ISSN 1870-4069

Early Corruption Detection System Using Machine Learning Algorithms

Maria Fernanda Heredia-Soto, Edgar Huaranga, Franci Suni-Lopez

Universidad de Lima,
Laboratorio de Inteligencia Artificial,
Peru

{mfheredi,ehuarang,fsuni}@ulima.edu.pe

Abstract. This work addresses the development of an early corruption detection system using advanced data analysis and artificial intelligence algorithms. The system uses open data collected from the Anti-Corruption Observatory of Peru's Comptroller General's Office. A dataset of 2815 public entities evaluated through the *"Índice de Riesgos de la Corrupción e Inconducta Funcional"* (INCO), with scores ranging from 0 to 100, was structured. The primary task involved classification into six risk levels, from very low to very high. Histogram-based Gradient Boosting (HGB) was applied, achieving an accuracy of 89.6%. Regression tasks on the raw INCO scores and experiments with Large Language Models (LLMs) were also conducted. The system represents an early-stage, yet scalable, tool to support public sector transparency. Future work proposes the integration of explainable AI for improved transparency and real-time policy support.

Keywords: Web scraping, data analysis, histogram-based gradient boosting, machine learning, anti-corruption, early detection model.

1 Introduction

Corruption is a pervasive phenomenon that affects societies at both local and global levels, discrediting public trust, weakening governmental institutions, and undermining sustainable development. It manifests in various forms such as bribery, fraud, extortion, and embezzlement, and is broadly defined by the misuse of official power for personal gain or to benefit third parties. According to the Code of Conduct for Public Officials, corruption involves any act involving illegal benefits in the exercise of public duties. Beyond damaging the integrity of government systems, corruption exacerbates poverty, increases inequality, and violates fundamental human rights [1,2] particularly in less developed democratic systems [3]. International organizations like the United Nations and the Council of Europe stress that national efforts alone are insufficient to tackle this complex issue, emphasizing the need for comprehensive and collaborative strategies to enhance transparency and accountability [4].

In this context, the present study focuses on the creation of an Early Corruption Detection System, leveraging machine learning algorithms to identify high-risk entities proactively.The system is built on open data published annually by the Anti-Corruption Observatory (OBANT) of Peru's Comptroller General's Office (CGR)[12],[13], incorporating artificial intelligence models such as Random Forest and HGB, which are well-suited for detecting patterns in large, heterogeneous datasets [5],[6],[7],[8]. The proposed approach not only aims to detect entities with potential misuse of public resources but also seeks to monitor early warning signals for timely interventions, contributing to improved transparency and efficiency in public management. Furthermore, the initiative aligns with Sustainable Development Goal 16, which promotes strong institutions and the rule of law [9]. The objective of this study is to develop and validate a robust, scalable system using advanced evaluation metrics such as F1-Score, accuracy, and recall to measure its effectiveness.

This paper is organized by outlining the dataset and its relevance, detailing the methodology and results of various machine learning models, discussing key implications for public governance, and concluding with future research directions.

2 INCO Data Description

INCO is a standardized tool created by Peru's Comptroller General's Office to assess the likelihood of corruption and misconduct within public institutions. It is calculated through a composite score derived from multiple officially reported indicators, each assigned a specific weight based on its relevance tocorruption risk.

Entities are assigned a score from 0 to 100 based on weighted evaluation of these indicators. For classification, the scores were mapped into six corruption risk levels: Very Low, Low, Moderate, Medium, High, and Very High. The final cleaned dataset included the 2023 INCO scores and the label distribution was as follows: Very High (205), High (248), Medium High (248), Medium (248), Moderate (248), Low (247), and Very Low (100). To address class imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was applied.

The dataset consists of data from 2815 Peruvian state entities, covering the years 2021 to 2023. It includes 24 standardized indicators grouped into two dimensions: *"Institutional Management and Public Integrity"* and "Operational Execution and Control Mechanisms". Key indicators include: (1) sanctions against public officials (I1), (2) administrative/civil/penal responsibilities (I14-I18), (3) irregular public contracting processes (I3, I4, I20), (4) control audit results (I19), and (5) integrity declaration compliance (I13).

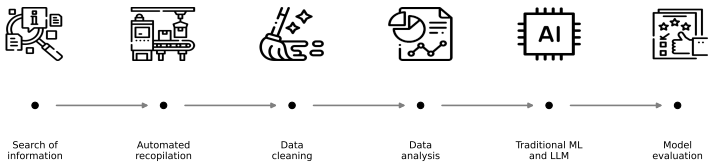

Fig. 1. Workflow of the Methodology.

3 Methodology

This section explains the process for developing the proposed solution. The workflow of the methodology to be followed is shown in Figure 1 from data collection to model evaluation.

