Data and Text Mining for the Detection of Fraud in Public Contracts: A Case Study of Ecuador's Official Public Procurement System

Yeferson Torres-Berru[1,2(✉)] and Vivian Felix López Batista[1]

[1] Universidad de Salamanca Plaza de la Merced, Salamanca, Spain
{ymtorresb,vivian}@usal.es
[2] Instituto Superior Tecnológico Sudamericano, Loja, Ecuador

Abstract. Corruption is present in different forms and typologies, directly affecting the execution of both public and private contracts. The doctoral thesis aims to establish a methodology to prevent and detect corruption automatically in public procurement. By using machine learning techniques and Natural Language Processing (NLP), algorithms for detecting and predicting favouritism and oligopoly are developed. In addition to detecting corruption and its types in the Ecuadorian Public Procurement System (SERCOP) and also visualising the results in an appropriate way, in order to detect and prevent future acts of corruption. In order to analyse the feasibility of the study, a mapping and systematic literature review was carried out, allowing the hypothesis and the methodology to be followed in order to execute and evaluate the developed algorithms. Finally, the detection of favouritism based on process qualification parameters and types of contracting is tested.

Keywords: Corruption · Public procurement · Data mining · Machine learning

1 Introduction

Transparency International estimates that the costs of corruption in public procurement amount to 20–25% of the value of the contract, and can sometimes reach 40–50% [1]. According to the Inter-American Development Bank (IDB), on average, public procurement accounted for 32.5% of general government spending, 29.8% in Latin American countries and 8.6% of gross domestic product (GDP) in the Caribbean. However, the size of expenditure on this item varies roughly from 15 to 47%, due to the higher share of capital expenditure in total expenditure. Ecuador is the third country in the region in capital expenditure,

Doctorado en Ingeniería Informática Universidad de Salamanca.

S. Berrezueta and K. Abad (Eds.): *Doctoral Symposium on Information and Communication Technologies - DSICT*, LNEE 846, pp. 116–127, 2022.
https://doi.org/10.1007/978-3-030-93718-8_10

with a figure close to 45%, as well as being the first country in the region in terms of public procurement in terms of GDP, with 16% [2,3].

In Ecuador, SERCOP is in charge of promoting citizen participation, increasing access to and use of public information by the population. Increasing transparency, combating fraud and corruption that could arise from malpractices in public procurement. The platform provided by SERCOP contains documents in PDF format for each contracting process, where Data on the Specifications (TDR in Spanish), invitations to suppliers, bids made, observations, in short, all the documentation generated by the purchase are stored. Among the processes present in SERCOP, the following are available:

- Execution of works.
- Procurement of products and services.
- Contracting of the consultancy.

Different forms and levels of corruption have emerged, namely bribery, embezzlement, fraud, extortion, breach of trust, collusion and favouritism [5,8]. The main corruption mechanisms according to [6,7] are: non-existence of contract, inappropriate contracting, fractioning, contract modifications. As can be seen, the study of corruption covers a wide range of social and human sciences, with theoretical and scientific contributions.

It is also established that for there to be corruption in a process, the following factors must be taken into account [9,10]:

- The type of product, the amount of contracting and the type of purchase.
- The bidding period and validity.
- Modifications during the execution of the process (sheets, parameters, questions and answers).
- Changes in the percentages in the qualifying parameters.
- The personal relationships between persons of the contracting companies.
- The specific experience required of a supplier.
- The detailed technical specifications.

The data mining plays an important role in corruption detection, and is most commonly used to find hidden information in large amounts of data [11]. The corruption case known as *Panama papers* [12] revealed tax and financial fraud to the public opinion; in Brazil the Observatory of Public Expenditures [13] reviewed more than 120,000 public contracts and uncovered more than 7,500 cases involving $ 104 million in financial operations of dubious legality. These examples illustrate the importance of data science in the fight against corruption. This doctoral thesis aims to implement a methodology to prevent and detect corruption automatically, in public procurement in SERCOP, with the use of Artificial Intelligence (AI) techniques such as machine learning and PLN.

