Minería de datos y texto para la detección de fraude en contratos públicos: Caso de estudio Sistema Oficial de Contratación Pública de Ecuador

TESIS DOCTORAL

UNIVERSIDAD DE SALAMANCA

Departamento de Informática y Automática

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Autor:

Yeferson Torres Berrú

Salamanca, 2023

UNIVERSITY OF SALAMANCA

DOCTORAL THESIS

Departamento de Informática y Automática
Facultad de Ciencias
Grupo de investigación:
MIDAS

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

Minería de datos y texto para la detección de fraude en contratos públicos: Caso de estudio Sistema Oficial de Contratación Pública de Ecuador

AUTOR:
Yeferson Mauricio Torres Berrú

DIRECTOR:
Vivian Félix López Batista, PhD.

Declaración de Autoría

La Dra. Vivian Félix López Batista, Profesora Titular de Universidad del Departamento de Informática y Automática de la Universidad de Salamanca.

HACE CONSTAR:

Que, el doctorando Yeferson Mauricio Torres Berrú ha desarrollado el trabajo titulado “Minería de datos y texto para la detección de fraude en contratos públicos: Caso de estudio Sistema Oficial de Contratación Pública de Ecuador” bajo su supervisión, y por ello autoriza su presentación para la obtención del título de Doctor.

Vivian Félix López Batista

Firmado digitalmente por Vivian Félix López Batista
Fecha: 2023.07.17 18:18:47 +02'00'

Vivian Félix López Batista, PhD.

DEDICATORIA

Dedicado a aquellos que han iluminado mi camino con su sabiduría y apoyo inquebrantable.

A mi padre, aunque ya no esté físicamente presente, su legado perdura en mi corazón y en cada paso que doy en este viaje. Esta tesis es un tributo a tu espíritu perseverante y al amor incondicional que siempre me brindaste.

A mi madre, por su sacrificio y por inculcarme la importancia del conocimiento y la perseverancia. A mis hermanos, por su aliento constante y por ser mi fuerza motivadora. A mi querida pareja, por su paciencia, comprensión y por ser mi apoyo en los momentos más desafiantes.

Esta tesis está dedicada a todos aquellos que han dejado una huella en mi vida y han sido parte fundamental de mi crecimiento académico y personal.

AGRADECIMIENTOS

A todos aquellos que, de alguna manera, han contribuido a este logro, ya sea con palabras de aliento, gestos amables o simplemente estando allí cuando más los necesitaba. Vuestra presencia ha sido fundamental para mi éxito.

A mi pareja, cuyo amor, paciencia y comprensión han sido mi ancla en los momentos de incertidumbre y cansancio. Agradezco por cada palabra de aliento y por celebrar cada logro junto a mí.

A mi familia, cuyo amor incondicional me ha dado la fuerza para perseguir mis sueños y superar cualquier obstáculo. Vuestra constante fe en mí ha sido mi mayor motivación, agradezco vuestra comprensión y por creer en mí cuando más lo necesitaba.

A mi querida tutora y directora, cuyo apoyo y guía experta han sido fundamentales en el desarrollo de este trabajo de investigación. Tu conocimiento profundo, tus consejos sabios y tu compromiso con mi crecimiento académico han sido una inspiración constante.

Agradezco también a las instituciones académicas y a los programas de investigación que han brindado los recursos necesarios para llevar a cabo este proyecto.

Con gratitud eterna.

Yeferson.

RESUMEN

En el mundo globalizado actual, la corrupción según afirma la ONU, está presente en todos los países, en diferentes formas y tipologías. Afectando directamente en la ejecución de contratos tanto públicos como privados. Los avances en las Tecnologías de la Información (TIC) y la Inteligencia Artificial (IA) han propiciado la lucha contra esta lacra a través del uso de la tecnología. Aportando y tomando como aliada el fomento de la transparencia en los actos de contratación, que constituyen la clave para lograr la integridad del proceso de adquisición gubernamental.

Las iniciativas actuales se centran en el análisis e identificación de riesgos de corrupción en las compras públicas, pues según afirma Transparencia Internacional, es este proceso el que alcanza los mayores costos de la corrupción. La detección de corrupción de forma automática requiere una gestión avanzada de importantes cantidades de información con técnicas típicas de minería de datos, aprendizaje automático y analítica avanzada. Pero la mayoría de las propuestas recogidas en la literatura, en la exploración de estos datos, solo están enfocadas a la resolución de problemas muy específicos. Evidenciándose que algunos tipos de corrupción como el favoritismo, han sido poco estudiados. Así como los sesgo o la combinación de varios tipos de corrupción. Por lo que en la actualidad se han venido desarrollando nuevos trabajos de investigación sobre el uso de los datos, asociadas al proceso de compra pública, para abordar estas carencias, que han llamado nuestra atención. En esta tesis doctoral se establece una metodología para prevenir y detectar corrupción de forma automática, desde la información generada en la contratación pública, mediante el uso de técnicas de aprendizaje automático y Procesamiento del Lenguaje Natural (PLN). Con el objetivo de comprobar el potencial de la metodología propuesta, se desarrollan algoritmos de detección y predicción de favoritismo, sobreprecio y sesgo en contratos elaborados con instituciones del estado. La investigación geográfica, también refleja que Latinoamérica, se considera un nicho de investigación sobre el tema, al no existir una gran cantidad de trabajos en el área. En consecuencia, se diseñaron diferentes casos de estudio a través del Servicio Nacional de Contratación Pública (SERCOP) de Ecuador. Permitiendo validar la metodología, para la detección de sobreprecio en compras de medicamentos durante la pandemia COVID 2019, el favoritismo basado en los parámetros de calificación de procesos y los tipos de contratación. Además, del favoritismo y sesgo basado en el texto generado en cada proceso de compra pública. En cada caso se emplean diversas técnicas de extracción de información, que se aplican a los algoritmos de aprendizaje automático y luego se comprueba su rendimiento a través de diferentes métricas. La metodología permite validar la hipótesis planteada: a través de la evaluación, de los datos y el texto generado en las fases de un proceso de compra pública, mediante la combinación de técnicas de minería de datos y PLN, se puede identificar la corrupción, con resultados muy alentadores para su detección y prevención en compras del

sector público. Se pudo comprobar que entre el 30% y el 35% de los procesos de compra pública presentaban indicios de corrupción independientemente de su tipo.

ABSTRACT

In the current globalized world, corruption, as stated by the United Nations, is present in all countries, in various forms and typologies, directly affecting the execution of both public and private contracts. Advances in Information Technology (IT) and Artificial Intelligence (AI) have facilitated the fight against this scourge using technology, serving as an ally in promoting transparency in procurement practices, which is key to achieving integrity in government procurement processes. Current trends focus on analyzing and identifying corruption risks in public procurement, as Transparency International asserts that this process incurs the highest costs of corruption. Automatic detection of corruption requires advanced management of significant amounts of information using typical techniques such as data mining, machine learning, and advanced analytics. However, most proposals found in the literature regarding the exploration of this data are only focused on solving very specific problems, highlighting those certain types of corruption, such as favoritism, have been under-studied, as well as biases or the combination of various types of corruption. Consequently, new research efforts have been carried out on the use of data associated with the public procurement process to address these gaps that have drawn our attention. This doctoral thesis establishes a methodology for the automatic prevention and detection of corruption based on information generated in public procurement, utilizing machine learning techniques and Natural Language Processing (NLP). To verify the potential of the proposed methodology, detection and prediction algorithms are developed for favoritism, overpricing, and bias in contracts with government institutions. Geographically, the research also reflects that Latin America is considered a research niche in this field due to the lack of a significant number of works in the area. Consequently, different case studies were designed through Ecuador's National Public Procurement Service (SERCOP) to validate the methodology, including the detection of overpricing in medication purchases during the COVID-19 pandemic, favoritism based on process qualification parameters, and types of procurement. Additionally, favoritism and bias based on the text generated in each public procurement process are examined. Various information extraction techniques are employed in each case, applied to machine learning algorithms, and their performance is assessed using different metrics. The methodology validates the hypothesis proposed: through the evaluation of data and text generated in the phases of a public procurement process, by combining data mining and NLP techniques, corruption can be identified, yielding highly encouraging results for its detection and prevention in public sector purchases. It was found that between 30% and 35% of public procurement processes showed indications of corruption regardless of their type.

INDICE

Parte I
MODALIDAD DE LA TESIS

1: MODALIDAD DE LA TESIS

La presente tesis doctoral de la Universidad de Salamanca está realizada bajo el formato de compendio de artículos. Esta tesis incluye cuatro contribuciones publicadas en revistas y dos capítulos de libro.

1.1. Autorización del supervisor

Vivian Félix López Batista, Profesora Titular del Departamento de Informática y Automática la Universidad de Salamanca y directora de la tesis doctoral de Yeferson Mauricio Torres Berrú.

Autoriza:

Que, Yeferson Torres Berrú presente y defienda su tesis doctoral en la modalidad de compendio de artículos.

Vivian Félix López Batista

Firmado digitalmente por Vivian Félix López Batista
Fecha: 2023.07.17 18:19:51 +02'00'

Vivian Félix López Batista, PhD.

1.2. LISTA DE CONTRIBUCIÓN

1.2.1. Contribución 1

Título: "Exceptional prices of medical and other supplies during the COVID-19 Pandemic in Ecuador". *American Journal of Tropical Medicine and Hygiene,* vol. 105, no. 1, pp. 81-87, 2021.

***DOI*:** https://doi.org/10.4269/ajtmh.21-0221

Autores:

- Esteban Ortiz Prado
- Raúl Fernández Naranjo
- Yeferson Torres Berrú
- Rachel Lowre
- Irene Torres

Revista*:* American journal of tropical medicine and hygiene (ISSN 0002-9637).

Índices de calidad:

- **WoS JCR Impact Factor 2021:** 3.707**. Rank** 7/24 (Q2**). Area:** Tropical Medicine.
- **SCOPUS Cite Score 2021:** 5.4**. Rank:** 128/295 (Q1). **Área:** Medicine.

1.2.2. Contribución 2

Título: "Data mining to identify anomalies in public procurement rating parameters". *Electronics*, vol. 10, no. 22, pp. 1–15, 2021.

DOI: https://doi.org/10.3390/electronics10222873

Autores:

- Yeferson Torres Berrú
- Vivian F. López Batista

Revista: MDPI Electronics (ISSN 2079-9292)

Índices de calidad:

- **WoS JCR Impact Factor 2021:** 2.690**. Rank:** 100/164 (Q3)**. Área:** Computer Science, Information Systems**.**
- **SCOPUS Cite Score 2021:** 3.7. **Rank:** 146/359 (Q2**). Área:** Computer Science.

1.2.3. Contribución 3

Título: "A Data Mining Approach to Detecting Bias and Favoritism in Public Procurement". *Intelligent Automation & Soft Computing*. Volume 36, number 3, pp. 3501-3516, 2023.

DOI: https://doi.org/10.32604/iasc.2023.035367

Autores:

- Yeferson Torres Berrú
- Vivian F. López
- Lorena Conde

Revista: Automation & Soft Computing

Índices de calidad:

- **WoS JCR Impact Factor 2022:** 2.0. **Rank**: 43/65 (Q3). **Área:** Automation & Control Systems.
- **Scopus:** 3.0 **Rank**: 69/165 (Q2). **Área:** Computer Science.

1.2.4. Contribución 4

Título: Data mining to detect and prevent corruption in contracts: Systematic mapping review. RISTI- *Revista Iberica de Sistemas e Tecnologias de Informacao*, vol. 2020, no. E29, pp.13–26, 2020.

Autores:

- Yeferson Torres Berrú
- Vivian F. López
- Pablo Vicente Torres-Carrión

Revista: RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao (ISSN 1646-9895).

***Índices de calidad*:**

- **SCOPUS Cite Score 2019:** 0.4. **Rank:** 197/221 (Q4). **Área:** General Computer.

1.2.5. Contribución 5

Título: Artificial intelligence techniques to detect and prevent corruption in Procurement: A Systematic Literature Review, *in Communications in Computer and Information Science*, vol. 1194, pp. 254–268, 2020.

DOI: https://doi.org/10.1007/978-3-030-42520-3_21

Autores:

- Yeferson Torres Berrú
- Vivian F. López
- Pablo Vicente Torres-Carrión
- María Gabriela Jiménez

Libro: Lecture Notes in Communications in Computer and Information Science (ISSN 978-3-030-42519-7).

Disponible en: https://link.springer.com/chapter/10.1007/978-3-030-42520-3_21

Índices de calidad:

- **SCOPUS Cite Score 2019:** 0.7**. Rank:** 180/221 (Q3). **Área**: General Computer Science.

1.2.6. Contribución 6

Título: Data and Text Mining for the detection of fraud in public contracts: A case study of Ecuador's Official public Procurement System, in Doctoral Symposium on Information and Communication Technologies - DSICT. *Lecture Notes in Electrical Engineering*, vol. 846, pp.116–127, 2022.

DOI: https://doi.org/10.1007/978-3-030-93718-8

Autores:

- Yeferson Torres Berrú
- Vivian F. López

Libro: Lecture Notes in Electrical Engineering (ISSN:1876-1100)

Disponible en: https://link.springer.com/chapter/10.1007/978-3-030-93718-8_10

Índices de calidad:

- **Scopus:** 0.6. **Rank:** 280/338 (Q4). **Área:** Engineering

CAPÍTULO 2

En el presente capítulo se describe el área de la ciencia en la cual se enfoca la tesis doctoral. La sección 2 se inicia con la introducción. En la sección 2.1 se detalla la motivación del trabajo, la sección 2.2 muestra la metodología de investigación, la sección 2.3 la hipótesis y la 2.4 los objetivos a desarrollar. Finalmente se resume la estructura de la memoria en la sección 2.5.

2. INTRODUCCIÓN

La corrupción tiene muchas formas de manifestarse, por lo tanto, no existe una definición universal legitimada. Sin embargo, podríamos resumir la corrupción como: *un acto viciado en el cual una persona obtiene ganancias ilegitimas (económicas o de otro tipo) al aprovecharse de una situación concreta para romper la ley y beneficiarse tanto a uno mismo como al otro participante en el acto (si lo hubiera)* [1]. Muchos investigadores han desarrollado diferentes trabajos para clasificarla. Como es el caso de *Democratic Transitions and Forms of Corruption* [2] y de *Research on Corruption A Policy Oriented Survey* [3], donde los autores coinciden en que la corrupción puede ser activa o pasiva dependiendo de la persona que tiene el poder de decisión. Dentro del sector estatal, la contratación pública [4] es una de las más vulnerables a la corrupción, por el volumen de recursos que maneja y los múltiples casos de corrupción que se denuncian anualmente. Considerándose que la "transparencia, responsabilidad y profesionalidad" son la clave para lograr la integridad del proceso.

Transparencia Internacional [5] estima que los costos de corrupción en las compras públicas alcanzan entre el 20 y el 25% del valor del contrato, e incluso pueden llegar hasta el 50% en algunos casos. En Latinoamérica, las compras públicas constituyen aproximadamente el 33% del presupuesto general de los países de la región. Esta situación ha impulsado tanto al sector público como al privado a incrementar sus esfuerzos para regular y garantizar la transparencia en estas transacciones. Como alternativa para detectar prácticas corruptas y evitar el control público, se han desarrollado nuevas técnicas y métodos para identificar la corrupción en las personas o empresas involucradas en las compras públicas. Estas prácticas se basan en estrategias como favorecer a ciertos proveedores a través de términos direccionados, utilizar contratos poco transparentes y fraccionar contratos para evitar la atención y supervisión adecuada. Además, se aplican sobreprecios y se sesga la selección de proveedores mediante la modificación de los términos del contrato [6, 7].

Los procesos de compras públicas carecen generalmente de transparencia y resultan difíciles de comprender para la población, lo que dificulta la detección de actividades corruptas [6]. Además, en muchos países que no priorizan la detección de la corrupción en las compras públicas, son limitados los recursos disponibles para que las autoridades encargadas investiguen y combatan este tipo de delitos [1]. A menudo, los casos de corrupción son denunciados por los medios de comunicación, las redes sociales, los funcionarios involucrados o encontrados en las auditorías. Sin embargo, los sistemas actuales de detección de corrupción en compras públicas suelen estar desactualizados y no utilizan tecnologías de

vanguardia, como la Inteligencia Artificial (IA), para analizar grandes cantidades de datos y detectar patrones de corrupción. Esto motiva un aumento de riesgo de irregularidades al hacerse prácticamente manual la validación de los casos de corrupción.

Existen varias técnicas mediante las cuales se puede detectar la corrupción en compras públicas, una de ellas es el análisis de patrones en los datos relacionados al proceso como: precios, empresas contratadas y fechas de las compras [8–10]. En [11] se implementan mecanismos de transparencia y rendición de cuentas en las compras públicas, como la publicación de información detallada sobre los procesos de contratación y las decisiones tomadas, para facilitar la detección de posibles casos de corrupción [12]. En los últimos años se han implementado diversas soluciones basadas en datos para el análisis de corrupción en compras públicas, en especial en los países europeos y asiáticos [13–20] Para la identificación y predicción de este delito, en especial en sobreprecio y fraude, se usan principalmente técnicas de aprendizaje supervisado y aprendizaje no supervisado. Recientemente se está empezando a trabajar con técnicas de Procesamiento de Lenguaje Natural (PLN) para el análisis de los documentos generados en el proceso de compras públicas [21]

El PLN ha ganado reconocimiento ante la creciente automatización de los procesos de compras públicas. En[22] se menciona que el empleo de estas técnicas y algoritmos de MD, permiten procesar grandes cantidades de textos y analizar las relaciones entre documentos y términos. Así mismo el uso de un mayor número de datos en la aplicación de técnicas como el aprendizaje automático, mejora el rendimiento del modelo. En [22] aprendizaje supervisado, analiza un conjunto de datos etiquetados para encontrar patrones que le permitan hacer predicciones en nuevos conjuntos de datos. En cambio, en el aprendizaje no supervisado, el modelo detecta patrones en los datos sin información previa. Esta técnica se utiliza principalmente para agrupar datos, encontrar asociaciones y detectar anomalías [23]. Puede argumentarse que la combinación de técnicas de IA, como el aprendizaje automático y el PLN, es muy efectiva para identificar posibles casos de corrupción en el proceso de compras públicas.

En esta tesis doctoral se establece una metodología para prevenir y detectar corrupción de forma automática, desde la información generada en la contratación pública, mediante el uso de técnicas de aprendizaje automático y PLN. Con el objetivo de comprobar el potencial de la metodología propuesta, se desarrollan algoritmos de detección y predicción de favoritismo, sobreprecio y sesgo en contratos elaborados con instituciones del estado. Como son escasos los trabajos de investigación en estos temas en Latinoamérica [24–26], se diseñaron diferentes casos de estudio a través del Servicio Nacional de Contratación Pública (SERCOP)[1] de Ecuador. Permitiendo validar la metodología en diferentes contextos (casos de estudio) para identificar la corrupción:

1. El sobreprecio en compras de medicamentos durante la pandemia COVID 2019.
2. El favoritismo basado en los parámetros de calificación de procesos y los tipos de contratación.
3. El sesgo de género y el favoritismo en cada proceso de compra pública, basándonos en la evaluación de igualdad de condiciones tanto en género como en oportunidades para el participante.

[1] https://portal.compraspublicas.gob.ec/sercop/la-institucion/

En cada caso se emplean diversas técnicas de extracción de información, que se aplican a los algoritmos de aprendizaje automático y luego se comprueba su rendimiento a través de diferentes métricas. La metodología permite validar la hipótesis planteada. Se pudo identificar otros tipos de corrupción menos explorados. Los resultados son muy alentadores para su detección y prevención en compras del sector público. Lo que contribuirá a mejorar la transparencia en las compras públicas, combatir el fraude y la corrupción no solo en Ecuador, sino también en toda Latinoamérica.

2.1. Motivación

En Ecuador el SERCOP es la entidad rectora del Sistema Nacional de Contratación Pública (SNCP), siendo la encargada de promover la participación ciudadana, incrementar el acceso y uso de la información pública por parte de la población, que podría evitar las malas prácticas en la contratación pública[2]. En el 2017, a través de ella se transaccionaron 5.800 millones de dólares, que corresponden al 19.6% del Presupuesto General del Estado y el 5.8% del Producto Interno Bruto. La participación por sectores de gobierno estuvo distribuida principalmente en la Administración del Estado (28,5%), en los gobiernos autónomos municipales (21,2%) y en las empresas públicas (18%). La plataforma provista por el SERCOP aloja por cada proceso de contratación documentos en formato PDF. En ellos se guardan datos de contratación y las especificaciones o Términos de Referencia (TDR), que incluyen invitaciones a proveedores, ofertas realizadas, observaciones, y toda la documentación generada por la compra. Entre los procesos presentes en el SERCOP, destacan: la ejecución de obras, adquisición de bienes y servicios y contratación de consultoría.

Los sistemas actuales para la detección de corrupción se basan en la revisión manual de documentos y registros financieros, con el fin de buscar comportamientos sospechosos, basados en denuncias o auditorías. Lo cual implica la identificación previa del acto corrupto. En algunos países, los documentos se obtienen mediante estrategias de datos abiertos, pero no es la norma más generalizada. Sin embargo, es necesario impulsar la democratización de los datos, para posibilitar una identificación más efectiva de la tipología de corrupción. Además, del análisis de los datos, se requiere incorporar el análisis del texto generado en cada proceso, lo que permitiría una valoración más profunda de la situación.

La investigación realizada destaca que existen varias estrategias para detectar la corrupción en compras públicas. Una de ellas es mediante el análisis de patrones en los datos relacionados con las compras públicas, como los precios, las empresas contratadas y las fechas de las compras. Para poder analizar dichos datos, se deben aplicar técnicas de minería de datos (MD). Por lo que se hace fundamental combinar técnicas de IA, como el aprendizaje automático y el PLN, para identificar posibles casos de corrupción. Además, es importante implementar mecanismos de transparencia y rendición de cuentas en las compras públicas, tales como la publicación de información detallada sobre los procesos de contratación y las decisiones tomadas, con el fin de facilitar la detección de posibles casos de corrupción y asegurar la integridad del proceso.

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

Con la revisión sistemática de la literatura sobre el tema de investigación, se pudo detectar que, en Latinoamérica, la cantidad de trabajos relacionados con la corrupción en datos asociados a compras públicas es mínima. Se evidenció la necesidad de desarrollar soluciones que permitan detectar corrupción en este tipo de datos. Especialmente los que procedan de fuentes que no cumplen con las buenas prácticas de datos abiertos. Además, potenciar la combinación de técnicas de minería de datos con PLN, para poder analizar la información del proceso con documentos de texto. Obteniendo de esta manera resultados más precisos para la detección de corrupción y sesgo. Lo que contribuirá a que las autoridades puedan tomar medidas para prevenir y combatir la corrupción en los procesos de compras públicas.

2.2. Metodología de investigación

El proceso de investigación y el método de trabajo en el que se ha fundamentado esta tesis es la metodología *Action Research* (AR) [27]. Esta metodología se basa en identificar un problema y formular a partir de él una hipótesis, partiendo de conceptos definidos dentro de un modelo cuantitativo de la realidad. De esta manera, se lleva a cabo el proceso de recopilación, organización y análisis de la información, continuando con el diseño de una propuesta de solución al problema. Finalmente, tras evaluar los resultados obtenidos de la investigación, se formulan las conclusiones respectivas. Por tal razón, se siguen las siguientes etapas:

- **Definir el problema:** Planteamiento del problema a solucionar con la investigación, lo que permite elaborar los objetivos y la hipótesis de investigación.
- **Revisar estado del arte:** Análisis de las soluciones planteadas hacia el problema, tanto en técnicas propias de nuestra investigación, como en métodos abordados por otros investigadores. Este proceso es constante a lo largo de la investigación.
- **Adquisición y limpieza de datos:** Obtención de datos de compras públicas, aplicando técnicas informáticas que permitan el almacenamiento y pre-procesamiento adecuados.
- **Elaboración de casos de estudio:** Desarrollo de diferentes escenarios para la detección de distintos tipos de corrupción en compras públicas, que han sido menos estudiados. Como son: la detección de sobreprecios en compra de medicamentos, el favoritismo basado en los parámetros de calificación de procesos y los tipos de contratación. Así como el favoritismo y sesgo basado en el texto que se genera en cada proceso de compra pública.
- **Estudio de algoritmos de clasificación supervisados y no supervisados:** Estudio comparativo de los algoritmos, selección de variables, reducción de dimensionalidad y evaluación con diferentes métricas de rendimiento, aplicadas a los algoritmos para distintos tipo de datos asociados a un proceso de compra pública.
- **Procesamiento de lenguaje natural:** Estudio de técnicas de PLN para la obtención de sesgo y análisis de sentimientos en datos textuales generados en un proceso de compra pública**.**
- **Publicación de resultados:** Presentación de resultados en conferencias y revistas de alto impacto que promuevan el intercambio de ideas y la validación de la propuesta de tesis doctoral.

2.3. Hipótesis

La hipótesis de esta tesis doctoral se basa en la aplicación de la minería de datos en el análisis e identificación de riesgos de corrupción en las compras públicas. Concretamente, que, a través de la evaluación de los datos y el texto generado en las fases de cada proceso, mediante la combinación de técnicas de aprendizaje automático y PLN, se pueda detectar si ha existido corrupción y clasificar su tipo (sobreprecio, favoritismo, sesgo). El proceso se inicia desde la obtención de datos en un portal que no cumple con los principios de datos abiertos, como ocurre en Ecuador.
Continúa con la realización del análisis exploratorio de los mismos, con el objetivo de encontrar la técnica adecuada para su limpieza y reducción de su dimensionalidad para su tratamiento. Con la base de datos depurada y el texto obtenido, se combinan algoritmos de aprendizaje automático supervisados y no supervisados, para la detección de corrupción en la asignación de contratos públicos.
Como el estudio preliminar sobre el tema refleja, que la mayoría de las propuestas recogidas en la literatura son insuficientes y solo están enfocadas a la resolución de problemas muy específicos. Se evidencian algunas necesidades:

- Detectar posibles sobreprecios, en base a las compras realizadas en periodos similares por otras instituciones públicas.
- Investigar tipos de corrupción como el favoritismo, que ha sido poco estudiado.
- La detección de sesgos a través del análisis del texto generado por cada proceso, para encontrar sesgo hacia determinados proveedores o de género.
- Poder detectar la combinación de varios tipos de corrupción en un mismo proceso.

Por lo tanto, la tesis se centra en establecer una metodología para prevenir y detectar corrupción mediante el uso de técnicas de aprendizaje automático y PLN, con el desarrollo de algoritmos de detección y predicción de sobreprecio, favoritismo y sesgo en compras públicas.

Para validar la metodología, se diseñaron diferentes casos de estudio en varias áreas poco exploradas como: el sobreprecio en compras de medicamentos durante la pandemia de COVID-19 en Ecuador, el favoritismo basado en los parámetros de calificación de procesos y los tipos de contratación. Además, se crearon otros casos de estudio, que permitieran analizar el texto generado en cada proceso y poder detectar el favoritismo hacia cierto proveedor y asó como el sesgo de género en cada proceso de compra pública.

Todos los resultados fueron evaluados con diferentes técnicas de rendimiento, empezando por el porcentaje de sobreprecio y enfatizando en el porcentaje de acierto de los algoritmos de clasificación. Además, los resultados obtenidos se contrastaron y validaron con las noticias de medios de comunicación y de redes sociales del país, sobre procesos de compras públicas con corrupción.

2.4. Objetivos

El objetivo general para validar la hipótesis planteada es el siguiente:

2.4.1. Objetivo General

Desarrollar una metodología para prevenir y detectar corrupción de forma automática, mediante el uso de minería de datos, técnicas de aprendizaje automático y PLN.

2.4.2. Objetivos Específicos

- Analizar los diferentes métodos para la detección de corrupción en compras públicas y determinar las variables que influyen en los casos de corrupción.
- Seleccionar las técnicas apropiadas de representación de datos, haciendo un estudio comparativo de los diferentes enfoques de análisis de datos, que permita obtener un conjunto óptimo de entrenamiento y prueba para los algoritmos utilizados.
- Recuperar, ordenar y limpiar los datos del portal SERCOP en el periodo entre el año 2010 y 2020.
- Desarrollar algoritmos de detección de corrupción en compras públicas, diseñando un enfoque apropiado para optimizar el coste computacional de los algoritmos de clasificación. Acorde con la representación del conjunto de entrenamiento de los diferentes casos de estudio.
- Seleccionar y aplicar las métricas de evaluación de los diferentes algoritmos para elegir los más adecuados en la clasificación de los datos y poder establecer comparativas de rendimiento.
- Con el empleo de PLN desarrollar un análisis de sentimiento, de las preguntas y respuestas proporcionadas por las entidades adjudicadoras, para la detección de sesgos y el favoritismo a los proveedores.

2.5 Estructura de la memoria

La presente tesis doctoral se encuentra en la modalidad de compendio de artículos científicos. El capítulo II, corresponde a la presentación de la modalidad de la tesis. Por su parte, el capítulo III muestra el contexto y el estado del arte correspondiente a la temática de la tesis doctoral. La sección IV, presenta la propuesta. La coherencia y relación directa entre los artículos presentados y los objetivos de la tesis doctoral. Donde se muestran la metodología, resultados y conclusiones de cada publicación. Finalmente, se presentan los anexos de cada contribución en el capítulo V.

CAPÍTULO 3

El presente capitulo, muestra las bases teóricas necesarias para el desarrollo de esta tesis doctoral, la cual se divide en dos partes empezando con la definición de corrupción como objeto de estudio, sus tipologías, la corrupción en compras públicas, variables de detección y formas actuales (secciones 3.1, 3.2, 3.3, 3.4). En la segunda parte se describen las técnicas usadas para la validación de la hipótesis como son: minería de datos, aprendizaje supervisado, no supervisado y PLN (secciones 3.5, 3.6, 3.7, 3.8).

3. CONTEXTO Y ESTADO DEL ARTE CORRUPCIÓN EN COMPRAS PÚBLICAS

3.1. Corrupción

Etimológicamente la palabra "*corrupción*" derivada del verbo latín "*corrumpĕre*", tiene muchos significados como: alterar y trastrocar; echar a perder, depravar, dañar, podrir; sobornar; pervertir; estragar, viciar. Como sustantivo, el diccionario de la Real Academia de la Lengua lo define con referencia a las instituciones: "*En las organizaciones, especialmente en las públicas, prácticamente consistente en la utilización de las funciones y medios de aquellas en provecho, económico o de otra índole, de sus gestores*". La corrupción infiere un abuso del poder de un empleado público para obtener "ganancias" beneficiando a entidades privadas [4]. Las "ganancias" para el corrupto no solo incluyen dinero también bienes materiales e inmateriales como status y poder. Contextualizada en el interés público, Pérez[28] la define como *"el mal uso del poder público o privado para obtener un beneficio indebido; comportamiento que se desvía de los deberes formales de la función pública para obtener ventajas privadas; como tal, constituye un problema público".*

Se puede argumentar que la corrupción es un acto viciado en el que una persona busca obtener ganancias ilegítimas ya sea en términos económicos u otros. Existe consenso en que se produce cuando se aprovecha un conflicto de intereses para satisfacer intereses personales, lo que se traduce en obtener beneficios infringiendo un marco legal existente.

3.2. Formas de corrupción

La clasificación propuesta por Vargas-Hernández [29] contempla la corrupción política, corrupción económica y corrupción administrativa.

En la Figura 1 se pueden observar las diferentes formas de corrupción presentados por Chan [30], quien al detallarlas menciona que el soborno es la forma más común de corrupción, impulsada por la idea de obtener beneficios por parte de las personas involucradas.

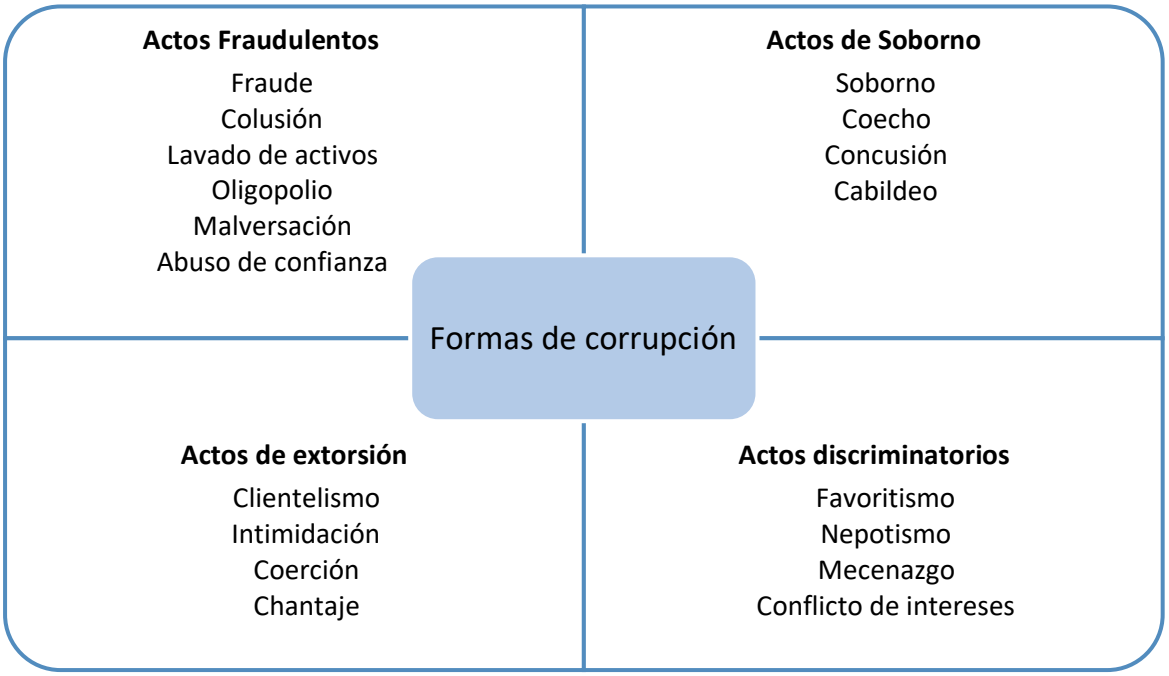


Figura 1: Marco de trabajo de detección de corrupción de Chan [30]

La corrupción puede manifestarse en varias formas que se detallan a continuación:

Soborno: Es la práctica corrupta en la que se ofrece dinero, bienes o servicios como parte de una relación corrupta. Esta forma de corrupción es la más extendida debido a los beneficios que se supone obtienen las partes involucradas. El soborno consiste en proporcionar una cantidad específica de dinero, un porcentaje del valor de un contrato u otros favores monetarios a un funcionario estatal que tiene la capacidad de realizar contratos en nombre del gobierno o distribuir beneficios a empresas, individuos y clientes.

Colusión: Se concreta en un pacto ilícito de daños a terceros; se refiere a un acuerdo entre dos o más individuos con la intención de perjudicar a otra persona. En el contexto de la colusión, las partes A y B se unen para manipular una situación contractual en su beneficio y obtener ganancias superiores a las normales. En el ámbito de la contratación pública, esto se evidencia cuando dos o más proveedores acuerdan aumentar los precios o disminuir la calidad de los productos con el fin de participar en un proceso de licitación. Esta práctica no solo causa perjuicio al estado, sino que también desalienta la participación de otros proveedores y socava la confianza del público en los procesos.

Malversación: Consiste en apropiarse de parte de los bienes que han sido confiados por alguien para su administración. En el ámbito privado, esto ocurre cuando los empleados abusan de la confianza de sus empleadores y se apropian indebidamente de recursos. En el caso del sector público, se refiere a la acción de un funcionario que desvía los recursos del estado que deberían ser utilizados para el beneficio público.

Fraude: Consiste en manipular o distorsionar información con el objetivo de engañar a alguien para que entregue voluntariamente una propiedad. En el ámbito público, esta manipulación de información se realiza con el fin de obtener beneficios personales. Según Pérez [28]pueden identificar causas

fundamentales en el fenómeno del fraude, conocidos como el "triángulo del fraude". Donde se incluyen la presión económica, que surge de la necesidad de resolver problemas financieros; la oportunidad, que se presenta cuando existe la posibilidad de aprovechar la posición de confianza y poder; y la racionalización, que es la manera en que se justifica o acepta el fraude.

Se puede postular que para racionalizar el fraude se buscan razones o motivos que permitan considerar el fraude como aceptable o necesario en determinadas circunstancias, a pesar de su carácter ilegal o inmoral. La racionalización del fraude puede involucrar la creación de falsas justificaciones éticas, la minimización de las consecuencias negativas, la comparación con comportamientos similares aceptados o la atribución de la responsabilidad a factores externos. En resumen, implica encontrar argumentos que justifiquen internamente el acto fraudulento para reducir la percepción de culpabilidad o responsabilidad.

Abuso de confianza: Se manifiesta cuando la administración pública aplica prácticas corruptas sin inducción externa. Funciona desde altos niveles nacionales hasta los niveles locales. Se establece por gobiernos corruptos con la premisa de permitir que los empleados públicos abusen de los derechos de los ciudadanos para conseguir beneficios personales [2].

