Automatic learning model to predict transparency indicators for effective management of public resources

Modelo de aprendizaje automático para pronóstico de indicadores de transparencia para la gestión efectiva de los recursos públicos

Modelo de aprendizado de máquina para previsão de indicadores de transparência para gestão eficaz de recursos públicos

Natalia Andrea Ramírez Pérez[1]
Ernesto Gómez Vargas[2]
Harold Vacca González[3]

Received: June 15th, 2023
Accepted: September 18th, 2023
Available: October 18th, 2023

How to cite this article:
N.A. Ramírez Pérez, E. Gómez Vargas, H. Vacca González, "Automatic learning model to predict transparency indicators for effective management of public resources", *Revista Ingeniería Solidaria,* vol. 19, no. 3, 2023.
doi: https://doi.org/10.16925/2357-6014.2023.03.09

Research article. https://doi.org/10.16925/2357-6014.2023.03.09

[1] Professor Institución Universitaria Pascual Bravo.
Email: natalia.ramirez@pascualbravo.edu.co, naaramirezp@correo.udistrital.edu.co
ORCID: https://orcid.org/0000000343897295
CVLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001700480

[2] Professor Universidad Distrital Francisco José de Caldas.
Email: egomez@udistrital.edu.co
ORCID: https://orcid.org/0000000349577313
CVLAC: http://scienti.colciencias.gov.co:8081/cvlac/visualizador/generarCurriculoCv.do?-cod_rh=0000472069

[3] Professor Universidad Distrital Francisco José de Caldas
Email: hvacca@udistrital.edu.co
ORCID: https://orcid.org/0000000170170070
CVLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000712353

Abstract

Introduction: This article is the product of the application of predictive analytical models to measures and indicators of corruption risk, as researched by the Pascual Bravo University Institution and the Francisco José de Caldas District University in 2022 for doctoral research on the risk model for state transparency.

Problem: From measurements of institutional capacities, it is possible to generate anticorruption measurements, such is the case of the National AntiCorruption Index (INAC for its Spanish acronym). However, there are improvements to be made in the indicators and the need to incorporate more and better measurements that support this scourge that has long been manifested in Colombia.

Objective: The objective of this research is to emphasize the need to take advantage of open data, to generate measurements of state institutional corruption and, therefore, metrics that support its transparency and integrity based on predictive analytical models to generate predictions about government indices.

Methodology: First, the importance of generating measurements for the management of corruption cases is pointed out. Then, the application of predictive analytical models to predict scores of the National AntiCorruption Index is evidenced, finding the best model to finally make a forecast based on the identification of the relevant variables.

Results: The implementation of higher levels of digital government (egovernment) can significantly contribute to the fight against corruption and the generation of better public policies that support controls and sanctions. It not only facilitates citizen access to state services, but also allows for more open and agile access to data. This constantly promotes transparency at all levels and at all times. The Huber regression that has been implemented, its smaller penalty function, and its linear rather than quadratic growth, make it more suitable for dealing with outliers. This improves the error meter estimates and provides a good estimate of the National AntiCorruption Index score.

Conclusion: It is essential to establish a framework that anticipates the behavior of INAC and directs public policy efforts towards transparency and the prevention of corruption. In addition, it is necessary to develop objective metrics, indicators, indices and risk models that promote and evaluate transparency in the fight against corruption. This implies generating early warnings, applying sanctions, implementing controls and designing improvement plans to promote recommendations based on data that can trigger actions and take advantage of free access to public information to support citizens and the country.

Originality: A predictive analytical model based on machine learning was trained to predict the future behavior of the National AntiCorruption Index, with the aim of supporting roadmaps for entities and creating improvement actions for national entities, in which it becomes necessary to explore open government data to create new indicators and improve current ones.

Limitations: The regression models on the historical data of free access for the INACs were selected, because in terms of measurement it is what is already consolidated and available for the generation of transparency policies, access to information and the fight against corruption. The challenge for future work is to have more historical data, and to create more indicators that support measurements with the creation of improvement actions per entity that is reflected in numerical measurements.

Keywords: Transparency, corruption, artificial intelligence, machine learning, risk.

Resumen

Introducción: Este artículo es producto de la aplicación de modelos analíticos predictivos a las medidas e indicadores de riesgo de corrupción, como investigación de la Institución Universitaria Pascual Bravo y la Universidad Distrital Francisco José de Caldas en 2022 para la investigación doctoral sobre el modelo de riesgo para la transparencia estatal.

Problema: A partir de mediciones de capacidades institucionales es posible generar mediciones anticorrupción, tal es el caso del Índice Nacional Anticorrupción (INAC's), pero hay mejoras en los indicadores y la necesidad de incorporar más y mejores mediciones que sustentan este flagelo que se manifiesta desde hace años en nuestro país.

Objetivo: El objetivo de esta investigación es enfatizar la necesidad de aprovechar los datos abiertos, para generar mediciones de corrupción institucional estatal y por ende métricas que sustenten su transparencia e integridad a partir de modelos analíticos predictivos para generar predicciones sobre índices gubernamentales.

Metodología: En primer lugar, se señala la importancia de generar mediciones para el manejo de casos de corrupción, luego se evidencia la aplicación de modelos analíticos predictivos para predecir puntajes del Índice Nacional Anticorrupción, encontrando el mejor modelo para finalmente realizar un pronóstico con base en la identificación de las variables relevantes.