2 Research Work Development

In order to form the state of the art on the investigation topic, we conducted a systematic bibliography search. We followed the methodology proposed by [14],

that divides the process into three phases: planning, carrying out the review and elaboration of the paper. As a product of this, two articles were written and indexed in SCOPUS (Quartile 3).

Work 1

Systematic mapping was developed [15] of scientific publications (2015–2019), centered on contractual corruption in its various forms, by applying data mining and machine learning techniques. We present six research questions to answer the analysis of 147 articles obtained from the Web of Science (WoS) and Scopus databases. The detection of fraud, financial fraud and corruption predominate in the investigations, the most common forms of corruption are fraud (72.72%) and overpricing (8.84%). The investigations were carried out in the United States (16.32%), China (10.88%), the United Kingdom (8.94%) and in Latin America, mainly in Brazil (3.4%), with minimal contributions from Colombia and Paraguay. Figure 1 shows a survey of the articles consulted according to the country in which they were published and the type of corruption.

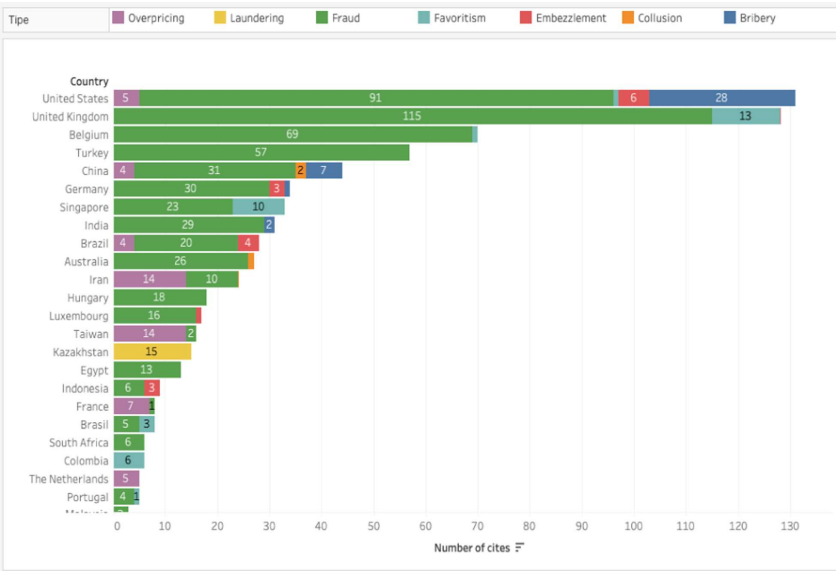

Fig. 1. Articles published by countries and type of corruption.

Work 2

In their study Torres-Berru et al. [16] analyzes different types of corruption (bribery, collusion, embezzlement, fraud, abuse of discretionary power, favoritism, nepotism) and six types of IA techniques (classification, regression, *clustering*, prediction, outlier detection and visualization). The methodology proposed by Torres-Carrión was used, [14], and four investigation questions were posed to find out the types of searches performed, the characteristics of the

organizations in which the investigations are carried out, the technological tools and the data mining methodologies and techniques. The review was conducted in the Scopus and WoS databases, obtaining 108 articles published between 2015 and 2019.

As a result of this first phase of documentation, as part of the development of the research work, we can summarize that the *Web Scraping* It is a little-used technique to obtain data on corruption studies in contracts. Its use can serve as a basis for future data collection. The few works related to contract analysis in public procurement use isolated data sets and do not consider documents as an initial basis for analysis. It is also evident that the software tools developed for the analysis of corruption in contracts, both in the public and private sectors, are not considered as computer security standards, and the percentage of tools in the web environment is very low.

The main techniques of AI found are logistic models, neural networks, Bayesian networks and support vector machine. The Fraud Score is proposed as a specific metric for assessing corruption risk. Also taken into account are the metrics used to evaluate classification within machine learning, based on the matrix of confusion and (*Receiver Operating Characteristic*) the ROC curves. In addition, supervised learning is the most widely used technique when applying machine learning models in this area.