Favoritismo y Nepotismo: El favoritismo es la propensión humana natural para favorecer a amigos, familiares y cualquier persona cercana y confiable dentro de un proceso público. Siendo el nepotismo el otorgar favores a personas con el lazo de parentesco al funcionario que desempeña un cargo público [31]. El favoritismo causa menos poder de compra para la institución pública, precios más altos que impactan la calidad del producto y generan competencia desleal.

3.3. Corrupción en compras públicas

La contratación pública consta de cuatro fases que incluyen: la preparatoria del proceso, la precontractual donde se evalúan los participantes, la contractual que se refiere a la ejecución del contrato de bienes o servicios y la post-contractual que se encarga de la evaluación del proceso. La corrupción puede darse en cualquiera de esas fases, por lo que se considera necesario indicar los mecanismos de corrupción que pueden ser utilizados para su implementación:

Inexistencia de contrato: La celebración de acuerdos verbales entre un funcionario público y un particular, en virtud de los cuales, el primero entrega fondos públicos al segundo a cambio de un favor sin contrato alguno[32] .

Contratación directa indebida: Mecanismo con una excepción del principio de concurrencia, pues permite la selección directa del contratista que ejecuta el contrato público. Por esta razón debe usarse en circunstancias excepcionales, previstas en la legislación del país, puesto que al no existir competencia no existen pautas para determinar si se trata o no de favorecer a un determinado contratista [32] Un caso típico de este tipo de contratación, se presentó en la pandemia de COVID-19, donde se permitió realizar compras de manera directa a ciertos contratistas, originando múltiples casos de corrupción [33].

Contratación indebida: La primera forma para manipular una licitación inicialmente legal es por medio de la tergiversación o contradicción de los informes técnicos o jurídicos sobre las ofertas presentadas por los interesados. Estos informes son conceptos. Consignar en ellos una ilegalidad puede constituir prevaricación o eventualmente una falsedad en documento público. Por otra parte, si se adjudica contradiciendo un informe técnico, podrá incurrirse en el delito de prevaricación, aunque si en virtud de este se celebra un contrato, la conducta se subsumiría en el delito de celebración de contrato, sin el cumplimiento de los requisitos legales esenciales. También podrá incurrirse en delito por falta de otros requisitos que impliquen la injusticia de la resolución, como la inscripción del contratista en el registro de proponentes [7].

Fraccionamiento: Constituye una práctica habitual de las administraciones. Fraccionar los contratos para disminuir su cuantía y eludir así los requisitos procedimentales de las leyes de contrataciones y adquisiciones por etapas, tramos, paquetes o lotes, posibles en función a la naturaleza del objeto de la contratación o adquisición. También para propiciar la participación de las pequeñas y medianas empresas en aquellos sectores económicos donde exista oferta competitiva [34]

Modificaciones abusivas: La modificación unilateral consiste en el poder de modificar, por razones de interés público, los contratos de bienes y servicios [35]. En múltiples oportunidades se utiliza el instrumento jurídico de la modificación para articular nuevos contratos con fraude, al principio licitatorio. Incluso en ocasiones las modificaciones duplican o triplican el precio inicial del contrato y se asignan al proveedor previamente destinado[35]

En base a los mecanismos detallados previamente, en la Figura 2 se identifican las variables más significativas para el tratamiento de datos en compras públicas [34]. Dichas variables, podrían ayudar a detectar las modificaciones abusivas de contratos, la dirección del ganador a un determinado proveedor y la contratación directa indebida evaluando la relación entre personas y empresas. Es importante señalar que no se consideran variables para la evaluación de inexistencia de contrato puesto que la información subida al SERCOP solamente incluye los contratos firmados.

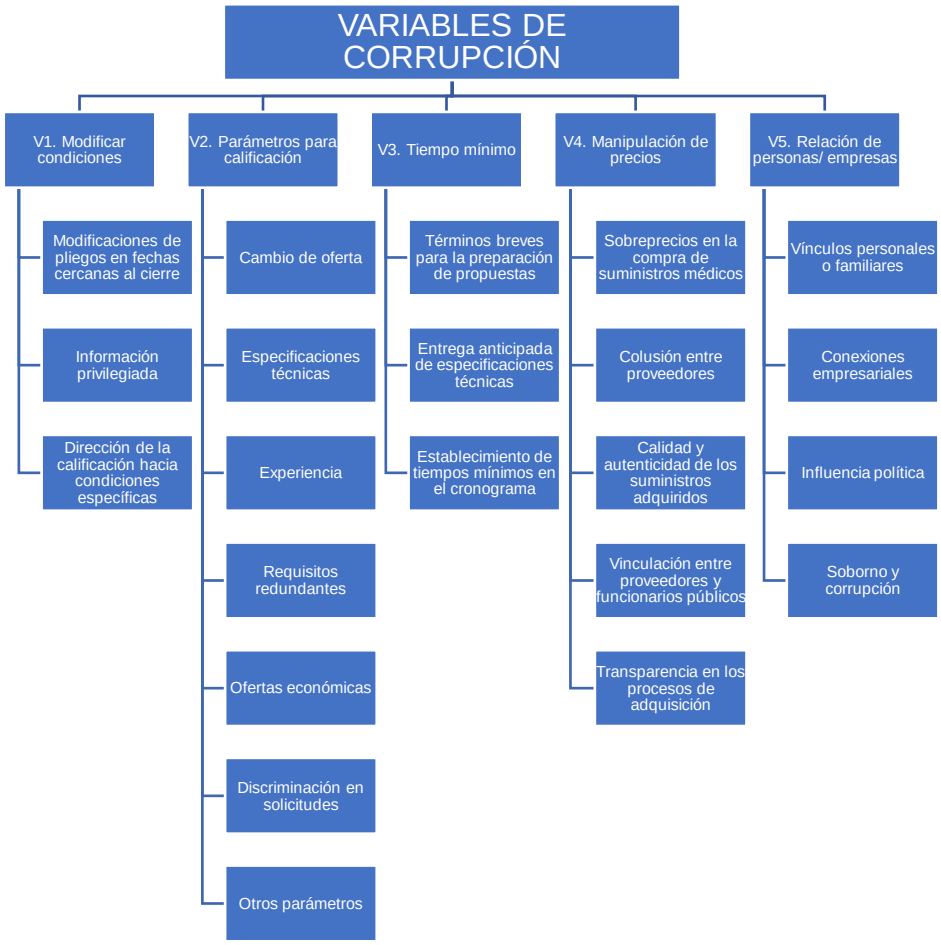


Figura 2: Variables que influyen en la corrupción en compras públicas

Cada variable cuenta con subvariables que se detallan a continuación.

Variable 1. Modificar condiciones

Modificar las condiciones se refiere a la modificación de los términos establecidos en los pliegos de contratación en una fecha cercana al cierre del proceso. Esta práctica tiene como objetivo favorecer a un único aspirante al permitirle anticipadamente conocer las nuevas exigencias y preparar previamente la documentación requerida. Durante la etapa de preguntas, respuestas y aclaraciones de un procedimiento de compra pública, las entidades encargadas de la contratación realizan modificaciones de último momento, dejando a todos los oferentes sin suficiente tiempo para ajustar sus propuestas según los nuevos requisitos. Esto automáticamente beneficia al oferente previamente informado, asegurando que sea el único que cumpla con todas las condiciones solicitadas.

Subvariables de modificar las condiciones

Modificaciones de pliegos en fechas cercanas al cierre: Cambios realizados en los términos y condiciones establecidos en los pliegos de contratación poco antes de la fecha límite. La falta de tiempo suficiente para que los demás oferentes se adapten a las nuevas exigencias.

Información privilegiada: Suministro anticipado de información sobre las modificaciones a un oferente específico. El beneficiario de esta información previa puede preparar sus documentos y propuesta en consecuencia.

Dirección de la calificación hacia condiciones específicas: Modificación de las disposiciones relacionadas con la metodología de calificación por puntaje durante la etapa de preguntas, respuestas y aclaraciones. Los cambios se realizan para favorecer a un proponente en particular, asegurando que solo su oferta cumpla con los nuevos criterios de calificación.

En resumen, la variable "Modificar condiciones" implica la alteración de los términos de contratación en beneficio de un oferente específico. Todas las acciones buscan de manera previa favorecer al ofertante privilegiado asegurando que este cumpla con los requisitos, alcance el mayor puntaje y logre la victoria en el proceso de contratación.

Variable 2. Parámetros para calificación

Los criterios de valoración de las ofertas han de ser elegidos previamente, con el objetivo de permitir seleccionar cuál es la oferta económicamente más ventajosa, para la entidad contratante de las que reciba en cada procedimiento [5]. Los actos de corrupción podrían presentarse, cuando se otorgan puntajes muy por debajo del mínimo establecido en los modelos de pliego para evaluar las ofertas económicas. Permitiendo que se adjudiquen ofertas más caras, en perjuicio del presupuesto de la entidad y del estado, al establecer un puntaje más bajo que el exigido en el modelo de pliegos para la oferta económica. La entidad contratante, da mayor importancia a otros parámetros. Los puntos que no fueron incluidos en la oferta económica los incluirá en otro que favorezca a un participante en particular. Como resultado, un oferente con un precio más alto, ganará, ya que a nivel técnico se le valora con más puntos y por tanto el parámetro oferta económica no es el dirimente.

También puede darse el caso, que se establezcan otros parámetros con condiciones que solo un oferente pactado podrá cumplir, y al cual se otorgarán puntajes muy altos en relación con el resto de los parámetros establecidos. Este parámetro tiene como finalidad solicitar un requisito específico relacionado con el objeto de la contratación; sin embargo, las entidades lo pueden utilizar, para incluir algún requisito que solo el oferente con el que llegaron a un acuerdo pueda cumplir, asegurando descalificar al resto de ofertas.

Por ejemplo, se solicita que el personal tenga una edad y título específico, sin tener justificación legal para respaldar dichas solicitudes. Dentro de los procedimientos de contratación, las entidades al conocer las características del personal que posee el proveedor a quien quieren beneficiar, solicitan dentro de los pliegos requisitos específicos, que solo dicho proveedor puede cumplir, afectando la participación de los demás posibles oferentes.

Subvariables de parámetros para la calificación

Cambio de oferta: En la etapa de preguntas y aclaraciones dentro de un procedimiento, se modifican las disposiciones contenidas en los pliegos de condiciones referentes a la metodología de calificación por puntaje, con el fin de direccionar la calificación hacia condiciones que solo un proponente presentará. La Entidad después de mantener conversaciones con el oferente que quiere que sea el ganador, modifica ciertos parámetros en la etapa de "Preguntas, Respuestas y Aclaraciones", con el fin de que la oferta de dicho proveedor obtenga los máximos puntajes y de esta manera asegurar que sea el ganador del procedimiento.

Especificaciones técnicas: Las especificaciones técnicas de los equipos no ostentan características acordes a las necesidades y funciones que estipule el objeto del contrato. Con el fin de direccionar los procedimientos, a proveedores con los que se ha llegado a un acuerdo, las entidades incluyen requisitos y/o equipos que solo cumple dicho proveedor. Evidenciándose que los solicitados no constituyen un factor importante dentro del procedimiento y que el único objetivo es direccionar.

Experiencia: En experiencia general y especifica. Se solicita la experiencia obtenida en cierto sector concreto del ámbito público. Las entidades con el fin de direccionar los procedimientos a los oferentes con los cuales han trabajado anteriormente solicitan experiencia especifica con la misma entidad, lo cual deja sin posibilidades que nuevos oferentes sean adjudicados.

Requisitos redundantes: Se refiere a la inclusión de requisitos innecesarios o repetitivos en los pliegos de condiciones de los procedimientos de contratación, sin que exista una justificación legal que justifique que estos están relacionados con el objeto de la contratación. Por tanto, estos requisitos no aportan valor ni cumplen una función clara relacionada con el objeto del contrato, pero se incluyen con el propósito de favorecer a un participante específico o limitar la participación de otros posibles oferentes.

Por ejemplo, en un proceso de contratación para la adquisición de equipos médicos, se incluyen requisitos adicionales que no son relevantes para el funcionamiento o la calidad de los equipos; podrían ser solicitar certificaciones específicas que solo el proveedor preferido tiene, exigir características técnicas no necesarias o imponer condiciones que no están directamente relacionadas con el cumplimiento de los objetivos del contrato. Estos requisitos redundantes se utilizan como una estrategia para favorecer a un proveedor específico que cumple con estos requisitos preestablecidos, excluyendo a otros posibles oferentes que podrían cumplir con los requisitos principales del contrato.

Ofertas económicas: Otorgar puntajes inadecuado, por debajo del mínimo establecido, en los modelos de pliegos para evaluar las ofertas económicas. Esto permite la adjudicación de ofertas más caras en detrimento del presupuesto de la entidad y del estado, al dar más importancia a otros parámetros en lugar de la oferta económica.

Discriminación en solicitudes: Las entidades solicitantes conocen las características del personal del proveedor preferido y establecen requisitos específicos en los pliegos que solo ese proveedor puede

cumplir. Esto afecta la participación de otros posibles oferentes al imponer requisitos discriminatorios y favorecer al proveedor preseleccionado.

Otros parámetros: Inclusión de otros parámetros favorecedores de un participante, que incluyan condiciones que solo un oferente preseleccionado puede cumplir. Otorgar puntajes altos a este participante en relación con el resto de los parámetros establecidos. El objetivo es descalificar al resto de las ofertas al incluir requisitos específicos que solo el oferente favorecido puede cumplir.

Puede argumentarse que la variable "Parámetros para calificación" involucra prácticas corruptas en la evaluación de las ofertas en los procedimientos de contratación. Esto incluye otorgar puntajes inadecuados a las ofertas económicas, priorizar otros parámetros sobre la oferta económica, y establecer requisitos específicos injustificados o discriminatorios para favorecer a un participante preseleccionado. Estas acciones perjudican la transparencia y equidad en el proceso de contratación al dar ventajas indebidas a ciertos oferentes en detrimento de otros.

Variable 3. Tiempo mínimo

Esta variable se refiere al establecimiento de términos muy breves para el diseño, preparación, elaboración y presentación de propuestas, cuando se abre un proceso de contratación, pero "*alguien*" que conozca anticipadamente todos los requisitos, estará ya listo para presentar su propuesta dentro de los términos establecidos. Las especificaciones técnicas o términos de referencia de un procedimiento de contratación son entregados con anticipación al oferente con el cual se ha llegado a un acuerdo. La Entidad se asegura que el mismo tenga el tiempo necesario para elaborar su oferta; posteriormente, establece los tiempos mínimos dentro del cronograma del procedimiento, con el fin de que los demás oferentes no puedan elaborar su oferta de manera adecuada, dejando vía libre al oferente que tuvo esta información con anterioridad.

Subvariables de tiempo mínimo

Términos breves para la preparación de propuestas: Establecer plazos muy cortos para el diseño, preparación, elaboración y presentación de propuestas en un proceso de contratación. Esto dificulta que los demás oferentes puedan completar adecuadamente sus propuestas dentro de los plazos establecidos, mientras que alguien que conozca anticipadamente los requisitos estará preparado para presentar su propuesta a tiempo.

Entrega anticipada de especificaciones técnicas: Proporcionar a un oferente preseleccionado las especificaciones técnicas o términos de referencia con anticipación. El oferente favorecido tiene tiempo suficiente para preparar su oferta antes de que se abra oficialmente el proceso de contratación.

Establecimiento de tiempos mínimos en el cronograma: La entidad contratante establece plazos mínimos en el cronograma del procedimiento de contratación. Esto impide que los demás oferentes tengan

suficiente tiempo para elaborar adecuadamente sus ofertas, dejando ventaja al oferente que recibió la información anticipada.

La variable "Tiempo mínimo" involucra prácticas corruptas relacionadas con los plazos establecidos en los procedimientos de contratación; incluye establecer términos breves para la preparación de propuestas, entregar anticipadamente las especificaciones técnicas a un oferente preseleccionado y establecer tiempos mínimos en el cronograma para perjudicar a los demás oferentes. Estas acciones garantizan que el oferente favorecido tenga una ventaja indebida al tener más tiempo para preparar su oferta, mientras que los demás se ven limitados y pueden ser incapaces de presentar propuestas adecuadas dentro de los plazos establecidos. Esto socava la competencia justa y la igualdad de oportunidades en el proceso de contratación.

Variable 4: Manipulación de precios

Esta variable se centra en investigar la manipulación de precios en la adquisición de suministros médicos durante emergencias sanitarias, como la pandemia de COVID-19. Se analiza cómo los precios de los suministros médicos pueden ser inflados de manera injustificada, beneficiando a ciertos proveedores o contratistas y generando un perjuicio económico para las entidades contratantes y el estado en general.

Subvariables de manipulación de precios

Sobreprecios en la compra de suministros médicos: Investiga casos en los que se han registrado sobreprecios significativos en la adquisición de suministros médicos durante la pandemia de COVID-19, comparando los precios pagados con los precios de mercado o los precios previos a la emergencia sanitaria.

Colusión entre proveedores: Analiza la existencia de colusión entre proveedores de suministros médicos, donde se ponen de acuerdo para fijar precios inflados de manera coordinada, limitando la competencia y generando un perjuicio económico para las entidades contratantes.

Calidad y autenticidad de los suministros adquiridos: Investiga si los suministros médicos adquiridos a precios excepcionales durante la pandemia cumplen con los estándares de calidad y autenticidad requeridos, evitando situaciones de fraude o adquisición de productos falsificados.

Vinculación entre proveedores y funcionarios públicos: Analiza posibles vínculos entre proveedores de suministros médicos y funcionarios públicos encargados de la adquisición, investigando si existen conflictos de interés, corrupción o tráfico de influencias que puedan influir en los precios acordados.

Transparencia en los procesos de adquisición: Evalúa la transparencia y la rendición de cuentas en los procesos de adquisición de suministros médicos durante la pandemia, investigando si se siguieron los procedimientos adecuados, si se realizaron licitaciones competitivas y si se publicaron los precios y las condiciones de contratación de manera accesible.

Esta variable se enfoca en investigar la manipulación de precios en la adquisición de suministros médicos durante emergencias sanitarias, como la pandemia de COVID-19. Se analizan aspectos como los sobreprecios, la colusión entre proveedores, la calidad y autenticidad de los suministros, la vinculación entre proveedores y funcionarios públicos, y la transparencia en los procesos de adquisición.

Variable 5: Relación de personas/ empresas

La variable "Relación de personas/empresas" se refiere a las interacciones y conexiones existentes entre personas o empresas involucradas en los procesos de contratación pública. Estas relaciones pueden ser utilizadas como una estrategia para facilitar actos de corrupción y favorecer a ciertos actores en detrimento de la transparencia y la competencia justa.

Subvariables de la variable Relación de personas /empresas

Vínculos personales o familiares: Relaciones basadas en vínculos familiares, amistad cercana o relaciones personales entre los funcionarios de la entidad contratante y los proveedores. Estos vínculos pueden influir en la toma de decisiones, favoreciendo a personas o empresas con las que tienen una relación cercana.

Conexiones empresariales: Relaciones comerciales o de asociación entre diferentes empresas participantes en los procesos de contratación. Estas conexiones pueden llevar a prácticas colusorias o a la manipulación del proceso de contratación para beneficiar a un grupo específico de empresas.

Influencia política: Relaciones o vínculos con actores políticos que pueden influir en la toma de decisiones relacionadas con los contratos públicos. La influencia política puede utilizarse para favorecer a ciertos actores o empresas afines a un partido político o grupo de interés.

Soborno y corrupción: Ofrecimiento o aceptación de sobornos y pagos ilícitos para establecer relaciones corruptas entre los funcionarios encargados de la contratación y los proveedores. Estas prácticas corruptas pueden incluir el pago de comisiones ilegales a cambio de la adjudicación de contratos o la manipulación de procesos de licitación.

La variable "Relación de personas/empresas" aborda las interacciones y conexiones existentes entre personas y empresas involucradas en los procesos de contratación pública. Esto puede incluir vínculos personales o familiares, conexiones empresariales, influencia política y prácticas de soborno y corrupción. Estas relaciones pueden facilitar actos de corrupción al favorecer a ciertos actores o empresas específicas, comprometiendo la transparencia y la equidad en los procesos de contratación.

A partir de la revisión y definición de las variables y subvariables relacionadas con la detección de la corrupción en compras públicas, se logra establecer un marco sólido que permitió el diseño de los casos de estudio para la experimentación. Estas variables abordan diversos aspectos que pueden influir en la aparición y detección de prácticas corruptas, como la manipulación de precios, la discrecionalidad en la

calificación, la influencia de la relación entre personas/empresas, el tiempo mínimo para la preparación de propuestas y la transparencia en los procesos de contratación. Además, se ha respaldado científicamente cada tipo de corrupción a partir de las contribuciones de reconocidos autores en el campo.

3.4. Detectar corrupción en compras públicas

La investigación científica está explorando diferentes enfoques para detectar la corrupción en las compras públicas. Tradicionalmente la estadística se ha considerado una herramienta útil para analizar datos relacionados con las compras públicas. En la búsqueda de patrones que indiquen posibles casos de corrupción, se ha utilizado la estadística para comparar los precios de compras similares realizadas por diferentes entidades y detectar posibles sobreprecios [18]. También para analizar la frecuencia de compras realizadas y detectar posibles casos de favoritismo. Sin embargo, es importante tener en cuenta que en la actualidad con el volumen de datos que manejan las administraciones públicas, la detección de corrupción en compras públicas no se puede basar únicamente en el análisis estadístico de los datos.

Se hace necesario buscar método avanzados que empleen nuevas tecnologías para detectar y prevenir casos de corrupción. Específicamente el uso de Minería de Datos (MD) e IA, para identificar patrones y tendencias relacionadas con la corrupción en las compras públicas. Se han utilizado también modelos de análisis de riesgo para identificar y priorizar los procesos de compra pública más propensos a ser afectados por la corrupción[36]. Más recientemente, se ha observado la necesidad de una investigación interdisciplinaria, donde colaboren diferentes áreas, como la economía, la sociología y la psicología, para entender mejor los motivos y factores que contribuyen a la aparición de corrupción en las compras públicas [18]. Estas investigaciones ayudan a mejorar la eficacia en la detección de la corrupción en las compras públicas y a desarrollar soluciones más eficientes y efectivas para abordar este problema.

Para que un proceso de compras públicas sea transparente intervienen tres actores fundamentalmente, como se observa en la Tabla 1:

1. El estado de cumplimiento, en el levantamiento de los procesos de las compras y de auditorías.
2. Los proveedores, para posibilitar una competencia legal y la alerta temprana cuando observen irregularidades.
3. La sociedad civil, para la toma de decisiones colaborativas y la vigilancia del proceso mediante redes sociales.

Con la participación de los tres actores principales del sistema, se puede lograr que se realicen compras públicas transparentes.

Tabla 1: Resumen de variables y su incidencia en la alerta de corrupción

	Estado	**Proveedores**	**Sociedad civil**
Desarrollo y eficiencia	Retroalimentar decisiones de compra, planificación y gestión de contratos.	Visualizar y acceder a oportunidades de negocios (menores costos de transacción, mayor	Retroalimentar participación ciudadana (decisiones colaborativas, mayor

		participación y competencia, mayor innovación).	legitimidad, mejores políticas de compras).
Transparencia	Auditorias Visitas *in-situ*. Análisis de indicadores peligrosos.	Controlar mediante análisis de indicadores (impugnaciones y de otras formas participación).	Controlar y denunciar mediante redes sociales.

Existen varias iniciativas del uso de las nuevas tecnologías para detectar corrupción en compras públicas. Tenemos como ejemplo el caso de Suecia [37], que, en el 2011, demuestra que el 58% de las veces el licitador que presenta la oferta más baja no es el ganador del proceso. Por otro lado, usando el análisis empírico *Random Utility Model*, se observa el caso de Paraguay en 2015 [25] utilizando datos acumulados durante 4 años y 47.615 procesos de contratación, realizó un estudio que estima, a través de la construcción de un modelo matemático, la correlación entre las empresas y su posibilidad de obtener un contrato. Detectando la existencia de una relación previa entre el proveedor y la entidad contratante, que produce corrupción cuando se realiza la contratación. No obstante, en estos sistemas no se aplican técnicas y procedimientos de MD para la obtención de corrupción, en especial favoritismo.

La solución SALER planteada en compras públicas españolas [24] combina modelos de aprendizaje automático descriptivos y predictivos con estadísticas y visualización avanzadas, pero todo en base a una definición específica de una serie de requisitos en cuanto a preguntas y análisis de datos. Así como indicadores de riesgo y otros patrones de anomalías. Otra iniciativa similar pero que solamente ayuda a analizar riesgo es Arachne [38] [39]. Desarrollada por la Comisión Europea, su principal objetivo es ayudar a las autoridades en la gestión en sus controles administrativos y la gestión en el ámbito de los Fondos Estructurales (Fondo Social Europeo y Fondo Europeo de Desarrollo Regional). Esta potente herramienta de puntuación de riesgos, genera más de 100 indicadores. Clasificados en categorías de riesgo específicas, para ayudar a las autoridades de gestión y a los organismos intermedios a prevenir y detectar errores e irregularidades entre proyectos, beneficiarios, contratos y contratistas.

ZIndex [10] es otra herramienta de evaluación comparativa de la contratación pública, para calificar a los poderes adjudicadores. Desarrollada en la República Checa, por investigadores de la Universidad Carolina de Praga. Gracias a esta herramienta, las instituciones públicas pueden compararse en función de cómo gestionan el dinero público lo cual permite medir el cumplimiento por parte de la autoridad contratante de las recomendaciones de buenas prácticas definidas por organizaciones internacionales.

Otro uso exitoso de *BigData* para la detección de la corrupción es el trabajo de Chu [40]. Donde se usa una metodología que integra técnicas de PLN, Algoritmo Genético Cuántico (QGA) y máquina de vectores de soporte (SVM) para desarrollar un método de detección de fraudes para las narrativas en los informes anuales, que mejore la precisión de la detección de fraudes y reduzca los riesgos de inversión.

En la detección de corrupción de compras públicas otro enfoque innovador fue el presentado por Gallego [41] al emplear modelos de aprendizaje automático y un modelo *Gradient Boosting Machine* (GBM), para la identificación de transacciones de tendencia problemática, mediante el reconocimiento de factores que impulsan corrupción. Estos modelos, describieron las variables que contribuyen a la probabilidad de la presencia de un contrato problemático y su contribución a la corrupción; lo que permite ofrecer una predicción de la situación corrupta.

En este orden de importancia se encuentra el aporte de [42]que emplea una metodología innovadora de *Big Data* para identificar los efectos de un cambio de gobierno, en los mercados de contratación en dos países: Hungría y el Reino Unido. Los resultados mostraron que las empresas favorecidas políticamente se aseguran entre el 50% y el 60% del mercado de contratación del gobierno central en Hungría, pero solo el 10% en el Reino Unido.

Las políticas y leyes de contratación pública deben aspirar a crear instituciones, procesos y sistemas impulsados por la necesidad de cumplir los objetivos del gobierno. Los procesos y sistemas deben basarse en valores de servicio público que defiendan el bien común y la armonía de la sociedad. El marco de gobernanza de la contratación pública [8] debe basarse en sistemas sólidos de supervisión y evaluación y debe contar con Tecnologías de la información y las comunicaciones (TIC) de confianza.

Como se ha descrito en los trabajos relacionados anteriormente, existen ciertos avances en el uso de aprendizaje automático y PLN en la detección de la corrupción en compras públicas. Estas investigaciones se basan principalmente en la identificación de unos pocos tipos de corrupción, en especial sobreprecio y fraude. En consecuencia, en la tesis doctoral que se presenta, se decide abordar el favoritismo y el sesgo, pues se detectó en el mapeo de la bibliografía que eran aspectos poco estudiados. Además, se consideró oportuno crear una metodología que permita detectar el reconocimiento de varios tipos de corrupción simultáneamente, mediante un modelo multifase. En cuanto a las técnicas seleccionadas, se prefiere generalmente el aprendizaje supervisado para abordar la clasificación, pero esto requiere una identificación previa del proceso de fraude, aspecto con el que no se puede contar siempre, lo que motiva que en este trabajo se investigue también la identificación de patrones con aprendizaje no supervisado. Finalmente se detectó, la necesidad de incrementar la investigación del uso del PLN en el análisis del texto generado dentro de cada proceso; así como la insuficiencia de trabajos en estos temas, en determinadas áreas geográfica. Tal es el caso de América Latina, que se considera un nicho de investigación al existir muy pocos trabajos sobre corrupción.

3.4.1. Compras públicas y sesgo

Existen minorías que han sido tradicionalmente excluidas de los esquemas de compras públicas, tales como mujeres, personas con discapacidad, poblaciones indígenas, comunidades de la diversidad sexual, entre otros. Existe la posibilidad de que el dinero público se utilice también para beneficiar a estos sectores de la población que generalmente son excluidos.

Es importante señalar la perspectiva de género en compras públicas [43, 44] mediante la inclusión de criterios de igualdad de género en los procesos de compra y contratación puede fomentar la participación

de las mujeres en el mercado laboral y mejorar su situación económica y social. Además, la equidad de género en las compras públicas puede contribuir a reducir la brecha salarial entre hombres y mujeres y promover la igualdad de oportunidades en el acceso a recursos y oportunidades. En definitiva, la integración de la perspectiva de género en las compras públicas no solo es un acto de justicia social, sino también una medida efectiva para promover el desarrollo sostenible e inclusivo.

3.4.2. Datos abiertos en compras públicas

La disponibilidad de datos abiertos en compras públicas es crucial para mejorar la transparencia, eficiencia y eficacia de estos procesos. La información abierta permite que la ciudadanía tenga acceso a datos relevantes sobre los procedimientos de contratación, lo que promueve la rendición de cuentas y la participación ciudadana en la toma de decisiones. En los últimos años, la Unión Europea promueve dos iniciativas separadas de datos abiertos para licitaciones históricas, para su descarga masiva como bases de datos cohesionadas y fáciles de usar, tratando los contratos por encima y por debajo de los umbrales de tamaño en diferentes momentos del tiempo. Durante varios años y mediante varios trabajos de investigación [11, 45] se ha demostrado que los datos abiertos permiten aumentar la transparencia. Esto ha permitido concluir que los datos sobre contratación pública abierta promueven la licitación competitiva. En Latinoamérica [6] y África principalmente, no se establecen aun iniciativas de datos abiertos adecuadas, por lo que es un reto importante a la hora de desarrollar de sistemas que combaten la corrupción. La falta de uso de datos abiertos en compras públicas puede llevar a una toma de decisiones ineficiente y opaca. Lo que a su vez puede aumentar el riesgo de corrupción y disminuir la confianza en el gobierno y sus procesos de adquisiciones.

Debido a que en Ecuador hasta el 2022 no se tiene disponible un portal de compras públicas se analiza el uso de las herramientas de Web *scraping* [46] que permiten extraer datos del sitio web de compras públicas, para utilizarlos en un formato más estructurado. Se obtiene por lo tanto información como precios de productos, tipos de compras, parámetros de calificación, documentos asociados, etc.

3.5. Minería de datos (MD)

Es una disciplina que se sitúa entre de la estadística y las ciencias de la computación. Intenta descubrir patrones en grandes volúmenes de conjuntos de datos de forma automática, utilizando los métodos de aprendizaje automático, estadística y sistemas de bases de datos. Se concreta en la extracción de conocimiento significativo a partir de información útil pero no evidente, que está oculta en grandes conjuntos de datos. Su objetivo es extraer información y transformarla en una estructura comprensible, para su uso posterior y toma de decisiones Por otro lado, el término aprendizaje automático se refiere a los métodos computacionales utilizados para gestionar grandes volúmenes de datos de manera inteligente, mediante el desarrollo de algoritmos, con el objetivo de obtener ideas prácticas y aplicables [47]

La MD combina técnicas de IA con técnicas de análisis estadístico, las cuales permiten analizar grandes cantidades de información de todo tipo como datos, texto, audio [48] obteniendo patrones, tendencias e

información útil, que ayude a la toma de decisiones. Cada técnica utilizada dentro de la MD tiene sus propias reglas y métodos que se adaptan al problema que se pretende resolver, la elección del modelo viene determinada básicamente por dos condicionantes: el tipo de los datos y el objetivo que se quiere alcanzar.

Existen distintos tipos de algoritmos dentro de la minería de datos, cada uno con sus propios requisitos y que retornan información de distintos tipos. Los mismos siguen dos tipos de modelos que se clasifican en:

- Predictivos: a su vez de distinguen algoritmos de clasificación (árboles de decisión, reglas de asociación, redes bayesianas, redes neuronales, SVM) y regresión. Estos a su vez se clasifican en dos grandes grupos, los que utilizan el aprendizaje supervisado y no supervisado.
- Descriptivos: encontramos los siguientes algoritmos: reglas de asociación, correlaciones, *clustering*.

3.5.1. Metodología para minería de datos

La *Cross Industry Standard Process for Data Mining* (CRISP-DM) [49] es una metodología de libre distribución que puede trabajar con cualquier herramienta para desarrollar cualquier proyecto, esta metodología estructura el ciclo de vida de un proyecto de Minería de Datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.
Una de las ventajas es que es un modelo de proceso iterativo y flexible que puede adaptarse a diferentes proyectos de minería de datos, incluyendo la detección de corrupción en compras públicas. CRISP-DM se enfoca en las tareas esenciales que deben ser abordadas durante todo el ciclo de vida del proyecto, incluyendo la comprensión del problema, la selección de datos, la preparación de estos, la modelización, la evaluación y la implementación. Por lo que CRIS-DM permite establecer un enfoque sistemático y estructurado para el análisis de datos, lo que puede mejorar la precisión y la eficacia del proceso de detección de corrupción en compras públicas. CRIS-DM se ha utilizado para elaborar el modelo multifase que identifique procesos con anomalías y posteriormente pueda detectar posible corrupción en la asignación de parámetros de calificación de procesos en la contratación pública.

3.6. Aprendizaje no supervisado

El aprendizaje no supervisado puede ser una herramienta útil para el análisis de corrupción en compras públicas, al permitir la identificación de patrones y relaciones ocultas en los datos [50] sin la necesidad de tener etiquetas o categorías predefinidas. Encontrar un buen algoritmo de aprendizaje no supervisado para analizar la corrupción en compras públicas es importante porque los datos relacionados con este tema son muy complejos, heterogéneos y grandes en volumen. Un algoritmo de aprendizaje no supervisado efectivo puede ayudar a identificar patrones y tendencias en los datos que podrían ser difíciles de detectar de otra manera. Además, el análisis no supervisado permite descubrir información valiosa sin necesidad de tener un conocimiento previo de las variables principales, ya que en muchos procesos puede ser difícil de detectar las variables adecuadas o se puede involucrar variables que no son evidentes a simple vista.

3.6.1. Mapas autoorganizados

En 1982 T. Kohonen presentó un modelo de red denominado mapas auto-organizados o SOM (Self-Organizing Maps). En la actualidad siguen siendo una herramienta muy útil en el campo del aprendizaje no supervisado. En particular, se utilizan para resolver problemas de agrupamiento de datos y de visualización de información. La principal ventaja de los SOM es su capacidad para representar de manera topológica los datos, es decir, preservando las relaciones de cercanía entre los nodos. De esta manera, el SOM son capaces de encontrar patrones y estructuras complejas en los datos, lo que resulta en una mejor comprensión de estos [51]permiten una fácil interpretación de los resultados y una visualización intuitiva de los mismos, lo que resulta en un mayor entendimiento y una mejora en la toma de decisiones. En el caso específico de esta tesis doctoral se han utilizado para encontrar las variables de mayor relevancia, para evaluar los ganadores del proceso de compras públicas. Mediante el entrenamiento de los mapas autoorganizados con datos de compras, usando la topología adecuada para identificar cuáles características tienen una mayor influencia en la distribución de los datos. Los nodos cercanos entre sí en el mapa indican que las características asociadas tienen patrones similares y podrían ser redundantes o altamente correlacionadas. De esta forma, es posible identificar patrones de alta dimensionalidad y reducir la dimensionalidad del conjunto de características de manera no supervisada. Este enfoque también permite visualizar los datos en el mapa y detectar agrupaciones o tendencias que pueden ser relevantes para la interpretación y el análisis posterior.

3.6.2. Algoritmos de agrupación

Se pueden utilizar técnicas de *clustering* para agrupar las compras públicas en diferentes categorías según su similitud en términos de características relevantes, lo que puede ayudar a identificar posibles casos de corrupción.
El algoritmo K-Means [52] es una de las técnicas de aprendizaje no supervisado más populares. Se utiliza para analizar datos y encontrar grupos dentro de esos datos, utilizando algún tipo de medida de similitud, como la distancia euclidiana. El algoritmo K-Means al depender del número de centroides definidos al inicio, puede presentar varios problemas, pero en diversos estudios se ha demostrado que se pueden solucionar complementando con varios métodos, de acuerdo con el tipo de datos y el problema a solucionar. En compras públicas se puede aplicar el *clustering* para agrupar las transacciones similares y analizar si existen patrones de comportamiento inusual, como precios extremadamente altos, empresas que siempre ganan los mismos contratos, tipos de compras que son favoritas por ciertas instituciones, fechas de levantamiento y adjudicación de procesos controversiales.