Resultados: La implementación de mayores niveles de gobierno digital (egoverment) puede contribuir significativamente a la lucha contra la corrupción y a la generación de mejores políticas públicas que apoyen los controles y sanciones. No solo facilita el acceso de los ciudadanos a los servicios estatales, sino que también permite un acceso más abierto y ágil a los datos. Esto promueve constantemente la transparencia en todos los niveles y en todo momento. La regresión de Huber que se ha implementado, su función de penalización más pequeña y el crecimiento lineal en lugar de cuadrático lo hacen más adecuado para tratar con valores atípicos. Esto mejora las estimaciones del medidor de errores y proporciona una buena estimación de la puntuación del Índice Nacional Anticorrupción.

Conclusión: Es fundamental establecer un marco que anticipe el comportamiento del INAC y permita orientar los esfuerzos de política pública hacia la transparencia y la prevención de la corrupción. Además, es necesario desarrollar métricas, indicadores, índices y modelos de riesgo objetivos que promuevan y evalúen la transparencia en la lucha contra la corrupción. Esto implica generar alertas tempranas, aplicar sanciones, implementar controles y diseñar planes de mejora para promover recomendaciones basadas en datos que puedan desencadenar acciones y aprovechar el libre acceso a la información pública para apoyar a la ciudadanía y al país.

Originalidad: Se entrenó un modelo analítico predictivo basado en aprendizaje automático para predecir el comportamiento futuro del índice nacional anticorrupción, con el objetivo de apoyar las hojas de ruta para las entidades y crear acciones de mejora para las entidades nacionales, en lo cual se hace necesario explorar datos abiertos de gobierno para crear nuevos indicadores y mejorar los actuales.

Limitaciones: Se seleccionaron los modelos de regresión sobre los datos históricos de libre acceso para los INAC, debido a que en términos de medición es lo que se encuentra consolidado y disponible para la generación de políticas de transparencia, acceso a la información y lucha contra la corrupción. El desafío para el trabajo futuro es tener más datos históricos, y por qué no crear más indicadores que apoyen la medición con la creación de acciones de mejora por entidad que se refleje en las mediciones numéricas.

Palabras clave: Transparencia, corrupción, inteligencia artificial, aprendizaje automático, riesgo.

Resumo

Introdução: Este artigo é produto da aplicação de modelos analíticos preditivos a medidas e indicadores de risco de corrupção, conforme pesquisa da Instituição Universitária Pascual Bravo e da Universidade Distrital Francisco José de Caldas em 2022 para pesquisa de doutorado sobre o modelo de risco. .

Problema: A partir de medições de capacidades institucionais é possível gerar medidas anticorrupção, como é o caso do Índice Nacional Anticorrupção (INAC's), mas há melhorias nos indicadores e a necessidade de incorporar mais e melhores medidas que apoiem este flagelo que se manifesta há anos no nosso país.

Objetivo: O objetivo desta pesquisa é enfatizar a necessidade de aproveitar dados abertos para gerar medições de corrupção institucional estatal e, portanto, métricas que apoiem sua transparência e integridade com base em modelos analíticos preditivos para gerar previsões sobre índices governamentais.

Metodologia: Primeiramente aponta-se a importância de gerar medidas para a gestão de casos de corrupção, em seguida fica evidente a aplicação de modelos analíticos preditivos para prever pontuações do Índice Nacional Anticorrupção, encontrando o melhor modelo para finalmente fazer uma previsão com base em a identificação de variáveis relevantes.

Resultados: A implementação de níveis mais elevados de governo digital (egogovernment) pode contribuir significativamente para a luta contra a corrupção e a geração de melhores políticas públicas que apoiem controlos e sanções. Não só facilita o acesso dos cidadãos aos serviços do Estado, mas também permite um acesso mais aberto e ágil aos dados. Isto promove constantemente a transparência em todos os níveis e em todos os momentos. A regressão de Huber que foi implementada, a sua função de penalidade menor e o crescimento linear em vez de quadrático tornam-na mais adequada para lidar com valores discrepantes. Isto melhora as estimativas do medidor de erro e fornece uma boa estimativa da pontuação do Índice Nacional Anticorrupção.

Conclusão: É essencial estabelecer um quadro que antecipe o comportamento do INAC e permita orientar os esforços das políticas públicas para a transparência e a prevenção da corrupção. Além disso, é necessário desenvolver métricas, indicadores, índices e modelos de risco objetivos que promovam e avaliem a transparência no combate à corrupção. Isto envolve gerar alertas precoces, aplicar sanções, implementar controlos e conceber planos de melhoria para promover recomendações baseadas em dados que possam desencadear ações e tirar partido do acesso gratuito à informação pública para apoiar os cidadãos e o país.

Originalidade: Foi treinado um modelo analítico preditivo baseado em machine learning para prever o comportamento futuro do índice nacional anticorrupção, com o objetivo de apoiar roadmaps para entidades e criar ações de melhoria para entidades nacionais, nas quais é necessário explorar dados governamentais abertos criar novos indicadores e melhorar os atuais.

Limitações: Os modelos de regressão foram seleccionados sobre os dados históricos de livre acesso ao INAC, porque em termos de medição é o que está consolidado e disponível para a geração de políticas de transparência, acesso à informação e combate à corrupção. O desafio para trabalhos futuros é ter mais dados históricos, e porque não criar mais indicadores que apoiem a medição com a criação de ações de melhoria por entidade que se reflitam em medições numéricas.

Palavras-chave: Transparência, corrupção, inteligência artificial, aprendizado de máquina, risco.