2.1 Thesis Objectives

Taking into account the theoretical basis discussed in the previous section, it is chosen to work on favouritism, for in [15] work sit is evident that out of 147 articles only 8 deal with this type of corruption.

The objective is focused on establishing a methodology to prevent and detect corruption automatically, through the use of algorithms for the detection of corruption in public procurement. Using machine learning techniques and PLN, the aim is to develop algorithms for detecting and predicting favouritism and oligopoly in public procurement.

In addition, the following **hypothesis** was formulated.

The evaluation of the data and text generated in the phases of a public procurement process facilitates the detection of the presence of corruption, its type, the phase in which it occurs and the detriment to the state.

3 Methodology for Intelligent Discovery of Corruption

After reviewing the different approaches that exist in the current literature on corruption, which attempt to provide an answer to the problem posed, this section details the proposal of the present work, designed to test the hypothesis of the present study. In general, we describe a data mining system that uses PLN to intelligently perform content analysis of contracts and automatically detect corruption. In particular, the two approaches that have been developed and

how the experimentation phase is planned to respond to different shortcomings detected in the field of study.

The work aims to assess favouritism and oligopoly as a common form of procurement, which leads to other forms of corruption such as price increases, irregular processes, bribery, etc. Favouritism is the natural human propensity to favour friends, family and anyone close and reliable within a public process, nepotism means the granting of favours to persons who are related to the official holding a public office [8]. To evaluate the favouritism, the variables listed in Table 1 are considered.

Table 1. Variables for assessing favouritism

Variable	Description
Economic offer	Awarding low scores for the evaluation of economic offers, causing high priced purchases to the detriment of the state
Other parameters	Conditions that only one agreed bidder will be able to meet, and to which they will award very high scores
Time	Very brief terms are stipulated for the design, preparation, drafting and submission of proposals. If all the requirements are fulfilled in anticipation, someone will submit the proposal within the established terms
Experience	Entities in order to direct the procedures to the bidders with whom they have previously cooperated before request specific experience with the same entity, which leaves no possibility for new bidders to be awarded
Technical specifications	Institutions include requirements and/or equipment that only satisfy The supplier with which it reached an agreement
Change conditions	The Entity modifies certain parameters in the "Questions, Answers and Clarifications" stage, in order to ensure that the selected supplier's offer obtains the highest scores by including requirements and/or equipment that only that supplier fulfills
Relations between individuals	A group of companies are involved in all tenders of an entity, agreeing beforehand, between them, who is to win the process. In addition, the entity and suppliers arrive at an agreement to generate a kind of oligopoly, where they seek to generate specifications for the benefit of this group of bidders

3.1 Machine Learning Algorithms

The literature review found that most of the published work uses supervised learning, as contracts with anomalies are correctly labelled, 79% of the research corresponds to detection and 21% to prediction. Ecuador's public procurement system does not have an anomalous procurement section. Therefore, lacking labelled data, in the initial phase of the research, the decision was made to

use unsupervised learning techniques to detect anomalous patterns in contracts. The proposed methodology is summarised in Fig. 2. Once the retrieved public procurement data is processed, it is analysed through a multi-stage model, which uses different algorithms, like *clustering* (K-Means), Self-organizing map (SOM), Support Vector Machine (SVM) y Deep Learning.