3.6.3. Métricas de evaluación aprendizaje no supervisado

Al no existir una salida de datos específica a comparar con los resultados obtenidos en el aprendizaje no-supervisado, es necesario medir la calidad de los agrupamientos o las relaciones encontradas entre las variables de entrada para la detección de corrupción. Por lo tanto, se utilizan métricas como la distancia intra e intergrupal, la cohesión y la separación entre los grupos y la tasa de reducción de dimensionalidad.

Además, es importante considerar la interpretación y relevancia de los agrupamientos encontrados y cómo estos pueden ser utilizados para identificar patrones de corrupción en las compras públicas.

3.7. Aprendizaje supervisado

El aprendizaje supervisado es una técnica de aprendizaje automático que se utiliza para detectar patrones y tendencias en los datos, utilizando un conjunto de datos etiquetados. Es capaz de inferir información con el empleo de algoritmos y un grupo de datos previamente etiquetado y como resultado transferir sus características a una predicción [53]. En el contexto de la detección de corrupción en compras públicas, el aprendizaje supervisado es importante porque puede utilizar datos históricos etiquetados como "*corruptos"* y "*no corruptos*" para entrenar un modelo de detección de corrupción. Una vez que se entrena el modelo, se puede utilizar para identificar posibles casos de corrupción en nuevas compras públicas. El aprendizaje supervisado permite la identificación de patrones complejos y sutilmente ocultos, que podrían ser difíciles de detectar de otra manera, lo que lo hace una herramienta poderosa en la lucha contra la corrupción en compras públicas.

3.7.1. Suported Vector Machine

Generalmente se usa *Suported Vector Machine* (SVM) para solventar problemas de clasificación binaria, el SVM encuentra el hiperplano óptimo que separa los datos de entrenamiento en dos clases [54]. Estos modelos pueden ser especialmente útiles cuando se trata de problemas de clasificación binaria, como la detección de transacciones corruptas. Cuando se aplica esta técnica la propiedad *nu* permite controlar el equilibrio entre los valores atípicos y los casos normales. Por lo que se asigna *nu= [1e-3, 1e-2, 1e-1, 1],* mientras que el parámetro que afecta al número de iteraciones utilizadas[55], cuando se optimiza el modelo se toma como *epsilon = [1e-4, 1e-3, 1e-2]*. Luego se determinan los hiperplanos óptimos para el aprendizaje automático mediante un modelo *Hyperparameters*. Este proceso se acompaña de validación cruzada para minimizar la dependencia de los datos sobre el resultado de la parametrización, permitiendo que el modelo tenga una mejor capacidad de generalización y sea menos propenso a sobre ajustarse a los datos de entrenamiento. Se usan varias métricas para determinar la precisión y eficacia del modelo. SVM es una herramienta valiosa para el análisis de corrupción en compras públicas, ya que es un modelo muy preciso y puede funcionar bien con diferentes tipos de datos y características.

3.7.2. Redes Neuronales

Las redes neuronales se inspiran en el funcionamiento del cerebro humano y su capacidad para procesar información de manera paralela. Están compuestas por múltiples capas de nodos interconectados que procesan los datos de entrada y producen una salida. Su capacidad para identificar patrones complejos y encontrar relaciones no lineales entre los datos las hace muy útiles en la detección de anomalías. Una vez realizado el entrenamiento, se puede detectar desviaciones de la norma en nuevos datos presentados a la red neuronal y señalarlos como posibles anomalías. Las redes neuronales se usan con éxito en diversas aplicaciones, como la detección de fraudes.

Pero las redes neuronales ordinarias ignoran la estructura de los datos de entrada. Esto se debe a que, para introducir las entradas en la red, los datos tienen que ser convertidos en un vector unidimensional. Lo que funciona bien para datos normales, pero no para datos complejos utilizados en aplicaciones de visión artificial o PLN. Aquí es donde entra en juego el aprendizaje profundo y las redes neuronales convolucionales o *Convolutional Neural Networks* (CNN). Si construimos una red neuronal con muchas capas, esta se convierte en red neuronal profunda. Este tipo de modelos está revolucionando el PLN haciendo posible que los ordenadores entiendan mejor el lenguaje humano.

Una red neuronal de aprendizaje profundo (*Deep Learning*) es una estructura compuesta por múltiples capas de neuronas que se conectan entre sí de manera distinta a las redes neuronales tradicionales y utiliza una función de activación para generar automáticamente atributos en las capas internas. La conexión entre capas se realiza en tres formas básicas: convolución, conexión total y agrupación [56]. El aprendizaje profundo es la actividad automática de adquisición de conocimiento, mediante el uso de redes neuronales u otro sistema de aprendizaje automático, donde se usan varios niveles de abstracción[56]. Este permite que un modelo informático realice tareas de clasificación utilizando imágenes, texto o sonido, lo cual lo hace especialmente útil en el PLN. Se ha demostrado que los modelos de aprendizaje profundo logran niveles de precisión que en ocasiones supera el rendimiento humano. Se logran modelos vanguardistas de aprendizaje profundo entrenándolos con conjuntos de datos etiquetados extensos y arquitecturas de redes neuronales con muchas capas [57].

En numerosas investigaciones se ha demostrado que el uso de redes neuronales puede permitir una clasificación y predicción adecuada de corrupción en compras públicas [58]. Sin embargo, también se nota que para lograr esto se necesita cierta calidad en los datos previos y un etiquetado de procesos sospechosos. El modelo de red incluye una capa de entrada con nodos correspondientes a las características de las compras públicas, una o varias capas ocultas que procesan la información y una capa de salida que indica si se ha detectado corrupción o no. Para el entrenamiento, se debe dividir el conjunto de datos en conjuntos de entrenamiento, validación y prueba. Posterior a la etapa de entrenamiento se utiliza el modelo entrenado para predecir la corrupción en nuevas compras públicas. Si el modelo indica que una compra puede ser corrupta, se puede investigar más a fondo para determinar si se requiere una acción adicional.

3.7.3. Métricas de evaluación aprendizaje supervisado

La evaluación de la detección de corrupción en compras públicas puede involucrar el uso de diversas métricas, dependiendo de la técnica o modelo utilizado. Para el caso puntual de la tesis doctoral y los algoritmos empleados, se usan: el *recall* (ecuación 1), la precisión (ecuación 2) y el F*1-score*, que permiten evaluar el rendimiento del modelo en términos de la detección de los casos de corrupción. El *recall*, también conocido como sensibilidad o tasa de verdaderos positivos, mide la proporción de casos positivos que el modelo clasifica correctamente. Su cálculo se basa en la división del número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos. La precisión es utilizada para poder saber qué porcentaje de valores que se han clasificado como positivos son realmente positivos.

En cambio, *F1-score* es una medida que combina precisión y el *recall* en una única métrica. Es útil cuando hay un desequilibrio entre las clases en los datos de entrada. Se calcula como la media armónica entre precisión y *recall*, y proporciona una medida equilibrada del rendimiento del modelo.

- *Recall*

$$P = \frac{TP}{TP + FN} \quad (1)$$

- Precisión

$$P = \frac{TP}{TP + FP} \ (2)$$

Donde,

Verdadero positivo (TP, True Positive): se refiere a una detección y clasificación correctas. Son los valores que el algoritmo clasifica como positivos y que realmente son positivos.

Falso positivo (FP, False Positive): valores que el algoritmo clasifica como positivo cuando realmente son negativos. Se trata de una detección y clasificación incorrectas.

Verdadero negativo (TN, True Negative): son valores que el algoritmo clasifica como negativos (0 en este caso) y que realmente son negativos.

Falso negativo (FN, False Negative): valores que el algoritmo clasifica como negativo cuando realmente son positivos.

Además, se pueden utilizar métricas específicas según la técnica utilizada, como el *AUC-ROC* (*Area Under Curve*), que es una medida de rendimiento que evalúa la capacidad de un modelo de clasificación para distinguir entre clases positivas y negativas. Representa el área bajo la curva trazada por la tasa de verdaderos positivos (sensibilidad) en función de la tasa de falsos positivos (1-especificidad). Un valor de *AUC-ROC* cercano a 1 indica un buen rendimiento del modelo en la clasificación, mientras que un valor cercano a 0.5 indica una clasificación aleatoria. En el contexto de la detección de corrupción un alto valor de *AUC-ROC* puede indicar que el modelo tiene la capacidad de identificar procesos sospechosos o fraudulentos con mayor precisión.

3.8. Procesamiento de lenguaje natural

El procesamiento de lenguaje natural (PLN) es una rama de la IA que se enfoca en la interacción entre las máquinas y el lenguaje humano[59] . Para lograr este objetivo se necesita representar las palabras y oraciones de una manera que las máquinas puedan entender. Lo que significa convertir un texto a datos numéricos para que puedan ser utilizados por los sistemas del aprendizaje automático. Así surgieron los vectores con codificación *one hot*. Los cuales tienen el mismo número de dimensiones que el vocabulario que se quiera representar. Esto supone un problema porque ocuparía demasiado espacio, además de que este tipo de codificación considera que todas las palabras están a la misma distancia

unas de otras. Por lo que se necesita comprimir estos vectores, utilizando las redes neuronales artificiales, ya que el vector de salida puede ser de cualquier dimensión y se puede entrenar a la red en función de lo que se busque predecir. Constituyendo el origen de los *word embeddings* o incrustación de palabra, que es el nombre de un conjunto de modelos de lenguaje y técnicas de aprendizaje en PLN en donde las palabras o frases del lenguaje natural son representadas como vectores de números reales. Conceptualmente implica el encaje o la incrustación matemática de un espacio con una dimensión por palabra a un espacio vectorial continuo con menos dimensiones [2]. Estos modelos son actualmente un componente esencial en la mayoría de las tareas de PLN. Uno de los más utilizados es Word2Vec [3], que es un algoritmo de aprendizaje autosupervisado en el que palabras iguales, tienen *word embeddings* similares.

En el contexto de la corrupción, el PLN tiene el potencial de ayudar a analizar grandes cantidades de texto, como contratos gubernamentales, informes de auditoría, noticias y redes sociales, para identificar patrones de corrupción y detectar posibles casos. Las técnicas de PLN también pueden utilizarse para construir sistemas de alerta temprana, que detecten y notifiquen a los responsables de la toma de decisiones cuando se detectan indicios de corrupción en los documentos analizados. Para utilizar PLN en este contexto, primero es necesario preprocesar los datos textuales, lo que implica cubrir diferentes etapas como: la eliminación de palabras irrelevantes, la *tokenización* y la lematización. Luego, se puede aplicar una técnica de modelado para identificar temas y patrones comunes en los datos, lo que puede ayudar a identificar anomalías y posibles casos de corrupción. En trabajos recientes se están aplicando técnicas novedosas basadas en *Transformers* [60]. De ahí que los modelos basados en *Transformer* se les denomine también *Embeddings* Contextuales. Los resultados de investigación sugieren, que las *word embeddings* de Word2Vec logran los mejores resultados basados en las métricas ROUGE-1 [61]y BLEU [62]. Las métricas de evaluación pueden incluir la precisión, el *recall* y la *F1-score*, que miden la capacidad del modelo para detectar la corrupción y evitar falsos positivos y falsos negativos.

Dentro del PLN, el análisis de los sentimientos puede ser útil para determinar si hay señales de corrupción, en los documentos de compras públicas, como la presencia de términos o frases negativas o sospechosas. Además, el PLN también puede ser utilizado para crear *chatbots* y asistentes virtuales que brinden información y orientación a los ciudadanos sobre concursos públicos. Así como brindar información a los empleados del gobierno sobre cómo denunciar casos de corrupción.

CAPÍTULO 4

El presente capitulo trata de dar respuesta a la problemática planteada, detallando la propuesta del presente trabajo, diseñada para contrastar la hipótesis. De manera general, se describe un sistema de minería de datos que utiliza PLN para realizar de forma inteligente el análisis de contenido de los contratos y detectar, de forma automática la corrupción. Además, se resumen las 6 contribuciones científicas que forman parte del trabajo doctoral.

4. CONTRIBUCIONES

4.1. Propuesta

El planteamiento ha sido diseñado para contrastar la hipótesis de esta tesis doctoral. La cual sostiene que, con el uso de la MD, el aprendizaje automático y el PLN, es posible detectar y prevenir la corrupción en los procesos de compras públicas. Se incluyen técnicas de análisis de datos que detectan patrones y anomalías en los procesos de compras; y evalúan la representatividad de los proveedores y la justicia de la evaluación de ofertas. La metodología se compone de diferentes módulos que son descritos en la propuesta; cada uno diseñado para abordar los distintos tipos de corrupción presentes en los procesos. Se pretende identificar los procesos con anomalías y posteriormente poder detectar posible corrupción en la asignación de parámetros de calificación, para poder distinguir y clasificar el sobreprecio, favoritismo y sesgo. Con la combinación de las técnicas que se emplean, se busca obtener resultados sólidos y confiables que contribuyan a la detención temprana y la prevención efectiva de la corrupción en el ámbito de las compras públicas.

4.1.1. Obtención de información

Tras analizar los trabajos publicados, donde se ha podido detectar de mejor manera la corrupción, se determina que, para la recopilación de datos, se deben cumplir con estándares específicos basados en datos abiertos, sin embargo, en el portal de compras públicas de Ecuador estos estándares no se cumplen; en consecuencia, en la propuesta se contemplan 3 fases para la obtención de información:

I. *Web scraping* sobre el portal SERCOP para obtener los datos textuales y documentos, mediante un *script* automático. A través de esta técnica, se obtienen datos no estructurados para convertirlos en datos estructurados y almacenarlos en una base de datos [59].

II. La limpieza de procesos con el Análisis Exploratorio de Datos o EDA (Exploratory data analysis) en cada conjunto de datos.

III. Almacenamiento de datos en una base de datos no relacional para mejorar el acceso a los mismos.

4.1.2. Obtención y limpieza de datos

Con el empleo del *Web scraping*, fue posible extraer información pública y oficial de la base de datos del SERCOP de Ecuador. Se obtuvieron datos de 993759 procesos de compras públicas efectuados en el país. Los datos obtenidos correspondían a procesos enmarcados entre los años 2010 y 2020. Incluían descripciones de los procesos, fechas, productos, parámetros de calificación, invitaciones, archivos y preguntas de los proveedores. Todos fueron guardados en una base de datos independientemente del estado en el que finalizaron (exitosos, o desiertos), del monto otorgado, de los parámetros de calificación emitidos y de las preguntas y aclaraciones realizadas en el proceso. No se discrimina ningún proceso y los datos fueron agrupados de la siguiente forma:

1. Datos de compras, sin importar distinción.
2. Datos de compras de medicamento*s* durante la pandemia.
3. Datos de procesos con parámetros de calificación.
4. Datos de procesos en los cuales se realizaron, preguntas, aclaraciones o modificaciones.

Para la limpieza de datos, en los grupos de datos se eliminaron los valores atípicos, se hizo la corrección de errores, el manejo de valores faltantes y la normalización de los datos antes de realizar el análisis. Para el análisis exploratorio de datos se empleó el preprocesamiento de datos para minería predictiva de datos [63]. En este sentido se siguió una limpieza filtrando datos de manera tal que quedaran descartados o corregidos, los valores que según criterios preestablecidos se consideraron sospechosos. De esta forma fue posible incluir el conjunto de datos que se tomaron como entrada al algoritmo de minería de datos seleccionado. Se adoptaron técnicas de preprocesamiento que ofrecen modelos de clasificación de alto rendimiento [64]

Esta metodología permitió buscar dentro de los datos los que tenían ciertos parámetros incorrectos. Por ejemplo, parámetros de calificación que no llegaban al 100%, los procesos considerados repetidos, o con valores inconsistentes, los valores atípicos e incorrectos. Estos fueron limpiados para trabajar con los procesos que sí aportan valor. Como se mencionó anteriormente, sin importar el estado en el que finalizaron (exitosos, o desiertos); solo se excluyeron los que tenían inconsistencias según el análisis exploratorio de datos.

Resumiendo, se realizó un proceso de limpieza para eliminar el ruido y las inconsistencias. Esto implicó identificar y corregir datos incorrectos, eliminar duplicados y normalizar los valores para garantizar la coherencia en los datos.

Los atributos o características relevantes utilizados para encontrar patrones fueron los siguientes:

1. Oferta económica

   Se evaluaron las puntuaciones de la oferta económica en los contratos, lo que permitió identificar procesos en los que se otorgaban puntuaciones bajas, lo que podría indicar posibles favoritismos o perjuicio para el gobierno.

2. Experiencia
Se consideró la experiencia de los licitadores como un atributo relevante para evaluar los procesos de contratación.
3. Equipos e instrumentos
Se evaluaron los requisitos de equipos e instrumentos en los procesos de contratación.
4. Producción nacional
Se tuvo en cuenta la producción nacional como un factor de calificación en los contratos.

4.1.3. Metodología planteada

Una vez analizadas las principales investigaciones recientes en relación con la hipótesis que plantea esta tesis, y tras haber detectado las carencias y necesidades relacionados con la detección automática de la corrupción en los sistemas existentes, se desarrolla la propuesta. Concretamente, se plantea la arquitectura de un sistema heterogéneo capaz de englobar diferentes metodologías y técnicas para el análisis e identificación de riesgos de corrupción en las compras públicas. A continuación, se describen detalladamente cada uno de los enfoques que se han considerado, para abordar los distintos tipos de corrupción presentes en los procesos. Constituyendo cada uno de ellos un marco de trabajo de esta propuesta.

La metodología inicia con la recopilación y procesamiento de datos relacionados con las compras públicas: descripciones de los procesos, fechas, productos, parámetros de calificación, invitaciones, archivos y preguntas de los proveedores.

Dentro de aprendizaje automático se emplea el aprendizaje supervisado que trabaja con datos etiquetados, el aprendizaje no supervisado para datos difíciles de visualizar, y que evita la necesidad de tener conocimiento previo de las variables principales que intervienen en el proceso y que son difíciles de detectar, pues se encuentran en varias dimensiones, lo que exige la reducción de la dimensionalidad. También se emplea el aprendizaje semisupervisado que tiene como objetivo la clasificación, con una entrada que contiene tanto datos etiquetados y como sin etiquetar.

Dada la heterogeneidad de la metodología y la independencia de cada uno de los enfoques desarrollados, se describen los tres casos de estudio bien diferenciados que contribuyen a la validación de la hipótesis:

- El primer caso de estudio, se investiga sobreprecio en compras de medicamentos durante la pandemia COVID 2019. En el mismo se utiliza el aprendizaje supervisado para establecer comparaciones de los precios de los bienes y servicios adquiridos en los procesos de compras públicas, pues en este caso se cuenta con datos etiquetados en relación con los precios de referencia, tales como precios de mercado o precios históricos de compras anteriores.

- El segundo caso de estudio se centra en el favoritismo, basado en los parámetros de calificación de procesos y los tipos de contratación. En este caso, se utilizan técnicas de aprendizaje no supervisado para detectar patrones anómalos en los contratos. En base a esta información se aplica aprendizaje supervisado para clasificar y prevenir procesos con anomalías y proyectar trabajos futuros. Además del análisis de proveedores, se aplican técnicas de *clustering* (como SOM) y *Deep Learning* para identificar patrones y agrupar procesos de compras públicas sospechosos. Esto ayuda a detectar comportamientos anómalos y potencialmente corruptos en los procesos de contratación. A partir de la información obtenida de los pasos anteriores, se implementa un enfoque de aprendizaje semisupervisado. Esto implica utilizar un conjunto de datos etiquetados (procesos con anomalías, anulados por conflictos de intereses o denuncias en redes sociales) y un conjunto de datos no etiquetados (procesos restantes) para entrenar modelos de detección de corrupción más precisos. Este enfoque se implementa porque el utilizar datos etiquetados y no etiquetados mejora la precisión del sistema de detección de corrupción.

- El tercer caso de estudio, basándonos en la evaluación de igualdad de condiciones tanto en género como en oportunidades para el participante, se trata de detectar el sesgo de género y el favoritismo en cada proceso de compra pública, a través del indicador preguntas, respuestas y aclaraciones. Una vez realizada la extracción de patrones, se realiza un análisis de los datos de compras públicas donde se involucra la identificación de palabras claves, temas y agrupación de documentos que permiten detectar sesgo y favoritismo en el proceso de contratación con el empleo del PLN. Esta técnica se emplea para identificar perfiles de proveedores que pueden representar diferentes tipos de comportamiento sospechosos o niveles de riesgo de corrupción, a partir de agrupar proveedores similares en función de características comunes.

La metodología descrita se concreta sobre la base de un sistema de MD (Figura 3), que resume las fases empleadas en cada caso de estudio. Donde para detectar sobreprecio se emplean técnicas de aprendizaje supervisado y para detectar patrones anómalos en los contratos se utiliza técnicas de aprendizaje no supervisado. Posteriormente en base a la información de salida de esta fase, se aplica nuevamente el aprendizaje supervisado para clasificar y prevenir procesos con anomalías. Finalmente se analiza el indicador de preguntas, respuestas y aclaraciones para detectar favoritismo y sesgo por parte de las entidades que realizan compras. Como consecuencia la metodología analiza tres tipos de corrupción con un enfoque hibrido de técnicas, además el diseño modular favorece la condición heterogénea del enfoque propuesto.

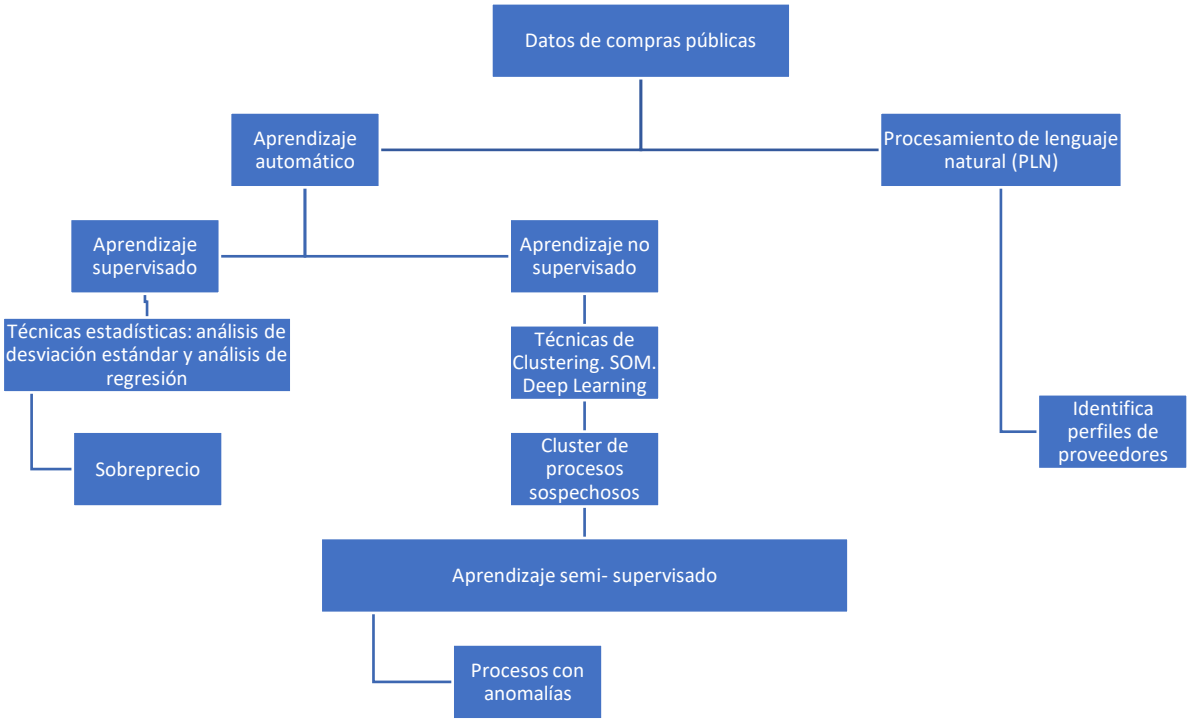


Figura 3: Metodología Propuesta para el segundo y tercer caso de estudio

Dentro del análisis de corrupción se consideran varios indicadores, entre ellos la publicación de la convocatoria, la toma de decisiones durante los procesos de selección y contratación, los precios, el número de ofertas presentadas, las competencias de la entidad compradora y el origen de los oferentes. Además, se tienen en cuenta las señales de alerta relacionadas con el proceso de licitación, como: un tiempo muy corto entre el llamado a licitación y la presentación de ofertas, o requisitos muy desafiantes para participar en los procesos de selección. En definitiva, la metodología propuesta consta de tres fases y abarca varios casos de estudio.

La Figura 4, resume las diferentes fases de la metodología para evaluar un proceso de compras públicas. Se inicia analizando el sobreprecio en los bienes o servicios comprados y generando un modelo de regresión lineal para futuras evaluaciones. Posteriormente, se toman en cuenta los parámetros para la calificación del ganador de la oferta. Si los parámetros no son adecuados se elabora un modelo de detección de anomalías basado en SVM y redes neuronales. Finalmente se evalúan las preguntas, respuestas y aclaraciones del proceso con PLN, para determinar favoritismo y sesgo en la compra obteniendo con esto un indicador final que determina la probabilidad de indicio de corrupción.

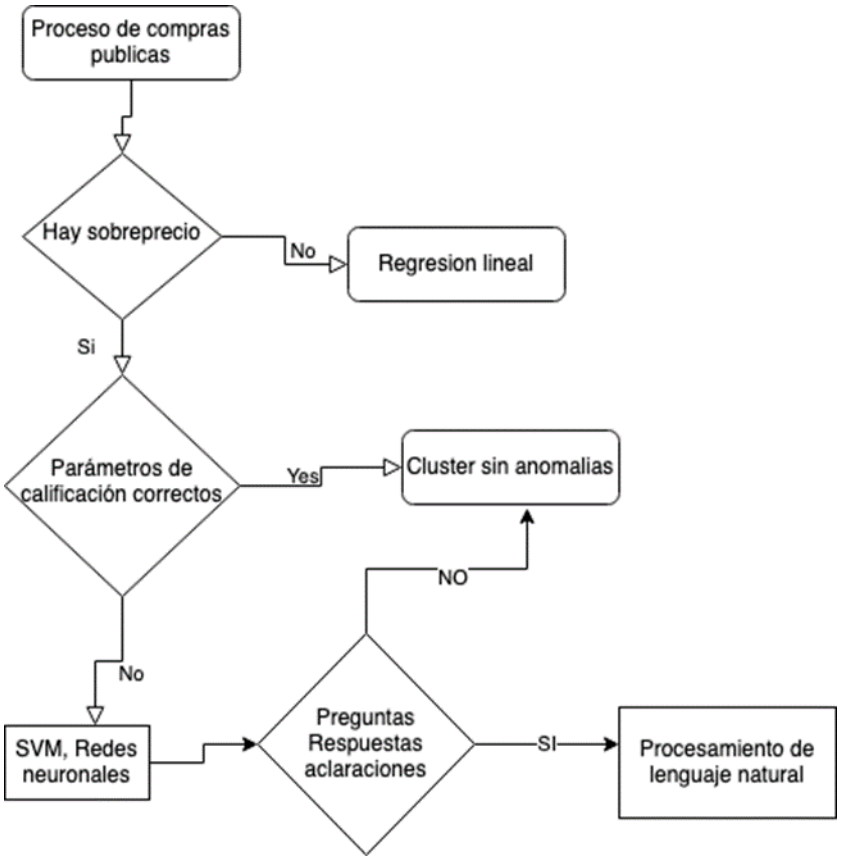


Figura 4: Fases de la metodología para evaluar proceso de compras públicas

4.1.4. Fase 1 (Sobreprecios)

La detección del sobreprecio en el contexto de compras públicas se lleva a cabo mediante un análisis exhaustivo que involucró técnicas estadísticas y de aprendizaje automático. En esta fase se emplea específicamente el aprendizaje supervisado, pues se realiza una comparación de los precios de los bienes y servicios adquiridos en los procesos de compra pública con precios de referencia, tales como precios de mercado o precios históricos de compras anteriores.

Para llevar a cabo el análisis, se aplican técnicas estadísticas avanzadas, como el análisis de desviación estándar y el análisis de regresión, que permiten identificar aquellos precios que se encuentren significativamente por encima de los precios normales. El análisis de desviación estándar se utiliza para evaluar la dispersión de los precios de los bienes y servicios adquiridos en los procesos de compra pública. Esto permite identificar cuán variados son los precios en relación con su precio promedio. Al encontrar una desviación estándar alta en comparación con los precios de referencia, se indica una posible existencia de sobreprecio en algunas adquisiciones. Es preciso destacar que un valor de desviación estándar alto indica que los datos están más dispersos, mientras que un valor bajo indica que los datos están más cercanos a la media.

En este orden de ideas, el análisis de regresión se utiliza para establecer una relación entre los precios de los bienes y servicios adquiridos en los procesos de compra pública y los precios de referencia, como son: los precios de mercado y los precios históricos de compras anteriores. Este análisis permite determinar la existencia o no de una asociación significativa entre precios y si los precios observados se desvían de los precios de las referencias esperados. En consecuencia, se emplea un modelo de regresión lineal tomando en cuenta que los datos analizados en esta fase son cuantitativos y continuos y se busca analizar la relación entre variables precios observados y los precios de referencia.

Asimismo, se utilizaron algoritmos de aprendizaje automático, como clasificadores y modelos de detección de anomalías, para identificar patrones y comportamientos sospechosos en los datos de precios. Entre los algoritmos utilizados se encuentran clasificadores como *Random Forest*, SVM y *Naive Bayes*, los cuales permiten categorizar las transacciones en normales o sospechosas de sobreprecio. Además, se emplearon modelos de detección de anomalías como *Isolation Forest* y *One-Class SVM*, los cuales identifican transacciones que se desvían significativamente del comportamiento típico. Estos algoritmos de aprendizaje automático se utilizaron en conjunto con técnicas de procesamiento de datos y extracción de características relevantes para identificar de manera eficiente y precisa los casos de sobreprecio en las compras públicas.

Una vez obtenidos los resultados del análisis, estos fueron sometidos a una verificación detallada, mediante la revisión de documentos, denuncias en redes sociales y comentarios de otros proveedores asociados al proceso, empleando las técnicas de PLN. Esta evaluación permitió confirmar y respaldar los hallazgos obtenidos a través del análisis estadístico y de aprendizaje automático.

Es importante destacar que este análisis se llevó a cabo mediante la implementación de una metodología rigurosa. Se recopilaron y se prepararon los datos pertinentes, los cuales incluyeron información detallada sobre los procesos de compra, los productos adquiridos y los precios asociados. Se aplicaron técnicas estadísticas y algoritmos de aprendizaje automático, como árboles de decisión y modelos de detección de anomalías basados en redes neuronales *autoencoder* para obtener una detección precisa y confiable del sobreprecio. La combinación de estas técnicas con el PLN permitió realizar una verificación adicional que corroboró la veracidad de los hallazgos.

4.1.5. Fase 2 (Parámetros de calificación)

Los parámetros de calificación son criterios que se utilizan para evaluar y comparar las ofertas de los proveedores. Es por ello por lo que, esta fase se encargó de evaluar si los parámetros de calificación habían sido definidos de manera clara y si se aplicaban de forma imparcial, evitando interpretaciones subjetivas y sesgadas que puedan aumentar el riesgo de corrupción.

Para llevar a cabo este análisis, se empleó un enfoque que permite identificar patrones y tendencias en la evaluación de las ofertas y en la selección de los proveedores. Se examinan detalladamente los criterios de calificación establecidos y posteriormente se realiza una agrupación de ofertas similares con el fin de evaluar el desempeño de los proveedores en relación con los parámetros de calificación.

Esto implicó analizar las puntuaciones obtenidas por los proveedores en cada criterio de calificación y compararlas entre sí. El objetivo de este análisis fue identificar posibles casos de corrupción o favoritismo que puedan haber influido en la selección de proveedores.

En los casos donde se detectan posibles desviaciones o anomalías en la evaluación de las ofertas, se recurre al aprendizaje supervisado como una herramienta para predecir los resultados esperados de la evaluación en función de los parámetros de calificación. Esto permite identificar discrepancias significativas entre las evaluaciones reales y las predicciones. Lo cual puede indicar la presencia de irregularidades en el proceso de selección.

Es importante destacar que este análisis de los parámetros de calificación se llevó a cabo siguiendo un enfoque meticuloso y basado en datos. Se utilizaron técnicas estadísticas y de aprendizaje automático. Haciendo un análisis de los resultados de clasificación, para asignar categorías a instancias de datos según ciertas características relevantes de los parámetros de calificación, que podrían destacar la presencia de corrupción o favoritismo. De manera que el modelo aprende a reconocerlas y detecta patrones y comportamientos sospechosos. Además, se aplica el modelo de detección de anomalías de varias fases que utiliza diferentes algoritmos, incluyendo *clustering* (K-Means, SOM), SVM y *Principal Component Analysis* (PCA). Por su parte, el modelo de regresión lineal fue empleado para evaluar la relación entre los parámetros de calificación y los resultados de la evaluación de las ofertas.

Las técnicas de PLN permitieron procesar y analizar documentos como convocatorias, decisiones de selección y contratación, y requisitos establecidos, con el objetivo de identificar patrones, tendencias y posibles irregularidades que podrían indicar casos de corrupción o favoritismo. El uso del PLN en esta fase contribuyó a obtener una comprensión más profunda de los datos textuales y a detectar indicios de posibles desviaciones o anomalías en los procesos de adquisición.

4.1.6. Fase 3 (Preguntas y aclaraciones)

En la tercera fase del trabajo, se puso énfasis en el análisis de las preguntas realizadas por los participantes en un proceso de compras. Estas preguntas desempeñan un papel crucial al permitir aclarar dudas cuando los términos de referencia no son claros y proporcionar un mecanismo para que los proveedores realicen observaciones cuando sospechan que el proceso está sesgado hacia un proveedor en particular. Además, esta sección permite a la entidad contratante realizar modificaciones al proceso, por lo que es fundamental analizar el contenido escrito por cada una de las partes involucradas.

En esta fase, se aplican fundamentalmente técnicas de PLN, como el análisis de sentimiento para evaluar el sesgo y los sentimientos expresados en las respuestas y modificaciones realizadas por la entidad contratante. El PLN permitió analizar el texto escrito y extraer información relevante, como la presencia de sesgos o favoritismos hacia ciertos proveedores. Asimismo, evaluó el sentimiento y

las actitudes implícitas en el texto. Lo que ayudó a detectar cualquier indicio de favoritismo en los procesos de contratación.

El uso de técnicas de PLN y análisis de sentimiento en esta fase permitió evaluar de manera objetiva y sistemática, si las respuestas y modificaciones realizadas por la entidad contratante están libres de sesgo y favoritismo. Con su aplicación es posible detectar en los procesos de contratación posibles irregularidades o comportamientos sospechosos. Estas técnicas implicaron enfoques específicos como el uso de algoritmos de clasificación, extracción de características lingüísticas y modelos de aprendizaje automático.

Se emplean técnicas específicas como Word2Vec y FastText para analizar el sesgo de género en preguntas y aclaraciones, así como hacia otros proveedores. Estas técnicas permitieron identificar patrones lingüísticos y representar de forma óptima las palabras en un espacio vectorial, lo que facilitó el análisis y la detección de sesgos. Este último resulta relevante en la detección de irregularidades y comportamientos sospechosos en los procesos de contratación, ya que revela acciones discriminatorias o favoritismos indebidos.

4.2. Resultados

Los resultados obtenidos a través de la aplicación de la metodología propuesta en la tesis doctoral cumplen con los objetivos planteados y contribuyen significativamente al estudio de la detección de corrupción, dando como resultado diferentes publicaciones de impacto indexadas.

La selección de las fases de la metodología se basó en la falta de exploración conjunta por parte de otros autores, así como en su potencial para evaluar problemas fundamentales que se habían detectado dentro del SERCOP: sobreprecios, alteración de parámetros de calificación en la contratación pública, posibles sesgos y favoritismos en la adjudicación de contratos. En la Tabla 2 se presenta un resumen de las variables de alerta detectadas en cada etapa, así como los datos requeridos para su evaluación.

Tabla 2: Resumen de variables y su incidencia en la alerta de corrupción

Fases del modelo	Variables de análisis	Variables de alerta
Sobreprecio	Precio referencial. Marca producto. Fecha de compra. Tipo de compra. Entidad contratante.	Precios fuera de la media. Monopolio de marcas Frecuencia de proveedores. Compras de otras entidades.
Parámetros de calificación	Tiempo de oferta. 32 parámetros de calificación. Tipo de proceso.	Requisitos extras solicitados. No se considera oferta económica. Procesos de consultoría direccionados.

Preguntas y aclaraciones	Preguntas de los proveedores. Respuestas de la entidad contratante. Cambios en los términos por parte de la entidad.	Sesgo de género en respuestas. Favoritismo en respuestas. Cambios de términos direccionados.