1. INTRODUCTION

The proper management of public resources has been an objective that states have always pursued. Through the promotion of policies of greater transparency, several efforts have been made to prevent corrupt behaviors from taking advantage of what should be public interest and redirecting it towards private interests. This has generated a discussion between politicians and academics on how to address the issue, since there is no single way to promote transparency to prevent corruption. However, there is a consensus on how promoting transparency and the fight against corruption must be measured to generate indicators and implement risk management. As a

matter of fact, the indicators depend on direct variables: those associated exclusively with transparently managing resources and, indirectly, with those defined by the context under which the policies are enforced.

Regarding risk management, academics and politicians differ on the most critical moment to apply the indicators. Should it focus on prior control? Should the resource execution processes be improved? Should it focus on increased sanctions that deter future corrupt behaviors?

Likewise, public data management has been crucial for analyzing the variables and the critical moment for promoting transparency. The importance will grow as citizens increasingly have easier access to information from the State's electronic platforms to follow how public resources are managed. In this sense, the effectiveness of the indicators for promoting transparency policies and corruption risk management will depend more on the correct analysis of the significant amount of existing information. For this purpose, data science, machine learning techniques, or Bigdata analysis represent a great opportunity.

Thanks to: the regression analysis carried out on a substantial amount of free access data on the INAC, an environment configuration for mathematical regression models, cross validation, hyper parameter tuning, graph analysis for the evaluation of regression metrics, and a prediction on a reserve sample drawn randomly as a test of the score on a new data set; it was possible to obtain the best predictive model, adjusted to the data, for the INAC to estimate the score and possible adjusted scores for the following year.

1.1 Background

The following article is a product of the investigation and implementation of mathematical regression models based on automatic learning to predict numerical values of the historical compilation of the Colombian National Anticorruption Index (INAC). It was conceived as an index of indices that integrates results from measurements generated and evaluated by different Colombian public entities with public access. This measurement is focused on anticorruption policies or tools to assess institutional capacities, aiming to identify the best regression models that generate predictions and produce future measurements based on the analysis results. It also strives to generate a subsequent risk model for state transparency based upon the construction of transparency indicators for the effective management of public resources.

Thus, a review of specialized academic publications was evidenced, including the relationship between transparency and corruption, to determine: Several analysis

models; the direct and indirect variables with which the indicators are built to measure the mentioned phenomenon, prioritizing new developments; the moment of corruption risk management and; it was observed how the increase of available information has become crucial, for which the application of advanced computational techniques represents an opportunity.

2. MATERIALS AND METHODS

A. Transparency or corruption?

There is academic and political ambiguity regarding how to approach transparency and the fight against corruption. Reviewing the available literature, the terms transparency and corruption are, in principle, related to the solution and the problem, respectively. However, the definitions of the second term are found more frequently in the reviewed publications [1].

Relevant organizations such as the NGO 'Transparency International' define corruption as "the abuse of power entrusted for private benefit" [2]. The 'Office of the High Commissioner for the United Nations' considers corruption "the greatest obstacle to the effective protection of human rights" [2]. The 'World Bank' assumes corruption is "the abuse of public office for private benefit" (World Bank, 1997). Following these organizations, several authors define the problem of corruption as the abuse of a public position to benefit private interests. Thus "a public official who misuses his power to obtain benefits for himself, family, friends or politicians is committing corruption" [3] or "the sale of state property, bribery in public procurement and the misappropriation of government funds" produce corruption [4].

In addition, it should be noted that in its 2022 annual report, 'Transparency International' pointed out that the levels of corruption compared globally "have stood still in the last ten years, amid an environment of human rights abuses and deterioration of human rights and democracy." Also, they noted that 21 countries obtained the lowest score since they were measured. In the case of Colombia, the country was ranked 87th out of 180 with 39 points out of 100. The same place as in the 2021 report; however, 52% of the people who were part of the study stated they had felt corruption increased in the last 12 months [5].

On the other hand, the concept of transparency is linked to how states create more open policies, plans, and indicators that avoid corruption. The 'United Nations Office on Drugs and Crime' (UNODC) states that transparency "is a situation in which information about a decision making process is made publicly available and can be

easily verified both in terms of rules and the identity of decision makers, increasing the probability of detecting corruption" [6]. Thus, public institutions either have the word transparency or anticorruption in their name, or the policies, plans, and indicators they develop are related to promoting transparency and the fight against corruption. For example, in Colombia, there is a 'Transparency Office of the Presidency of the Republic'. They maintain transparency and anticorruption indicators such as the National AntiCorruption Index (INAC). Spain has the 'Transparency Portal of the General State Administration'; Italy has the 'National Anti Corruption Authority of Italy' (ANAC); and International Transparency has its 'Corruption Perception Index' (IPC).

Additionally, in recent years, the promotion of transparency and anticorruption has been linked to the greater importance of more open data to citizens and new techniques for analyzing data. To mention two cases, globally, there is an association between more unrestricted access to data and data publication with significant transparency. For example, today, it is possible to find 'The Global Open Data Index' created by 'The Open Knowledge Foundation' or 'The Public, Open, Useful, Reusable Data Index' (OUR) data by the 'Organization for Economic Cooperation and Development' (OECD). Likewise, improving the transparency and accountability of good governance by applying Information and Communication Technologies (ICT) has become a powerful tool to combat corruption [7]. The transit of governments to electronic government increases flexibility and complexity of handling the enormous amount of multidimensional, structured, and unstructured data with conventional database systems [8]. Finally, using more technology can offer valuable opportunities to integrate legality controls and business processes consistent with the rising call to combat corruption at the institutional level [9].