3.1 Data Acquisition and Preparation

The primary data source was INCO published by OBANT of CGR [12],[13]. The dataset included information for 2,815 public entities across multiple years (2021–2023), structured around 24 corruption risk indicators. These indicators were grouped into two major dimensions: Institutional Management and Public Integrity and Operational Execution and Control Mechanisms. The features included sanctions against public officials, administrative and penal responsibilities, irregularities in procurement processes, audit results, and compliance with integrity declarations[12],[13].

To build the dataset, relevant information was gathered from public platforms using Web Scraping. Data cleaning involved standardizing variables, removing inconsistencies, and structuring a final dataset ready for machine learning analysis. No manual feature engineering or complex transformations were applied beyond basic consistency adjustments, as the INCO dataset provided pre-aggregated, labeled information suitable for supervised learning.

Feature Analysis To acquire relevant data that contain the necessary information for analysis and detection of potential acts of corruption, sources were identified. Open government data platforms that publish information on public procurement, government expenditures, audit reports, among others, were considered. In some cases, these data were not available on open platforms, so formal requests were also made to obtain the necessary datasets in accordance with transparency laws, allowing their subsequent download. As shown in Figure 2, the most relevant features for predicting corruption risk were selected through an ANOVA-based feature selection process, highlighting indicators related to sanctions, audits, and responsibilities.

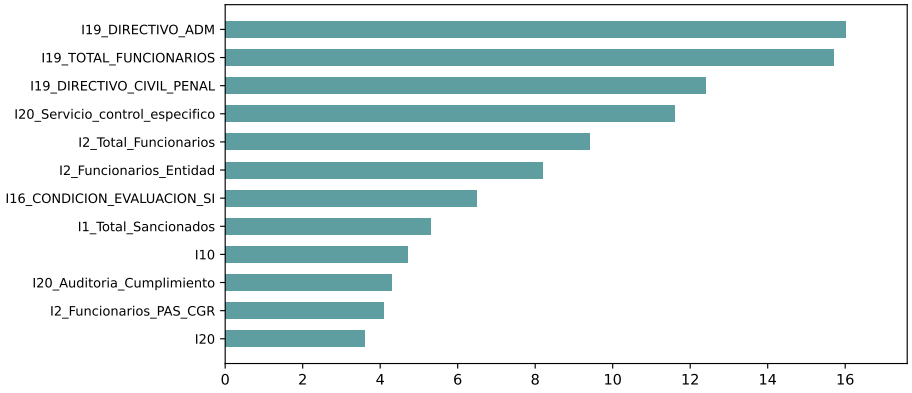

Fig. 2. Features with greater relevance.

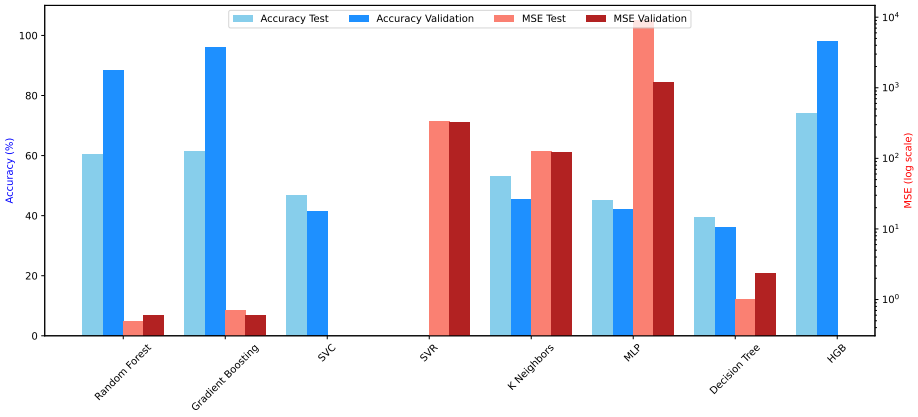

Fig. 3. Modeling.

3.2 Modeling

Two main predictive tasks were addressed: classification and regression. For the classification task, the goal was to categorize entities into six predefined corruption risk levels which are Very Low, Low, Moderate, Medium, High, Very High based on their INCO scores. Several machine learning algorithms were trained and evaluated, including Random Forest, Gradient Boosting, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees. Among these, the HGB classifier achieved the highest performance and was selected as the primary model. Hyperparameter tuning was conducted using RandomizedSearchCV, and evaluation relied on 5-fold cross-validation.