Debido a la gran cantidad de datos obtenidos, el proceso de análisis se realizó utilizando computación en la nube y los datos se almacenaron en una base de datos no relacional. A través del análisis de los datos recopilados, se identificaron los resultados para detectar diversas alertas de corrupción en cada etapa de la metodología:

1) **En la fase de sobreprecio** se observaron precios fuera de la media, así como la presencia de monopolios de marcas y una alta frecuencia de determinados proveedores en ciertos procesos de compra.
   Se realizó un análisis estadístico descriptivo para extraer medidas relevantes para cada producto y determinar la variabilidad entre todos los artículos. Para la visualización y el análisis se utilizó el *software* Tableau. Con el análisis estadístico descriptivo también se pudo obtener medias, tendencias y cambios porcentuales:

   a) Se pudo observar los datos relacionados con las instituciones. Por ejemplo, los que más gastos realizaron durante el pico de la pandemia en 2020, fueron el Gobierno Municipal de Guayaquil (US$19 millones) y el Gobierno Provincial del Guayas (US$5 millones). Seguidos por el Ministerio de Salud Pública (US$18 millones). Los hospitales que gastaron las sumas más altas también se ubicaron en Guayaquil, aproximadamente US$5 millones cada uno.

   b) Datos por producto o servicio, observándose los productos comprados con mayor frecuencia en contratos públicos. Estos fueron dispositivos médicos (ventiladores), suministros de laboratorio (*kits* de ensayo RT-qPCR), todo tipo de mascarillas faciales, medicamentos (paracetamol, piperacilina, hidroxicloroquina), desinfectante (amonio cuaternario), HRP y kits de alimentos.

   c) Los precios de los productos variaban según el contrato y el número de unidades adquiridas. En la Figura 5 se muestra la distribución de precios de todos los productos principales. Durante el período de estudio, se analizaron los precios por artículo, institución y mes. Encontramos una amplia gama de precios. Por ejemplo, algunos productos individuales tenían precios tan bajos como US$0.03 (pastillas de 500 mg de acetaminofén), mientras que otros, como los ventiladores mecánicos, alcanzaban los US$140,000. Para representar estas variaciones de precios en relación con la mediana/precio medio, utilizamos una escala logarítmica en la gráfica. Los puntos que se encuentran fuera de las barras y cajas representan valores extremos y atípicos en la distribución de precios.

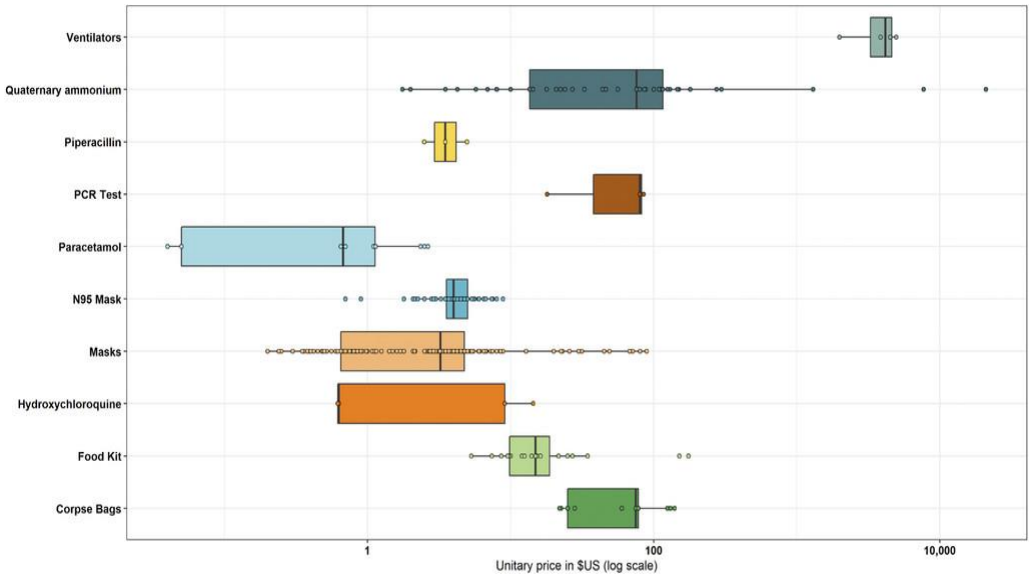


Figura 5: Distribución de precios de insumos básicos relacionados con la respuesta al COVID-19 en Ecuador (1 de marzo 31 de julio de 2020)

d) Datos por fechas, en cuanto al comportamiento de los precios se comprobó que los precios de los *kits* de alimentos e insumos médicos como las mascarillas experimentaron fuertes aumentos durante los primeros meses de la pandemia (Figura 6). Si bien los precios de las mascarillas oscilaron mucho desde al menos 2019, es evidente el aumento en abril y mayo de 2020.

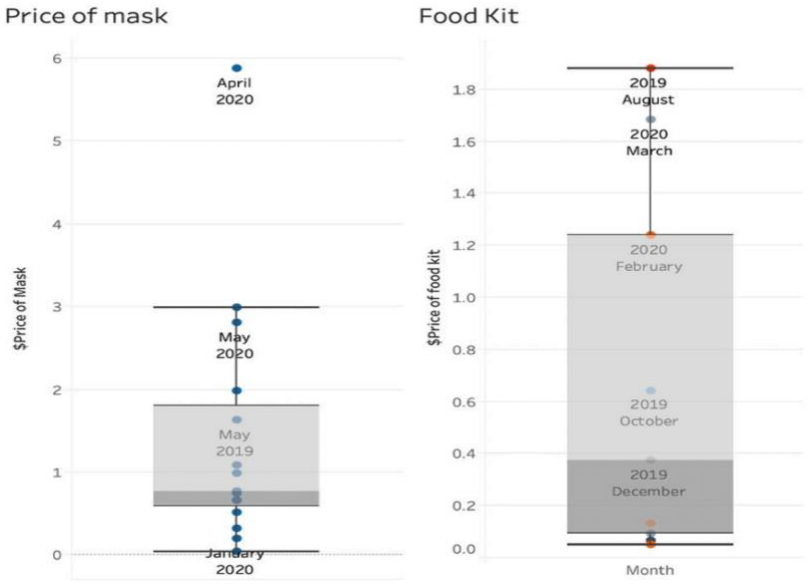


Figura 6: Distribución de precios de mascarillas y kits de alimentos en Ecuador por mes (1 de agosto de 2019 – 31 de mayo de 2020).

El estudio encontró diferencias significativas en los precios de los productos y servicios adquiridos por diferentes instituciones gubernamentales en Ecuador durante la pandemia de COVID-19. Confirmando que la Fase 1 de la metodología es efectiva y permite controlar el comportamiento de

otros actores involucrados en el proceso como las instituciones, productos o servicios, precios y fechas.

2) **En la fase de parámetros de calificación**, se detectaron requisitos extras solicitados que podrían indicar posibles sesgos, así como la exclusión de la oferta económica en ciertos procesos o la dirección de procesos de consultoría hacia proveedores específicos.

Siguiendo las dos etapas principales que componen la Fase 2 (ver Figura 4) de la metodología desarrollada, se inicia la primera con la identificación de contratos con anomalías, utilizando aprendizaje no supervisado con el Algoritmo K-Medias. Realizada la validación interna de los grupos, se obtienen los resultados esperados:

a) Los principales parámetros que tienen la mayor influencia en la selección del ganador del proceso. Los mismos se obtienen a través de un aprendizaje no supervisado, con el uso del SOM (Figura 7). Se identifican los principales parámetros que intervienen en la calificación del proceso y su influencia en la clasificación del clúster. Se pudo observar la influencia de cada parámetro de calificación en el grupo, siendo los de color azul, los que tienen la menor influencia y la escala de otros colores los que representan la mayor influencia. Por lo tanto, los principales parámetros de calificación encontrados al utilizar SOM son: oferta económica, cumplimiento de especificaciones, otros parámetros de calificación, experiencia general, experiencia específica, equipo propuesto, garantía técnica, instrumentos y equipos y obras similares.

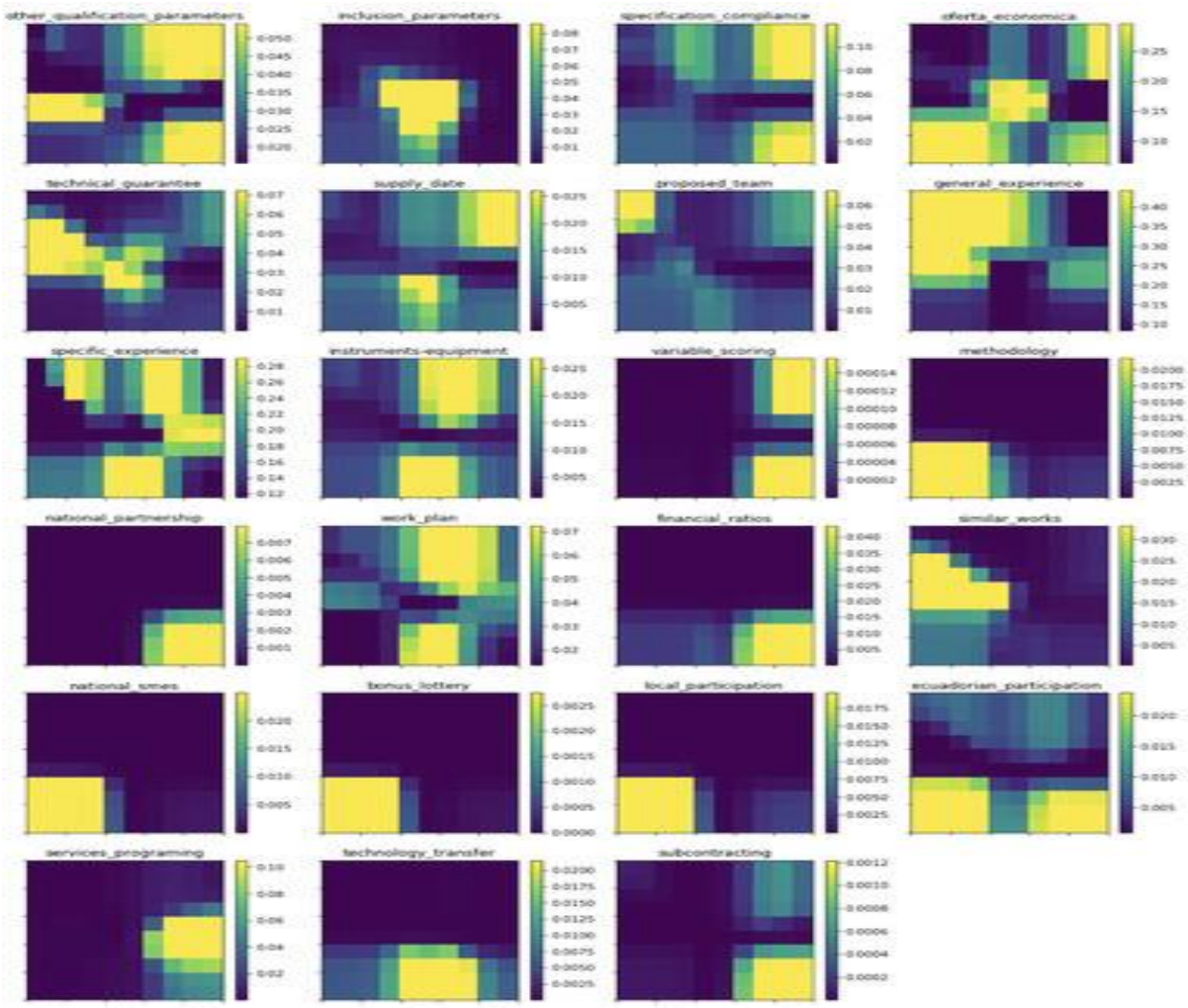


Figura 7: Evaluación de la influencia de los parámetros de calificación con SOM

Para evaluar la fiabilidad de los resultados de la aplicación del SOM, en la Figura 8 se muestra la evolución del error de cuantización y el error topográfico, con 1000 iteraciones. Observándose que a partir de la iteración 600 el mapa se estabiliza y alcanza valores óptimos para el modelo, obteniendo un error de cuantificación de 0,2878 y un error topográfico de 0,30796, asegurando así una correcta fiabilidad de los mapas.

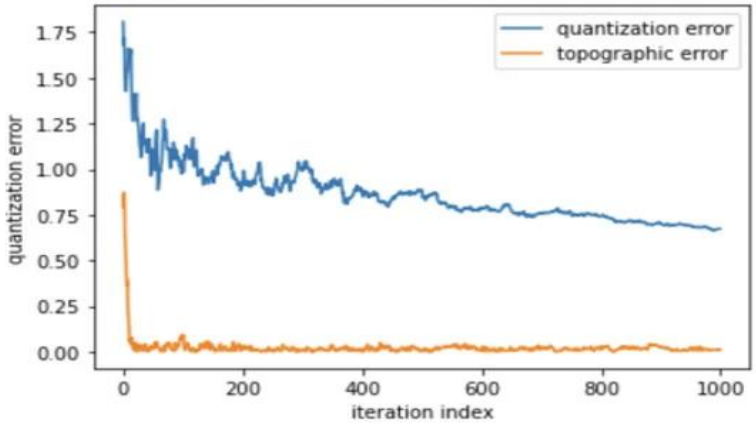


Figura 8: se muestra la evolución del error de cuantización y el error topográfico

b) Con la identificación de los grupos y la obtención la influencia de cada parámetro en la calificación se requiere para la segunda fase del modelo, la detección de anomalías con el uso de SVM y PCA. Posteriormente, se realiza un entrenamiento con las métricas descritas en la metodología para evitar sobre entrenamiento. El modelo se evalúa con datos que no están presentes en el entrenamiento (en este caso, datos de 2021). Obteniéndose una precisión del modelo del 95% y una exactitud del 92%. Además, se detectan cuatro grupos. En tres de ellos se espera que la oferta económica determine el ganador del proceso y en uno no (Clúster 4). Pudiéndose identificar: contratos regulares y contratos con anomalías. Para una mejor comprensión, los clústeres se nombran en función de la influencia de la oferta económica: Clúster 1 = Oferta económica moderada, Clúster 2 = Oferta económica alta, Clúster 3 = Oferta económica baja, Clúster 4 = Oferta económica nula.

Figura 9: Relación de la oferta económica con otros parámetros de calificación.

La Figura 9, muestra la influencia de los seis principales parámetros de calificación, que están relacionados con la oferta económica. Se puede ver gráficamente, la nula participación de la Oferta Económica en el Clúster 4, una participación moderada en el Clúster 1, participación alta en el Clúster 2 y participación débil en el Clúster 3. El algoritmo K-Means permitió identificar los tres grupos (clústeres) en los que el parámetro de calificación "Oferta Económica" predominaba entre los parámetros de puntuación. Pero también detectó, un grupo en el que la "Oferta Económica" era menor al 1% ("nula") de la puntuación total, sobre el 100% de la calificación. Evaluando "otros parámetros" con el mayor peso.

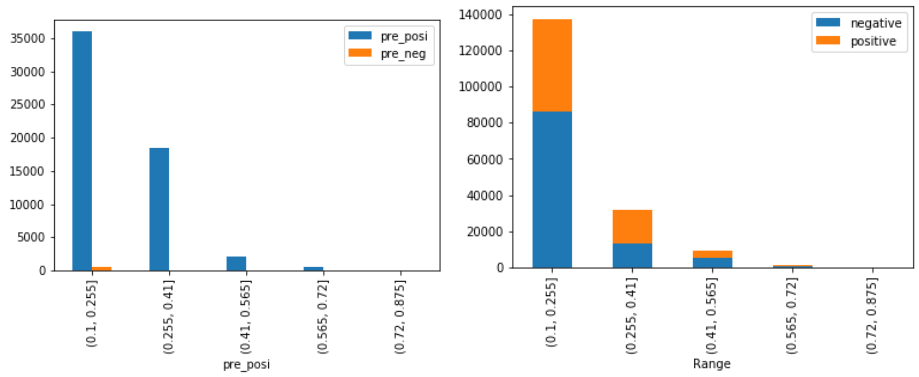
Lo que nos confirma que la Fase 2 de la metodología, permite detectar si se piden requisitos extras, que podrían indicar posibles sesgos, así como la exclusión de la oferta económica en ciertos procesos o la dirección de procesos de consultoría hacia proveedores específicos. Además, la identificación de los tipos de contratación predominantes, como: contratación directa, preselección y publicaciones especiales. Así como el tipo de compra. Para el caso específico analizado, la mayoría de las compras de este Clúster 4 son "Consultorías". Por lo tanto, 88.358 de los procesos evaluados presentes en el *clúster* "Oferta Económica Nula" podrían presentar anomalías en los parámetros de evaluación para la adjudicación de contratos. Lo que es equivalente al 32,11% del total analizado.

Se detectaron requisitos extras solicitados que podrían indicar posibles sesgos, así como la exclusión de la oferta económica en ciertos procesos o la dirección de procesos de consultoría hacia proveedores específicos.

3) **En la fase de preguntas y aclaraciones**, se encontraron evidencias de sesgo de género en las respuestas de la entidad contratante, así como indicios de favoritismo en dichas respuestas. Además, se identificaron cambios en los términos de los procesos de adquisición, que parecían estar direccionados hacia ciertos proveedores.

   Mientras que las Fase 1 y 2 de la metodología propuesta se centran en la identificación de sobreprecios y la detección de anomalías en los procesos de contratación pública, respectivamente, la Fase 3 se enfoca en la detección de procesos sospechosos de favoritismo y sesgo de género utilizando técnicas de PLN. Al abordar este aspecto específico, se enriquece la metodología general propuesta en la tesis doctoral, brindando un enfoque más completo con el empleo de técnicas de PLN. Se trata de identificar patrones y características lingüísticas en el texto generado, que podrían indicar posibles irregularidades o comportamientos sospechosos. Así como detectar compras sesgadas utilizando un corpus en idioma español, analizando el texto de la plataforma de registro de preguntas y respuestas de los solicitantes en un proceso de contratación pública en Ecuador. Se detectan sesgos de género, partiendo del hecho que tanto hombres como mujeres puedan participar en las mismas condiciones. Para detectar sesgos de género y favoritismo hacia determinados proveedores por parte de las entidades contratantes, se aplica un modelo híbrido, que combina algoritmos de Inteligencia Artificial y PLN. Para evaluar el sesgo de género y el favoritismo se utilizó el modelo Word2-Vec con incrustación de palabras.

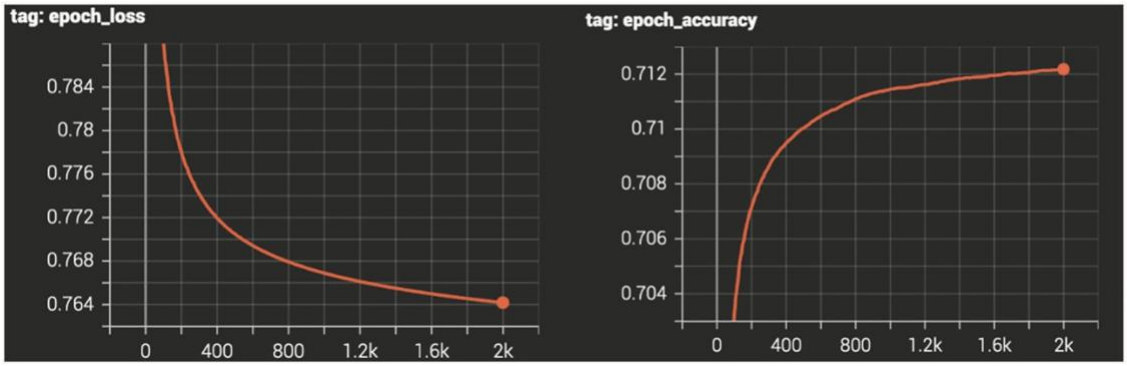
   Para el análisis de sentimiento de las preguntas y respuestas el algoritmo VADER como se observa en la Figura 10. En el lado izquierdo de la Figura 10, las escalas de evaluación de sentimientos se trazan en un intervalo de 0,15 (eje X), mientras que el número de preguntas se traza en un eje Y. En las preguntas realizadas por los proveedores a las entidades contratantes encontramos un mayor porcentaje de opiniones positivas. En el lado derecho de la Fig. 10, también se muestra la clasificación del análisis de sentimiento, en las respuestas de las instituciones públicas, resaltando solo las positivas y negativas (descartando las neutras). Se evidencia un mayor número de respuestas negativas (105.547) frente a 73.665 respuestas positivas, siendo esto un indicio de un posible sesgo. También se evaluó el sentimiento negativo en las respuestas de las instituciones públicas, destacando que la mayoría se encuentran en el rango de 0.1 a 0.41. Siendo estas respuestas “parcialmente negativas”, sin embargo, es importante resaltar la cantidad de respuestas con un alto porcentaje de negativismo, que son aquellas con puntuaciones superiores a 0,41 a 0,875.



La Figura 10: Distribución de preguntas y respuestas

Los resultados del análisis mostraron que en el 32% de los casos, hubo favoritismo o sesgo de género. El modelo propuesto proporciona índices de precisión del 88% para detectar favoritismos y del 90% para detectar sesgos de género. En consecuencia, un tercio de los procesos de contratación llevados a cabo por el Estado de Ecuador tienen indicios de corrupción y parcialidad.

Para seleccionar el modelo adecuado en las incrustaciones de palabras, se evaluaron los algoritmos Word2-Vec y FastText, utilizando métricas como la curva ROC, *accuracy* y la función de pérdida, para medir los porcentajes de precisión de todos los modelos. La Figura 11 muestra los resultados de la evaluación para 2000 iteraciones de uno de los modelos entrenados (FastText). En el lado izquierdo donde se muestra la Especificidad (0.71) y en el lado derecho la función de pérdida (0.78). Se puede observar que el modelo evaluado no puede mejorar significativamente sus resultados, justificando así la elección de Word2Vec en lugar de FastText.



La Figura 11: Curva ROC con 2000 interacciones, precisión y pérdida del algoritmo FastText

En la fase de preguntas y aclaraciones, se encontraron evidencias de sesgo de género en las respuestas de la entidad contratante, así como indicios de favoritismo en dichas respuestas. Además, se identificaron cambios en los términos de los procesos de adquisición que parecían estar direccionados hacia ciertos proveedores.

4.3. Conclusiones

Con la tesis doctoral se han podido comprobar las diferentes fases de la metodología propuesta para identificar procesos con anomalías y generar el modelo de detección de corrupción. Lo que permitió validar la hipótesis planteada: a través de la evaluación de los datos y el texto generado en las fases de un proceso de compra pública, mediante la combinación de técnicas de minería de datos y PLN, se puede identificar la corrupción, con resultados muy alentadores para su detección y prevención en compras del sector público. Además, en esta investigación se han alcanzado otros objetivos propuestos:

- Al analizar los diferentes métodos para la detección de corrupción en compras públicas, se han identificado las variables que influyen en los casos de corrupción, lo que permite enfocar el análisis y detección con mayor precisión en un sistema más completo.

- Con el estudio comparativo de diferentes enfoques de análisis de datos se llegaron a seleccionar para el estudio técnicas estadísticas y de aprendizaje automático, que incluyeron el aprendizaje supervisado, el aprendizaje no supervisado y el PLN. Contribuyendo con la eficiencia y precisión de los algoritmos utilizados en la detección de corrupción.

- Con la recuperación, ordenamiento y limpieza de los datos del portal SERCOP, en el periodo de tiempo específico (2010 hasta 2020) se detecta que entre un 30 y 35 % de procesos de compras públicas tienen indicios de corrupción, lo que constituye una base sólida para el análisis y detección de corrupción en compras públicas.

- Al seleccionar y aplicar métricas adecuadas de evaluación, para los diferentes algoritmos utilizados, se logran establecer comparativas de rendimiento, lo que hace posible elegir los algoritmos más adecuados en la clasificación de datos. Los desarrollados en esta investigación demostraron ser eficaces en la clasificación y detección de casos de corrupción.

- El empleo de PLN brinda una visión adicional en la detección de comportamientos irregulares en los procesos de compras públicas, al realizar un análisis de sentimiento de las preguntas y respuestas proporcionadas por las entidades adjudicadoras, que permite detectar sesgos y favoritismo hacia los proveedores.

- Se demuestra que la metodología científica basada en MD, técnicas de aprendizaje automático y PLN desarrollada en la investigación es efectiva, pues con la ejecución de sus fases se detecta de forma automática la corrupción y se ofrecen indicios de corrupción en un porcentaje significativo de procesos de compras públicas. Su implementación puede ayudar a optimizar los procesos de compra, lo que aumenta la eficiencia y la eficacia en el empleo del gasto público.

  Este trabajo representa un gran avance en la investigación de la corrupción con herramientas tecnológicas en América Latina porque como ya se ha comentado, no se emplean este tipo de técnicas para combatir la corrupción.

4.4. Trabajos futuros

Como línea de trabajo futura, se propone la construcción de una herramienta de alerta temprana que permita evaluar, detectar y predecir el favoritismo en los procesos de compras públicas, incluso antes de que el proceso sea publicado por la entidad contratante o cuando se realicen preguntas y aclaraciones por parte de los contratistas. La información obtenida en las contribuciones, que incluye datos como: la institución gubernamental responsable del proceso de compra, la ubicación del contrato, el tipo de producto o servicio, el precio unitario y el tipo de contrato, resultará valiosa para este trabajo futuro.

Además, se propone la integración del aprendizaje profundo (*Deep learning*) a la metodología actual. El uso de técnicas de aprendizaje profundo, como las CNN, permitirá capturar patrones y características más complejas en los datos, lo que puede mejorar la precisión y eficiencia en la detección de favoritismo en los procesos de compras públicas.

Otro enfoque prometedor es la utilización de técnicas basadas en grafos u otras técnicas de representación y análisis de relaciones. Esto permitirá explorar las interacciones entre los contratistas y las entidades, y analizar de manera más completa las posibles relaciones asociadas a la corrupción. El uso de grafos puede proporcionar una visión más amplia y estructurada de las conexiones y vínculos entre los diferentes actores involucrados en los procesos de compras públicas. Lo que ayudará a identificar patrones y comportamientos sospechosos de manera más efectiva.

4.5. Contribución 1

4.5.1. Titulo

Ortiz-Prado, Esteban; Fernandez-Naranjo, Raul; Torres-Berru, Yeferson; Lowe, Rachel; Torres, Irene; “Exceptional Prices of Medical and Other Supplies during the COVID-19 Pandemic in Ecuador”. *American Journal of Tropical Medicine and Hygiene*, 105 (1). pp. 81-87, 2021.

4.5.2. Objetivos

En Ecuador durante la pandemia de COVID 2019, en redes sociales y diferentes medios de comunicación, se evidenciaron denuncias de corrupción asociadas a la compra de insumos para la lucha contra el coronavirus, muchas de estas denuncias incluso derivaron en casos judiciales. Constatando que el organismo encargado de supervisar las compras públicas (SERCOP), no había realizado el debido control respecto a estos hechos. Por ejemplo, se dan algunas situaciones típicas que evidencian estos hechos: no se encuentra explicación de por qué un proveedor vende al hospital *X* un determinado producto a $10.00 y al Instituto Ecuatoriano de Seguridad Social a $50.00. En otros casos se lanzan unos TDR dirigidos a un proveedor específico, el resto de los proveedores hace

diferentes ofertas y el único que cumple los TDR accede a vender esos productos al estado. Incluso se evidencian compras de productos por parte de instituciones a pesar de que sus competencias por ley no permiten realizan compras de medicamentos.

Con esos antecedentes la investigación tiene como objetivo comparar la variabilidad de los precios e identificar sobreprecios de los productos adquiridos a través de contratos públicos, para su uso en entornos clínicos de la población en general en respuesta a la pandemia.

4.5.3. Metodología

La metodología planteada, parte del empleo de técnicas estadísticas para comparar los precios de compras similares realizadas por diferentes entidades y detectar posibles sobreprecios. Se realizó un análisis estadístico descriptivo para extraer medidas relevantes para los productos de compra común en emergencias de salud, como dispositivos médicos, medicamentos farmacéuticos y otros bienes. Los datos se obtuvieron en el período comprendido entre el 1 de marzo y el 31 de julio de 2020, de la base de datos del SERCOP. Para su adquisición se empleó la técnica de *web Scraping (*Selenium*)*, debido a que en ese período no se contaba con un portal de datos abiertos. Los datos se almacenaron en una base de datos no relacional *Mongodb*. Para la implementación se utilizó el lenguaje Python. El código fuente está disponible en el siguiente enlace: https://github.com/torresyeferson/DattaMinigCorruption/blob/main/Scraping.py. Para el análisis se cumplen los siguientes pasos:

(i) Se seleccionan las variables a analizar, como: institución gubernamental responsable del proceso (Ministerio, Gobierno Central, Gobiernos Autónomos Descentralizados), lugar donde radicaba el contrato (municipio y provincia), tipo de producto o servicio, precio unitario y tipo de contrato (contrato directo o subasta pública).

(ii) Mediante coincidencia aproximada (*Fuzzy Matching*) un algoritmo para la concordancia aproximada de secuencias y expresiones regulares en cadenas de texto, se identifican los productos más frecuentes en las licitaciones y contratos durante el periodo evaluado.

(iii) Para la visualización y el análisis de resultados se utilizó el programa Tableau[3], el cual permite encontrar las tendencias en precios en productos relacionados.

(iv) Posteriormente se aplica la técnica de regresión lineal, comparando los precios actuales con los precios de los productos en años anteriores.

4.5.4. Resultados

Se pudo observar que, los productos adquiridos con más frecuencia eran: las mascarillas, paracetamol, *kits* de detección de COVID, desinfectante (amonio cuaternario) y *kits* de alimentos. Además, que los precios variaron mucho dependiendo de cada contrato individual y del número de unidades adquiridas.

[3] https://www.tableau.com/

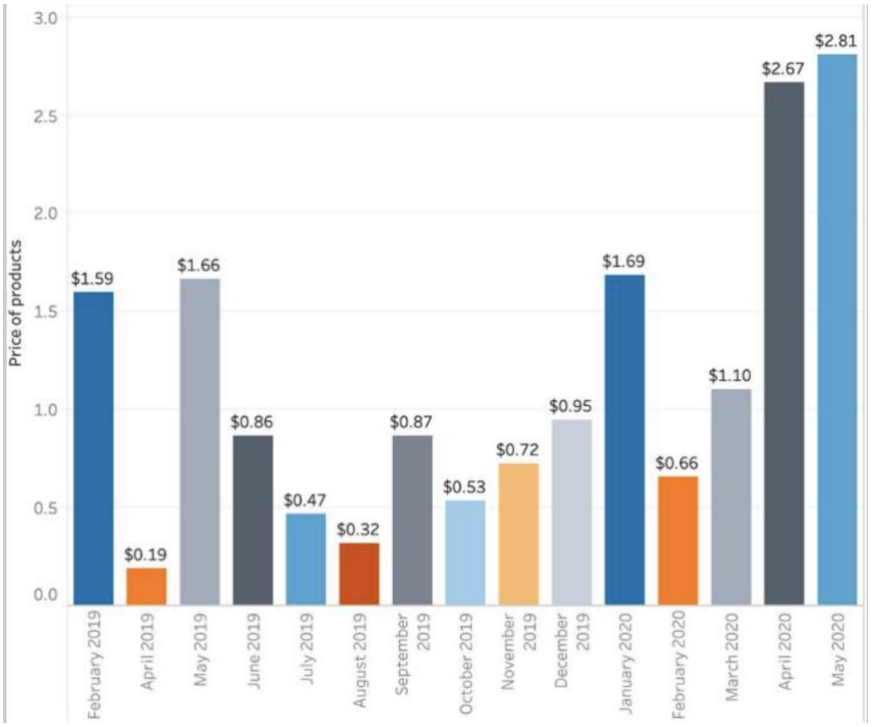


Figura 12: Precio mensual de las mascarillas faciales en los contratos públicos (no había información para enero y marzo de 2019)

La Figura 12, tomada del artículo *¨Exceptional Prices of Medical and Other Supplies during the COVID-19 Pandemic in Ecuador¨*, representa la variabilidad de precios de los suministros médicos y otros productos esenciales adquiridos a través de contratos gubernamentales en Ecuador durante la respuesta temprana a la pandemia de COVID-19. En ella se evidencia la diferencia porcentual entre el precio de adquisición y el precio de referencia, lo que indica la magnitud de la variabilidad de precios.

Algunos fueron excepcionalmente superiores a su valor de mercado en comparación con el año previo a la pandemia o con compras realizadas por otras instituciones públicas del mismo medicamento. Por ejemplo, en algunos casos, el precio medio de los guantes de examen médico aumentó hasta un 1.307%, el de las pastillas de paracetamol de 500 mg, hasta un 796% y el de los frascos de oxígeno, un 30,8%.

Otra compra excepcional detectada, fue la realizada por la Secretaría Nacional de Gestión de Riesgos que adquirió 7.000 *kits* de alimentos para familias vulnerables, a $150,82 por cesta. Posteriormente, la Contraloría General de Ecuador estableció que el valor de mercado de cada *kit* de alimentos era de $95,16.

4.5.5. Conclusiones

La pandemia de COVID-19 ha puesto de manifiesto las deficiencias del sector gubernamental en el control de los precios en Ecuador. Lo que trae como consecuencias que merme la eficacia de los recursos públicos con la especulación de los precios y que deriven en posibles actos de corrupción. El estudio identifica la variabilidad en los precios de los productos adquiridos a través de contratos

impulsados por el gobierno, y cómo la especulación y los aumentos de precios pueden haber influído negativamente el acceso a estos suministros

Mediante el trabajo se evidencia que algunos precios eran excepcionalmente más altos que su valor de mercado o del pagado por otro hospital de la misma red de salud. Pudiendo llegar hasta a un 600% de diferencia entre compras. Lo que evidencia que los sobreprecios, en procesos de compra pública, llegaron a alcanzar un total de 257 millones de dólares, afectando directamente al sector de la salud pública durante la pandemia.

Con el artículo se evidencia científicamente lo que se venía denunciando en las redes sociales y medios digitales: la corrupción asociada a los sobreprecios en la contratación pública, por las autoridades competentes y que se justificaba a través de compras especiales necesarias por la pandemia.

4.6. Contribución 2

4.6.1. Titulo

Torres-Berru, Y.; López Batista, V.F. “Data Mining to Identify Anomalies in Public Procurement Rating Parameters*”*. *Electronics*, 10, 2873, 2021.

4.6.2. Objetivos

La adjudicación de los procesos de contratación pública es una de las principales causas de corrupción en los gobiernos, debido a que en muchos casos los contratos se adjudican a proveedores previamente acordados (favoritismo). Para seleccionar el ganador de un contrato, los parámetros de calificación de la oferta ganadora juegan un papel fundamental, por lo que la manipulación de estos parámetros está directamente relacionada a la aplicación de prácticas corruptas, ocasionando que ganen oferentes con precios elevados, causando perjuicio al estado. El objetivo principal del trabajo es hacer un modelo multifase que identifique procesos con anomalías y posteriormente pueda detectar posible corrupción en la asignación de parámetros de calificación de procesos en la contratación pública. Aplicando técnicas de minería de datos, se podrá detectar la existencia de patrones anómalos, que faciliten en un futuro, encontrar la existencia de nuevos casos de corrupción, con el uso de técnicas de aprendizaje supervisado.

4.6.3. Metodología

Se evalúan 275.730 contratos públicos en el periodo 2010 - 2020, por lo que la propuesta del trabajo diseñada para contrastar la hipótesis basada en la metodología CRISP-DM consta de las siguientes etapas:

(i) Se procesa el conjunto de datos, eliminando los valores erróneos correspondientes a procesos con parámetros de calificación que tenían errores, ya que éstos deben sumar un valor igual al 100% y en algunos casos tenían valores inferiores, como el 98% o superiores, como el 105%.

(ii) Se evalúan los 21 parámetros numéricos para determinar el ofertante ganador de un proceso de compras públicas, los parámetros de calificación varían según el proceso y tienen impacto en la puntuación final.

(iii) Se emplean los Mapas Auto-organizados, los cuales permiten de manera gráfica determinar el nivel de influencia de los 21 parámetros de evaluación. Permitiendo encontrar los más influyentes. Para su implementación se codifica los mapas en Python usando las librerías *Mini-som* [4]*y Sompy*[5].

(iv) Se identifican grupos con parámetros de calificación relacionados, con el uso del algoritmo de agrupamiento K-Means. Posteriormente se usan métricas de evaluación como cohesión, *cluster acurracy* y *normalized mutual information*. Para la implementación se usa Python con la librería Sticky-Learn

(v) Con los procesos clasificados en grupos, se aplica Aprendizaje Semisupervisado con el empleo de SVM y PCA, para generar un modelo de detección de anomalías en los procesos de contratación, usando el servicio de AZURE para los conjuntos de datos de entrenamiento y prueba. Se consideran como métricas de evaluación: Curvas ROC, exactitud, precisión, F*1-score* y *Recall.*

4.6.4. Resultados

Con el trabajo experimental, se pudieron verificar las diferentes fases de la metodología propuesta para detectar procesos con anomalías y generar el modelo de detección de corrupción.