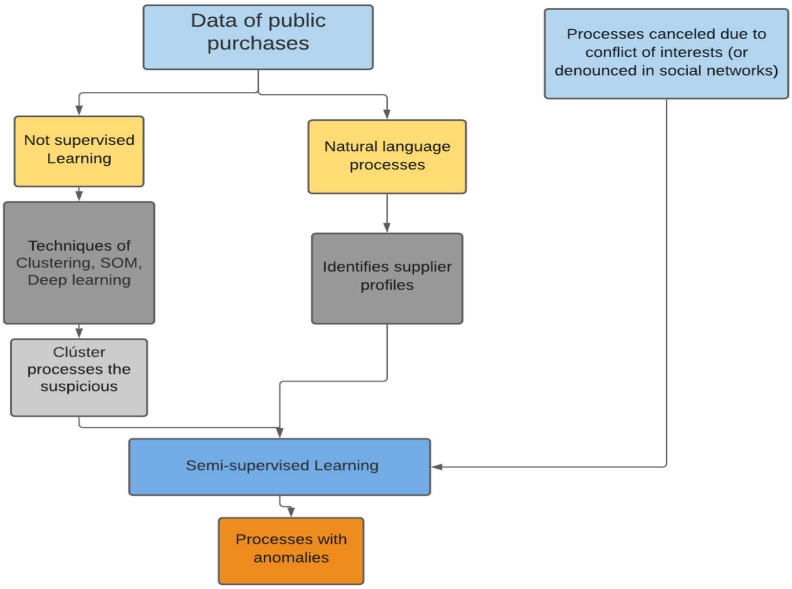

Fig. 2. Anomaly detection model

4 Experimentation

4.1 Data Collecting

In accordance with the proposed methodology, the first step is to retrieve data on public procurement. Because there is no open access portal to the data, the web-based *scraping* technique is applied [19] on the data provided on the website of the National Public Procurement Service of Ecuador[1]. Through this process, information is obtained regarding public processes from 2010 to 2020, as well as the documents (attachments) of each process. Information was retrieved for a total of 1276867 procurement processes.

Considering the process *URL* as input, the different fields of each process are: its description, dates, products, qualification parameters, invitations, files and supplier questions. Each section was extracted according to its equivalent html tag through *scraping* and stored in a non-relational database (MongoDB). Once the data is obtained, it is sorted to remove noise and inconsistency and reduce dimensionality.

[1] https://www.compraspublicas.gob.ec/ProcesoContratacion/compras/.

4.2 Training Phase

Clustering is one of the most popular unsupervised learning techniques. It is used to analyse data and find groups within that data using some kind of similarity measure, such as Euclidean distance. For this we should assume that the number of clusters is known in advance, so we start by setting the cluster number to 8 to be possible to classify our data accordingly. There is no universal similarity metric that works for all cases (it depends on the problem itself). So we iterated to update the centroids until they stopped changing and have been placed in optimal locations to cluster the observations into 8 different centroids. The *elbow* method in $k = 4$ suggests that it is the optimal value for the number of clusters with 10000 iterations to obtain the best result in the evaluation metric.

Self-organising maps (SOM) are unsupervised artificial neural networks based on the winner-takes-all principle [22]. A typical SOM consists of a layer of input neurons, an array of nodes as an output map and an array of connections between each unit of the output layer and all units of the input layer. Input nodes are propagated to a set of output nodes, which are organised into topographic maps, which determine how the spatial location of an output node on the topographic map corresponds to a particular feature of the input data pattern [20]. A rectangular topology consisting of 20 rows and 20 columns is used, allowing individual features to connect to a node and set weights. Each input neuron i is connected to each of the output neurons j by a weight W_{ji}. In this way, the output neurons have an associated vector of weights W_j called the reference vector, allowing the map to make a projection from a multidimensional data space to a two-dimensional map of neurons. To find the proximity between the data, the neighbourhood is evaluated *gaussian* which causes the change of values to decrease with distance and a bubble neighbourhood, which changes all vectors belonging to the neighbourhood to the same shape.

With the 4 clusters obtained (representing the types of contract), semi-supervised learning is applied for anomaly detection (Fig. 3) which is detailed below: the first step is to separate the datasets into a training set (80%) and a test set (20%). For this technique, the processes associated with the clusters where the economic factor is the main factor for the qualification are defined as "normal", subsequently two models are compared:

1. **One-Class Support Vector Machine**. SVM [23] is a supervised type classifier, which is defined by a hyperplane between classes. Given labelled training data and a binary classification problem, the SVM finds the optimal hyperplane that separates the training data into two classes. The algorithm requires training data with two labels, belonging to one of the two classes. The problem appears when you want to apply the algorithm on data where there is a lot of information for one class, but not for the other. In these cases, SVM can be used as an anomaly detector, as a classifier of a class. The model is trained on the data of the class that is considered normal, and data that is different can be predicted. When this version of the algorithm is applied, we have used the property *nu* which allows to control the balance between outliers and normal cases, and therefore assigned nu= [1e−3, 1e−2, 1e−1, 1], while the parameter

