies, which may suggest they have limited experience or lack the capacity to manage larger or more varied contracts.
Period	Represents the total number of days a company has been active within the dataset. This metric helps distinguish between established and 'one-day' companies, which typically indicate a lack of market presence.
Duration	Measures the period between the opening and the homologation date of a tender. Serves as an indicator of the tender process's efficiency. However, an unusually short duration may suggest that the process was rushed, potentially limiting competitive bidding.
Unique	Reflects the proportion of single winners in tenders managed by a specific Managing Unit. A low diversity of winners, where a single company repeatedly wins tenders (indicated by a Unique score of 0), could suggest potential favoritism within the Managing Unit.
Num of Participants	Quantifies the total number of bidders involved in a tender process. This figure helps assess the level of competition within tenders. A higher number of participants typically indicates a healthy competitive environment.
Unit Price	Indicates the unit price of the equipment being procured. This feature is crucial for assessing the cost-effectiveness and price competitiveness of bids. Significant deviations from average market prices may signal potential issues such as overpricing or bid rigging.

this dataset, we prepared the data by normalizing values and selecting specific feature subsets that are most indicative of irregular activities. This careful preparation ensured that our GMM model could effectively leverage nuanced insights from the data, highlighting the critical importance of connecting dataset characteristics directly with preparation techniques for a more targeted fraud detection approach.

The data preparation phase involves multiple steps to ensure a robust dataset for the GMM-based methodology. These steps are designed to collect, enrich, and structure the data for optimal analysis.

Step 1 *Fraudulent Tenders Data Acquisition.* The initial step involved gathering a dataset of 92 fraudulent tenders from the "Operação Patrola". These tenders were identified through consultations with the Special Anti-Corruption Group (GEAC) of the Prosecutor's Office of Santa Catarina (MPSC). This dataset formed the foundation for subsequent analysis and model development aimed at detecting and understanding fraud patterns in public procurement.

Step 2 *Dataset Enrichment.* Information about bids and company participation in tenders conducted in Santa Catarina state between 2009 and 2015 was collected using a database containing the history of 986,516 tenders. During this stage, the tenders identified in the previous step were labeled fraudulent. This data enrichment process added context and depth to the analysis, enabling a more detailed understanding of the characteristics and behaviors associated with fraud in public procurement.

Step 3 *Product Categorization.* To further refine the analysis, the identified frauds related to heavy equipment were categorized into four product categories: road roller, crawler tractor, excavator, and motor grader. This division was achieved by applying keyword filters to the tendered items, ensuring that the dataset was organized by relevant product types.

Step 4 *Feature Development.* Key features were critical to evaluate procurement processes and identify potential irregularities comprehensively. These features, listed in Table 3, encompass various aspects of the tendering process, such as the proportion of winning bids from a particular company, prior meetings with the Managing Unit, and the number of tenders a company has participated in. Additionally, features like the tenders' duration, the winners' uniqueness, and the number of participants involved in the tender process were detailed. We analyzed these features to uncover patterns that might indicate favoritism, limited competition, or other procurement anomalies.

Step 5 *Model Fitting and Hyperparameter Optimization.* To find the optimal GMM configuration, we tested components ranging from 1 to 7 and explored four types of covariance: Spherical, Tied, Diag, and Full. This process, known as hyperparameter optimization, used the Bayesian Information Criterion (BIC) to determine the best model configuration. The BIC helps balance the model fit to the data and its complexity, ensuring that the most appropriate model was selected for detecting potential irregularities in tenders.

Step 6 *Feature Subset Selection.* Feature subset selection was a critical step where all 255 combinations of features were generated and evaluated to determine their effectiveness in grouping labeled fraud tenders. Tenders were filtered to highlight those grouped into high-fraud clusters, identified based on a predefined threshold. The values 20%, 40%, 60%, 80%, and 100% were tested as thresholds.

Step 7 *Leave-one-out Cross Validation.* To evaluate the selected model's performance, an adapted leave-one-out cross-validation technique was employed. For each feature subspace, a known fraudulent tender was left out. The model was then run with the test set to check whether the left-out tender was grouped with at least another known fraudulent tender. This method results in a relative recall since we only know the positively labeled cases. All other tenders are unlabeled, meaning their status as fraudulent or non-fraudulent is unknown.