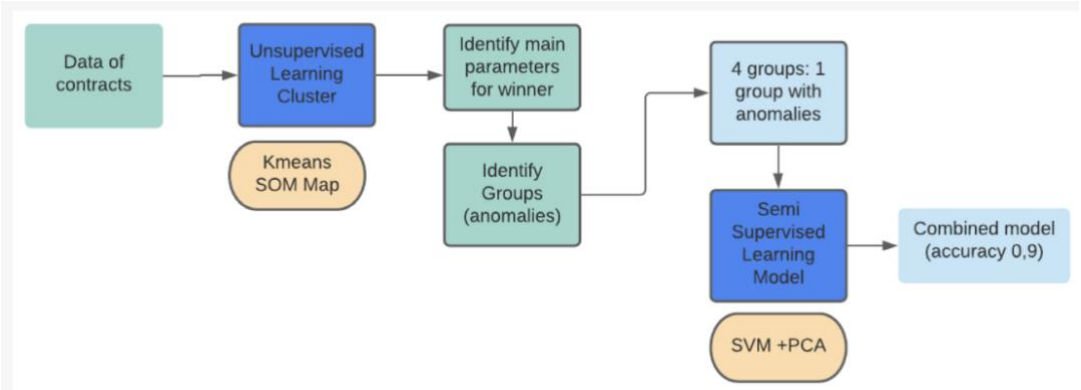


Figure 13: Diagrama de flujo de las dos principales fases que componen el modelo

La Figura 13, tomada del artículo *¨Data Mining to Identify Anomalies in Public Procurement Rating Parameters¨* muestra las dos fases principales que componen el modelo desarrollado para detectar posibles anomalías en los procesos de contratación pública en Ecuador. La primera fase implica la

[4] https://www.kaggle.com/code/garginirmal/selforganizingmaps-with-minisom
[5] https://anaconda.org/bioconda/sompy

identificación de contratos con anomalías utilizando aprendizaje no supervisado con el algoritmo K-Means, una vez que se ha logrado la validación interna del clúster. En esta fase, se identifican cuatro grupos, de los cuales en tres se espera que la oferta económica determine al ganador del proceso y en uno no.

La segunda fase implica la evaluación de los principales parámetros que tienen una mayor influencia en la determinación del ganador del proceso utilizando los mapas SOM. Como resultado de esta fase, se identifican dos tipos de contratos: contratos regulares y contratos con anomalías. Puede resumirse que la Figura 13, muestra el proceso de identificación de contratos con anomalías y la evaluación de los parámetros clave que influyen en la determinación del ganador del proceso en los procesos de contratación pública en Ecuador.

Con el algoritmo SOM se identifican los 12 principales parámetros que intervienen en la calificación de los adjudicatarios de la compra pública. El algoritmo K-Means permitió identificar los tres grupos (clústeres) en los que el parámetro de calificación “Oferta Económica” predomina entre los parámetros de puntuación. Además, un grupo en el que la "Oferta Económica" era menor al 1% (“nula”) de la puntuación total sobre el 100% de la calificación y que "otros parámetros" fueron evaluados con el mayor peso.

Estas evaluaciones predominan en los tipos de contratación: contratación directa, preselección y publicaciones especiales. En cuanto al tipo de compra, la mayoría de las compras de este clúster son "Consultorías". Por lo tanto 88.358 de los procesos evaluados presentes en el *clúster* “Oferta Económica Nula” podrían presentar anomalías en los parámetros de evaluación para la adjudicación de contratos. Lo que es equivalente al 32,11% del total analizado. A partir de los hallazgos (el *clúster* "Oferta Económica Nula") para la detección de anomalías, los procesos asociados a los otros tres *clústeres* se definen como "normales", donde se respeta el indicador económico como factor preponderante para la calificación y determinación del ganador.

Utilizando las anomalías identificadas, los investigadores construyeron un modelo de aprendizaje semisupervisado para la detección de anomalías (SVM y PCA), el cual logró una precisión del 95%, y una exactitud de un 92%. Este modelo puede ayudar a detectar procedimientos donde el objetivo es beneficiar a un proveedor en particular, manipulando los parámetros de asignación. Por lo que podemos verificar la hipótesis que guía esta investigación.

4.6.5. Conclusiones

La investigación propone una metodología para identificar procesos de contratación pública con anomalías y desarrollar un modelo de detección de corrupción basado en técnicas de aprendizaje automático. Se identificaron los parámetros principales en la calificación de los ganadores en una licitación pública utilizando el algoritmo SOM. El algoritmo K-Means, permitió identificar grupos donde la oferta económica no se consideró adecuadamente. A partir de estos hallazgos, se desarrolló

un modelo de detección de anomalías basado en SVM y PCA que demostró resultados confiables superiores al 90%.

Los resultados experimentales mostraron que, con el uso de técnicas de MD, este modelo también puede aplicarse en otros países para identificar procesos con anomalías en la calificación y ajustar el modelo en consecuencia. En general, la investigación muestra que la MD es una herramienta útil para identificar casos de corrupción en la contratación pública y ayudar a prevenir futuros casos de corrupción. El estudio representa un avance en la investigación de la corrupción en América Latina utilizando herramientas tecnológicas.

Como trabajos futuros se plantea comparar los procesos con anomalías, con los datos del organismo rector (SERCOP). Además, hacer pública la base de datos de procesos con anomalías, para poder utilizar otras técnicas de aprendizaje supervisado como Redes Convolucionales, *Random Forest* o utilizar técnicas combinadas.

4.7. Contribución 3

4.7.1. Titulo

Torres-Berru, Yeferson, López-Batista, Vivian F. and Conde Lorena. “A Data Mining Approach to Detecting Bias and Favoritism in Public Procurement”. *Intelligent Automation & Soft Computing* Volume 36, number 3, 2023.

4.7.2. Objetivos

Esta investigación se centra en el desarrollo de un modelo basado en IA, que utiliza técnicas de PLN para detectar procesos de compras públicas sospechosos en lengua española.
Tradicionalmente, la información generada por proveedores y entidades públicas, orientada a la identificación de actividades sospechosas de corrupción en un proceso de contratación pública, no se suele utilizar. Debido a que las denuncias generadas por los proveedores o las respuestas sospechosas realizadas por las instituciones financieras no son revisadas ni tomadas en cuenta por ningún ente de control. Esto genera malestar por parte de los proveedores que se sienten perjudicados por el Estado, creyendo que existe una predilección (favoritismo) hacia un determinado proveedor o grupo de proveedores.

El objetivo de la investigación es desarrollar un modelo de aprendizaje automático que reconozca los sesgos de género y el favoritismo en los procesos de contratación pública para hispano hablantes, basándonos en una evaluación de igualdad de condiciones, tanto en género como en oportunidades para el participante.

4.7.3. Metodología

Este modelo complementa y valida la metodología de detección de corrupción en compras públicas propuesta por los autores en trabajos anteriores, puesto que analiza el texto generado en la fase de contratación por las entidades contratantes y los licitantes en el proceso [65] lo que determina una nueva fase dentro del modelo planteado. Esta contribución es referencia y un punto de partida para completar la investigación. Complementa las contribuciones anteriores; mientras que las Contribuciones 1 y 2 se centran en la identificación de sobreprecios y la detección de anomalías en los procesos de contratación pública, respectivamente, la Contribución 3 se enfoca en la detección de procesos sospechosos de favoritismo y sesgo de género utilizando técnicas de PLN. Al abordar este aspecto específico, se enriquece la metodología general propuesta en la tesis doctoral, brindando un enfoque más completo para prevenir y detectar la corrupción. Con el empleo de técnicas de PLN se identifica patrones y características lingüísticas que podrían indicar posibles irregularidades o comportamientos sospechosos.

Es importante resaltar la potencial aplicabilidad de los resultados y la metodología propuesta en contextos y países hispanohablantes, ya que podrían ser aplicados en entornos donde el español es el idioma predominante. Esto permitiría adaptar y utilizar la metodología propuesta en países hispanohablantes, contribuiría así a la lucha contra la corrupción en dichos contextos.

En esta contribución se evalúa el favoritismo siguiendo las siguientes frases:

(i) Se consideran los siguientes indicadores: la fecha de publicación de la convocatoria de propuestas de los proveedores, las decisiones de selección y contratación, los precios, el número de ofertas presentadas, el tipo de responsabilidades de las instituciones y el origen de los licitadores.

(ii) Se evalúan 303076 procesos de compras públicas ejecutados desde el año 2010 hasta el 2020, dentro de cada proceso se realizan preguntas o aclaraciones/modificaciones que son respondidas o emitidas por la entidad contratante.

(iii) En total se evalúan 1009739 preguntas y respuestas con las siguientes características:
- Pregunta emitida por el contratante, donde se consultan aspectos necesarios para la licitación o se pretende informar de alguna novedad en el proceso.
- Respuesta emitida por la entidad que realiza la contratación, sirve para aclarar las dudas de los contratantes. Además, se puede usar para emitir aclaraciones o modificaciones del contrato. Se entiende que las respuestas y aclaraciones deben ser parciales y poner la referencia. Por lo que no deben tener sesgo ni sentimientos a favor o en contra del proveedor.

(iv) Los textos están escritos en español y dentro de cada proceso de compra se añade información adicional, como: código, cantón, monto, fecha, hora, dirección web del proceso, estado del proceso, forma de pago, tipo de compra, tipo de contratación.

4.7.4. Resultados

Se evaluaron técnicas como *Word2Vec*, para crear los *word embedding*, así como *FastText*, que es un modelo de *Gensim* que sirve además para hacer clasificación. Ambas herramientas se utilizaron para para analizar sesgo de género (en preguntas y respuestas) y hacia otros proveedores.

Figura 14: Nube de palabras de las preguntas de los proveedores

La Figura 14, tomada del artículo *¨A Data Mining Approach to Detecting Bias and Favoritism in PublicProcurement¨*, muestra la nube de palabras correspondiente a los términos más utilizados para la elaboración de preguntas por parte de los proveedores. Palabras como: tarde, sí, confirmar, validar, dinero (plata en español coloquial), etc. Con este resultado reflejado en la Figura 14, se pretende tener una idea de la terminología utilizada por los proveedores.

Se constata que en las preguntas de los proveedores se menciona frecuentemente que existen procesos que están direccionados o limitados hacia ciertos contratistas o profesionales; por ejemplo, en una de las nubes de palabras, que se utilizan para hacer el análisis, la palabra *direccionándose* se relaciona con: *existen, defectos, contratista, concuerda, pliego*. Cabe destacar que se denomina "pliego" a los términos de condiciones realizados por la entidad contratante, por lo que se entiende que los proveedores en sus preguntas la mencionan directamente, cuando existe favoritismo por parte de la institución pública.

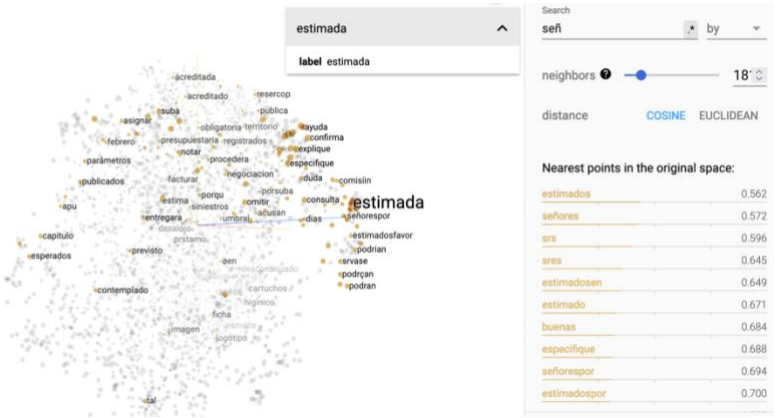


Figura 15: Nube de palabras de las respuestas de la entidad pública

La nube de palabras correspondiente a la terminología frecuentemente utilizada para preparar respuestas (Figura 15), por parte de las entidades contratantes, muestra el sesgo de género que se observa al predominar palabras orientada al género masculino como: *señores, estimados, etc.*, respecto a las palabras equivalentes en el género femenino o en un lenguaje neutro. De esto se deriva que las entidades contratistas como las instituciones públicas asumen que la persona con la que están interactuando son hombres.
En las respuestas utilizadas por las entidades contratantes, aunque se espera que sean imparciales, hay términos sesgados que aparecen con frecuencia, como:
- "Experiencia": respuestas orientadas a un proveedor con el que la institución ha trabajado previamente.
- "Referencia": respuestas orientadas a que el producto cumpla con ciertas referencias a un modelo o marca en particular.
- "Señores": respuestas dirigidas a una audiencia masculina.

El modelo utiliza algoritmos como Word2Vec y FastText para analizar el sesgo de género en las preguntas y respuestas. Además, se consideran los términos sesgados que se mencionan: "experiencia", "referencia" y "señores". Se realizó un análisis de sentimientos de las preguntas y respuestas proporcionadas por las entidades contratantes. Los resultados mostraron un nivel de precisión del 88% en la detección de favoritismo entre contratistas, lo que indica que este análisis cumplió su objetivo de complementar el modelo y mejorar la precisión en el análisis de las preguntas y respuestas.

Se observa una tendencia evidenciada en trabajos previos de la tesis doctoral, que existen entre un 30% y un 35% de los procesos de contratación pública con posibles anomalías, en los que se pretende favorecer a un determinado contratista, mediante favoritismo utilizando las aclaraciones y las modificaciones de los términos de contratación.

4.7.5. Conclusiones

La investigación demuestra que el análisis de las preguntas y respuestas en un corpus de texto de contratación pública en español permite detectar compras sesgadas por parte de las entidades contratantes. Mediante un modelo de detección de sesgos de género, basado en Word2vec, se alcanza una precisión del 90% en sesgo de género, mientras que en la evaluación del modelo de sesgo entre contratistas (favoritismo) se obtiene una precisión del 88%. El modelo muestra que un tercio de los procesos de contratación llevados a cabo por el estado tienen indicios de corrupción y sesgo, lo que resulta en perjuicios para el estado y discriminación contra los proveedores.

El análisis de texto es una herramienta valiosa para detectar este tipo de problemas, ya que los documentos relacionados con los procesos de contratación contienen mucha información que puede ser analizada, por lo tanto, el trabajo contribuye a que las contrataciones públicas sean dirigidas a adquirir bienes con los criterios de inclusión, de igualdad y de equidad social.

4.8. Contribución 4

4.8.1. Titulo

Torres Berrú, Y., V. Batista, and Pablo Torres-Carrión. "Data mining to detect and prevent corruption in contracts: systematic mapping review." *RISTI-Revista Iberica de Sistemas e Tecnologias de Informacao* (2020): pp 13-25.

4.8.2. Objetivos

Previo al inicio de un trabajo doctoral es importante tener en cuenta el nicho de investigación sobre el que se va a trabajar. Es necesario hacer una investigación exploratoria de la bibliografía existente en las diferentes bases de datos científicas, que nos permita conformar el marco teórico, para desarrollar la investigación. En esta primera etapa, se encontraron revisiones sistemáticas de literatura científica que relacionan la lucha contra la corrupción mediante el uso de tecnología y se enfocan en encontrar algoritmos de detección de fraude en transacciones en línea. Así como en las diferentes técnicas de aprendizaje automático, que se emplean para la detección de fraude en los diferentes campos del ámbito financiero. Dichos trabajos han aportado parte del aval necesario para guiar esta tesis doctoral, pues se detectó que era necesario abordar la corrupción desde una perspectiva más amplia, debido a que el tipo de corrupción más estudiado era el conocido como fraude. Destacándose la necesidad de investigar en otras formas de corrupción.

El objetivo del estudio fue identificar de manera general, el estado de la investigación relacionada con la minería de datos en la detección y prevención de la corrupción. Más concretamente, conocer el número de artículos publicados, los autores, las revistas en las que se publica, los países donde se investiga y las temáticas relacionadas.

4.8.3. Metodología

Con base a lo expuesto, el presente trabajo analiza la información científica publicada desde el 2015 hasta el 2019, aplicando la metodología de Torres-Carrión [66]para la búsqueda sistemática de la bibliografía y García-González [67]para la organización del mapeo. Se plantearon seis preguntas de investigación a las que se da respuesta desde el análisis de 147 artículos obtenidos de las bases de datos Web of Science *(WoS) y Scopus*.
Las preguntas analizadas son las siguientes:

- RQ1: ¿Cuántos estudios hay en las bases de datos WOS y Scopus entre 2015 a 2019?
- RQ2: ¿Quiénes son los autores de los artículos más citados?
- RQ3: ¿Cuál es la distribución geográfica de los autores?
- RQ4: ¿Cuáles son las revistas con más publicaciones sobre esta línea de investigación?
- RQ5: ¿En qué contextos se desarrollan estos estudios?
- RQ6: ¿Cuáles son los principales tipos de corrupción en esta línea de investigación?

4.8.4. Resultados

Se describen los resultados por cada una de las preguntas planteadas para el mapeo sistemático de la literatura:

1. Existen 147 artículos publicados en el periodo de evaluación, 61 en *WoS* y 86 en *Scopus*, de los cuales solamente el 16% son artículos de acceso abierto.
2. El estudio más citado (60 citas) se refiere a la detección de fraude bancario utilizando minería de texto, y hay otros estudios relacionados con corrupción en el sector bancario con citas significativas superiores a 30.
3. Se pudo evidenciar que Estados Unidos (16,32%), China (10,88%) y Reino Unido (8,94%) son los países con mayores publicaciones sobre MD y corrupción. En Latinoamérica sobresale Brasil (3,4%), y aportes mínimos desde Colombia y Paraguay.
4. Se han considerado todas las revistas científicas con al menos dos artículos y se han organizado por cuartil, según *Scopus*. *Expert Systems with Applications* es la revista con el mayor número de artículos.
5. Según la tipología de corrupción, la mayor cantidad de artículos corresponden a fraude (72.72%), sobreprecio (9,09%) y malversación (6,8%).
6. Los principales temas referentes a corrupción han sido detección de fraude (14,28%), fraude financiero (5,61%), corrupción de cualquier tipo (4,64%) y fraude de tarjetas de crédito (4,64%).

4.8.5. Conclusiones

El texto indica que Estados Unidos, China y el Reino Unido tienen la mayor cantidad de publicaciones sobre MD y corrupción, pero China y Bélgica son los países más citados. En América Latina, Brasil es el país con la mayor cantidad de publicaciones, pero se considera un área de investigación limitada. En términos de publicaciones y citas, el fraude es el tipo de corrupción que más interesa, no se consideran por lo tanto otros tipos como el favoritismo. En cuanto a tipo de técnicas empleadas, se están utilizando herramientas emergentes como la MD, el aprendizaje automático y el agrupamiento en este campo. Por último, se señala que se necesitan futuras investigaciones sobre otros tipos de corrupción y sus mecanismos, complementados con la MD.

4.9. Contribución 5

4.9.1. Titulo

Torres Berru, Y., López Batista, V.F., Torres-Carrión, P., Jimenez, M.G. (2020). “Artificial Intelligence Techniques to Detect and Prevent Corruption in Procurement: A Systematic Literature Review” . In: Botto-Tobar, M., Zambrano Vizuete, M., Torres-Carrión, P., Montes León, S., Pizarro Vásquez, G., Durakovic, B. (eds) *Applied Technologies. ICAT 2019. Communications in Computer and Information Science*, vol 1194. Springer, Cham

4.9.2. Objetivos

Esta revisión sistemática, pretende hacer un análisis más detallado sobre la corrupción. Aborda los estudios sobre los distintos tipos de fraude, a nivel privado y gubernamental. También profundizar en las diferentes técnicas de aprendizaje automático, minería de datos, en el proceso de detección y prevención de fraude que se están utilizando. Analizar los diferentes tipos de corrupción en la contratación pública, explorar el uso de técnicas de IA para detectar y prevenir la corrupción. Así como, identificar los métodos de investigación más comúnmente utilizados, las características de las organizaciones estudiadas y las herramientas tecnológicas y metodologías de minería de datos utilizadas.

El objetivo del estudio se centra en resolver las siguientes preguntas de investigación:
RQ1: ¿Que métodos se están aplicando para investigar sobre corrupción en contratos relacionados a compras públicas?
RQ2: ¿Características de las organizaciones en las que se ha desarrollado la investigación?
RQ3: ¿Qué herramientas tecnológicas se están usando para investigar sobre detección y prevención de la corrupción?
RQ3: ¿Qué algoritmos, metodologías y herramientas de análisis de datos se usan para detección de corrupción?

4.9.3. Metodología

En el artículo también se sigue la metodología de Torres-Carrión, que involucró tres fases: la planificación, la realización de la revisión y la presentación de la revisión utilizando el método PRISMA. La fase de planificación implicó definir criterios generales y específicos de inclusión y exclusión para el proceso de búsqueda. La fase de revisión implicó buscar artículos en las bases de datos *Scopus* y WoS para analizar los datos utilizando técnicas de minería de datos. La fase de presentación implica presentar los hallazgos del estudio de manera clara y concisa.

4.9.4. Resultados

Se pudo observar que la investigación cuantitativa experimental es la más usada, así como el método estadístico de correlación. La mayor parte del estudio se realiza con datos que abarcan un periodo entre 1-3 años, con principal aporte en las ciencias de la computación. La actividad comercial principal de las organizaciones es la prestación de servicios, específicamente el sector bancario. En el ámbito gubernamental destaca el fraude de impuestos y con presencia en menor grado de procesos de compras públicas.

Además, que web *scraping* es un método poco usado para la obtención de datos en estudios de corrupción en contratos, pudiendo ser usado como base para futuros trabajos. Se evidencia que los pocos trabajos relacionados al análisis de contratos en compras públicas usan *datasets* y no se consideran documentos como base inicial para el análisis de datos. Las herramientas informáticas

creadas para realizar análisis de corrupción en contratos tanto en el sector público y privado, no se consideran estándares de seguridad informática y el porcentaje de herramientas en entorno web es muy bajo.

Las principales técnicas de IA encontradas son: regresión logística, redes neuronales, redes bayesianas y SVM. En cuanto a herramientas informáticas para la programación, tratamiento y almacenamiento de datos usadas, Java, MatLab, Weka y Python ocupan el mayor porcentaje destacando en menor medida otras herramientas como: R, RapidMiner, Hadoop, Spark, Neo4j, Casandra, Kafta, Visual Studio.

4.9.5. Conclusiones

El artículo concluyó que pueden ocurrir diferentes tipos de corrupción, como soborno, colusión, malversación, apropiación indebida, fraude, abuso de discreción, favoritismo y nepotismo. El estudio también encontró que las técnicas de minería de datos, redes neuronales, redes bayesianas, SVM y árboles de decisión, pueden utilizarse para detectar y prevenir la corrupción de manera efectiva.

Con la revisión de literatura se pudo determinar que la mayoría de las investigaciones son de tipo descriptivo (detección), mientras que sólo el 21% son de tipo predictivo (prevención). Se pudo identificar el nicho de investigación al constatar la poca explotación del favoritismo como forma de corrupción en las investigaciones previas. También se evidencia la necesidad el uso del PLN como técnica de análisis. Por lo que estos datos permiten establecer estrategias de investigación para solventar la corrupción en compras públicas y evitar desgaste en investigaciones duplicadas.

4.10. Contribución 6

4.10.1. Titulo

Torres-Berru, Y., López Batista, V.F. (2022). Data and Text Mining for the Detection of Fraud in Public Contracts: “A Case Study of Ecuador’s Official Public Procurement System”. In: Berrezueta, S., Abad, K. *Doctoral Symposium on Information and Communication Technologies - DSICT. Lecture Notes in Electrical Engineering*, vol 846.

4.10.2. Objetivos

El objetivo del trabajo es mostrar los avances y fundamentos del trabajo doctoral, el cual se centra en proponer una metodología para prevenir, detectar la corrupción en la contratación pública utilizando técnicas de aprendizaje automático y PLN. Los objetivos específicos son:

- Desarrollar algoritmos para detectar y predecir el favoritismo y oligopolio en la contratación pública.

- Detectar y clasificar diferentes tipos de corrupción en el Sistema de Contratación Pública Ecuatoriano (SERCOP).
- Visualizar los resultados de manera adecuada para detectar y prevenir futuros actos de corrupción.
- Probar la viabilidad del estudio mediante la realización de un mapeo y revisión sistemática de la literatura.
- Probar la detección de favoritismo basada en los parámetros de calificación del proceso y los tipos de contratación.

4.10.3. Metodología

De manera general, se describe un sistema de minería de datos que utiliza PLN para realizar de forma inteligente el análisis de contenido de los contratos y detectar de forma automática la corrupción. De manera particular, los dos enfoques que se han desarrollado:

- Un enfoque de aprendizaje supervisado que utiliza un conjunto de datos de contratos etiquetados para entrenar un modelo para detectar el favoritismo y el oligopolio.
- Un enfoque de aprendizaje no supervisado que utiliza algoritmos de agrupamiento para identificar patrones de corrupción en contratos no etiquetados.

Esta contribución realiza un exhaustivo mapeo sistemático de investigaciones científicas sobre corrupción en contratos en diferentes bases de datos. Con este enfoque riguroso se identifica y analiza la literatura existente sobre el tema, proporcionando una visión global de los estudios realizados en un período específico. Analizando los diferentes tipos de corrupción abordados en la literatura científica, centrándose en el fraude. Proporcionando además información detallada sobre la prevalencia de cada tipo de corrupción y su representación en las investigaciones.

Se examina la distribución geográfica de las investigaciones sobre corrupción en contratos. Se destaca los países con mayor número de publicaciones y aquellos que reciben mayor cantidad de citas. Además, resalta la importancia de América Latina como un contexto de investigación en este tema.

El estudio identifica las herramientas y tecnologías emergentes utilizadas en el análisis de la corrupción en contratos, como la minería de datos, el aprendizaje automático y el agrupamiento. Señala las áreas de investigación futura que requieren una mayor atención científica, como los otros tipos de corrupción distintos al fraude y sus mecanismos, así como el uso de técnicas avanzadas de análisis de datos. Estas áreas ofrecen oportunidades para profundizar en el conocimiento y desarrollar estrategias más efectivas para combatir la corrupción en contratos.

4.10.4. Resultados

En el artículo se muestra la metodología seguida para detectar corrupción en compras públicas, la cual se basa en:

- Dos trabajos bibliográficos con los que se sustenta la investigación (Scopus) sin impacto JCR, así como se justifica la elección de las diferentes técnicas, además de definir posibles experimentos para su análisis.
- Dos trabajos con impacto JCR Q2, en los cuales la metodología se probó en el Sistema de Contratación Pública Ecuatoriano (SERCOP) y los resultados se visualizaron de manera adecuada para detectar y prevenir futuros actos de corrupción. El estudio también probó la detección de favoritismo basado en parámetros de calificación de procesos y tipos de contratación.
- En general, el artículo proporciona una metodología útil para prevenir y detectar la corrupción en la contratación pública. Finalmente se indican los trabajos futuros para validar y culminar la metodología planteada, a través del refinamiento de los algoritmos desarrollados en el artículo y la experimentación con diferentes conjuntos de datos. Además, el artículo sugiere explorar el uso de otras técnicas de aprendizaje automático y fuentes de datos para mejorar la detección y prevención de la corrupción en la contratación pública

4.10.5. Conclusiones

Se resume la metodología en desarrollo para prevenir y detectar corrupción en compras públicas, con el uso de algoritmos de aprendizaje automático. Se han establecido las bases teóricas del trabajo doctoral al constatar la poca explotación del favoritismo como forma de corrupción. Se ha definido una metodología científica acorde a la hipótesis planteada, basada en un modelo hibrido que incluye diferentes fases con aprendizaje supervisado y no supervisado, PLN y redes neuronales.

La metodología desarrollada en el artículo puede ser utilizada por gobiernos y organizaciones para detectar y prevenir automáticamente la corrupción en la contratación pública. Esto puede ayudar a asegurar que los procesos de contratación pública sean justos y transparentes, y que los fondos públicos se utilicen de manera eficiente y efectiva.

Bibliografía

1. Porter, L.E., Graycar, A.: Hotspots of corruption: Applying a problem-oriented policing approach to preventing corruption in the public sector. Security Journal. 29, 423–441 (2016). https://doi.org/10.1057/sj.2013.38.
2. Moran, J.: Democratic transitions and forms of corruption. Crime Law Soc Change. 36, 379–393 (2001). https://doi.org/10.1023/A:1012072301648.
3. Andvig, J.C., Fjeldstad, H., Amundsen, I., Sissener, T., Søreide, T.: Research on Corruption A policy oriented survey. (2000).
4. OECD: Public Sector Integrity Management Framework. OECD Publishing (2005).
5. Martinez Fernandez, J.M.: Transparencia versus corrupción en la contratación pública. Medidas de transparencia en todas las fases de la contratación pública como antídoto contra la corrupción., (2015).
6. Volosin, N.A.: Datos abiertos, corrupción y compras públicas. (2015).
7. Castañeda Rodríguez, V.M.: Una investigación sobre la corrupción pública y sus determinantes. Rev Mex Cienc Polit Soc. 61, 103–135 (2016). https://doi.org/10.1016/S0185-1918(16)30023-X.

8. Basheka, B.C.: Public Procurement Governance: Toward an Anti-corruption Framework for Public Procurement in Uganda. Public Procurement, Corruption and the Crisis of Governance in Africa. 113–141 (2021). https://doi.org/10.1007/978-3-030-63857-3_7.
9. Ferreira, I., Camões, P.J., Cunha, S., Amaral, L.A.: Electronic platforms and transparency in public procurement. In: Proceedings of the 30th International Business Information Management Association Conference, IBIMA 2017 - Vision 2020: Sustainable Economic development, Innovation Management, and Global Growth. pp. 3898–3906 (2017).
10. Chvalkovská, J., Skuhrovec, J.: Measuring transparency in public spending: Case of Czech Public e-Procurement Information System. IES working paper. 1–20 (2010).
11. Duguay, R., Rauter, T., Samuels, D.: The Impact of Open Data on Public Procurement. SSRN Electronic Journal. (2022). https://doi.org/10.2139/SSRN.3483868.
12. Põlluste, M.K.: Detecting corruption in public procurement through open data analysis. (2019).
13. Porter, L.E., Graycar, A.: Hotspots of corruption: Applying a problem-oriented policing approach to preventing corruption in the public sector. Security Journal. (2016). https://doi.org/10.1057/sj.2013.38.
14. Cieślik, A., Goczek, Ł.: Control of corruption, international investment, and economic growth – Evidence from panel data. World Dev. (2018). https://doi.org/10.1016/j.worlddev.2017.10.028.
15. Bhattacherjee, A., Shrivastava, U.: The effects of ICT use and ICT Laws on corruption: A general deterrence theory perspective. Gov Inf Q. (2018). https://doi.org/10.1016/j.giq.2018.07.006.
16. Grace, E., Rai, A., Redmiles, E., Ghani, R.: Detecting fraud, corruption, and collusion in international development contracts: The design of a proof-of-concept automated system. Presented at the (2016). https://doi.org/10.1109/BigData.2016.7840752.
17. Nisa, H., Dwisetia, P., Firmansyah: Corruption in Banten Province, Indonesia. Adv Sci Lett. (2017). https://doi.org/10.1166/asl.2017.9311.
18. Ferwerda, J., Deleanu, I., Unger, B.: Corruption in Public Procurement: Finding the Right Indicators. Eur J Crim Pol Res. (2017). https://doi.org/10.1007/s10610-016-9312-3.
19. de Michele, R., Prats Cabrera, J.O., Losada Revol, I.: Effects of Corruption on Public–Private Partnership Contracts: Consequences of a Zero-tolerance Approach. Inter American Development Bank. 10, 1–15 (2018).
20. Lee, M.-H., Lio, M.-C.: The impact of information and communication technology on public governance and corruption in China. Information Development. 32, 127–141 (2016). https://doi.org/10.1177/0266666914529293.
21. Hamisu, M., Mansour, A.: Detecting advance fee fraud using NLP bag of word model. Proceedings of the 2020 IEEE 2nd International Conference on Cyberspace, CYBER NIGERIA 2020. 94–97 (2021). https://doi.org/10.1109/CYBERNIGERIA51635.2021.9428793.
22. Porta Zamorano, J., Sancho Sánchez, J.L.: Procesamiento de lenguaje natural aplicado a datos masivos generados en medios sociales. Revista Española de Lingüística. 2, 111–124 (2021). https://doi.org/10.31810/rsel.51.2.7.
23. Núñez Torres, F., Pérez Cabello de Alba, M.B.: Desarrollo de un sistema de aprendizaje automático supervisado para la desambiguación léxica automática utilizando DAMIEN (Data Mining Encountered). Revista Electrónica de Lingüística Aplicada. 21, 150–178 (2023). https://doi.org/10.58859/rael.v21i1.504.
24. Alzate, C., Monreale, A., Assem, H., Bifet, A., Sandra Buda, T., Caglayan, B., Drury, B., García-Martín, E., Gavaldà, R., Kramer, S., Lavesson, N., Madden, M., Molloy, I., Nicolae, M.-I., Sinn, M.: SALER: A Data Science Solution to Detect and Prevent Corruption in Public Administration. 103–117 (2019). https://doi.org/10.1007/978-3-030-13453-2.
25. Auriol, E., Straub, S., Flochel, T.: Public Procurement and Rent-Seeking: The Case of Paraguay. World Dev. 77, 395–407 (2016). https://doi.org/10.1016/j.worlddev.2015.09.001.