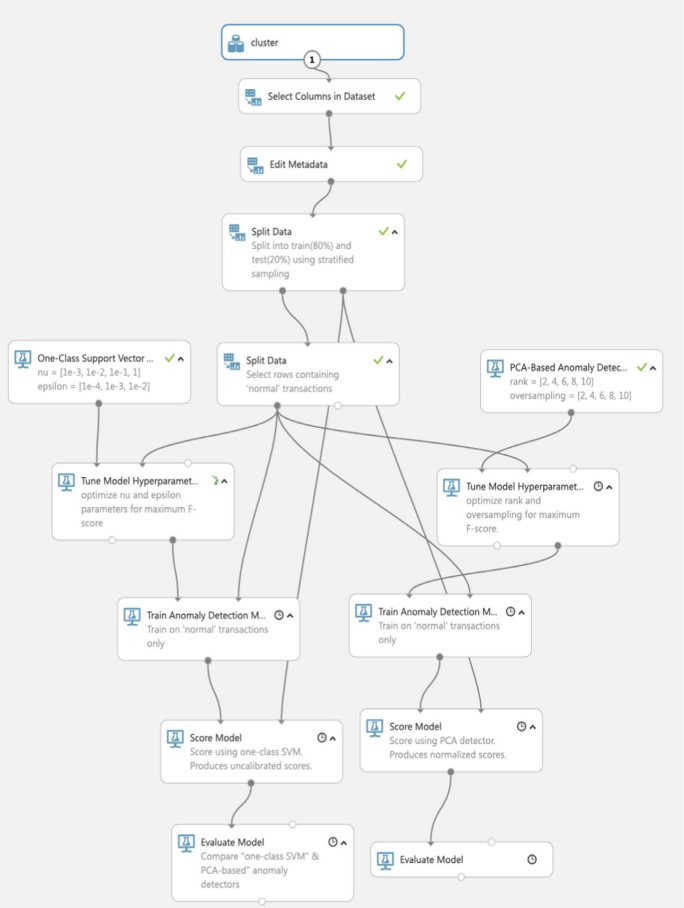

Fig. 3. Azure anomaly detection model

affecting the number of iterations used, when optimising the model, is defined as *epsilon* = [1e−4, 1e−3, 1e−2]. The optimal hyperplanes for machine learning are then determined using a *Hyperparameters* Model. Finally, the model is trained and evaluated using the ROC and *accuracy* metrics.

2. **Principal Component Analysis (PCA)**. The anomaly detection module based on PCA analyses the available features to determine what constitutes a "normal" class, and applying distance metrics to identify the cases that represent anomalies, therefore used with a range of parameters (*rank*) y *oversampling* de [2, 4, 6, 8, 10]. Finally, the model is trained and evaluated using the Score Model, the ROC and *accuracy* metrics.

4.3 Study Cases

In order to validate the functioning of the proposed method, two case studies have been carried out: one to detect anomalies in the qualification parameters in contracts (in the process of publication), allowing the identification of processes in which favouritism exists, and another to investigate price speculation in medical products generated by COVID-19 in Ecuador (published).

Anomaly Detection in Contract Qualification Parameters. To create the dataset for the first case study, the bid scoring parameters were evaluated in 275,730 public procurement contracts in Ecuador, between 2010 and 2020. Twenty-three variables were evaluated to determine the winner of each contract process, among them: economic offer, experience, equipment and instruments, national production, "Other parameters", etc. In this way, it is possible to determine in which processes low scores are awarded to evaluate the economic offer, causing prejudice to the government.

As a result of the experimentation phase, applying SOM and Kmeans, we can highlight that in 3 of the 4 clusters the economic offer is respected as an outstanding qualification parameter.

Note that the semi-supervised learning model applying SVM and PCA can be applied in the evaluation of the regression model and for the detection of anomalies in the processes, taking into account that they mostly belong to the *cluster 4*. As metrics for evaluating the success of the applied algorithms, the following were used: precision (0.95%) and *accuracy* (0.85%).