Step 8 *Distance Calculation.* Distance calculation involved measuring each tender's similarity to known fraud cases. The Euclidean distance between a suspect tender and each known fraudulent tender was calculated for each subset of features. The average of these distances was computed to quantify how similar a suspect tender was to fraudulent ones. This calculation provided a metric for assessing the likelihood of fraud based on proximity to known fraudulent patterns.

Step 9 *Evaluation Metric Establishment.* Finally, an evaluation metric for each tender was established, incorporating the inverse of the average Euclidean distance (since lower distances indicate a higher risk), the number of feature subsets where the tender appeared, and the average relative recall value. These variables were combined by weighting each normalized metric equally.

5. Results and Discussion

The threshold selection for evaluating the risk of fraudulent tenders was driven by the balance between high recall and average cluster size, ensuring a moderated number of cases will be flagged as risky for further manual analysis. As shown in Table 4, a threshold of 80 was selected due to its high recall of 0.844 and a manageable average cluster size of 137. This threshold represents a good trade-off, maximizing the correct identification of fraudulent cases while keeping the cluster sizes reasonable for detailed analysis. The recall value, which measures the ability to list the known fraudulent cases kept out from the analysis as suspect cases, remained relatively high across thresholds. However, based on the combination of high recall and average cluster size, we chose the threshold of 80% to perform a more in-depth analysis of the cases flagged as suspect.

Threshold	**20**	**40**	**60**	**80**	**100**
Distance for Fraud	0.332	0.353	0.366	0.396	0.437
Recall	0.891	0.838	0.831	0.844	0.749
Average Cluster Size	194	153	131	137	114
Average Subspaces	504	504	453	339	85

Table 4. Summary of metrics by Threshold

In the analysis of the sorted risk indicators for threshold 80, illustrated in Figure 3, the ”elbow method” can be conceptually applied to identify the most significant shift in the data distribution, indicating a natural breakpoint for high-risk tenders.

Visual inspection of the risk indicator graph reveals an inflection point where the curve of risk scores starts to rise noticeably. This increase becomes pronounced at a risk indicator value greater than 0.65. The risk scores continue to escalate sharply towards the dataset’s end, peaking significantly at the final tenders. This visual observation suggests that the elbow, or the point of greatest curvature change, occurs just as the risk indicator exceeds 0.65.

The results from applying the elbow method are illustrated in Figure 4, which displays the distribution of suspect cases across product categories at a threshold of 80 and a risk indicator above 0.65. The chart reveals a disproportionately higher number of suspect cases in Motor Graders and Crawler Tractors, at 20.32% and 15.38% respectively, compared to Excavators and Road Rollers. This pattern underscores the need for focused investigative efforts on specific equipment categories where the risk indicators are notably higher.

Table 5 presents a detailed view of tenders with the highest risk indicators and their top 3 frequent features and corresponding values. The risk indicators for these tenders range from 0.855169 to 0.900709, indicating a consistently high level of potential risk. The frequent features *num*, *duration*, and *unit_price* appear prominently across most tenders, suggesting their significant role in risk assessment. Specifically, the features *num* and *duration* are in the top 3 frequent features for most tenders, highlighting their strong correlation with high-risk cases.

The highest risk indicator of 0.900709 is associated with tender ID 121, with feature values of 0.008013 for *num*, 0.038698 for *duration*, and 0.839503 for *unit_price*.