26. van Erven, G.C.G., Carvalho, R.N., de Holanda, M.T., Ralha, C.: Graph database: A case study for detecting fraud in acquisition of Brazilian Government [Banco de Dados em Grafo: Um Estudo de Caso em Detecção de Fraudes no Governo Brasileiro]. Iberian Conference on Information Systems and Technologies, CISTI. 1–6 (2017). https://doi.org/10.23919/CISTI.2017.7975974.
27. Acero López, A.E., Ramirez Cajiao, M.C., Peralta Mejia, M., Payán Durán, L.F., Espinosa Díaz, E.E.: Participatory Design and Technologies for Sustainable Development: an Approach from Action Research. Syst Pract Action Res. 32, 167–191 (2019). https://doi.org/10.1007/S11213-018-9459-6/METRICS.
28. Pérez Mundaca, A.: Corrupción en las contrataciones públicas: investigaciones recientes y tendencias de investigación. Ciencia Latina Revista Científica Multidisciplinar. 6, 1652–1670 (2022). https://doi.org/10.37811/cl_rcm.v6i4.2686.
29. Vargas-Hernández, J.G.: The Multiple Faces of Corruption: Typology, Forms and Levels. SSRN Electronic Journal. 43 (2009).
30. Chan, A.P.C., Owusu, E.K.: Corruption Forms in the Construction Industry : Literature Review. 143, 1–12 (2017). https://doi.org/10.1061/(ASCE)CO.1943-7862.0001353.
31. Bramoullé, Y., Goyal, S.: Favoritism. J Dev Econ. 122, 16–27 (2016). https://doi.org/10.1016/j.jdeveco.2016.04.006.
32. Castro Cuenca, C.G.: La corrupción pública y privada: causas, efectos y mecanismos para combatirla. (2017).
33. 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.
34. SERCOP: Manual De Buenas Prácticas En La Contratación Pública Para El Desarrollo. 1–46 (2015).
35. Cassagne, J.Carlos., Rivero Ysern, Enrique.: La contratación pública. Hammurabi (2007).
36. Ponce, H.G., Gil, M.T.N., Durán, M.P.: Responsible public procurement. Design of measurement indicators. CIRIEC-Espana Revista de Economia Publica, Social y Cooperativa. 253–280 (2019). https://doi.org/10.7203/CIRIEC-E.96.12627.
37. Hyytinen, A., Lundberg, S., Toivanen, O.: Politics and Procurement: Evidence from Cleaning Contracts. SSRN Electronic Journal. (2011). https://doi.org/10.2139/ssrn.1082817.
38. Commission, E.: ARACHNE risk scoring tool, https://ec.europa.eu/social/main.jsp?catId=325&intPageId=3587&langId=en.
39. European Commission: Preventing fraud and corruption in the European Structural and Investment Funds—taking stock of practices in the EU Member States. (2019).
40. Chen, Y.J., Wu, C.H., Chen, Y.M., Li, H.Y., Chen, H.K.: Enhancement of fraud detection for narratives in annual reports. International Journal of Accounting Information Systems. 26, 32–45 (2017). https://doi.org/10.1016/j.accinf.2017.06.004.
41. Gallego, J., Rivero, G., Martínez, J.: Preventing rather than punishing: An early warning model of malfeasance in public procurement. Int J Forecast. 37, 360–377 (2021). https://doi.org/10.1016/j.ijforecast.2020.06.006.
42. Dávid-Barrett, E., Fazekas, M.: Grand corruption and government change: an analysis of partisan favoritism in public procurement. Eur J Crim Pol Res. 26, 411–430 (2020). https://doi.org/10.1007/s10610-019-09416-4.
43. Pierri, G., Jarquin, M.J., de Michele, R.: Transparencia y género: el impacto de las compras electronicas en el acceso a licitaciones públicas de las pyMe lideradas por mujeres, (2021).
44. PNUD: Perspectiva de Género en las compras públicas. Llamado a la acción.
45. Csáki, C., Prier, E.: Quality issues of public procurement open data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in

Bioinformatics). 11032 LNCS, 177–191 (2018). https://doi.org/10.1007/978-3-319-98349-3_14/COVER.
46. Mahto, D.K., Singh, L.: A dive into Web Scraper world. In: Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016. pp. 689–693 (2016).
47. Han, Jiawei., Kamber, Micheline., Pei, Jian.: Data mining : concepts and techniques. Elsevier Science (2011).
48. van Erven, G.C.G., Holanda, M., Carvalho, R.N.: Detecting evidence of fraud in the Brazilian government using graph databases, (2017). https://doi.org/10.1007/978-3-319-56538-5_47.
49. Wirth, R., Hipp, J.: CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. pp. 29–39. Springer-Verlag London, UK (2000).
50. Solorio-Fernández, S., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: A review of unsupervised feature selection methods. Artif Intell Rev. 53, 907–948 (2020). https://doi.org/10.1007/s10462-019-09682-y.
51. Chang, H.Y., Thomson, J.A., Chen, X.: Microarray Analysis of Stem Cells and Differentiation. Handbook of Stem Cells, Second Edition: Volume 1-2. 1, 399–407 (2012). https://doi.org/10.1016/B978-0-12-385942-6.00034-2.
52. Kapil, S., Chawla, M.: Performance evaluation of K-means clustering algorithm with various distance metrics. In: 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, ICPEICES 2016. Institute of Electrical and Electronics Engineers Inc. (2017). https://doi.org/10.1109/ICPEICES.2016.7853264.
53. Kotsiantis, S.B.: Supervised Machine Learning: A Review of Classification Techniques. (2007).
54. Rajak, I., Mathai, K.J.: Intelligent fraudulent detection system based SVM and optimized by danger theory. IEEE International Conference on Computer Communication and Control, IC4 2015. 2–5 (2016). https://doi.org/10.1109/IC4.2015.7375705.
55. Guo, J., Li, T., Li, Y.: SVM Based on Gaussian and Non-Gaussian Double Subspace for Fault Detection. IEEE Access. 9, 66519–66530 (2021). https://doi.org/10.1109/ACCESS.2021.3075273.
56. Computación, A.M.D.E.: El Reconocimiento de Patrones y su Aplicación a las Señales Digitales María del Pilar Gómez Gil. (2019).
57. Guzmán Lembo, A., Mayorga Alvarado, C.D., Dávila Vázquez, J.F., Martínez Reyna, J., Rodriguez-Liñan, A., Torres-Treviño, L.M.: Clasificador de objetos en MATLAB® con redes neuronales de aprendizaje profundo. Ingenierias. 24, 41–54 (2021). https://doi.org/10.29105/ingenierias24.90-16.
58. Wang, Q., Xu, W., Huang, X., Yang, K.: Enhancing intraday stock price manipulation detection by leveraging recurrent neural networks with ensemble learning. Neurocomputing. 347, 46–58 (2019). https://doi.org/10.1016/j.neucom.2019.03.006.
59. Moreira, D., Cruz, I., Gonzalez, K., Quirumbay, A., Magallan, C., Guarda, T., Andrade, A., Castillo, C.: Análisis del Estado Actual de Procesamiento de Lenguaje Natural Analysis of the Current State of Natural Language Processing. Iberian Journal of Information Systems and Technologies. 126–136 (2021).
60. Caparrós-Laiz, C., García-Díaz, J.A., Valencia-García, R.: Evaluating Extractive Automatic Text Summarization Techniques in Spanish. 79–92 (2021). https://doi.org/10.1007/978-3-030-88262-4_6.
61. Conroy, J.M., Schlesinger, J.D., O'Leary, D.P.: Nouveau-ROUGE: A novelty metric for update summarization. Computational Linguistics. 37, 1–8 (2011). https://doi.org/10.1162/coli_a_00033.
62. Callison-burch, C., Callison-burch, C., Osborne, M.: Re-evaluating the role of BLEU in machine translation research. IN EACL. 249--256 (2006).
63. Alexandropoulos, S.-A.N., Kotsiantis, S.B., Vrahatis, M.N.: Data preprocessing in predictive data mining. Knowl Eng Rev. 34, e1 (2019). https://doi.org/10.1017/S026988891800036X.

64. Alshdaifat, E., Alshdaifat, D., Alsarhan, A., Hussein, F., El-Salhi, S.M.F.S.: The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. Data (Basel). 6, 1–23 (2021). https://doi.org/10.3390/data6020011.
65. Torres-Berru, Y., López Batista, V.F.: Data and Text Mining for the Detection of Fraud in Public Contracts: A Case Study of Ecuador's Official Public Procurement System. 116–127 (2022). https://doi.org/10.1007/978-3-030-93718-8_10.
66. Torres-Carrion, P.V., Gonzalez-Gonzalez, C.S., Aciar, S., Rodriguez-Morales, G.: Methodology for systematic literature review applied to engineering and education. IEEE Global Engineering Education Conference, EDUCON. 2018-April, 1364–1373 (2018). https://doi.org/10.1109/EDUCON.2018.8363388.
67. García-González, A., Ramírez-Montoya, M.-S.: Systematic Mapping of Scientific Production on Open Innovation (2015–2018): Opportunities for Sustainable Training Environments. Sustainability. 11, 1–15 (2019). https://doi.org/10.3390/su11061781.

Anexo de contribuciones

Am. J. Trop. Med. Hyg., 105(1), 2021, pp. 81–87
doi:10.4269/ajtmh.21-0221

Exceptional Prices of Medical and Other Supplies during the COVID-19 Pandemic in Ecuador

Esteban Ortiz-Prado,[1,2*] Raul Fernandez-Naranjo,[1] Yeferson Torres-Berru,[3,4] Rachel Lowe,[5,6] and Irene Torres[7*]

[1]*One Health Research Group, Faculty of Medicine, Universidad de las Americas, Quito, Ecuador;* [2]*Department of Cell Biology, Physiology and Immunology, Universidad de Barcelona, Barcelona, Spain;* [3]*University of Salamanca, Salamanca, Spain;* [4]*Instituto Superior Tecnológico Sudamericano, Loja, Ecuador;* [5]*Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom;* [6]*Centre on Climate Change and Planetary Health, London School of Hygiene & Tropical Medicine, London, United Kingdom;* [7]*Fundacion Octaedro, Quito, Ecuador*

Abstract. Shortages of essential supplies used to prevent, diagnose, and treat COVID-19 have been a global concern, and price speculation and hikes may have negatively influenced access. This study identifies variability in prices of products acquired through government-driven contracts in Ecuador during the early pandemic response, when the highest mortality rates were registered in a single day. Data were obtained from the National Public Procurement Service (SERCOP) database between March 1 and July 31, 2020. A statistical descriptive analysis was conducted to extract relevant measures for commonly purchased products, medical devices, pharmaceutical drugs, and other goods. Among the most frequently purchased products, the greatest amounts were spent on face masks (US$4.5 million), acetaminophen (US$2.2 million), and reverse transcriptase quantitative polymerase chain reaction assay kits (US$1.8 million). Prices varied greatly, depending on each individual contract and on the number of units purchased; some were exceptionally higher than their market value. 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%. In a context of budgetary constraints that actually required an effective use of available funds, speculative price hikes may have limited patient access to health care and the protection of the general population and health care workers. COVID-19 vaccine allocations to privileged individuals have also been widely reported. Price caps and other forms of regulation, as well as greater scrutiny and transparency of government-driven purchases, and investment in local production, are warranted in Ecuador for improved infectious disease prevention.

INTRODUCTION

Shortages of essential supplies used to prevent, diagnose, and treat COVID-19 have been documented around the world, with the pandemic exacerbating the impact of systemic weaknesses on availability in places with high transmission rates.[1] Lack of medical devices such as mechanical ventilators, reverse transcriptase quantitative polymerase chain reaction (RT-qPCR) test kits, and a full range of personal protection equipment (PPE) for health care workers to safely comply with clinical duty has not been the only cause for concern; limited access to basic food supplies may increase exposure and vulnerability to the disease.[2]

Breaches of anticorruption standards, including surcharges, speculation, collusion, cutting corners in procurement processes, and even politicians taking advantage of the crisis to increase their private benefits, have been recorded in various countries during the COVID-19 pandemic.[3] Government transparency in decision-making has been described as a major problem in response efforts worldwide,[4] together with weak regulations and accountability mechanisms and low wages in the heath sector.[5] In countries such as the United Kingdom, a National Health Service official was found selling PPE on the side, and the U.S. administration paid extraordinary prices to third-party vendors.[6]

Ecuador is a South American country of 17 million people, which was severely hit early in the pandemic, registering more than 40,000 deaths in 2020, one of the highest rates per capita in the world. Although the government declared a health emergency in March followed by a national lockdown, social protection measures and health system resources have been limited. Vulnerable populations have also relied, for example, on food kits from the World Food Program[7] and provisionally lower-priced food and drug outlets supported by the peasant movement (Movimiento Social Campesino, FECAOL, in Spanish).[8,9]

Most medical equipment, devices, and medicines are imported into Ecuador, and the public health sector's resource allocation has forced populations to pay out of pocket for medication in the past, placing those who cannot pay at greater risk.[10] As the country's health system became overstretched to the point of collapse in Guayas, the first COVID-19 hotspot,[11,12] major corruption scandals emerged in relation to the procurement of medical devices such as face masks or human remains pouches (HRPs), commonly known as body bags.[13] Even though the law requires civil society organizations' participation in Emergency Operations Committee meetings, they have not been present during the sessions convened for the pandemic response, which has limited transparency in decision-making.[14] In addition, as Ecuador was seeing a steady rise in the number of confirmed cases early in the pandemic, its diagnostic capacity did not increase to meet demands.[15]

Ecuador mandated face masks on April 7, 2020, for the general population,[16] but these were not freely distributed by the government nor was a price cap established. In practice, this implied that only people with enough means could purchase them to remain protected.[17] Similarly, in countries with limited resources, those who can afford expensive private testing may presumably be at an advantage over those who cannot. With only 14% of medicines being produced locally, and 86% of this revenue going to a single company,[18] market prices are not necessarily competitive. It has been reported

*Address correspondence to Esteban Ortiz-Prado, One Health Research Group, Universidad de las Américas, Quito, Ecuador Calle de los Colimes y Avenida De los Granados, Quito 170137, Ecuador, E-mail: e.ortizprado@gmail.com or Irene Torres, Fundación Octaedro, El Zurriago E8-28, 170804, Quito, Ecuador, E-mail: irene.torres@octaedro.edu.ec.

that at least 202 public contracts signed during the pandemic response have irregularities; a majority of them involve the purchase of medical supplies, masks, suits, antibacterial gel, PCR tests, medicines, food kits, and other products and services.[19] The comptroller general of Ecuador found that food kits were purchased by public institutions for distribution among vulnerable groups across the country at a much higher cost than regular store prices.[20,21] Furthermore, tocilizumab vials donated by a pharmaceutical company to the government were sold on the black market for at least 5 and up to 10 times their market price in April 2020.[22] Finally, oxygen shortages and exceedingly high prices were also reported by the media in early April of the same year.[23]

This study aims to compare variability in prices of products acquired through public contracts in Ecuador for use in clinical settings or as protection for the general population in response to the COVID-19 pandemic.

METHODS

Study design. This is an observational descriptive study of the publicly available data on government-driven purchases for the public health system in Ecuador from January 1, 2020 to July 31, 2020.

Setting. The study was conducted in Ecuador, a country located in South America, with an estimated population of 17.68 million. The National Public Procurement Service (SERCOP) is in charge of promoting citizen participation, increasing access to and use of public information by the population, increasing transparency, and fighting fraud and corruption that could arise from bad practices in public procurement. Our study quantifies differences in prices of commonly purchased devices, medicines, and supplies in public contracts registered in SERCOP that have been identified as relating to the COVID-19 pandemic.

Variables. Considering the Uniform Resource Locator (URL) of each contracting process as input, the following sections were identified: description, dates, products, qualification parameters, invitations, files, and questions from suppliers. Each section was extracted through scraping according to its equivalent identification in html and was stored in an unrelated database. The variables analyzed were as follows: governmental institution responsible for the purchase process (Ministry, Central Government, Decentralized Autonomous Governments), location where the contract was based (Municipality and Province), type of product or service, unit price, and type of contract (direct contract or public auction).

Data sources. Data were obtained from the publicly available database from SERCOP. Data on government-driven purchase processes between March 1 and July 31, 2020, were extracted using Selenium with Python. Using fuzzy matching and regular expressions, we identified the most frequent products in the bids and contracts. The source code of python script is available in the following link: https://github.com/torresyeferson/DattaMinigCorruption/blob/main/Scraping.py

Statistical method. A statistical descriptive analysis was conducted to extract relevant measures for each product and determine variability across all items. Tableau software was used for visualization and analysis. A descriptive statistical analysis was performed to describe means, trends and percentage changes.

RESULTS

Data by institution. The highest spenders during the peak of the pandemic in 2020 were the Guayaquil Municipal Government (US$19 million) and the Guayas Provincial Government (US$5 million), followed by the Ministry of Public Health (US$18 million). Hospitals that spent the largest sums were also located in Guayaquil—Sagrado Corazon de Jesus and Abel Gilbert Ponton—which spent approximately US$5 million each (Table 1).

As a reference, Ecuador has 168 public hospitals; its 19 highest spenders used 1.4% of Ecuador's 2018 total public health budget (US$2.665 million) in only 3 months, between March 1 and May 31, 2020.[24]

Data by product or service. The most frequently purchased products in public contracts were medical devices (ventilators), laboratory supplies (RT-qPCR assay kits), all types of facial masks (including N95 respirators, surgical masks, and face masks), medicines (acetaminophen, piperacillin, hydroxychloroquine), disinfectant (quaternary ammonium), HRPs, and food kits.

Data by prices. Prices varied by contract and depended on the number of units purchased. Distribution of prices for all main products are included in Figure 1. During the study period, prices were analyzed by item, institution, and month, ranging from individual products as cheap as US$0.03 (acetaminophen 500 mg pills) to US$140,000 (mechanical ventilators). A total of US$4.5 million was spent on all types of masks (including N95 respirators, surgical masks, and face masks), ranging from US$0.50 to US$90 per unit. The government spent at least US$2.2 million on acetaminophen, with prices ranging from US$0.04 per 500-mg pill to US$2.65 per oral suspension bottle (different sizes). In terms of diagnostic supplies, at least US$1.8 million were spent on RT-qPCR essay kits. The lowest price for an individual primer was US$18 (per reaction) and the highest was US$85 per reaction when including the extraction kit. The price of an individual HRP reached exceptional high values per unit. For instance, the Ecuadorian Institute of Social Security (IESS) bought 4,000 units at US$148.5 per HRP, a value 493% higher than the market price of US$25.[13] Prices of individual items per unit can be found in the link shared in the Availability of Data section.

In another exceptional purchase, the National Secretariat for Risk Management acquired 7,000 food kits for vulnerable families, at US$150.82 per hamper. The Comptroller General of Ecuador has subsequently established that the market value of each food kit is US$95.16, one of the reasons being that only eight of the 18 products are taxable, but the vendor had included a tax value for all items.[21] Furthermore, the report indicates prices were also inflated.

Data by dates. The prices of food kits and medical supplies such as face masks experienced sharp increases during the first few months of the pandemic (Figure 2). Although the prices of face masks oscillated greatly since at least 2019, the increase in April and May 2020 is evident (Figure 3).

Most of the increase was on products that were highly needed during the health emergency caused by the COVID-19 pandemic. Compared with the previous year, medical exam gloves increased by 1,307%, acetaminophen 500-mg pills by 796%, and oxygen flasks (different sizes) by 30.8% in the

TABLE 1
Expenditures of local hospitals (March 1–May 31, 2020)

Hospital	March (in US$)	April (in US$)	May (in US$)	Total (in US$)	% Of total
Hospital Sagrado Corazon de Jesus	2,845,134	2,519,196	–	5,364,330	14%
Hospital Abel Gilbert Ponton	4,581,303	249,207	–	4,830,510	13%
Hospital Quito-Sur	2,799,047	–	–	2,799,047	7%
Hospital Martin Icaza	1,555,057	705,002	153,566	2,413,625	6%
Hospital General de Machala	1,669,934	391,820	–	2,061,754	5%
Hospital General Enrique Garcés	1,965,252	–	–	1,965,252	5%
Hospital de Especialidades Carlos Andrade Marin	1,826,454	58,500	–	1,884,954	5%
Hospital del Niño	1,840,205	24,754	–	1,864,959	5%
Hospital General Monte Sinaí	1,421,120	426,689	–	1,847,809	5%
Hospital General de Milagro	1,529,479	188,787	–	1,718,266	5%
Hospital General Guasmo Sur	1,666,462	–	–	1,666,462	4%
Hospital de Especialidades Eugenio Espejo	1,600,469	–	–	1,600,469	4%
Hospital de Especialidades Fuerzas Armadas No. 1	557,525	682,018	117,467	1,357,010	4%
Hospital General Marco Vinicio Iza	588,664	656,694	–	1,245,358	3%
Hospital General de Manta	1,196,829	45,298	–	1,242,127	3%
Hospital de Especialidades Jose Carrasco Arteaga	989,979	220,553	–	1,210,532	3%
Hospital Vicente Corral Moscoso	979,148	45,269	–	1,024,417	3%
Hospital Rodríguez Zambrano	963,248	–	–	963,248	3%
Hospital General de Ambato	848,625	–	–	848,625	2%
Total	31,423,934	6,213,787	271,033	37,908,754	100%

corresponding month during which data were available (Table 2).

DISCUSSION

The data in this study show great variability of prices and extreme price hikes in government-driven purchases related to the COVID-19 response in Ecuador—in particular, during the peak of the pandemic in the country's initial hotspot of Guayas. Although it is true that greater worldwide demand for medical equipment and supplies drove price increases in both high-income and low- and middle-income countries, government-based contracts in Ecuador diverted from market prices. Furthermore, the national and local governments, public hospitals, and the National Secretariat for Risk Management do not appear to have coordinated bulk negotiations

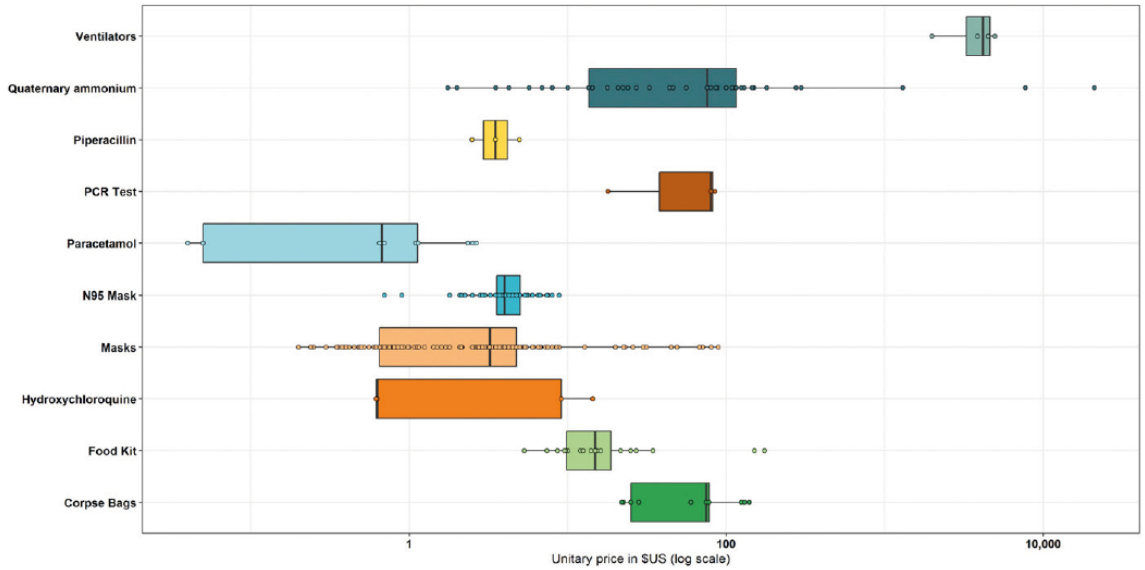


FIGURE 1. Price distribution of basic supplies related to the COVID-19 response in Ecuador (March 1–July 31, 2020). The plot depicts variations in price by item compared with the median/mean price, using as reference outliers in distribution in a logarithmic scale. Dots outside bars and boxes represent extreme and atypical values.

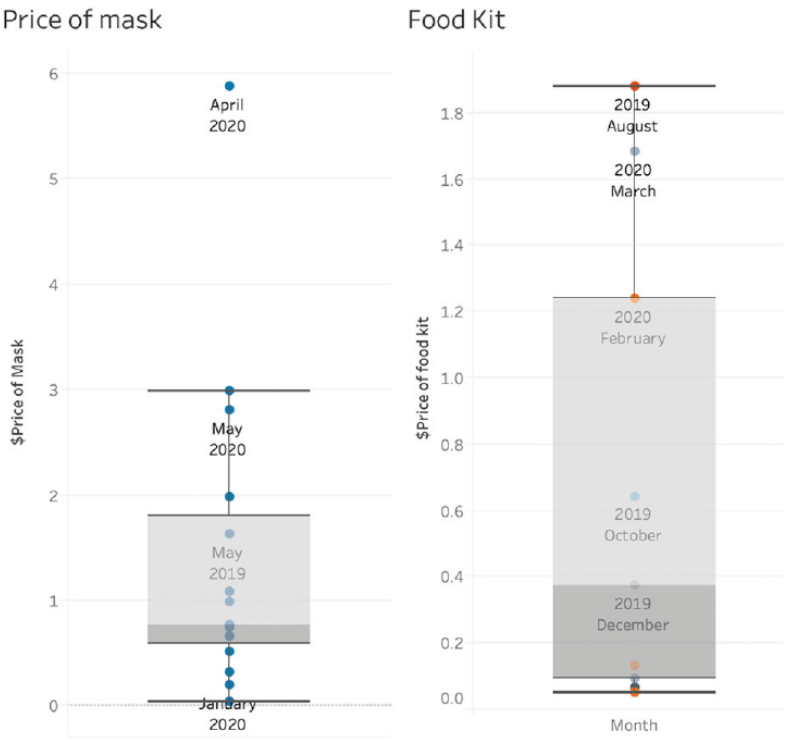


FIGURE 2. Price distribution of face masks and food kits in Ecuador by month (August 1, 2019–May 31, 2020). We selected these two items for comparison because contracts most frequently included them; data for other items were not available for the 2 years. This figure appears in color at www.ajtmh.org.

to ensure better prices. In addition, exceptional purchases of ammonium quaternary and hydroxychloroquine by the Ecuadorian government show that decisions were not necessarily based on need or the best available evidence.[25] As a consequence, government expenditures appear to have been costly and also wasteful.

Irrespective of limits in diagnostic capacity, which are a reality in Ecuador, high prices of RT-qPCR primers and extraction kits in government-driven purchases may have had a negative impact on the number of tests that the Ministry of Health was able to afford.[15] Accordingly, an increase in the public health budget probably would not have led to greater access to testing through added purchases but instead to further speculation and price hikes. This may help to explain the high price caps set by the government: US$80 per publicly funded and US$120 per privately paid RT-qPCR assay. Unsurprisingly, there was severe undertesting in the country during the study period, with RT-qPCR positivity rates (total confirmed cases as a share of the total number of people tested) between 48% on May 1, 2020[26] and 42% on July 31, 2020.[27]

Although the current analysis concerns only government-driven purchases, private demand for medical supplies and medicines in a country reliant on imports and with highly concentrated local production of medicines may have also influenced price increases. However, there is no evidence that the government was trying to keep prices down through control mechanisms or at least greater scrutiny and accountability of public purchases. Media reports on corruption in government-driven purchases pointed toward contracts remaining largely unmonitored as the health emergency

TABLE 2
Comparison of mean prices between January 1, 2019 and May 30, 2020*

	Month	2019	2020	% Increase
Medical exam and sterile gloves (100 units)	Jan	$0.03	$0.07	144%
	April	$0.19	$2.67	1,307%
	May	$1.66	$2.81	69%
Acetaminophen (500-mg pill)	Jan	$0.002	$0.014	796%
	April	$0.002	$0.004	166%
Oxygen (flasks, different sizes)	April	$2.82	$3.69	30.8%

* Values are shown only for the month in which purchases occurred in the database. Prices of individual items per unit can be found in the link shared in the Availability of Data section.

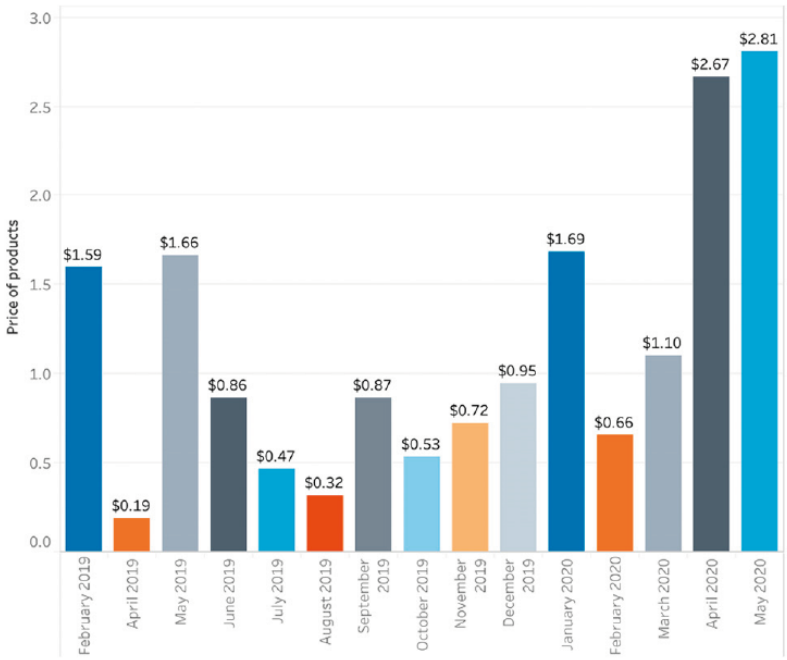


FIGURE 3. Monthly price of face masks in public contracts (there was no information for January and March 2019). This figure appears in color at www.ajtmh.org.

unfolded,[28,29] and investing in local production was not a high priority in the government agenda.

Speculative price hikes amid constraints in public spending may have limited patient access to diagnosis and treatment of COVID-19, and protection of both the general population and health workers in clinical settings. Prices of PPE and medicines, including hydroxychloroquine (antimalarial medication that was briefly suggested for COVID-19 treatment) experienced increases in several countries.[30] Lack of PPE may help to explain why, early in the pandemic (February 2–April 18, 2020), the most impacted occupational sector in Ecuador was health care (19% of total COVID-19 confirmed cases).[31] Although moderate increases in prices have not been found to negatively influence purchase of PPE, extreme increases such as 1,307% on the price of medical gloves may have had a limiting effect.

In countries such as Brazil, oxygen shortages and price speculation led to the death of vulnerable patients who were not able to procure cylinders that had been reserved for sale to the wealthy.[32] Similarly, the city of Guayaquil reported shortages and prices up to US$50 for the refill of an oxygen cylinder[23] in early April, when a record number of people died.

Although the excessive cost of food kits illustrates widespread speculation in COVID-19 related government-driven contracts, it is important to note that nutritional status may have been affected by the number of people who could actually receive this type of aid. Moreover, under such circumstances, requests for people to stay at home or quarantine were difficult to meet when they had to acquire their means of subsistence through informal work or personally procure foodstuffs, further undermining proper prevention and treatment. Direct purchases by the government from farming associations that were already helping in the response could have helped reach more people, together with keeping the economy moving, with the funds that had been allocated for emergency food procurement.[8,9]

Finally, information on COVID-19 vaccine vials is not available through SERCOP or other data sources, and the prices paid for them are unknown at the time of this publication. Unplanned, unethical allocation of doses privileging public officials and their spouses, journalists, and businesspeople has been widely reported,[33] making it even more relevant than before to correct obscure practices in the use of public health resources.

Study limitations. This study has limitations. First, data collection and analysis focused on the most frequently identified products in public contracts citing COVID-19. We do not know what the case is in other, including essential, supplies or equipment. Second, we do not have data on patient access to health care, such as diagnosis and treatment of COVID-19, so we cannot provide conclusions on the impact of price variation. Third, we do not know the factors that may have influenced product pricing, such as global increase in the cost of materials and transportation due to disruptions in supply chains. Nevertheless, the study brings to the fore public purchase practices that may have a detrimental effect on health outcomes.

CONCLUSION

The COVID-19 pandemic has exposed flaws in health system governance around the world, highlighting the potential for price speculation and unjustified hikes in prices to undermine the effectiveness of country responses. At the beginning of the pandemic, Ecuador suffered one of the most

aggressive outbreaks of COVID-19 worldwide, and many questions have remained unanswered regarding the extremely high spike in confirmed cases and excessive deaths in provinces such as Guayas. An unfulfilled need for continuous access to face masks for the general population, as well as food supplies, may have played a role in the increased transmission in the country. Supplies allocation during the COVID-19 response depended on budgetary constraints due to Ecuador's ongoing financial crisis and consequently required effective use of available funds. Furthermore, additional dependance on out-of-pocket payments warranted price caps and other forms of regulation, as well as greater scrutiny of public purchases.

Improvements in the health system and pandemic preparedness efforts should focus on ensuring adequate investment of public resources and planning for availability of supplies for the prevention of infectious diseases. Anticipating bulk purchases across hospitals, at the least, and across other institutions and sectors, should be prioritized to help guarantee affordable prices not just within the public health sector but also for patients who may be forced to pay out of pocket or use their private insurance. Especially in the case of an inexpensive but apparently highly effective measure such as mandatory or voluntary mask wearing, basic price controls may induce better compliance with self-care activities. Investing in local production of medical supplies and complementary but essential resources such as food should also be considered.

Received February 23, 2021. Accepted for publication April 17, 2021.

Published online May 20, 2021.

Acknowledgment: The American Society of Tropical Medicine and Hygiene has waived the Open Access fee for this article due to the ongoing COVID-19 pandemic.

Disclosure: Data are publicly available at www.compraspublicas.gob.ec, National Public Procurement Service (SERCOP). Web scraping was conducted at the following link: https://www.compraspublicas.gob.ec/ProcesoContratacion/compras/EP/BusquedaProveedorCpc.cpe. Data corresponding to all institutions responsible for the purchases recorded for this study are available here: https://drive.google.com/file/d/1SfEJA_zxCsH81UR3EbI5jW6ebTI9iPUS/view?usp=sharing.

Authors' addresses: Esteban Ortiz-Prado, One Health Research Group, Faculty of Medicine, Universidad de las Americas, Quito, Ecuador, and Department of Cell Biology, Physiology and Immunology, Universidad de Barcelona, Barcelona, Spain, E-mail: e.ortizprado@gmail.com. Raul Fernandez-Naranjo, One Health Research Group, Faculty of Medicine, Universidad de las Americas, Quito, Ecuador, E-mail: raul.fernandez.n@gmail.com. Yeferson Torres-Berru, University of Salamanca, Salamanca, Spain, and Instituto Superior Tecnológico Sudamericano, Loja, Ecuador, E-mail: yeferson.torres11@gmail.com. Rachel Lowe, Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom, and Centre on Climate Change and Planetary Health, London School of Hygiene & Tropical Medicine, London, United Kingdom, E-mail: rachel.lowe@lshtm.ac.uk. Irene Torres, Fundacion Octaedro, Quito, Ecuador, E-mail: irene.torres@octaedro.edu.ec.

REFERENCES

1. Ranney ML, Griffeth V, Jha AK, 2020. Critical supply shortages—the need for ventilators and personal protective equipment during the COVID-19 pandemic. *N Engl J Med 382:* e41.
2. UN, 2020. COVID-19 worsening food insecurity, driving displacement, warn UN agencies. *UN News*. Available at: https://news.un.org/en/story/2020/11/1077272. Accessed January 10, 2021.
3. U4, 2020. *Corruption in the Time of COVID-19: A Double-threat for Low Income Countries.* Available at: https://www.u4.no/publications/corruption-in-the-time-of-covid-19-a-double-threat-for-low-income-countries. Accessed January 10, 2021.
4. Rajan D et al., 2020. Governance of the COVID-19 response: a call for more inclusive and transparent decision-making. *BMJ Glob Health 5:* e002655.
5. Teremetskyi V, Duliba Y, Kroitor V, Korchak N, Makarenko O, 2021. Corruption and strengthening anti-corruption efforts in healthcare during the pandemic of COVID-19. *Med Leg J 89:* 15–28.
6. Kohler JC, Wright T, 2020.The urgent need for transparent and accountable procurement of medicine and medical supplies in times of COVID-19 pandemic. *J Pharm Policy Pract 13:* 1–4.
7. Ecuador United Nations, 2020. *Kits Puerta a Puerta en Ecuador. Naciones Unidas en Ecuador.* Available at: https://ecuador.un.org/es/48771-kits-puerta-puerta-en-ecuador. Accessed January 10, 2021.
8. Alainet, 2020. *Ecuador: "Brigadas Campesinas Abastecen a Guayaquil" (XV).* https://www.alainet.org/es/articulo/206824. Accessed April 4, 2021.
9. FECAOL, 2020. *Brigadas Campesinas—Movimiento Nacional Campesino.* Available at: https://movimientocampesinoec.org/index.php/category/brigadascampesinas/. Accessed April 4, 2021.
10. Ortiz-Prado E et al., 2017. Analysis of health and drug access associated with the purchasing power of the ecuadorian population. *Glob J Health Sci 9*. doi: 10.5539/gjhs.v9n1p201.
11. Ortiz-Prado E, Cevallos-Sierra G, Henriquez-Trujillo AR, Lowe R, Lister A, 2020. COVID-19 in Latin America. *BMJ.* Available at: https://blogs.bmj.com/bmj/2020/08/13/covid-19-in-latin-america/. Accessed October 3, 2020.
12. Torres I, Sacoto F, 2020. *Ecuador's Fragile Response to COVID-19. IHP*. Available atL https://www.internationalhealthpolicies.org/blogs/ecuadors-fragile-response-to-covid-19/. Accessed February 22, 2021.
13. Primicias, 2020. *Por la Emergencia se Compraron 7.824 Bolsas para Cadáveres*. Available at: https://www.primicias.ec/noticias/politica/emergencia-compraron-bolsas-cadaveres/. Accessed January 10, 2021.
14. Torres I, López-Cevallos D, 2021. In the name of COVID-19: legitimizing the exclusion of community participation in Ecuador's health policySpecial call: Health Promotion Perspectives on the COVID-19 pandemic. *Health Promot Int*. Available at: https://doi.org/10.1093/heapro/daaa139.
15. Torres I, Sippy R, Sacoto F, 2021. Assessing critical gaps in COVID-19 testing capacity: the case of delayed results in Ecuador. *BMC Public Health 21:* 637.
16. Emergency Operations Committee, 2020. *Resolutions—April 7, 2020.* Available at: https://www.gestionderiesgos.gob.ec/resoluciones-coe-nacional-07-de-abril-2020/. Accessed April 4, 2021.
17. Tucho GT, Kumsa DM, 2021. Universal use of face masks and related challenges during COVID-19 in developing countries. *Risk Manag Healthc Policy 14:* 511.
18. Iturralde P, 2015. *Concentración de Capital en el Sistema de Salud* [Concentration of capital in the health system]. Quito, Ecuador: Centro de Derechos Económicos y Sociales—CDES.
19. El Comercio, 2020. *Observan 202 Contratos Suscritos en la Pandemia por Anomalías en Ecuador | El Comercio*. Available at: https://www.elcomercio.com/actualidad/contratos-pandemia-posibles-anomalias-ecuador.html. Accessed January 10, 2021.
20. Primicias, 2020. *Contraloría Confirma Sobreprecio de 40% en Compra de Kits Alimenticios.* Available at: https://www.primicias.ec/noticias/politica/contraloria-irregularidades-contrato-kits-alimentos/. Accessed January 10, 2021.
21. Comptroller General of Ecuador, 2020. *Press Bulletin 017. Comptroller Establishes Grounds for Criminal Liability in Contract with National Service of Risk Managemen*t [Boletín de prensa N° 017. Contraloría establece indicios de responsabilidad penal en contrato de Servicio Nacional de Gestión

de Riesgos]. Published online May 11. Available at: https://www.contraloria.gob.ec/CentralMedios/BoletinesPrensa/23842. Accessed April 4, 2021.
22. Redacción Revista, 2020. *¿Cómo Vendía Abraham Muñoz, amigo de Daniel Salcedo, Medicamentos Donados al IESS Contra la COVID-19?* Vistazo. Revista Vistazo. Available at: https://www.vistazo.com/seccion/actualidad-nacional/como-vendia-abraham-munoz-amigo-de-daniel-salcedo-medicamentos-donados. Accessed January 10, 2021.
23. El Universo, 2020.*Coronavirus in Ecuador: Oxygen Demand in Guayaquil Grows due to Home-Treated Patient*s [Coronavirus en Ecuador: Demanda de oxígeno crece en Guayaquil por pacientes tratados en casa]. Published online April 9. Avalable at: https://www.eluniverso.com/noticias/2020/04/09/nota/7808535/demanda-oxigeno-crece-pacientes-tratados-casa. Accessed April 4, 2021.
24. Ministerio de Economia y Finanzas, 2018. *Informe de Rendición de Cuentas 2018*. Available at: https://www.finanzas.gob.ec/wp-content/uploads/downloads/2019/03/Informe-de-Rdc-2018-final.pdf. Accessed April 4, 2021.
25. Vinetz JM, 2020. Lack of efficacy of hydroxychloroquine in COVID-19. *BMJ 369:* m2018.
26. Secretariat of Risk Management, 2020. *Infographic 065*. Published online May 1. Available at: https://www.gestionderiesgos.gob.ec/wp-content/uploads/2020/05/INFOGRAFIA-NACIONALCOVI-19-COE-NACIONAL-01052020-08h00.pdf. Accessed April 4, 2021.
27. Secretariat of Risk Management, 2020. *Infographic 155*. Published online July 31. Available at: https://www.gestionderiesgos.gob.ec/wp-content/uploads/2020/07/INFOGRAFIA-NACIONALCOVI-19-COE-NACIONAL-08h00-31072020.pdf. Accessed April 4, 2021.
28. Fiscalía Realiza Allanamientos por Indagación sobre compra de bolsas para Cadáveres en Hospital de Guayaquil, 2020. *El Comercio*. Available at: https://www.elcomercio.com/actualidad/fiscalia-allanamientos-investigacion-compra-bolsas.html. Accessed January 10, 2021.
29. News teleSUR English, 2020. *Ecuador: New Arrests in COVID-19 Supplies Corruption Case*. Available at: https://www.telesurenglish.net/news/Ecuador-New-Arrests-in-COVID-19-Supplies-Corruption-Case-20200713-0010.html. Accessed July 18, 2020.
30. Haque M et al., 2020. Availability and price changes of potential medicines and equipment for the prevention and treatment of COVID-19 among pharmacy and drug stores in Bangladesh; findings and implications. *Bangl J Med Sci 19:* S36–S50.
31. Ortiz-Prado E et al., 2021. Epidemiological, socio-demographic and clinical features of the early phase of the COVID-19 epidemic in Ecuador. *PLoS Negl Trop Dis 15:* e0008958.
32. Malta M, Strathdee SA, Garcia PJ, 2021. The Brazilian tragedy: where patients living at the "Earth's lungs" die of asphyxia, and the fallacy of herd immunity is killing people. *EClinicalMedicine 32:* 100757.
33. La Lista de Vacunados VIP es una de 96 Causas Abiertas en la Pandemia por COVID-19 en Ecuador, 2021. *El Comercio*. Available at: https://www.elcomercio.com/actualidad/salud-lista-vacunados-vip-investigaciones.html. Accessed April 4, 2021.

electronics

MDPI

Article

Data Mining to Identify Anomalies in Public Procurement Rating Parameters

Yeferson Torres-Berru [1,2,3,*] and Vivian F. Lopez Batista [1]

[1] Department of Computer Science and Automatics, University of Salamanca, 37008 Salamanca, Spain; vivian@usal.es
[2] Departamento de Investigación, Instituto Tecnologico Superior Sudamericano, Loja 1101608, Ecuador
[3] Escuela de Ingeniería en Tecnologías de la Información, Universidad Internacional del Ecuador, Loja 1101608, Ecuador
* Correspondence: yeferson.torres11@gmail.com

Abstract: The awarding of public procurement processes is one of the main causes of corruption in governments, due to the fact that in many cases, contracts are awarded to previously agreed suppliers (favouritism); for this selection, the qualification parameters of a process play a fundamental role, seeing as due to their manipulation, bidders with high prices win, causing prejudice to the state. This study identifies processes with anomalies and generates a model for detecting possible corruption in the assignment of process qualification parameters in public procurement. A multi-phase model was used (the identification of anomalies and generation of the detection model), which uses different algorithms, such as *clustering* (K-Means), Self-Organizing map (SOM), Support Vector Machine (SVM) and Principal Component Analysis (PCA). SOM was used to determine the level of influence of each rating parameter, K-Means to create groups by clustering, semi-supervised learning with SVM and PCA to generate a model to detect anomalies in the processes. By means of a case study, four groups of processes were obtained, highlighting the presence of the group "null economic offer" where the values for the economic offer do not exceed 1%, and a greater weight is given to other qualification parameters, which include direct contracting. The processes in this cluster are considered anomalous. Following this methodology, a semi-supervised learning model is built for the detection of anomalies, which obtains an accuracy of 95%, allowing the detection of procedures where the aim is to benefit a particular supplier by means of the qualification assignment parameters.