Exceptional Prices of Medical Supplies During the COVID-19 Pandemic in Ecuador. An exploratory data analysis explores the prices of procurement of supplies, through public procurement contracts in Ecuador, for use in clinical settings or as protection for the general population in response to the COVID-19 pandemic. The study [24] quantifies the differences in the prices of commonly procured medicines and commodities in public procurement contracts identified as related to the COVID-19 pandemic. Statistical analysis was performed to extract relevant measures for each product and to determine variability over all products.

As a result, Ecuador was found to have spent $257 million on public procurement of basic supplies related to COVID-19. Among the most purchased products were masks, paracetamol and PCR tests. Prices varied widely, depending on the individual contract and the number of units purchased. Some prices were exceptionally higher than their market value and as much as 1300% difference with similar purchases. Compared with 2019, the mean price of medical examination gloves increased up to 1,307%, acetaminophen 500 mg pills, up to 796%, and oxygen flasks, 30.8%.

Figure 4 shows the significant price increase in procurement of medical products in April 2020, despite not being as high in demand as in previous months.

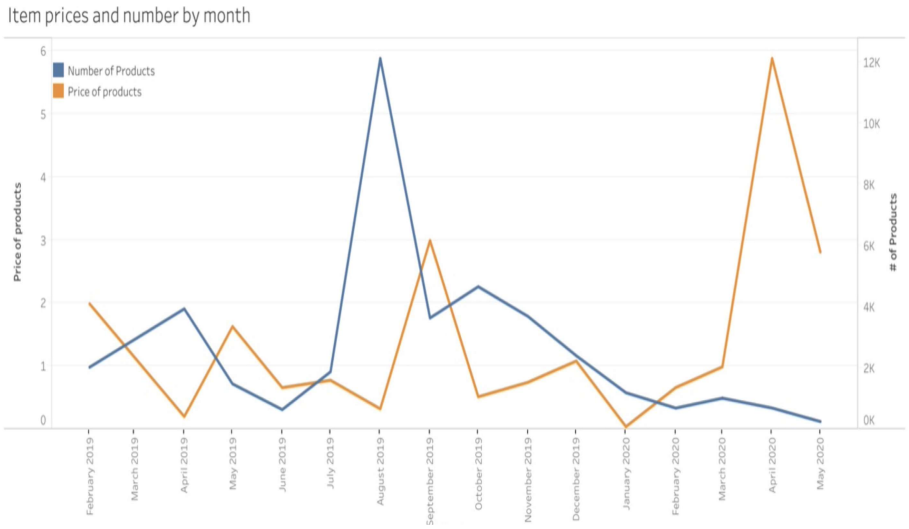

Fig. 4. COVID-19 pandemic medicine price developments.

5 Conclusions

In this article we have summarised the methodology we are developing to prevent and detect corruption in public procurement, using machine and deep learning algorithms. The theoretical basis of the doctoral work has been established, which has led to the publication of three scientific articles. In addition, the few exploitation of favouritism as a form of corruption in current research has been noted. A scientific methodology has been defined according to the hypothesis, based on a hybrid model that includes different phases with supervised and unsupervised learning, PLN and neural networks. This methodology is being tested in two case studies. One related to contract qualification and the other to investigate price speculation in medical products generated by COVID-19 in Ecuador.

6 Future Work

As a future line of work, the aim is to build a *framework* that evaluates, detects and helps in the prediction of favouritism in public procurement processes. In addition to incorporating *Deep Learning* algorithms in the methodology, as well as the NLP for the classification of contractors and relations with the entities, assessing award times.