As can be observed, there is a consensus that the digression of a public resource or using a public position to benefit private interests is corruption. Transparency aims to mitigate corruption. The more significant the amount of open data available, the more transparency within states, while creating both the need to analyze more data and making an opportunity to apply new information technologies that support transparency policies.

B. Indicator variables

Under this scenario, the literature on the subject has been progressing. Increasingly, promoting transparency has been linked to the generation of indicators supported by a broader and more rigorous analysis of the available data.

Traditionally, the measurement of transparency and anticorruption indicators has been established based on direct variables that involve consolidated data; That is, from previous years in which compliance with transparency policies is measured, and especially the contractual processes of the State are evaluated. "Public contracting is a particularly critical area for corruption" because "almost all activities that involve the public sector imply the need to acquire goods, services or works, from construction to education, from health to innovation" [10].

For example, for the Colombian case, the INAC index collects information based on data that national, departmental, and municipal entities report regarding access to public information, contractual issues, financial processes, and accountability (Transparency Secretariat of the Republic of Colombia, 2020). Hence, it is possible to identify a posteriori of which entities obtained the best performance and maintain and reinforce their transparency activities, as well as to focus and implement corrective actions on those with lower performance. The INAC was only established in 2020, but the Colombian State has been using the described logic based on data consolidated in other previous indicators. For example, the 'Accountability Composite Indicator', the 'Quality Composite Indicator of Corruption Risk Maps', the 'Composite Indicator of the Culture of Legality', and the 'Supply and Demand Indicator of Public Information'.

In the same way, the analysis considers the measurement of indirect variables mainly associated with the perception of third parties about corrupt manners. The best example is the perception surveys of NGOs and recognized international organizations. A well-recognized indicator is the 'Corruption Perception Index' (CPI) carried out by Transparency International. Annually, they estimate "the perceived levels of corruption in the public sector of each country... ...rated, according to the opinions of experts and businessmen." They also suggest that perception is important as "normally, corruption involves illegal and deliberately hidden activities that only come to light through scandals and judicial processes. This makes it very difficult to calculate its real impact" [11]. Another example is the World Bank; they generate their Governance Indicators considering the issue of corruption based on perception [12].

As can be seen, although the development of indicators, both direct and indirect, are based on consolidated data, corruption continues to be present in a significant way, especially in developing countries such as Colombia. "Corruption occurs in many developing countries and is very difficult to detect due to low legal awareness, lack of good governance and integrity. Corruption often cannot be detected because those who work together enjoy its benefits in mutual symbiosis" [13]. It means that the lack of institutional clarity in societies leads to a favorable scenario for the spread of corruption.

Then, are the variables measured correctly? Because even though there are more and more controls: there is more interest from citizens, unions, and NGOs for transparency; there are more measurements like the CPI made by Transparency International; and different surveys demonstrate the relevance of the issue. For example, a poll conducted prior to the 2022 Colombian presidential election showed that one of the most significant concerns of Colombians was corruption (see Figure 1).

Hence, it is proposed that the analysis of direct and indirect variables go deeper and involve a more significant amount of analyzed data. It is also suggested that a technique be incorporated for analyzing massive data such as data science, machine learning, or big data.

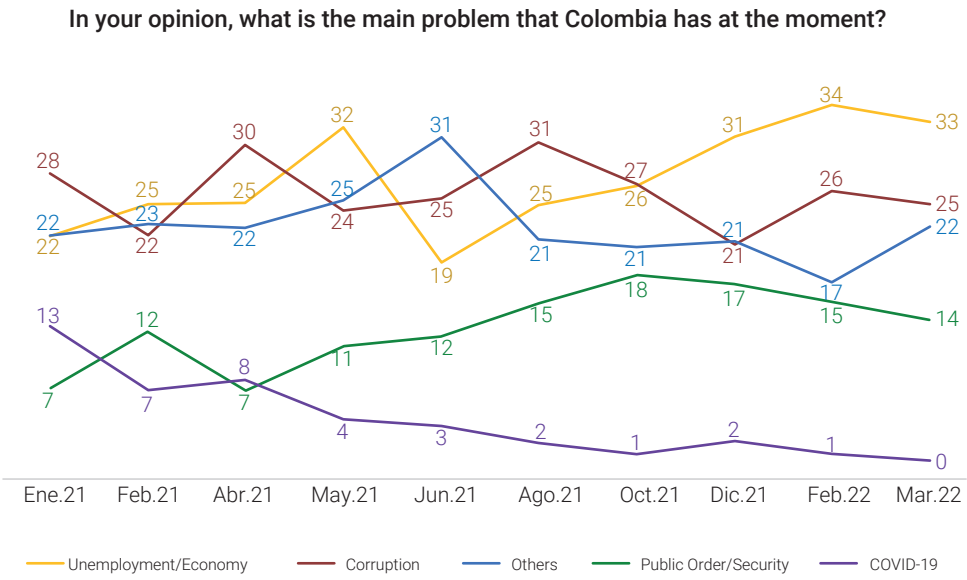

Figure 1:
Source: YanHaas Survey March 2022 [38].