Figure 3 presents a side-by-side comparison of classification accuracy and regression error (MSE, log-scaled) across multiple models. HGB demonstrates consistently high performance across both tasks, especially in classification

where it significantly outperforms alternatives. Models like MLP and SVC show limitations due to poor handling of class imbalance and feature interactions.

sfSeparately, a regression task was explored to predict the raw INCO scores directly. Models such as Gradient Boosting Regressor and Support Vector Regressor (SVR) were trained; however, regression results did not offer substantial advantages over classification models in terms of interpretability and practical application. In parallel, experiments were conducted using Large Language Models (LLMs) to assess their ability to perform structured corruption risk classification. Models such as Qwen, Llama, Gemma, Hermes, and Mistral were deployed locally via LM Studio. A few-shot prompting strategy was designed, where the LLMs received key indicators as input (e.g., number of sanctioned officials, audit outcomes, pending civil responsibilities) and were asked to predict the corresponding corruption risk level. The outputs were structured to match the classification labels, allowing direct performance comparison with traditional machine learning models.

3.3 Evaluation

Model performance was evaluated using standard classification metrics: Accuracy, F1-Score, and Recall. For regression models, Mean Squared Error (MSE) was used as the primary evaluation criterion. Cross-validation ensured that models generalized effectively and prevented overfitting. The Histogram-Based Gradient Boosting classifier achieved an overall classification accuracy of 89.6%, significantly outperforming the LLM-based approach, which reached 69% accuracy under the same evaluation conditions. The use of Synthetic Minority Over-sampling Technique (SMOTE) contributed to improving detection in minority classes and balancing the overall predictive performance across all corruption risk levels.

4 Results

The features in Figure 2 demonstrated the highest statistical correlation with the target label and were used to guide model training. Their strong relevance underscores the significance of institutional control and sanction-related data in identifying corruption patterns.

The HGB model achieved the best classification accuracy shows in Figure 3 while maintaining low prediction error in regression. Other models, such as SVC and MLP, underperformed due to limited handling of non-linear feature interactions and imbalanced data. These results confirm the robustness of tree-based ensemble methods in heterogeneous public sector datasets.

HGB reached an overall accuracy of 89.6%, with excellent F1-Scores above 98% for the Very High and Very Low classes, where the model's precision and recall were nearly perfect. Table 1 summarizes the classification performance of the HGB and LLM models across corruption risk levels. SMOTE clearly improved detection in minority classes.

Table 1. Classification Results by Risk Level.

Class	Model	Accuracy	Recall	F1-Score	Support
Low	HGB	97.6%	99.2%	98.4%	247
	LLM	74%	74%	74%	100
Moderate	HGB	87.4%	89.1%	88.2%	248
	LLM	64%	64%	64%	150
Medium	HGB	83.1%	81.5%	82.3%	248
	LLM	50%	50%	50%	150
Medium High	HGB	88.9%	87.5%	88.2%	248
	LLM	60%	60%	60%	200
High	HGB	80.7%	81%	80.9%	248
	LLM	65%	65%	65%	205
Very High	HGB	100%	99.6%	99.8%	248
	LLM	80%	80%	80%	100
Accuracy	HGB	-	-	89.6%	1487
	LLM	-	-	69%	905
Macro Avg	HGB	89.6%	89.7%	89.6%	1487
	LLM	68%	68%	68%	905
Weighted Avg	HGB	89.6%	89.6%	89.6%	1487
	LLM	70.5%	70.5%	70.5%	905

Other models like Random Forest and Gradient Boosting (non-HGB) achieved between 88% and 95% accuracy but were more prone to overfitting based on test/validation variance. Support Vector Machines and MLPs underperformed due to data imbalance and lack of deep feature interaction. In the regression task, models achieved moderate results (MSE 0.5–0.7), with no clear added value over classification.

5 Discussion

The empirical results obtained position the Histogram-Based Gradient Boosting (HGB) model as the most effective alternative for the classification of institutional corruption risk based on structured administrative data. Its superior and consistent performance across all predefined risk levels—particularly in the extremes (Very Low and Very High)—underscores its reliability in operational contexts that demand both accuracy and stability. This behavior is indicative of the model's capacity to capture intricate, non-linear relationships among heterogeneous indicators, such as sanctions, audit outcomes, and declarations of integrity.

The HGB model also demonstrates notable robustness in the presence of class imbalance and missing values, which are common characteristics in real-world public sector datasets. In contrast, algorithms such as Support Vector Classifiers (SVC) and Multilayer Perceptrons (MLP) showed diminished effectiveness, likely

due to their reduced capacity to model inter-variable dependencies and to adapt to uneven label distributions.

Regression-based approaches, while offering continuous predictions, were empirically less effective and substantively less actionable. From an institutional perspective, discrete risk categories are not only more interpretable but also more operationally aligned with audit planning and resource allocation. Furthermore, regression models exhibited higher variance, particularly in mid-level risk predictions, thereby compromising the reliability of their outputs in critical decision-making scenarios.