Keywords: corruption; public procurement; self-organizing map; support vector machine; machine learning; data mining

Citation: Torres-Berru, Y.; López Batista, V.F. Data Mining to Identify Anomalies in Public Procurement Rating Parameters. *Electronics* **2021**, *10*, 2873. https://doi.org/10.3390/electronics10222873

Academic Editors: Amir Mosavi

Received: 15 October 2021
Accepted: 19 November 2021
Published: 22 November 2021

1. Introduction

Favouritism in the state sector is the natural human propensity to privilege friends, relatives and any close and trustworthy person in a public procurement process [1]. When a public purchase is made to favour an entity or company with preliminary agreement with the contracting entity, the bidder with the best offer is not being awarded the contract; in this bad practice, usually in the winner's qualification parameters, lower scores are established than those required for the economic offer, therefore, in order to complete the remaining score, the contracting entity includes additional parameters which privilege a particular participant. In this sense, the economic offer is not the decisive parameter; instead, new technical parameters are used, allowing the bidder with the higher price to win the process [2]; another bad practice is to focus the procedures on bidders who have previously worked with the institutions, requesting previous work experience, excluding new (inexperienced) bidders [3,4]. It is also usual to establish in the section "Other Parameters", specific conditions and requirements with high scores that only an agreed bidder can satisfy, ensuring the disqualification of the rest of the proposals; the "technical specifications" of the object of the tender are not in accordance with the needs and functions

stipulated in the object of the contract, with the aim of directing the procedures to a supplier. Favouritism is also based on the characteristics of the staff that will be part of the project, and that only the agreed company possesses; thus, in the parameter "Compliance with specifications", a certain age, title, experience in a specific area are included, without a legal justification to support such requests. In other words, favouritism causes less purchasing power for the public institution, higher prices that have an impact on the quality of the product and generate unfair competition.

In Ecuador, the Public Procurement System (SERCOP) is in responsible for promoting access to and use of public information, increasing transparency, combating fraud and corruption that could originate from bad practices in public procurement. In 2017, 5.8 billion dollars were transacted through public procurement portal or 19.6% of the general state budget and 5.8% of the Gross Domestic Product (GDP). The participation by government sector was distributed mainly in state administration (28.5%), autonomous municipal governments (21.2%) and public agencies (18%). In 2019, public procurement accounted for 17% of the state's general government budget [5], also showing in Figure 1 a decrease in public investment from 2011 to date.

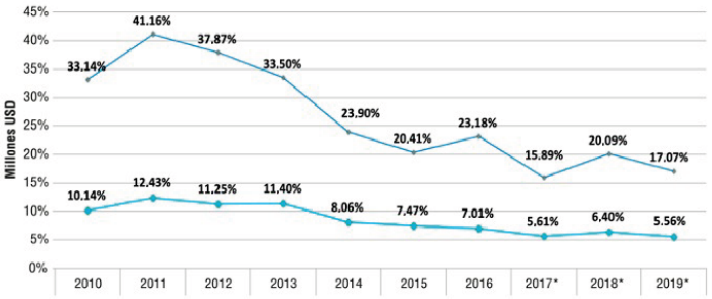


Figure 1. Percentage of public procurement of millons in relation to the General State Budget (public procurement portal Ecuador).

SERCOP hosts documents in PDF format for each contracting process, where data on the specifications are stored: Terms of Reference (TDR), invitations to suppliers, offers submitted, and observations, and in summary, all the documentation generated by the purchase. The types of procurement processes carried out by state entities and available in the SERCOP database are as follows:

- Execution of works.
- Purchase of products and services.
- Consultancy contracting

As part of the information for each process, in parallel to the qualification parameters, the following conditions are considered to evaluate the relationship of each purchase executed, and all the conditions (Table 1) identified are important and must comply with the execution of the contracting by the public entity; therefore, their importance is highlighted:

Table 1. Conditions considered of each process.

Condition	Description
Timeline of the procedure	It emphasises important dates in the process.
Duration of the offer	Item used to determine the number of days the process will remain in effect.
Purchase price	Is the price of the process (purchase), which the institution lists on the public procurement portal.
Type of purchase	The classification used by the institution for the purchase carried out can be: goods, consultancy, work, insurance and service.
Recruitment Types	It is the method used to contract the acquisition is classified in: bidding, quotation, special publication, short list and direct contracting.
Payment method	The forms of payment are: advance payment, remaining value of the contract and at the end of the contract.
Status of the process	Is the state in which the contracting process is currently running two general statuses are obtained: **correct** (to be awarded, awarded, finalised and in execution) and **not executed** (unilaterally terminated, terminated by mutual agreement, cancelled and deserted).

As a technological resource and with the objective of discovering favouritism, Data Mining (DM) has a fundamental role to contribute with its tools and methods to find hidden information in the massive volumes of data [6]. The use of this technique in public procurement is used as a critical tool, facilitating the monitoring of information, as well as the control of contracting processes [7]. Applying DM, it was established that in Sweden 58% of time the bidder who submits the lowest bid is not the winner of the process [8]; in Paraguay [9], using data from 4 years and 47,615 procurement processes, this study estimates, through the construction of a mathematical model, the correlation between the companies and their possibility of obtaining a contract, detecting the existence of a previous relationship between the supplier and the contracting entity, which produces corruption when the procurement is made. SALER [10] applying DM, analyses contracts and groups them by contract object, procurers, amount, number of contracts and total contract amount, determining characteristics of groups with corrupt practices and their relationship to a risk index for each process. The study conducted by Kehler [11] evaluates anomalies in public contracts using Isolation Forest algorithm [12] based on the modifications undergone by the contracts during the process to determine the corruption originated by these modifications to benefit a particular supplier.

With this background, the hypothesis is proposed: it is possible to develop a composed model to identify processes with anomalies in public procurement qualification parameters. The main objective of the work is to generate a model to identify patterns in the awarding of qualifications to public procurement contracts through the use of data mining techniques and then predict contracts where anomalies exist based on the reviewed data with the use of unsupervised learning techniques.

To simplify the reading of this document, after this introduction, Section 2 describes the data, models and techniques used; Section 3 presents the main results obtained, divided into two sub-sections: Section 3.2 related to unsupervised learning and Section 3.3 to semi-supervised learning, as techniques to validate the hypothesis. In the final part, the conclusions of the study are provided.

2. Methodology

After analysing the various approaches existing in the current literature on favouritism that attempt to provide an answer to the problem posed, this section details the proposal of the present work, designed to test the hypothesis based on the CRISP-DM methodology for data mining [13]. In the literature review, it was found that most of the published works

use supervised learning, as contracts with price anomalies are labelled [14]. About 79% of the research corresponds to detection and 21% to prediction. This is not the case in Ecuador, which still lacks labelled data; therefore, in the initial phase of the research, we decided to use unsupervised learning techniques to detect anomalous patterns in contracts.

2.1. Data Set Description

As Ecuador's public procurement does not have an open data website, a web scraping technique was applied [15] on the data provided on the website of the SERCOP (https://www.compraspublicas.gob.ec/ProcesoContratacion/compras/) (accessed on 11 November 2021). Through this technique, the information is obtained on public processes from 2010 until 2020, as well as the documents (attachments) of each process.

We approach our research through an experiment using publicly available datasets (https://bit.ly/PametersCorruption) (accessed on 11 November 2021), and the parameters for the qualification of bids were evaluated in 275,730 public procurement contracts in Ecuador. A total of 21 numeric parameters were assessed to determine the winner of each process, which is detailed in Table 2. The rating parameters vary according to the process, and they are all considered for the evaluation without excluding for the subsequent evaluation of the impact of each parameter on the final score, in addition to the fundamental aspects of the process such as the following: the type of purchase, status of the process and type of procurement.

Table 2. Parameters for qualification of an individual process.

Parameters	Description
General experience	Experience of the bidder in the general domain.
Specific experience	Experience in specific projects in the area of sourcing.
Similar works	Number of similar projects executed by the supplier.
Subcontracting	The supplier is able to partially subcontract the execution of the project.
Financial ratios	The solvency and debt ratios of the participating companies are assessed.
Methodology, Work Plan	Parameters for evaluating the bidder's presentation of the project.
Supply date	Estimated delivery date stated in the offer.
Economic offer	Value submitted by the bidder
Proposed team	Characteristics of the team that executes the work.
Inclusion parameters	They aim to include people and companies with disabilities.
Instruments equipment	Referring to models and brands of the products available.
Specification compliance	Technical product specifications and characteristics
Technical guarantee	Technical product guarantee.
National partnership	Priority is given to international suppliers who partner with local producers.
National SMEs	Priority to national micro-enterprises
Local participation	Priority to suppliers from the place of purchase
Ecuadorian participation	Priority to national companies
Bonus awarded by lottery	Bonus awarded by lot in case of a tie between bidders
Other qualification parameters	Defined by the procuring entity
Variable scoring	According to the requirements presented.
Technology transfer	Added value to processes that are born as a technology transfer from educational institutions.

2.2. Data Pre-Processing

For qualitative data review, it employs a technique proposed by Chu [16] which helps to find errors in the data and to scale or normalise them for use. Firstly, the data set is processed, eliminating erroneous values corresponding to processes with qualification parameters that had errors, as these must add up to a value equal to 100% and in some cases had lower values, such as 98% or bigger, such as 105%.

Using the pandas tool (https://pandas.pydata.org/) (accessed on 11 November 2021), the missing values are replaced, assigning 0 to the null fields, since the same qualification parameters are not met in all the processes. Finally, the data obtained are scaled. As this is unsupervised learning, it is decided to use the entropy measure for each attribute, so that more variability can be obtained. Therefore, the data for the 21 rating parameters are normalised.

2.3. Proposed System

The followed methodology and techniques are summarised in Figure 2. The process is initiated when data are collected from SERCOP, and once the data retrieved on public procurement are processed through web scraping, they are analysed through a multi-phase methodology, which uses different machine learning algorithms for the detection and prediction of favouritism in public procurement such as: *clustering* (K-Means), Self-Organizing Map (SOM), Support Vector Machine (SVM) and Principal Component Analysis (PCA). Following an analysis of different techniques for *clustering*, the K-Means algorithm was chosen [17] to group the data according to the type of recruitment. Leveraging the advantages of class visualisation provided by the SOM was used to identify the impact of every variable, and to compare the *clusters* obtained from K-Means based on distance and density, it can be used to analyse the data for possible *clusters* [18]. This allows for the identification of the clusters where the contracts with possible anomalies are located and is the input for the construction model.

Finally, with a semi-supervised learning model, anomaly detection is performed using PCA and SVM.

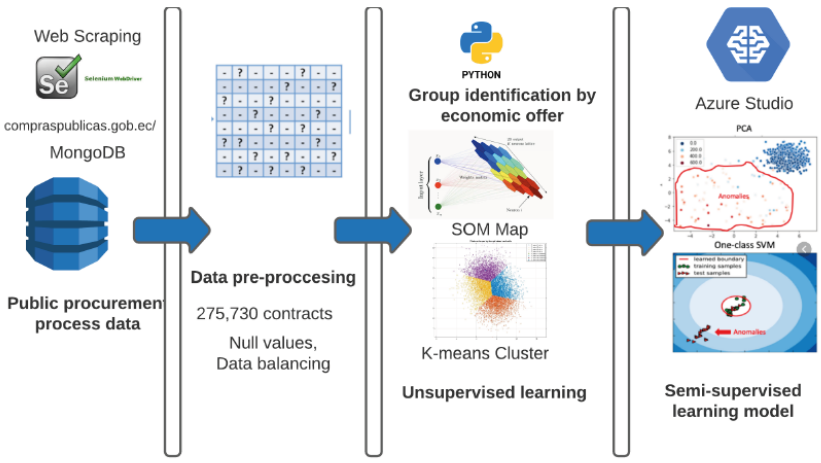


Figure 2. Diagram of the proposed methodology for find anomalies in qualification parameters.

2.4. Training and Learning Phase

The choice of learning algorithms is justified based on the number of data in the data set, the number of parameters to evaluate; starting from metrics such as the Clustering Accuracy (ACC) and the Normalised Mutual Information (NMI), based on the work of [19], it is understood that most of the unsupervised feature selection methods (filter, wrapper

or hybrids) require the specification of hyper-parameters such as the number of features, number of clusters or other parameters inherent to the feature selection technique used by each method, and the quality of the feature extraction of data directly affects the detection performance of SVM. Describing the autocorrelation among data is an important factor that affects the fault detection performance [20]. The use of machine learning techniques to classify public procurement processes according to their qualification parameters and generate the detection model is described.

2.4.1. Self-Organizing Maps

In Table 2, as many as 21 parameters are evaluated to determine the winner of a public process, but these parameters are not repeated in all processes; therefore, it is necessary to determine the main parameters common in most processes, which is why the SOM maps [21] were chosen. A rectangular topology was implemented, consisting of 10 input rows and 10 input columns [18]. The Gaussian neighbor' is selected, and the quality of the SOM map is influenced by the initial weights of the training map [17] we chose random. Finally, the number of training iterations is set to 1000, and finally, two types of metrics Quantification and topographic error were taken into consideration for the evaluation of SOM maps.

2.4.2. Clustering Algorithm

As described in Section 2.3, it is necessary to identify the processes with anomalies in the ratings, which is why clustering is used in combination with SOM maps. The K-Means clustering algorithm makes it possible to analyse data and find groups within that data using some kind of similarity measure, such as Euclidean distance. No one metric of universal similarity works for all cases [22] (depending on the problem itself). Therefore, starting at eight different centroids and using the elbow technique, the optimal number of clusters was determined ($k = 4$), and metrics such as ACC and NMI were evaluated. Once the cluster with anomalies was identified, semi-supervised learning was applied to detect anomalies in public procurement processes.

2.4.3. Support Vector Machine

SVM classifies the data, if the data are linearly separable, SVM classifies it linearly for the training and identification of anomalies with SVM, and the contracts of the groups where the economic offer has a greater weight in determining the winner are considered as normal (class 1), and data that are different can be predicted as anomalies (class 2).

When this version of the algorithm is applied, we use the property [23] *nu*, which allows us to control the balance among the outliers and normal cases, and therefore assigns $nu = [1e-3, 1e-2, 1e-1, 1]$, while the parameter affecting the number of iterations used, when optimising the model, is taken as $epsilon = [1e-4, 1e-3, 1e-2]$. The optimal hyperplanes for machine learning are then determined using a *Hyper-parameters*,the model is trained and evaluated using the ROC and *accuracy* metrics. The values of the minimum and maximum metrics are [0.9, 0.97] equivalent to a very good test.

2.4.4. Principal Component Analysis

The accuracy of PCA-based anomaly detection depends on a good choice of principal components, which is achieved with the use of SOM Maps being the main characteristic for the choice of the algorithm. Distance metrics are applied to identify the cases that represent anomalies; therefore, they are used with a range of parameters (*rank*) and *oversampling* of [2, 4, 6, 8, 10]. Finally, the model is trained using the Score Model and ROC; for this method, 80% of the data is used for training and 20% for testing.

2.5. Tools

For the programming job, the Python language is used with the Selenium test environment. https://selenium-python.readthedocs.io (accessed on 1 November 2021) and Jupyter

(https://jupyter.org/) (accessed on 1 November 2021), in addition to libraries Stick-learn (https://sklearn.org/) for the application of the machine learning algorithms and [24] the Python libraries Minisom https://pypi.org/project/MiniSom/ (accessed on 1 November 2021) and Sompy https://github.com/sevamoo/SOMPY (accessed on 1 November 2021) were used for Self-Organizing Maps .

It also uses the *machine learning* service provided by AZURE (https://studio.azureml.net) for training and testing data sets, due to the size of the data evaluated. It is assessed using the metrics: ROC curves, *accuracy*, *precision*, *FScore*n and *Recall*. The ROC curve shows the ratio between false positives and false negatives.

3. Experimental Results

To build the case study, information was retrieved considering the URL of the purchase process as input, fields such as: description, dates, products, qualification parameters, invitations, documents and questions from the suppliers. Each section was extracted according to its equivalent identification (tag) in HTML through scraping and stored in a non-relational database (MongoDB).

3.1. System Implementation

Figure 3 details the two main phases that composed the developed model, starting with the identification of contracts with anomalies using unsupervised learning with *K-Means* once the internal validation of the cluster was accomplished, and the following results are obtained: four groups, of which in in three, the economic offer is expected to determine the winner of the process and in one not; at the same time, the main parameters that have the greater influence on the determination of the winner of the process are evaluated with the use of *SOM maps*; therefore, two types of contracts are identified: regular contracts and contracts with anomalies.

With the identification of the groups and the influence of the variables on the rating, the following is required for the second phase of the model the detection of anomalies with the use of SVM and PCA; in the second phase, training is performed with the metrics described in the methodology to avoid overtraining, and the model is evaluated with data not present in the model (in this case, 2021 data). Therefore, it is suggested that the accuracy of the model obtained is between 85% and 97%.

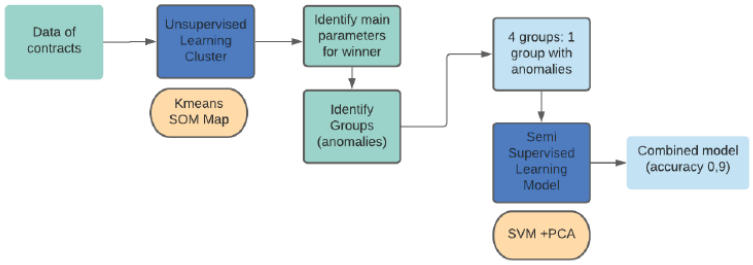


Figure 3. Detailing two main phases that composed the model.

3.2. Unsupervised Learning Cluster

Using the SOM, the main parameters influencing the process rating and their influence on the cluster classification are identified.

Figure 4 shows the influence of each rating parameter on the cluster, with those in blue having the least influence and the colour scale representing the greatest influence; therefore, the main rating parameters found by using SOM Maps are: economic offer, specification compliance, other qualification parameters, general experience, specific experience, proposed team, technical guarantee, instruments and equipment and similar works.

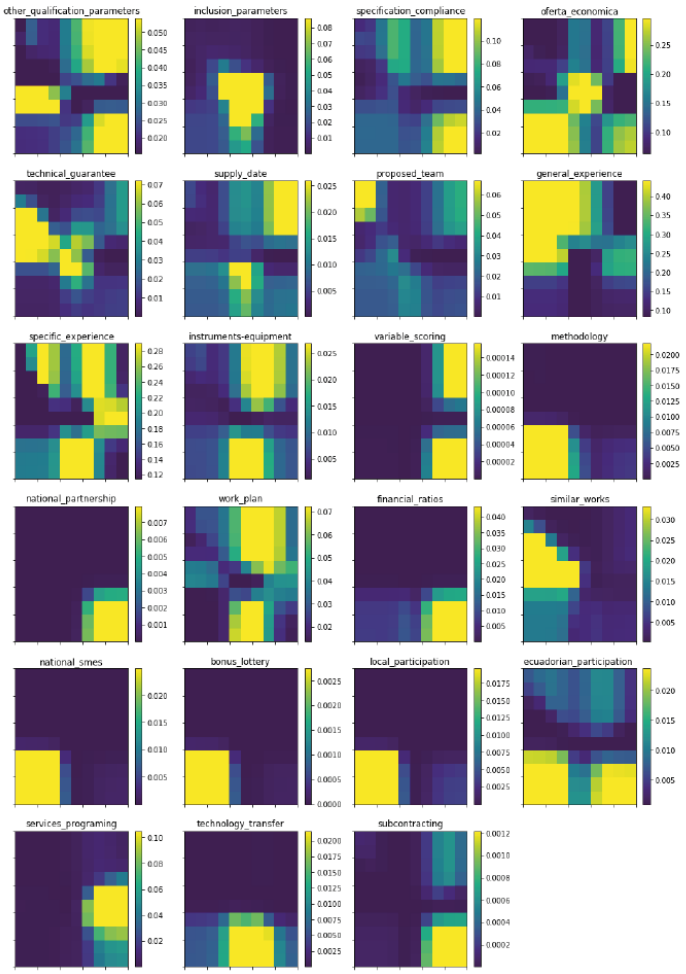


Figure 4. Evaluation of influence of qualification parameters with SOM.

A heat map (Figure 5) shows the assignment of the processes to each cluster represented in green, light blue, orange and red for each cluster, and the dark-blue values represent a small number of elements and are assigned to the nearest cluster. A colour scale from zero (withe) to 60,000 (dark green) represents the number of elements associated with the cluster.

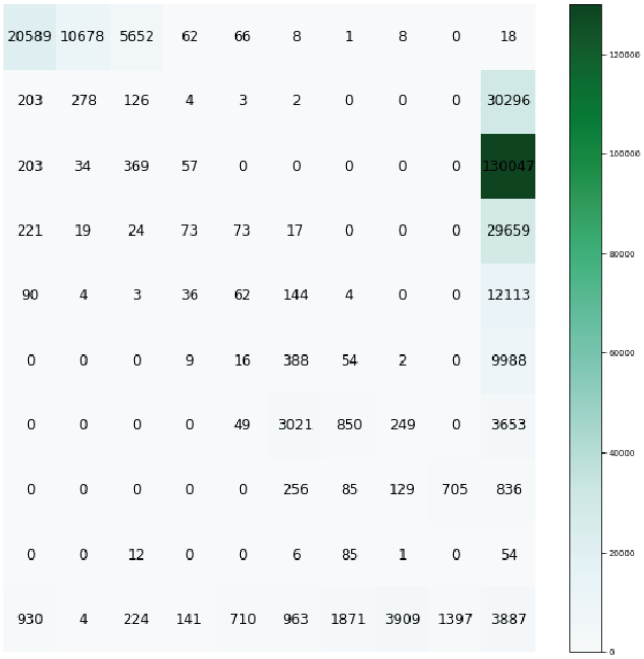


Figure 5. Heat map of distances to individual clusters.

Figure 6 shows the evolution of the quantisation and topographic error with 1000 iterations, observing that from iteration 600 it stabilises and reaches optimal values for the model, obtaining a quantisation error of 0.2878 and a topographic error of 0.30796, ensuring in this way a correct reliability of the maps.

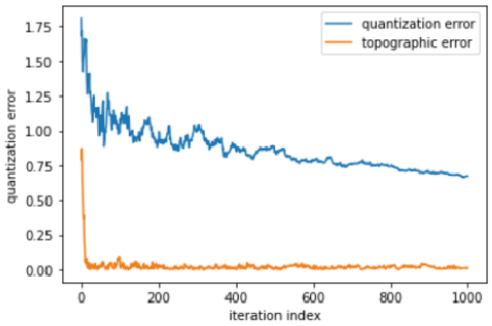


Figure 6. Iterations index.

By applying the K-Means algorithm with four centroids, four different clusters were obtained. Table 3 shows the 12 main characteristics associated with the variables related to the type of purchase. For example, *general experience* is predominant in the *cluster 3*, *specific experience* is predominant in the *cluster 4*, and other qualification parameters and specification compliance are predominant in *cluster 1*. The last row details the number of records (processes) belonging to each *cluster*.

Taking into consideration the state of the process, it can be classified as follows: correct or non-executed, the percentage of non-executed processes was 4.71% in cluster 1, 15.80%

in cluster 2, 39.93% in cluster 3 and 26.69% in cluster 4.0%. It is therefore determined that: in the cluster 1, the number of non-executed processes is under the average, and compliance with specifications and the economic offer have a greater influence. In cluster 2, the number of non-executed processes is equal to the average and is more influenced by the economic offer and an equal distribution among the other variables. The cluster 3 is below the average number of non-executed processes and is more influenced by overall experience and economic offer. Finally, at the cluster 4, the number of non-executed processes is above average, and general experience, specific experience and the work plan are more influential. This indicates that cluster 4 is the cluster with "anomalies".

Table 3. *Clusters* K-Means.

Parameter	1	2	3	4
Instruments-equipment	0.18%	0.21%	1.63%	4.94%
Specification compliance	46.66%	1.61%	5.22%	3.09%
Other qualification parameters	2.55%	7.16%	4.71%	5.61%
Specific experience	1.83%	10.73%	2.66%	51.03%
Similar works	0.27%	0.78%	1.09%	0.20%
General experience	2.12%	4.37%	46.46%	16.36%
Economic offer	33.99%	47.43%	18.63%	0.45%
Proposed team	2.68%	5.75%	2.93%	1.12%
Technical guarantee	2.22%	3.72%	3.42%	0.59%
Supply date	5.21%	5.85%	2.79%	0.50%
Methodology and work plan	0.17%	0.95%	6.00%	12.70%
Number of records	81,261	71,959	34,141	88,358

Figure 7 shows the influence of the six main qualification parameters, which are related to the economic offer. It can be seen graphically, the null participation of the Economic offer in cluster 4, a moderate involvement in the cluster 1, high participation in cluster 2 and weak participation in cluster 3. Therefore, for a better understanding for the reader, in the next sections, the clusters are renamed based on the influence of the economic offer and are as follows: Cluster 1 = Moderate economic offer, Cluster 2 = High economic offer, Cluster 3 = Low economic offer, Cluster 4 = Null economic offer

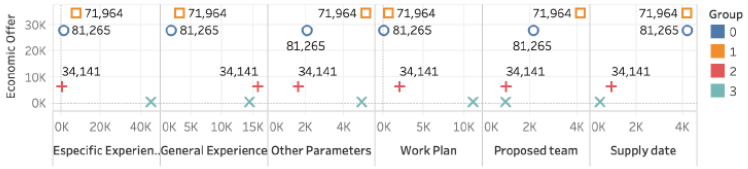


Figure 7. Relationship of the Economic offer to other qualifying parameters.

3.3. Cluster Analysis for Process Variables Not Involved in Purchasing Qualifications

The clusters obtained are matched with the type of purchase made and the type of procurement with which the process was performed.

With respect to the relationship between the rating parameters and the type of purchase made, Table 4 shows that the *"Moderate economic offer" cluster*, the Economic Offer rating is higher for the purchase of products and services, and as highlighted in the table, the compliance with technical specifications is higher for the purchase of services (specifications are usually given for products).

In the *"High economic offer" cluster* , the predominant procurement of products, services and works, with a high percentage is given to Economic offer in all processes; however, in works and services processes, a high value is given to experience between 15% and 10%, respectively, and in the procurement of services a value of 11% is assigned to other parameters. *"Low economic offer"* cluster purchase of services, products and consultancy

predominates, in the respective order of the main qualification parameters, the General experience, Economic offer and to a lesser extent the specific experience. Finally, in the *"Null economic offer" cluster*, the purchase of consultancy and services predominates, with a high influence of the parameters of qualification of specific experience, general experience and compliance with specifications.

Table 4. *Clusters* related type of purchase.

Type of Purchase	Equipment	Technical Guarantee	Work Plan	Supply Date	Proposed Team	Other Parameters	General Experience	Specification Compliance	Specific Experience	Economic Offer	Number of Process
"Moderate Economic offer" Cluster											
Product	0.1%	2.7%	0.0%	5.7%	3.4%	2.0%	1.24%	46.8%	0.9%	35%	46,928
Consultancy	0.4%	1.1%	1.4%	4.6%	3.5%	2.7%	4.6%	47.3%	7.2%	23.4%	102
Work	0.2%	1.7%	0.1%	4.1%	7.6%	2.1%	2.5%	43.3%	3.6%	28.6%	388
Assurance	0.0%	0.4%	0.1%	0.2%	0.5%	8.6%	4.6%	34.2%	3.8%	33.2%	823
Service	0.2%	1.3%	0.3%	4.7%	1.5%	3.1%	3.3%	47.3%	3.0%	32.7%	29,187
"High Economic offer" Cluster											
Product	0.2%	8.9%	2.3%	11%	10.5%	8.9%	2.2%	1.79%	2.9%	46.8%	23,024
Consultancy	0.9%	1.6%	3.0%	10.2%	8.9%	16.0%	1.9%	1.02%	8.5%	37.6%	59
Work	0.0%	0.0%	0.0%	0.4%	2.0%	1.9%	5.5%	0.3%	17.7%	47.6%	14,814
Assurance	0.0%	0.2%	0.0%	0.0%	0.8%	11.8%	9.1%	2.4%	15.1%	50.4%	2,313
Service	0.3%	2.3%	0.5%	7.0%	5.7%	10.4%	4.0%	2.2%	10.5%	47.3%	20,373
"Low Economic offer" Cluster											
Product	0.2%	6.9%	0.3%	4.8%	5.0%	2.4%	45.5%	6.34%	0.4%	23.4%	9,932
Consultancy	5.4%	0.0%	22.5%	0.0%	0.0%	9.8%	48.5%	0.24%	8.2%	0.2%	5,799
Work	1.1%	1.3%	0.5%	4.3%	3.7%	1.3%	42.6%	7.58%	2.2%	23.8%	384
Assurance	0.0%	1.2%	0.0%	0.8%	2.0%	14.3%	32.0%	11.8%	4.4%	19.6%	72
Service	0.3%	3.0%	0.7%	2.0%	2.9%	3.4%	46.4%	7.10%	0.9%	25.7%	14,563
"Null Economic offer" Cluster											
Product	0.6%	4.6%	0.8%	3.1%	7 %	3.3%	9.9%	**17.8**%	48.0%	1.3%	4,866
Consultancy	5.7%	0.0%	15.0%	0.0%	0.0%	5.8%	18.0%	0.05%	51.7%	0.0%	49,523
Work	0.9%	0.8%	0.8%	1.6%	7.3%	1.9%	6.6%	7.10%	45.7%	11.5%	458
Assurance	0.0%	1.1%	0.0%	1.1%	2.9%	18.6%	10.3%	13.8%	37.6%	9.7%	34
Service	1.3%	2.7%	1.9%	2.6%	5.3%	5.4%	8.6%	17.0%	47.6%	2.3%	9,895

The type of procurement performed influences the qualification parameters in Table 5; therefore, we observe that in the cluster "Moderate Economic offer" special publication processes predominate with 93.8% of the total number of processes in this cluster and 47.35% impact of compliance with specifications as a qualification parameter, while in the cluster "High economic offer", the quotation and special publication processes predominate. In the cluster "Low Economic offer", we have only direct contracting and special publication processes, with the latter being predominant. Finally, cluster "Null Economic offer" contains direct contracting processes and special publication highlighting the influence of specific experience reaching up to 60% of the total qualification of the process.

Table 5. *Clusters* Type of purchase.

Recruitment Type	Equipment	Technical Guarantee	Work Plan	Supply Date	Proposed Team	Other Parameters	General Experience	Specification Compliance	Specific Experience	Economic Offer	Number of Process
"Moderate Economic offer" Cluster											
Quotation	0.2%	1.6%	0.1%	1.7%	1.3%	1.9%	3.0%	30.7%	3.0%	28.9%	2,423
Bidding	0.0%	0.4%	0.0%	0.3%	0.5%	8.2%	4.6%	34.8%	3.8%	33.2%	1,105
Special	0.1%	2.2%	0.1%	5.4%	2.7%	2.4%	2.0%	47.4%	1.6%	34.3%	76,216
"High Economic offer" Cluster											
Quotation	0.1%	0.1%	0.0%	0.4%	1.6%	1.9%	6.5%	1.5%	16.3%	48.0%	27,233
Bidding	0.1%	0.1%	0.0%	0.5%	1.4%	2.9%	6.4%	0.8%	22.8%	49.0%	5,950
Assurance bidding	0.0%	0.2%	0.0%	0.1%	0.8%	12.0%	9.3%	2.4%	14.7%	50.1%	2,913
Special	0.2%	7.3%	1.8%	11.0%	10.0%	11.0%	1.9%	1.7%	4.0%	46.5%	35,570
"Low Economic offer" Cluster											
Direct contracting	5.6%	0%	22.9%	0.0%	0.0%	9.9%	48.6%	0.0%	8.1%	0.0%	8,120
Special	0.3%	4.6%	0.6%	3.5%	3.7%	3.0%	46.2%	6.8%	0.8%	24.8%	24,800
"Null Economic offer" Cluster											
Direct contracting	5.9%	0.0%	16.0%	0.0%	0.0%	6.0%	18.1%	0.0%	50.3%	0.0%	61,722
Short List	5.2%	0.0%	9.0%	0.0%	0.0%	2.5%	17.2%	0.0%	60.9%	0.0%	8,960
Special	1.1%	3.3%	1.8%	2.6%	5.9%	4.6%	9.3%	17.0%	47.8%	1.9%	15,536

3.4. Semi-Supervised Learning Model

As previously described in the cluster called "Null Economic offer", processes with anomalies are identified. For the detection of anomalies, the processes associated with the clusters are defined as "normal", where the economic indicator is respected as a preponderant factor for the qualification and determination of the winner of the process. For semi-supervised learning, a training data set (80%) and a test data set (20%) are separated. As detailed in the methodology, a semi-supervised learning model is applied using SVM and PCA that can be applied in the evaluation of the regression model and for the detection of anomalies in the processes. As metrics to evaluate the success of the applied algorithms, we use: ROC curves, where the blue line represents SVM and the red line PCA, which allows us to evaluate the influence of each technique on the model Figure 8. Analysing the results, we have that the precision of the model is (0.9%) and *accuracy* is (0.92%), indicating an acceptable detection rate.