C. Direct Variables

The direct variables included as an entry point can be classified as follows:

1. Transparency Measurements:

These correspond to local and national data under which indicators are generated for promoting transparency and anticorruption. Additionally, the different public entities report traditional consolidated data to the respective control entities.

e-ISSN 2357-6014 / Vol. 19, no. 3 / september-december 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

2. Framework Policies:

Policies associated with a legal framework for promoting transparency, such as laws, decrees, resolutions, and other legislation, as well as, the same public policies, the nature of public entities and their objectives [14].

Each country has a legal and public policy framework to promote transparency and anti-corruption. That is why it is crucial to review the content of how policies are written to identify patterns that could show when policies are more effective in the fight against corruption. Here, technological developments such as natural language processing offer an opportunity since "if integrated egovernment services are executed in a single big data environment, it will be easy to manipulate all the data related to all State services. Also, the government will be able to make the best decision based on the highspeed data processing system" [15].

3. Economical and Contractual Conditions:

These factors refer to the specific data of the contracting processes. It includes data analysis related to the variety of tenders, the number of participants, the time of publication of the processes, and the award. Likewise, the amount of the contracts, costs, and overtime in executing contracts are considered.

These variables represent an excellent opportunity for developing new technologies. Although academic discussions have been established about the subject, because early warnings or predictions can be generated from this information to prevent possible corruption cases, the approach has been mainly through traditional statistical methods.

For example, establishing so-called red flags has helped acquire data from tender documentation that can be used to detect suspicious bids. "Due to the prevention of suspicious activities from the beginning of the bidding procedure, we can speak of an early warning system" (Rabuzinand and Modrusan, 2019). However, the aforementioned use of traditional statistical models has limited data analysis to information that has always been analyzed, which does not allow for the inclusion of massive data analysis [16].

For this reason, the incorporation of big data technologies has been contemplated. However, it is noted that there is little systematic evidence, a reason that lies in the scarcity of reliable measures of the results of corruption. For example, in Italy, an exercise was established in which data were sought outside of those statistics (variables) that the Italian AntiCorruption Authority did not precisely track, even though it collected the necessary data to compute them. As established, "thanks to the

combination of both human and automatic textual analysis of these tenders, we obtain a broader set of indicators. There are several derived political implications; improving the data collection process on public contract tenders can be a useful strategy to limit corruption" [17].

It has been suggested to include more data; not collected just by analyzing the bidding process, such as information on contractors (contracts, type of company, directors), ongoing judicial investigations for corruption (bribery, extortion, fraud, embezzlement), and others—but detailed data associated with each contracting group.

4. Associated Criminal Behaviors with Money Laundering:

Money laundering is the main ally of corruption because it represents a two-way criminal symbiosis. Hence, on the one hand, public corruption can generate demand for money laundering. Furthermore, money laundering can serve as an effective way of cleaning up corruption proceeds for reinvestment as a multiplier effect. Finally, corruption can influence the probability that organized crime discovers money laundering activities as an accelerator effect [18].

The relationship between money laundering and corruption has led governments and non-governmental organizations to implement various international initiatives. Several of these have been vital, including those launched by the United Nations (UN), the Organization for Economic Cooperation and Development (OECD) and the Financial Action Task Force (FATF), the World Bank, the International Monetary Fund (IMF), and Interpol.

In this framework, since it is a transnational crime, the data included must be mainly information that national authorities publish on money laundering. Largely, the risk analyses these carry out must be considered, as well as the reports of companies and people sanctioned and/or prosecuted for this crime. With this information, both patterns are identified where money laundering can encourage corruption, and limits are also established.

For the Colombian case, the data derives from the national AntiMoney Laundering and Terrorist Financing (AML/CFT) system of the Financial Analysis and Control Unit (UIAF). There, suspicious operations are monitored to detect assets and identify criminal networks and structures, as well as the relevant natural and legal persons in each criminal network. It also includes knowing the modus operandi or typologies used to generate resources and insert them into the legal economy, and dissemination and effective feedback of the information provided to the authorities

to advance processes of domain forfeiture, captures, and convictions, among other measures of freezing or blocking assets [19].

D. Indirect Variables

As previously established, the indirect variables have been limited to third parties' perceptions of corrupt behavior. Although they continue to be the primary source for obtaining data beyond the state's official data, their limitation is reflected in what people believe about corruption.

Thus, to better understand the context in which anticorruption policies perform, it is necessary to include more objective data with external particularities to corruption. For this study, it is suggested:

1. Cultural:

Especially those referring to trust in institutions and sovereigns. These data are mainly found in surveys and polls, which reflect how citizens perceive the authorities. It can be complemented through data mining sentiment analysis to help to access what citizens feel and think about State services directly.

2. Social:

Analysis of demographic characteristics such as poverty indicators, unsatisfied needs, and unemployment. As well as security characteristics like public security figures including common crime, theft, or the presence of organized crime (data on drug trafficking, smuggling, or human trafficking) [20].

In this sense, "environmental analysis with big data technology allows researchers to develop prediction models with minimal data, and that becomes more accurate over time; it also facilitates the performance of anticorruption agencies in determining projects that are indicated as poorly executed and corrupt" [21].

E. Regression Analysis

Regression analysis is a set of statistical methods for estimating relationships between a dependent variable and one or more independent variables. It can be used to assess the strength of the relationship between variables and to model the future relationship between them [22].

Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used for more complicated data sets where the dependent and independent variables show a nonlinear relationship.

Regression analysis offers numerous applications in various disciplines. It depends on the correct interpretation of the dependent and independent variables.