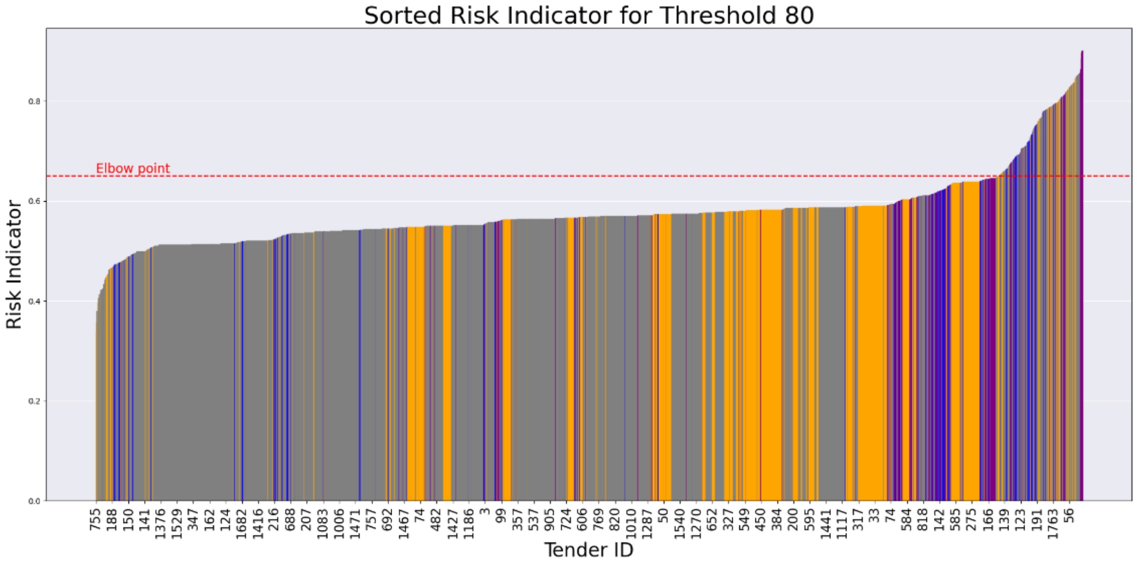


Figure 3. Risk indicator ranking

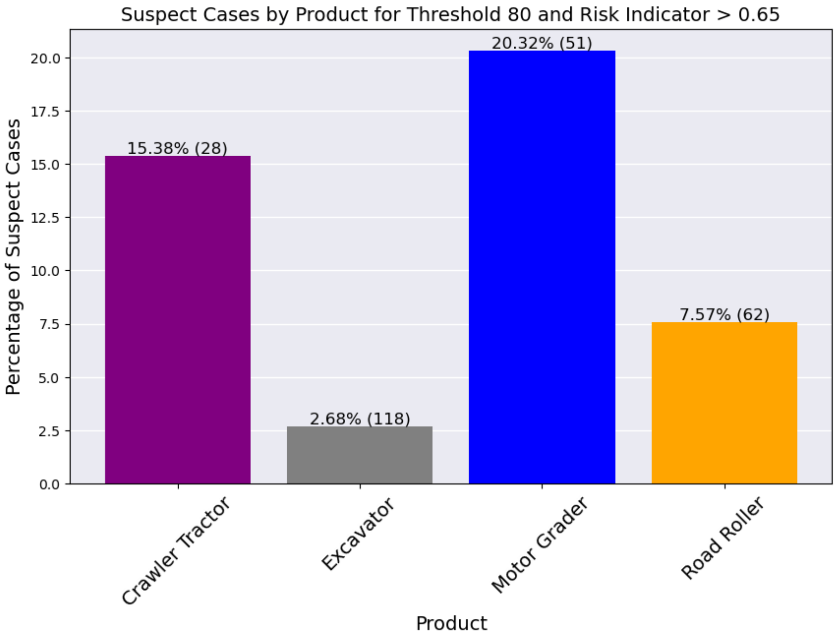


Figure 4. Suspicious tenders by product

ID	Risk Indicator	Top 3 Frequent Features	Feature 1	Feature 2	Feature 3
121	0.900709	num, duration, unit_price	0.008013	0.038698	0.839503
138	0.900331	num, duration, unit_price	0.000069	0.042770	0.862699
116	0.900187	num, duration, unit_price	0.000570	0.032571	0.833177
144	0.896936	num, duration, unit_price	0.008616	0.037241	0.880506
134	0.884527	num, duration, num_partic	0.000480	0.057920	0.858482
60	0.862939	num, duration, period	0.014854	0.052885	0.640347
53	0.857681	num, duration, period	0.000207	0.058116	0.596298
1643	0.857205	duration, num_partic, unique	0.000015	0.050629	0.677117
44	0.856101	num, duration, period	0.000220	0.057312	0.574977
1810	0.855169	duration, num_partic, unique	0.000643	0.028817	0.832709