A conceptual and methodological contrast arises when comparing this work with unsupervised anomaly detection frameworks such as the corruption-based self-supervised model (CAIT)[15]. Whereas CAIT relies on the effect of single-variable corruption on anomaly scores to infer variable relevance, the proposed system utilizes statistical feature selection (ANOVA) and model-specific attribution techniques (e.g., Shapley values) to identify key predictors. This allows for more transparent and reproducible analysis within governance frameworks that require justification and traceability of model outputs.

6 Conclusion

In this paper, we propose a scalable and accurate early corruption detection system for public sector entities, based on machine learning techniques and structured government data. The Histogram-Based Gradient Boosting (HGB) classifier demonstrated the highest performance among all tested models, achieving an overall accuracy of 89.6% and strong F1-Scores across all corruption risk levels. This validates its suitability for institutional monitoring and supports its use in real-world governance applications.

Our results highlight the value of supervised learning algorithms in handling complex, heterogeneous data typical of public administration. While regression models and LLMs were also explored, their practical utility remains limited in this specific classification context. However, future integration of LLMs and explainable AI (XAI) techniques may expand the system's capabilities in terms of interpretability and human oversight.

This approach not only enables proactive identification of high-risk entities but also contributes to policy design, resource allocation, and transparency monitoring. Further work will focus on improving adaptability through real-time data sources, expanding the dataset beyond INCO, and enhancing the model's ability to detect structural patterns of misconduct using causal inference.

Ultimately, the system proposed in this work is a step forward in operationalizing AI for public accountability and offers a practical framework for early detection and intervention in corruption-prone environments.

References

1. Kalienichenko, L. I., Slynko, D. V.: Concept, features and types of corruption. Law and Safety 84(1),pp. 39–46 (2022)
2. Oriolo, A.: The Contribution of the European Court of Human Rights to the Construction of a Corruption-Free Society. International Criminal Law Review **1**,pp. 1–28. Brill Nijhoff (2024)
3. Mubangizi, J.C., Sewpersadh, P.: A human rights-based approach to combating public procurement corruption in Africa. African Journal of Legal Studies **10**(1),pp. 66–90 (2017)
4. Caruso, S., Bruccoleri, M., Pietrosi, A. : Artificial intelligence to counteract "KPI overload" in business process monitoring: the case of anti-corruption in public organizations. Business Process Management Journal {29(4),pp. 1227–1248 (2023)
5. Aguilera-Martos, I., García-Barzana, M., García-Gil, D.,: Multi-step histogram based outlier scores for unsupervised anomaly detection: ArcelorMittal engineering dataset case of study. Neurocomputing 544, pp. 126228 (2023)
6. Lima, M. S. M., Delen, D.: Predicting and explaining corruption across countries: A machine learning approach. Government Information Quarterly 37(1), 101407 (2020)
7. Supriya, M., Adilakshmi, T.: Intrusion Detection in Wireless Sensor Networks Using Histogram Gradient Boosting Classifier. In: International Conference on Data Science, Machine Learning and Applications, pp. 473–480. Springer (2023)
8. Rajesh, P.K., Shreyanth, S., Sarveshwaran, R.: Bayesian Optimized Random Forest Classifier for Improved Credit Card Fraud Detection: Overcoming Challenges and Limitations. In: XVIII International Conference on Data Science and Intelligent Analysis of Information, pp. 205–214, Springer (2023)
9. Abdelaal, M., Ktitarev, T., Städtler, D. : SAGED: Few-Shot Meta Learning for Tabular Data Error Detection. In: EDBT, pp. 386–398 (2024)
10. Köbis, N.C., Starke, C., Edward-Gill, J.: The corruption risks of artificial intelligence. https://knowledgehub.transparency.org/assets/uploads/kproducts/The-Corruption-Risks-of-Artificial-Intelligence.pdf, last accessed 2025/03/10
11. Decarolis, F., Giorgiantonio, C.: Corruption red flags in public procurement: new evidence from Italian calls for tenders. EPJ Data Science 11(1), 16 (2022)
12. Contraloría General de la República del Perú: Guía Metodológica del Índice de Corrupción e Inconducta Funcional. Observatorio Anticorrupción, Contraloría General de la República del Perú (2023).
13. Contraloría General de la República del Perú: Índice de Riesgos de Corrupción e Inconducta Funcional (INCO) - Guía Metodológica 2024. Contraloría General de la República (2024)
14. Abdelaal, M., Koparde, R., Schoening, H.: Autocure: Automated tabular data curation technique for ML pipelines. In: Proceedings of the Sixth International Workshop on Exploiting Artificial Intelligence Techniques for Data Management, pp. 1–11 (2023)
15. Mok, C., Kim, S.B.: Corruption-Based Anomaly Detection and Interpretation in Tabular Data. In: Pattern Recognition, 159, 111149 (2024). https://doi.org/10.1016/j.patcog.2024.111149