1. Huber Regression

It is a regression technique with the main idea of using a different loss function instead of the traditional least squares. It is solved under:

$$minimize \sum_{i=1}^{m} \phi(y_i - x_i^T \beta)$$

Minimizing $\beta \in n$, where the loss function Φ is the Huber function with $M > 0$ and:

$$\phi(u) = \{u^2 \; if \; |u| \leq \; M; \; 2Mu \; - \; M^2 \; if \; |u| > \; M\}$$

This function is identical to the least squares penalty for small residuals; on large residuals, its penalty is minor and increases linearly rather than quadratically. Therefore, it is more forgiving of outliers.

3. RISK MANAGEMENT: PRIOR, ONGOING OR SUBSEQUENT CONTROL?

A risk analysis is contemplated that strives to overcome the traditional paradigms of:

- Sampling methods and regular audits with "ex ante" approvals which became irrelevant in the age of big data accumulation and continuous analysis, and instant decision making [23].
- Under the traditional audit paradigm, auditors perform risk assessments for different procedures and choose a sample of data to investigate corruption on a material basis [24].

e-ISSN 2357-6014 / Vol. 19, no. 3 / september-december 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

New risk maps should be continuous:

- Continuous auditing differs from the traditional approach in frequency, focus on automated processes, and unique exception analysis concepts. It improves the relevance and timeliness of audit results. However, although continuous auditing is advantageous, its use is still infrequent, as companies are not ready to actively adopt the method because concepts still need development [25].
- Automating auditing procedures makes the cost of verifying more transactions relatively small, with less demand on human resources.
- Another advantage to including new technologies for risk analysis is that auditors receive not only manually entered data but also metadata or data about data recorded automatically and independently of people [26].

4. ANALYSIS AND RESULTS

INAC's objective is to identify how the entities work and how far they need to reach an optimal level of transparency. Implementing the regulatory guidelines includes actions that should be promoted to improve the gaps identified in the measurement through a roadmap containing recommendations. This measurement does not present or promote rankings, classifications, or positions [27][28]. The INAC is a retrospective measurement based on the data and results of the immediately preceding period or year. Currently, there are measurements for 2018, 2019, 2020, and 2021, which are subsequently analyzed and integrated to issue an annual report of recommendations and advancement actions [29].

Data is extracted from open publications from the website of the Secretariat of Transparency of the Republic of Colombia [30]. These are unified for all years by the variables that accompany the INAC score. These include Transparency and Open State, Institutional Integrity, Surrender of Accounts, Public Purchases, Budget, and Integrity for each national entity [31][32].

Data cleaning consists of filling in missing data for the ITA Score and Public Purchases data attributes. Indicators oversee indices published by the Office of the Attorney General of the Nation and Colombia Compra Eficiente Office. Hence, cleaning was done by incorporating zero as a value for missing data to obtain gauges for the parameters of the implemented regressors. Sample retention of 10% was performed on the total data set to validate the final predictive model on new data [33].

The five best candidate models for predictions are displayed below, which are qualified by the usual regression error metric R^2 for comparative purposes:

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
en	Elastic Net	0.6292	1.3310	1.0067	0.9882	0.0156	0.0094	0.0060
br	Bayesian Ridge	0.6241	1.4597	1.0264	0.9871	0.0157	0.0093	0.0070
lasso	Lasso Regression	0.6279	1.4533	1.0280	0.9869	0.0157	0.0093	0.0060
huber	Huber Regressor	0.5771	1.5041	1.0342	0.9860	0.0154	0.0084	0.0130

Figure 2. Comparison of mathematical regression models for INAC.
Reference: Own work– Python 3.7

For the Huber regressor, it has an α = 0.0001 as regularization strength L2, with penalty equal to $\alpha * \|w\|^2$ in the range of [0,∞]. Additionally, it has ε = 1.35 which is the parameter that controls the number of samples that should be classified as outliers. It is left default since it is a low value, thus being more robust for outliers and with tolerance equal to 1e−05 which stops iterations when *max|pro jgi|i = 1, ...,n <=tol*; where *pgi* is the i−ésimo component of the projected gradient [34].

Performing hyperparameter tuning through the grid search method, usually implemented for mathematical regression models based on machine learning, optimizes the R^2, so:

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.5756	0.6746	0.8214	0.9943	0.0127	0.0084
1	0.5776	0.7055	0.8399	0.9891	0.0122	0.0081
2	0.5902	0.6431	0.8019	0.9979	0.0303	0.0142
3	0.4551	0.6043	0.7774	0.9952	0.0133	0.0070
4	0.5541	0.6433	0.8021	0.9964	0.0128	0.0085
5	1.0547	7.7944	2.7918	0.9303	0.0395	0.0143
6	0.4007	0.2600	0.5099	0.9977	0.0073	0.0057
7	0.6108	0.9655	0.9826	0.9794	0.0143	0.0089
8	0.5608	0.5295	0.7277	0.9964	0.0116	0.0085
9	0.3788	0.2197	0.4687	0.9968	0.0065	0.0051
Mean	0.5758	1.3040	0.9523	0.9873	0.0161	0.0089
Std	0.1776	2.1731	0.6301	0.0197	0.0099	0.0030

Figure 3. Model optimization.
Reference: Own work– Python 3.7

e-ISSN 2357-6014 / Vol. 19, no. 3 / september-december 2023 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

With an α = 0.9, an ε = 1.9 and with tolerance equal to 1e−05.