References

1. Náez Gómez, J.E.: Relación entre el Índice de Control de la Corrupción y algunas variables sociales, económicas e institucionales. Nómadas. Rev. Crítica Ciencias Soc. y Jurídicas 38 (2013)

2. Brito-Gaona, L.F., Iglesias, E.M.: Inversión privada, gasto público, presión tributaria en América Latina. Estududios de Economia. **44**, 5–30 (2017)
3. Izquierdo, A., Pessino, C., Vuletin, G.: Mejor gasto para mejores vidas: Cómo América Latina y el Caribe puede hacer más con menos, vol. 10. Inter-American Development Bank (2018)
4. Servicio Nacional de contratacion publica: Rendicion de cuentas (2018)
5. Moran, J.: Democratic transitions and forms of corruption. Crime, Law Soc. Chang. **36**, 379–393 (2001).https://doi.org/10.1023/A:1012072301648
6. Castro Cuenca, C.G.: La corrupción pública y privada: causas, efectos y mecanismos para combatirla - Google Play (2017)
7. Cassagne, J.C., Rivero Ysern, E.: La contratación pública, Hammurabi (2007)
8. Vargas-Hernández, J.G.: The Multiple Faces of Corruption: Typology, Forms and Levels. SSRN Electron. J. (2009). https://doi.org/10.2139/ssrn.1413976
9. Ponce, H.G., Gil, M.T.N., Durán, M.P.: Responsible public procurement. Des. meas. indicators. CIRIEC-Espana Rev. Econ. Publica, Soc. y Coop. **44**, 253–280 (2019)
10. Subdirección General de Control Coordinación Técnica de Controversias: Manual De Buenas Prácticas En La Contratación Pública Para El Desarrollo Del Ecuador. 1-46 (2015)
11. Alvarez-Jareño, J.A., Badal-Valero, E., Pavia, J.M.: Aplicación de métodos estadísticos, económicos y de aprendizaje automático para la detección de la corrupción. (2019)
12. Woodie, A.: Inside the Panama Papers: How Cloud Analytics Made It All Possible. https://www.datanami.com/2016/04/07/inside-panama-papers-cloud-analytics-made-possible/. Accessed 12 Aug 2019
13. Controladoria-Geral da União: Observatório da Despesa Pública - Controladoria-Geral da União. http://www.cgu.gov.br/assuntos/informacoes-estrategicas/observatorio-da-despesa-publica. Accessed 12 Aug 2019
14. Torres-Carrión, P.V., Gonzalez-Gonzalez, C.S., Aciar,S., Rodriguez-Morales, G.: Methodology for systematic literature review applied to engineering and education. In: IEEE Global Engineering Education Conference, EDUCON 2018-April, pp. 1364–73 (2018)
15. Torres-Berru, Y., López-Batista, V.F., Torres-Carrión, P.: Data mining to detect and prevent corruption in contracts: Systematic mapping review. RISTI - Rev. Iber. Sist. e Tecnol. Inf. **2020**, 13–26 (2020)
16. Torres Berru, Y., López Batista, V.F., Torres-Carrión, P., Jimenez, M.G. : Artificial Intelligence techniques to detect and prevent corruption in procurement: a systematic literature review. In: Botto-Tobar M., et al. (eds) Applied Technologies. ICAT 2019. Communications in Computer and Information Science, vol. 1194. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-42520
17. Hotelling, H.: A generalized T test and measure of multivariate dispersion. In: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, pp. 23–41 (1951)
18. Ultsch, A., Mörchen, F.: ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM, pp. 1–7. Tech. Rep. Dept. Math. Comput. Sci. Univ. Marburg Ger (2005)
19. Saurkar, A.V., Gode, S.A.: An overview on web scraping techniques and tools. Int. J. Futur. Revolut. Comput. Sci. Commun. Eng. **4**, 363–367 (2018)
20. Merkl, D.: Text classification with self-organizing maps: some lessons learned. Neurocomputing **211–3**, 61–77 (1998)

21. Vettigli, G.: MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map
22. Kohonen, T.: Self-organizing Maps. Springer-Verlag, Berlin (1995)
23. Kotsiantis, S.B.: Supervised machine learning: a review of classification techniques. Informatika (2007)
24. Ortiz-Prado, E., Fernandez-Naranjo, R., Torres-Berru, Y., Lowe, R., Torres, I.: Exceptional prices of medical and other supplies during the COVID-19 pandemic in Ecuador. Am. J. Trop. Med. Hyg. **105**, 81–87 (2021). https://doi.org/10.4269/ajtmh.21-0221