Table 5. Top 10 tenders with highest Risk Indicators and their top 3 frequent features

The *num* value suggests a high frequency of bids from the same company, while the *duration* value indicates the total time span of the tender process. The *unit_price* reflects the pricing consistency across bids, playing a crucial role in detecting anomalies. These values are close to those found in fraudulent cases, reinforcing their significance in risk assessment.

6. Conclusion

The findings of this study underscore the effectiveness of a Gaussian Mixture Model (GMM)-based approach in flagging tenders as suspicious within public procurement tenders. By leveraging known fraud cases and exploring various feature subspaces, this methodology demonstrated a robust capability to highlight tenders that exhibit patterns similar to those identified as fraudulent. Applying this approach in the case study of ”Operação Patrola” validates its practical utility, offering a significant tool for public agencies to enhance transparency and accountability in procurement processes.

Our methodology's strength lies in its adaptability and scalability across different datasets and procurement contexts. The unsupervised learning model, coupled with feature engineering and normalization, ensures that the analysis is both comprehensive and precise. Using the Bayesian Information Criterion (BIC) for model selection, leave-one-out cross-validation, and Euclidean distance calculations using different feature subspaces provides a reliable framework for calculating the risk associated with tenders.

The results highlight the potential of this approach to uncover fraudulent patterns and aid in combating corruption in public procurement. By identifying tenders that are grouped with known fraudulent cases, this methodology facilitates early detection and investigation, which are critical for preventing financial losses and ensuring fair competition.

Future research should enhance the model's feature set by incorporating additional variables derived from tender documents and other relevant sources. Experts' qualitative analysis can refine the detection criteria, ensuring the methodology captures a wider array of fraudulent patterns. Additionally, exploring the integration of this approach with real-time monitoring systems could provide continuous oversight, thereby improving the responsiveness and efficacy of anti-corruption measures in public procurement.

The continual evolution of data mining techniques and machine learning models promises further improvements in fraud detection methodologies. By staying at the forefront of these advancements, public agencies can better safeguard public funds and maintain the integrity of procurement processes.

References

Goryunova, N., Baklanov, A., and Ianovski, E. (2021). Detecting corruption in single-bidder auctions via positive-unlabelled learning. *HSE University*.

International, T. (2022). Corruption perception index. Acesso em: 26 abril 2023.

Ministério Público de Santa Catarina (2022a). Mpsc desarticula esquema de propina para compra de máquinas pesadas em santa catarina. `https://www.mpsc.mp.br/noticias/mpsc-desarticula-esquema-de-propina-para-compra-de-maquinas-pesadas-em-santa-catarina`. Accessed: 30 Sept 2023.

Ministério Público de Santa Catarina (2022b). Operação patrola: ex-prefeito de tangará é condenado a 13 anos de prisão. `https://www.mpsc.mp.br/noticias/operacao-patrola-ex-prefeito-de-tangara-e-condenado-a-13-anos-de-prisao`. Accessed: 30 Sept 2023.

Rabuzin, K. and Modrušan, N. (2019). Prediction of public procurement corruption indices using machine learning methods. In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, pages 333–340. SCITEPRESS.

Rothstein, B. (2011). *The Quality of Government: Corruption, Social Trust, and Inequality in International Perspective*. University of Chicago Press.

Tribunal de Contas do Estado de Santa Catarina (2024). e-sfinge: Sistema de fiscalização integrada de gestão. `https://www.tcesc.tc.br/esfinge`. Accessed: 26 Apr 2024.

Velasco, R. B., Carpanese, I., Interian, R., Paulo Neto, O. C. G., and Ribeiro, C. C. (2021). A decision support system for fraud detection in public procurement. *International Transactions in Operational Research*, 00:1–21.