It obtains R2, in the test/hold set of 0.9843 compared with 0.9873 on the total data set, which does not have a significant difference indicating a moderate fit. Finally, with a precision score of R^2 de 0.9939, it leads to a new data set with the following results:

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Huber Regressor	0.4042	0.3772	0.6141	0.9939	0.0088	0.0055

Figure 4. Regression for new data.
Reference: Own work– Python 3.7

The importance of variables recharged on the attributes of the data mentioned as Institutional Integrity and Transparency and Open State is highlighted. It corresponds to the integrity of an institutional form, covering those issues of particular importance such as the budget, contracting, politics of transparency, access to information, and the fight against corruption and integrity policy. In that context, to obtain better results regarding the execution of public resources and their coherence from what was planned, compared with what was executed, it should be shared with stakeholders through the different communication channels arranged for transparency purposes [35]. It should measure the degree of compliance or implementation of the Law of Transparency and Access to Public Information [36], as well as the elements of the Open State and the institutional conditions to develop scenarios of citizen participation, accountability, innovation, and technology:

Figure 5. Importance of variables on INAC regression analysis.
Reference: Own work– Python 3.7

A comparison of models trained on time series was implemented to make future value forecasts for INAC. Specifically for the following year (2022), regardless, there are not enough data points to achieve an acceptable error. The Naive predictor and the Huber trend removal and deseasonalization stand out as candidate models for this process. Those eliminate the effects of the accumulation of seasonality and trend

data sets to show only the absolute changes in the values, and allow potential cyclical patterns to be identified after removing the general deviation and the trend found to be constant due to the lack of historical data points.

	Model	MAE	RMSE	MAPE
naive	Naive Forecaster	487.0286	728.2224	0.0447
huber_cds_dt	Huber w/ Cond. Deseasonalize & Detrending	487.0286	728.2224	0.0447

Figure 6. Metrics for forecast.
Reference: Own work– Python 3.7

5. DISCUSSION AND CONCLUSIONS

It is crucial to establish a model to predict the behavior of INAC that allows focusing public policy efforts that promote transparency and prevent corruption. It is also necessary to create metrics, indicators, indices, and objective risk models that promote and measure transparency towards zero corruption. It generates early warnings, sanctions, controls, and improvement plans to promote intelligent recommendations based on data that could generate actions and take advantage of free access to public information to endorse the citizens and the country.

In the same way, it is necessary to continue increasing the quality and access of public data through which measurements are made to promote transparency and prevent corruption. Higher levels of digital government (egovernment) could help in this reference, thus not only making state services more easily accessible to citizens but also allowing access to more open and agile data. This always promotes transparency at all levels and times.

In the implemented Huber regression predictive modelling, since the penalty is smaller and increases linearly instead of quadratically, it is more sympathetic to outliers, improving the estimates for the error meter and providing good estimates on the National Anti Corruption Index score.

A data set can have outliers in either the input variables or the target variable, both of which can cause problems for a linear regression algorithm. Outliers in a data set can prejudice the summary statistics computed for the variable, such as the mean and standard deviation, which can favor the model toward outliers away from the central mass of observations. This results in models that attempt to balance good performance on outliers and average data and poorer performance on both overall.

Instead, the solution for the dataset case that has been successfully queried for INAC so far is to use modified versions of linear regression that specifically address the expectation of outliers in the dataset.

Better leveraged time series predictions induce constant measurements over time due to the absence of historical data. Nevertheless, the Naive predictor produces measurements that are not very acceptable for forecasting ratings far into the future. It is crucial to continue obtaining these measurements that promote transparency, fight against corruption, and encourage studies incorporating data science and new technologies.

It makes it possible, for example, to forecast and conduct future research to promote continuous improvement plans and roadmaps, such as actions to augment current information.

ACKNOWLEDGMENTS

Python 3.10.7, Institución Universitaria Pascual Bravo, Universidad Distrital Francisco José de Caldas, and Research groups on Electrical Sciences and Informatics (GICEI), and Basic Sciences (S)

6. REFERENCES

[1] F.N.D. Piraquive, O.S. Martínez, E.V. Pérez, R.G. Crespo, "Knowledge management model for project management: KM+PMTIC," in Construction projects: improvement strategies, quality management and potential challenges, K. Hall Ed. New York, NY, USA: Nova Publishers, 2017, pp. 2345.

[2] IEEE Reference Guide, IEEE Periodicals, Piscataway, NJ, USA, 2018. [Online]. Available: https://ieeeauthorcenter.ieee.org/wpcontent/uploads/IEEEReferenceGuide.pdf, pp. 1014.

[3] Transparency International, El ABC del CPI: cómo se calcula el índice de percepción de la corrupción (IPC), 2021. [Online]. Available: https://www.transparency.org/es/news/how cpiscoresarecalculated

[4] Índice Nacional Anticorrupción, "El cambio es cero corrupción" Resultados generales. [Online]. Available: : https://tariatransparencia.gov.co/observatorioanticorrupcion/Paginas/mediciones.aspx

[5] R. Vaitkus, A. Vasiliauskaite, "An Assessment of the Impact of Legal Regulation on Financial Security in OECD Countries," *Journal of Risk and Financial Management*, vol. 15, no. 2, pp. 5592

[6] A. Dwivedi, R. Pant, S. Pandey, M. Pande, A. Mittal, "Benefits of using big data sentiment analysis and soft computing techniques in Egovernance," *International Journal of Recent Technology and Engineering*, vol. 8, no. 3. doi: https://doi.org/10.3390/jrfm15020086 , pp. 1732

[7] Unidad de Inteligencia y Análisis Financiero, Compendio de Notas ALA/CFT, vol. III. Febrero 2017. [Online]. Available: https://www.uiaf.gov.co/salaprensa/publicaciones/notasalacft/compendionotasalacftvolumeniii, pp. 1783

[8] A. Cardoni, E. Kiseleva, F. De Luca, "Continuous auditing and data mining for strategic risk control and anticorruption: Creating "fair" value in the digital age," *Business Strategy, and the Environment*. doi: https://doi.org/10.1002/bse.2558 pp. 30723085 28(1)

[9] R. Barone, D. Masciandaro, F. Schneider, "Corruption and money laundering: You scratch my back, I'll scratch yours," *Metroeconomica,* vol. 73, no. 1.

[10] F. Decarolis, C. Giorgiantonio, "Corruption red flags in public procurement: new evidence from Italian calls for tenders," *EPJ Data Science*, vol. 11, no.1, pp. 5592.

[11] (n.d.). Data mining and its impact on business decision making in the context of crm.

[12] E. Fletcher, C. Larkin, S. Corbet, "Countering money laundering and terrorist financing: A case for bitcoin regulation," *Research in International Business and Finance*, pp. 5592

[13] A. Purwanto, A. Emanuel, Data Analysis for Corruption Indications on Procurement of Goods and Services, In 3rd International Conference on Information and Communications Technology, ICOIACT 2020, pp. 5660

[14] D. Kaufmann A. Kraay, M. Mastruzzi, Governance Matters IV: Indicadores de Gobernabilidad para 1996–2004. Draft, May 9, 2005. [Online]. Available: http://web.worldbank.org/archive/website00818/WEB/GOVMAT 2.HTM.

[15] Secretaría de Transparencia y Anti corrupción, Indicadores de Transparencia y Anticorrupción. 2020. [Online]. Available: http://2020.anticorrupcion.gov.co/Paginas/IndicadoresdeTransparenciaold.aspx

[16] K. Rabuzin, N. Modrusan, 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, 2019, pp. 333-340.

[17] Z. Chang, V. Rusu, J. Kohler, "The Global Fund: why anticorruption, transparency and accountability matter," *Globalization and Health*, vol. 17, no. 1.

[18] B. Craenen, A. Eiben, Computational Intelligence. In Encyclopedia of Life Support Sciences, EOLSS Publishers Co.

[19] P. Wilmott, Machine Learning: An Applied Mathematics Introduction. 2019. Panda Ohana Publishing.

[20] A. Müller, S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists. 2017. O´Reilly Media, pp. 5592

[21] J. Kaplan, Artificial Intelligence. What everyone needs to know. Editorial Teell.

[22] R. Kruse, C. Borgelt, C. Braune, S. Mostaghim, M. Steinbrecher, (n.d.), pp. 5592

[23] B. Craenen, A. Eiben, Computational Intelligence. In Encyclopedia of Life Support Sciences, EOLSS Publishers Co, pp. 5592.

[24] A. Engelbrecht, Computational Intelligence: An Introduction (2nd ed.), John Willey Sons, pp. 5592

[25] S. Russell, P. Norwig, Artificial Intelligence: A Modern Approach, (2nd ed.), Prentice Hall, pp. 5592.

[26] P. Wilmott, Machine Learning: An Applied Mathematics Introduction. Panda Ohana Publishing, pp. 5592

[27] R. Gonzalez, R. Woods, Digital Image Processing. 4th Edition. Pearson Education Limited, pp. 5592.

[28] S. Agarwal, "Trust as a missing link between quality of work life and subjective wellbeing," *Revista Ingeniería Solidaria*, vol. 16, no. 1, Jan. 2020, pp. 5592

[29] A. Müller, S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists. 2017. O´Reilly Media, pp. 5592

[30] S. Anand, V. Verma, A. Gupta Aggarwal, "Dimensional MultiRelease Software Reliability Modelling considering Fault Reduction Factor under imperfect debugging", *Revista Ingeniería Solidaria*, vol. 14, no. 25, pp. 1–12, pp. 5592

[31] Huber, J. Peter, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, pp. 5592

[32] Huber, J. Peter, Wiley series in probability and statistics. Robust statistics (huber/robust statistics), pp. 294–300, 1981, pp. 5592

[33] Wang, Yue, B. Wang, C. Peng, X. Li, H. Yin, "Huber Regression Analysis with a SemiSupervised Method" *Mathematics*, vol. 10, no. 20, pp. 3734. Doi: https://doi.org/10.3390/math10203734, pp. 5592

[34] S. Anand, V. Verma, A. Gupta Aggarwal, "Dimensional MultiRelease Software Reliability Modelling considering Fault Reduction Factor under imperfect debugging," *Revista Ingeniería Solidaria*, vol. 14, no. 25, pp. 1–12, pp. 5592

[35] H. Tong, "Functional linear regression with Huber loss," *Journal of Complexity*, pp. 101696. doi: https://doi.org/10.1016/j.jco.2022.101696, pp. 5592

[36] H. Tong, "Functional linear regression with Huber loss," *Journal of Complexity*, pp. 101696. doi: https://doi.org/10.1016/j.jco.2022.101696, pp. 5592

[38] YANHAAS, La gran encuesta, Elecciones 2022. Advanced market research. [Online]. Available: https://yanhaas.com/