We can observe 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. Table 6 indicates the evaluation metrics for each technique in the detection of anomalies model.

Table 6. Models metrics.

Technique	Precision	Recall	AUC	Accuracy	FScore
SVM	0.95%	0.92	0.90	0.92	0.96
PCA	0.92%	0.89	0.91	0.90	0.93

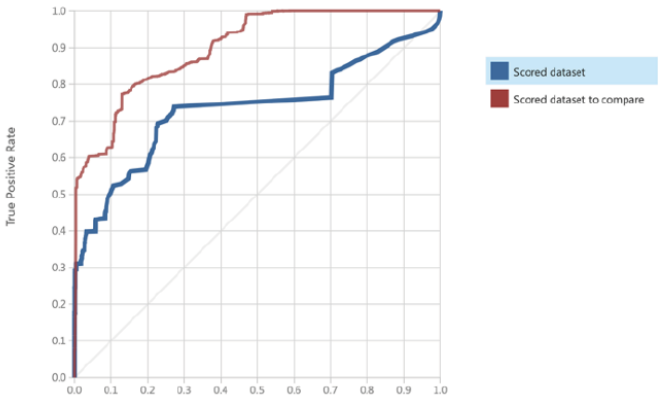


Figure 8. Semi-supervised models evaluation.

4. Discussion and Conclusions

With the experimental work, we have been able to verify the different phases of the proposed methodology to identify processes with anomalies and generate the corruption detection model. With the SOM algorithm, the main parameters involved in the qualification of winning bidders in a public procurement were identified. The K-Means algorithm allowed the identification of the three main groups where Economic Offer represented the main scoring parameter and also a group, *"Null Economic offer" Cluster*, where only 0.45% of the total rating was considered out of 100%. In this group, "other parameters" were evaluated with the greatest weight, with direct contracting, shortlisting and special publications predominating. Regarding the type of purchase, most of the purchases in this cluster are "Consultancies". It is therefore concluded that 88,358 (equivalent to 32.11%) of the processes evaluated could present anomalies in the evaluation parameters for the adjudication of contracts.

Based on the findings ("Null Economic Offer" cluster) obtained from the use of unsupervised learning, an anomaly detection model based on SVM and PCA was developed, obtaining results higher than 90% reliability; therefore, we can verify the hypothesis that guides this research.

The results of the application of the model created, in the case study, allow us to be optimistic. We consider, that through the use of data mining, anomalies can be identified, and new corruption cases can be detected. Specifically, in the definition of qualification parameters in a public procurement process which does not consider the Economic offer and causes prejudice to the government, permitting one to indicate in which cases the qualification parameters are correctly established and in which cases they are not. Experimental results are in concordance with the work of Hyytinen et al. [8], since the municipalities have the highest number of cases with anomalies in the qualification of contracts. The bidder with the lowest economic offer does not win but presents better results in terms of evaluation metrics, due to the machine learning techniques used. It also shows a difference in results with the SALER platform [10] which, while considering various parameters such as relationships between companies, does not rank contracts by the value of the economic offer in the qualification. The model is consistent and demonstrates what the previously reviewed literature points out [2], in that in order to favour certain suppliers, the contracting entity lowers the qualification of the economic offer so that the supplier with certain "special" conditions wins the process and not the provider that submits the most beneficial offer for the state. This research shows that with the use of data mining techniques, this model can be applied in several countries because in each public procurement process, qualification parameters are established to determine the winner, considering that the most important thing is to identify the processes with anomalies in the qualification, in order

to adjust the model. This work represents a breakthrough in corruption research with technological tools in Latin America because as already defined in [14], there has been no progress except for in three countries.

To continue with the present work, it is important to determine the present findings with the SERCOP portal, in addition to providing a base of processes with anomalies in their qualification, new techniques for supervised learning RandomForest, Convultional networks, etc., or new combined models can be tested to determine future anomalies, such as those of cluster 4.

5. Future Work

As a future line of work, it is intended to integrate the *deep learning* in the methodology with natural language processing for the classification of contractors and relations with entities, evaluating award times. In addition, it is planned to build a *framework* that evaluates, detects and helps in the prediction of favouritism in public procurement processes.

Author Contributions: Conceptualization, Y.T.-B. and V.F.L.B.; methodology, Y.T.-B. and V.F.L.B.; formal analysis, Y.T.-B.; investigation, Y.T.; writing—original draft preparation, Y.T.; writing—review and editing, Y.T.-B. and V.F.L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are available in https://bit.ly/PametersCorruption (accessed on 1 November 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bramoullé, Y.; Goyal, S. Favoritism. *J. Dev. Econ.* **2016**, *122*, 16–27. [CrossRef]
2. Martinez Fernandez, J.M. Transparencia Versus Corrupción en la Contratación pública. Medidas de Transparencia en Todas las Fases de la Contratación Pública Como Antídoto Contra la Corrupción. 2015. Available online: https://dialnet.unirioja.es/servlet/dctes?codigo=50035 (accesed on 1 October 2021)
3. Cordova Vinueza, J.; Vaca Ojeda, P.; Hernandez Jaramillo, M. *Las Compras Gubernamentales como Política Pública*; Servicio Nacional de Contratación Pública-SERCOP: Quito, Ecuador, 2015; pp. 43.
4. Dávid-Barrett, E.; Fazekas, M. Grand corruption and government change: an analysis of partisan favoritism in public procurement. *Eur. J. Crim. Policy Res.* **2020**, *26*, 411–430. [CrossRef]
5. Servicio Nacional de Contratacion Publica del Ecuador. Análisis Anual de Contratación Pública. 2020. Available online: https://portal.compraspublicas.gob.ec/sercop/wp-content/uploads/2020/01/analisis_anual_2019_2.pdf (accesed on 1 October 2021)
6. Hermawati, F.A. Data Mining . *Min. Massive Datasets* **2005**, *2*, 5–20. [CrossRef]
7. Ferreira, I.; Camões, P.J.; Cunha, S.; Amaral, L.A. Electronic platforms and transparency in public procurement. In Proceedings of the 30th International Business Information Management Association Conference, IBIMA 2017-Vision 2020: Sustainable Economic Development, Innovation Management, and Global Growth, Madrid, Spain, 8–9 November 2017; Volume 2017, pp. 3898–3906.
8. Hyytinen, A.; Lundberg, S.; Toivanen, O. Politics and Procurement: Evidence from Cleaning Contracts. *SSRN Electron. J.* **2011**, 233; [CrossRef]
9. Auriol, E.; Straub, S.; Flochel, T. Public Procurement and Rent-Seeking: The Case of Paraguay. *World Dev.* **2016**, *77*, 395–407. [CrossRef]
10. Alzate, C.; Monreale, A.; Assem, H.; Bifet, A.; Sandra Buda, T.; Caglayan, B.; Drury, B.; García-Martín, E.; Gavaldà, R.; Kramer, S.; et al. *SALER: A Data Science Solution to Detect and Prevent Corruption in Public Administration*; Springer: London, UK, 2019; pp. 103–117. [CrossRef]
11. Kehler, M.E.K.; Paciello, J.; Fernandez, J.I.P. Anomaly Detection in Public Procurements using the Open Contracting Data Standard; In Proceedings of the 2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG), Buenos Aires, Argentina, 22–24 April 2020.
12. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the IEEE International Conference on Data Mining, ICDM, Sorrento, Italy, 15–19 December 2008; pp. 413–422. [CrossRef]
13. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*; Springer: London, UK, 2000; pp. 29–39.
14. 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 *Communications in Computer and Information Science*; Springer: London, UK, 2020; Volume 1194, pp. 254–268. [CrossRef]

15. Saurkar, A.V.; Gode, S.A. An Overview On Web Scraping Techniques And Tools. *Int. J. Future Revolut. Comput. Sci. Commun. Eng.* **2018**, *4*, 363–367.
16. Chu, X.; Ilyas, I.F.; Krishnan, S.; Wang, J. Data cleaning: Overview and emerging challenges. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*; Association for Computing Machinery: New York, NY, USA, 2016; Volume 26, pp. 2201–2206. [CrossRef]
17. Akinduko, A.A.; Mirkes, E.M. Initialization of Self-Organizing Maps: Principal Components Versus Random Initialization. A Case Study. 2012. Available online: http://xxx.lanl.gov/abs/1210.5873 (accesed on 1 October 2021).
18. Ultsch, A.; Mörchen, F. *ESOM-Maps: Tools for Clustering, Visualization, and Classification with Emergent SOM*; Technical Report Dept. of Mathematics and Computer Science, University of Marburg: Marburg, Germany, 2005; pp. 1–7.
19. Solorio-Fernández, S.; Carrasco-Ochoa, J.A.; Martínez-Trinidad, J.F. A review of unsupervised feature selection methods. *Artif. Intell. Rev.* **2020**, *53*, 907–948. [CrossRef]
20. Guo, J.; Li, T.; Li, Y. SVM Based on Gaussian and Non-Gaussian Double Subspace for Fault Detection. *IEEE Access* **2021**, *9*, 66519–66530. [CrossRef]
21. Kohonen, T. The Self-Organizing Map. *Proc. IEEE* **1990**, *78*, 1464–1480. [CrossRef]
22. Kapil, S.; Chawla, M. Performance evaluation of K-means clustering algorithm with various distance metrics. In Proceedings of the 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, ICPEICES 2016, Delhi, India, 4–6 July 2016. [CrossRef]
23. Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques; *Informatica* **2007**, *160*, 3–24.
24. Vettigli, G. MiniSom, a minimalistic and Numpy based implementation of the Self Organizing Maps Giuseppe. *J. Open Source Softw.* **2021**, 1–2. Available online: http://xxx.lanl.gov/abs/1806.02199 (accessed on 1 October 2021).

Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.035367
Article

Tech Science Press

A Data Mining Approach to Detecting Bias and Favoritism in Public Procurement

Yeferson Torres-Berru[1,2,*], Vivian F. Lopez-Batista[1] and Lorena Conde Zhingre[3]

[1]University of Salamanca, Salamanca, 37008, Spain
[2]Instituto Tecnológico Superior Sudamericano, Loja, 1101608, Ecuador
[3]Universidad Internacional del Ecuador, Loja, 1101608, Ecuador
*Corresponding Author: Yeferson Torres-Berru. Email: yeferson.torres11@gmail.com
Received: 18 August 2022; Accepted: 29 November 2022

Abstract: In a public procurement process, corruption can occur at each stage, favoring a participant with a previous agreement, which can result in over-pricing and purchases of substandard products, as well as gender discrimination. This paper's aim is to detect biased purchases using a Spanish Language corpus, analyzing text from the questions and answers registry platform by applicants in a public procurement process in Ecuador. Additionally, gender bias is detected, promoting both men and women to participate under the same conditions. In order to detect gender bias and favoritism towards certain providers by contracting entities, the study proposes a unique hybrid model that combines Artificial Intelligence algorithms and Natural Language Processing (NLP). In the experimental work, 303,076 public procurement processes have been analyzed over 10 years (since 2010) with 1,009,739 questions and answers to suppliers and public institutions in each process. Gender bias and favoritism were analyzed using a Word2-vec model with word embedding, as well as sentiment analysis of the questions and answers using the VADER algorithm. In 32% of cases (96,984 answers), there was favoritism or gender bias as evidenced by responses from contracting entities. The proposed model provides accuracy rates of 88% for detecting favoritism, and 90% for detecting gender bias. Consequently one-third of the procurement processes carried out by the state have indications of corruption and bias. In Latin America, government corruption is one of the most significant challenges, making the resulting classifier useful for detecting bias and favoritism in public procurement processes.

Keywords: Favoritism; bias; natural language processing; Word2vec; sentiment analysis; word embeddings

1 Introduction

The Ecuadorian government has been working with new technologies to integrate its processes, one of which is public procurement. According to Thi Nguyen et al. [1] several reports propose that the modernization of the state be undertaken as indicated by organizations such as the United Nations and the

Organization for Economic Cooperation and Development (OECD). In order to simplify processes, improve productivity, gain transparency, reinforce good governance, and strengthen coordination and communication in institutions, the development and implementation of electronic initiatives were promoted. For community participation in government to increase, public services should be improved, policies should be effective, and citizen confidence must be increased [2]. A public procurement system is used by institutions to procure goods, services, and works for the community; therefore, it should meet three essential objectives: human development, efficiency, and transparency.

The 2008 constitution of Ecuador and the National Public Procurement System organic law created a set of operational and regulatory conditions for government purchases, modifying the scenarios for government purchases. Monitoring and controlling processes are key to achieving this productivity improvement. The National Public Procurement Service [3] promotes transparency and publicity of contracting acts in 2018. For example, it is prohibited to benefit from purchase orders generated through the "inclusive dynamic catalogue". Public authorities decide to purchase goods, works, and services, but It is possible at this point that the decision does not follow a policy rationale or an existing need but rather the desire to channel benefits to an individual or/and organization.

It has been shown that the promotion of transparency and publicity of acts in contracting, should focus on an analysis to identify risks of corruption in public purchases, which requires advanced management of substantial amounts of information with typical techniques of big data [4], sentiment analysis [5], advanced analytics and Artificial Intelligence (AI) [6]. In previous research, we have highlighted the need to use different machine learning techniques to assess favoritism and therefore corruption in public procurement. The literature on this topic has been reviewed, and we have proposed a model which combining supervised and unsupervised learning, that allowed the detection of corruption in the allocation of public procurement contracts [7–10]. By analyzing text generated by each public process, we want to extend their study to finding other defined terms, such as gender bias and favoritism, in questions and answers.

As a result of corruption in public procurement, Transparency International estimates costs ranging from 20%–25%, and sometimes up to 40%–50% [11]. Some issues encountered include various kinds of corruption like bribery, collusion, embezzlement, misappropriation, fraud, abuse of discretion, and favoritism. Based on our previous theoretical discussions, we chose to focus on favoritism in this paper because only eight of the 147 reviewed articles address this type of corruption [10].

Favoritism is defined as the natural human propensity to favor friends, relatives and any close or reliable person involved in a public process [12] once a tender process is started the tender provider can still dissuade competitive bidders by keeping the contracting process non-transparent and by circulating private information to favor a particular supplier. For this reason, the aim of machine learning systems is to analyze the relationship between the behavior of providers and contracting entities. The unilateral modification consists of the authority to modify administrative contracts for reasons of public interest [13]. Conflict of interest is a clear red flag for corruption, this can be due to family, business o political ties. On multiple occasions, the legal instrument of the modification is used to articulate new contracts with fraud at the beginning of the bidding process [14]. On other occasions, the modifications double or even triple the initial price of the contract. Also, during the "*questions, answers, and clarifications*" stage in a public process, the entities knowing the deadline for making clarifications do them at the last minute; leaving all the participants without the necessary time to generate an offer with the requested changes automatically benefiting the bidder who was given the information beforehand, to direct the qualification towards conditions that only one proponent will present. The entity after holding conversations with the bidder which it wants to be the winner, modifies certain parameters in the "questions, answers and clarifications" stage, so that the offer of this supplier obtains maximum scores and this way to ensure that he is the person who wins the procedure. There is a risk that evaluation criteria aren't clearly stated in

tender documents, leaving no grounds to justify awarding the tender to a corrupt supplier. Therefore, red flags are accumulations of clues that suggest corruption.

Section 2 shows the related works to highlight the importance of the current study, Section 3 details the techniques and methods used also describes the dataset and the technological tools such as programming languages. Section 4 shows the results obtained from the experimentation, the following section discusses the results and conclusions and finally presents future work of the paper.

2 Related Works

Natural language processing (NLP) is a branch of Artificial Intelligence (AI) that enables machines to understand the human language and is used in many fields. Pant et al. [15] in their research address NLP to detect subjective biases through models focused on a classifier based on Bidirectional Encoder for Transformers (BERT) which uses a bidirectional analysis model. This means that words that are both to the right and the left of a keyword are being analyzed. The method takes advantage of automatic denoising auto-encoders and a token-weight loss function considering bias when associations between gender and certain concepts are captured in processes, with the use of word embedding or in the parameters of the model. Bias is introduced into natural language through specific words and phrases. It can also be defined as the inclination or prejudice of a decision made by an AI system, unfairly for or against a person or group. For example, gender bias is the preference of one gender over another [16], which manifests itself in particular contexts or between social roles mainly, having prejudices against the less dominant gender. This occurs in [17] when there is a correlation between the representation of a neutral word concerning a word with gender (male or female) through a calculation called indirect inequality and the degree of bias with any gender [18].

Modrusan et al. [6] utilize AI to develop a model capable of identifying and extracting terms that determine the prediction of suspicious offers based on the quality and volume of the data under analysis, with a classification accuracy of 0,76%. Only 1,500 documents provided by the competent authority to the authors were reviewed. Hamishu et al. [19] use a model based on Bag-of-Words (BoW) designed to detect and classify fraud in tax advances, through the identification of frequently used words used by scammers, when processing substantial amounts of data. Also, it is important to note that these two parsers are based on English grammar and syntax. The NLP allows the implementation of systems for the detection of fraud taking text classification, information retrieval and information extraction as a reference.

The public procurement marketplace is an important instrument for achieving the inclusiveness of women and improving their purchasing power because it has the potential to promote their empowerment as well as its inclusion in non-traditional sectors such as science and technology by providing funds for women's companies to maintain and diversify employment [20]. For this reason, it is extremely necessary to detect if there are gender biases within this sector, as the complexity of natural language constructs makes this task even more challenging [21]. The work of Akhter et al. [22] indicate that more research is required on the evaluation of abstraction techniques for text summaries in Spanish especially novel techniques based on transformers. The results obtained by the authors suggest that Word2vec word embeddings achieved the best results based on the ROUGE-1 [23] and BLEU [24] metrics. As Tunyan et al. [25] describe, NLP can be used to detect subjective biases at the level of sentences, sentiments, opinions, as well as bias mitigation, but the vast majority of recent work on bias is focused on identifying hate speech on social media. In effect, as a result of the selection bias, these studies fail to differentiate those indicators that can effectively distinguish corrupt public procurements from the rest. However, reducing bias is a new challenge for the NLP and AI research community.

Traditionally, the information generated by suppliers and public entities oriented to the identification of suspicious activities of corruption in a public procurement process is not used, given that the complaints

generated by suppliers or suspicious responses made by financial institutions are not reviewed or taken into account by any controlling entity. This causes uneasiness on the part of suppliers who feel disadvantaged by the state, believing that there is a predilection (favoritism) towards a certain supplier or group of suppliers. Consequently, this research focuses on developing AI based model, which uses NLP techniques for detecting suspicious processes in the Spanish Language. The research objective is to develop a machine-learning model that recognizes gender biases and favoritism in the public procurement process for Spanish speakers.

The main novelty of the work is the evaluation of the validity of the evidence extracted from the questions and answers made in a public purchase process to detect biased (fraudulent) purchases in the Spanish Language, based on an evaluation of equal conditions, in terms of gender and chances of winning the process.

3 Materials and Methods

In this section, the detection of favoritism and gender bias in public procurement for Spanish speakers is described. This model complements the methodology for detecting corruption in public purchases proposed by the authors in previous work [10]. For the evaluation, the following indicators are considered: the publication date of the call for vendors' proposals, the selection and contract decisions, the prices, the number of bids submitted, the kind of institutions' responsibilities, and the origin of the bidders.

In addition, red flags related to the bidding process, such as a really short time between the call for tenders and the submission of the bid, or requirements that are excessively difficult to achieve by the participant in the selection processes. Moreover, suspicious indicators related to the results of the selection processes, such as the high participation of a supplier or contractor in the public contracts of a given institution. On the other hand, the model runs an analysis of the opinions of the questions and answers made by the contracting entities to complement the evaluation of favoritism.

3.1 Dataset Description

The experimental work 303,076 of public procurement processes executed from 2010 to 2020 were evaluated. Data collection is explained in our previous works [7,8]. The dataset in each process includes questions or clarifications made and answered by the contracting entity and vendors. In total 1,009,739 questions and answers are evaluated with the following characteristics:

- Questions issued by the contracting party where necessary aspects for the bidding are consulted or it is intended to report any novelty in the process.
- Responses issued by the entity that conducts the contracting, which serve to clarify the doubts of the contracting parties. Furthermore, it can be used to issue clarifications or modifications to the contract. It is understood that the answers and clarifications must be partial and include the references. Therefore, there should not be biases or sentiments for or against the provider.

In Fig. 1 we show an example of the questions and answers made in each process. The texts are written in Spanish Language and within each purchase process additional information is added, such as: code, town, amount, date, time, web address of the process, the status of the process, payment method, type of purchase and the type of hiring. The questions, answers, dates, and attached files can be seen in the questions section.

3.2 Proposed Methodology

Fig. 2 shows a proposed model for detecting favoritism and gender bias in public procurement processes. The model is trained based on the questions and answers made by the suppliers and contracting entities. As an initial step, the pre-processing of the data is conducted as explained in Section 3.3, for its treatment using Word2vec [26] based on word embeddings [27] (Section 3.4) What is

more, sentiment analysis is performed to detect favoritism, with a pre-trained classifier Balance Aware Dictionary for Sentiment Reasoning (VADER) at the same time. It is evaluated if there is any bias in the answers given by the institutions, to benefit a specific type of contracting party.


```
"CODIGO": "SIE-002-DPLD-2020",
"CANTON": "QUITO",
"MONTO": -1,
"FECHA_HORA": "5/8/20 13:00",
"LINK": "https://www.compraspublicas.gob.ec/ProcesoContratacion/compras/PC/informacionProce
"Estado del Proceso:": "Desierta\nRazon: Art. 1.- DECLARAR DESIERTO El Proceso de CONTRATAC
"Forma de Pago:": "Anticipo: 0% Saldo: Pago contra entrega de bienes obras o servicio 100.0
"Tipo Compra:": "Servicio",
"Tipo de Adjudicación:": "Total",
"Tipo de Contratación:": "Subasta Inversa Electrónica",
"preguntas": [
  {
    "pregunta": " Señores considerar plazo de entrega 60 dias",
    "respuesta": "Estimado oferente NO es posible porque el tiempo estimado es de 30 días c
    "fechaPre": " 2020-05-11 00:09:09",
    "archivos": "archivos"
  },
  {
    "pregunta": " Señores por favor indicar si se puede realizar una visita tecnica",
    "respuesta": "Estimado oferente si es posible para que realice una inspección técnica y
    "fechaPre": " 2020-05-11 00:10:17",
    "archivos": "archivos"
```


Figure 1: An example of the questions and answers that are taken as input data

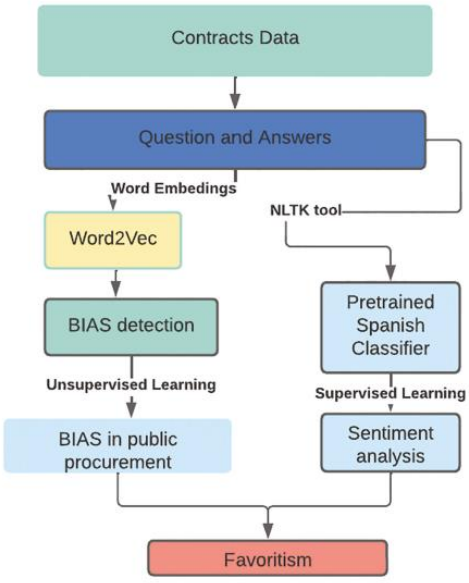


Figure 2: Proposed methodology

3.3 Data Pre-Processing

To precisely label each of the terms and normalize the text, stop words are eliminated, the text is converted to lowercase, punctuation marks, question marks, extra spaces, and tabulations are eliminated. Despite the fact that term root extraction (stemming) is common in NLP, some results were observed that showed how it affected sensitivity and precision, so we decided not to apply it. For this phase, the NLTK [28] and PANDAS [29] Python libraries are used.

3.4 Data Processing

Word2vec [26] word embeddings were used to analyze gender bias and favoritism. With Word2vec, a large amount of text corpus is taken as input and used to generate word embeddings. The word embeddings are represented in a high-dimensional vector space and the semantic information of the words is represented. As a result, similar words sharing the same context are grouped together in this vector space. For word embedding generation, the FastText algorithm [27] was also evaluated. In both algorithms, the questions and answers of each public procurement process can be projected from the original space to the new multidimensional space. In this way, word embeddings learn relationships derived from concurrency statistics. We also test Dipper throated optimization (DTO) algorithm's with is suitability for solving complex real-world issue [30].

The evaluation of connected words is obtained through the conditional probability that a word, see Eq. (1), (represented by w) is part of the context (w_i) and also w_j (next word) is given by (1) according to the matrix W and W' (the inverse) and its matrix of transformation (T) so that the equation allows to obtain the similarity between the words denominated as w_i, w_j, w_k thus substantively.

$$p(wj|wi) = \frac{\exp\left(\upsilon_{wi}\ T\quad w_j\right)}{\exp\left(\upsilon'_k T \upsilon_{w_i}\right)} \tag{1}$$

3.4.1 Sentiment Analysis

The questions and answers are analyzed as individual sentences to extract the opinion contained in each of them (positive, negative, or neutral). A value is assigned to each opinion, whose sum between the three sentiments must be equal to 1, for example (neutral: 0.1, positive: 0.0, and negative: 0.9). The algorithm VADER combines lexical features with the consideration of five general rules, which incorporate grammatical and syntactic conventions to express and emphasize sentiment and tension of the same. This algorithm is available through the NLTK *Sentiment Intensity Analyzer function* and is implemented in Google Colab.

VADER is used because it allows for obtaining a better accuracy than other techniques analyzed [31]. To evaluate the results, standard machine learning metrics [32] are used, such as precision, accuracy, F1 score, area under the receiver operating characteristics curve (AUC), and the loss function (loss) which establishes the penalty for not achieving the expected result. If the deviation from the expected value by our model is large, then the loss function returns the highest number and if the deviation is much smaller, it is moved closer to the expected value.

3.4.2 Data Bias Detection

In order to detect gender bias and favoritism in the datasets, the generated word embedding models were used for the questions and answers respectively. As we mentioned these models are based on Word2vec because this technique can be used to treat a large amount of data in the corpus. It is based on the application of Boolean type logical operators using a smooth transformation matrix in a linear subspace, in which word embedding obtains the highest variance [33,34]. The bias detail assessed is summarized in Table 1.

Table 1: Description of bias type to be evaluated

Bias type	Verification reason	Where to evaluate
Gender	Male terms are used to address suppliers and to request staff for each contract.	• Questions • Answers
Favoritism	The responses of the contracting entities must be the same for all providers, without directing or favoritism.	• Answers

For implementation, the Keras and TensorFlow tools were used provided by Google Colab. Due to the amount of data used, a convolutional neural network is trained in Tensorflow with the representation model "Continuous Skip-gram" [35], which predicts the words that are within a certain range before and after the current parsed word. In this phase, the following parameters are established for each sequence: *vocab_size = 4096* (range between words) and *sequence_length = 10* (sequence length). The source code and implementation are available in Google Colab. You can also download the *vector.tsv* (vectorized words) and *medatada.tsv* files (approximation of each word) to allow the visualization in the *TensorBoard Embedding Projector*.

4 Results

Fig. 3 shows the word cloud corresponding to the most used terms for the elaboration of questions by providers. Some of them in the image are in Spanish Language. Words such as: late, yes, confirm, validate, money (*plata* in colloquial Spanish), etc. In Fig. 3 is intended to get an idea of the terminology used by vendors.

Figure 3: *Word cloud* of vendor questions

Fig. 4 gives the word cloud corresponding to the terminology frequently used for preparing responses by the contracting entities. It should be noted that the answer provided by the state entities is expected will not be biased. However, frequently appearing biased words like experience (responses oriented to a provider with the institution previously worked), reference (response-oriented to the product must meet certain references to a particular model or brand), and gentlemen (answers that are aimed at a male audience).

4.1 Bias Analysis

Algorithms such as Word2Vec and FastText were evaluated to analyze gender bias (in questions and answers) towards other providers.

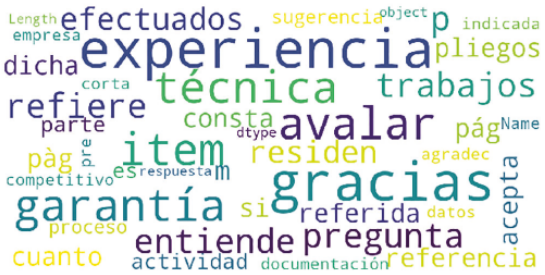


Figure 4: *Word cloud* of public entity answers

4.1.1 Bias in Questions

The process begins with the creation of the vocabulary of the obtained *tokens*, allowing the inverse vocabulary to be created to finally vectorize and obtain the transformation matrix of the words so that the equivalent vector is obtained in each word.

Table 2 analyzes the results obtained for the gender bias analysis through the comparison between Word2vec and FastText algorithms. Accuracy was used as a metric, which measures the percentage of cases in which the model was correct and also the function of loss (loss) which evaluates the discrepancy between the prediction and the expected result, showing Word2vec provides better results for the model.

Table 2: Gender bias results in questions

Algorithm	Metric	Bias in gender
FastText	Accuracy	0.71
Word2vec	Loss	0.78
	Accuracy	0.89
	Loss	0.59

Fig. 5 elaborates the results of evaluation SCORE with the ROC curve with 2000 iterations of one of the trained models (FastText) on the left side where the Specificity (0.71) is shown, and on the right, the loss function (0.78). It can be seen that the evaluated model cannot significantly improve its results, thus justifying the choice of Word2vec instead of FastText. The ROC curve and the loss function were used to measure the accuracy percentages of all the models. It is worth noting that these graphs were generated for all the biases in both questions and answers and the results are shown in Tables 2–4.

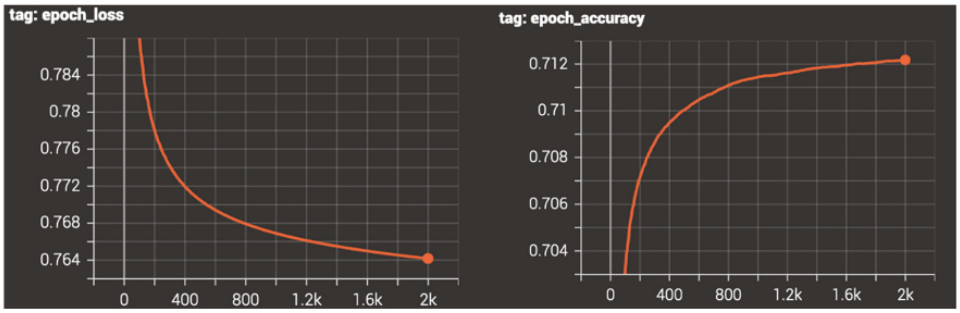


Figure 5: ROC curve with 2000 interactions, accuracy, and loss of the FastText algorithm

Table 3: Results of the response bias assessment

Technique	Metric	Gender bias	Favoritism
FastText	Accuracy	0.80	0.77
Word2vec	Loss	0.65	0.65
	Accuracy	0.90	0.88
	Loss	0.37	0.42

Table 4: Example of analysis elaborated by the model

Answer	Prediction	Probability
Dear Sir or Mrs., If the case of an interview or news is presented in an international media, the awarded supplier will be notified at least 24 h in advance so that the corresponding recording can be made, and the respective information can be sent to us immediately.	Without bias	0.9
Dear mrs bidder, it is not possible to present a woman professional with electronic engineering skills to replace the telecommunications professional, they must have a telecommunications degree, and the bidder must comply with what is requested in the specifications.	With bias (gender)	0,8
Dear supplier, form No. 7 will only be requested from the awarded supplier as a document prior to the execution of the contract. Form 7 is not part of the technical offer. At the time of including, it in it, it would be disqualified.	With bias (favoritism)	0,88

In the questions from suppliers, it was frequently mentioned that there are processes that are directed or limited to certain contractors or professionals which are previously chosen by the public institution. So, for example the word "*dirigido (biased)*" is related to the words: "*exist, defects, contractor, agree, specifications*". It should be noted that the terms of conditions made by the contracting entity are called "specifications", therefore it is understood that the providers in their questions directly indicate when there is favoritism on the part of the public institution, but these "complaints in many cases are not taken care of by the public institution". An example of this is the following question asked by a supplier: "*In a biochemical place, can we deliver professional mechanized dilutors for a correct dilution of chemicals? since a biochemist, chemical engineer, and pharmaceutical chemist prepare or manufacture and indicate their technical datasheets. This is very biased, and even more so with the 30% advance payment, please check the specifications*".

Fig. 6 analyzes the closeness of words, according to their similarity for bias detection using Word2vec, it is observed that in the responses made by the contracting entities, there are words such as: "*unfortunately*" which is directly related to the words "v*endors*" and "*number*" referring to the fact that the contracting entities use phrases such as "*unfortunately the process has such a number of invited vendors*" which suggests favoritism towards directly invited suppliers by the contracting entity.

In addition, words such as "*validate*" and "*require"* are common words with regard to the answers issued by the contracting party, referring to the number of suppliers invited to the public procurement processes, since the entity always asks if the supplier complies with the specific and on many occasions complex requirements.

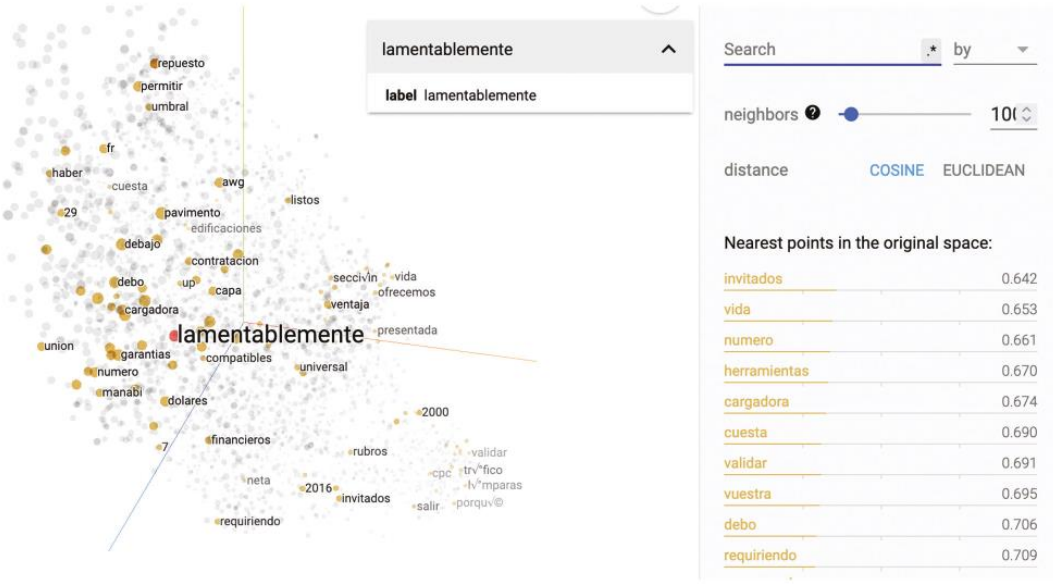


Figure 6: Word visualization questions

4.1.2 Bias in Answers

Table 3 shows the results of the evaluation of the metrics accuracy and loss function corresponding to gender bias and favoritism in the answers issued by the contracting entities in a public procurement process. It might be seen that the Word2vec technique presents better results, reaching 90% in the detection of gender bias and 88% in favoritism.

Fig. 7 shows the ROC curve, where the classificatory potentiality of the proposed model is indicated, showing the accumulation on the X-axis and the number of iterations for model training on the Y-axis.

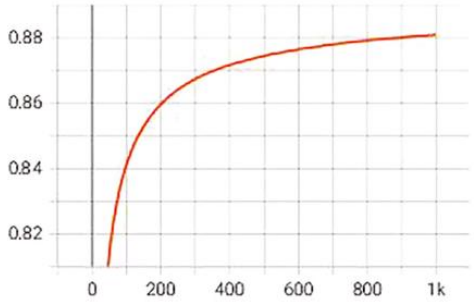


Figure 7: ROC accuracy answers

It can be observed in Fig. 8 that the visualization of words associated with gender bias, is predominantly oriented towards the masculine gender such as: gentlemen, Dear Sir/Mr., etc., with respect to equivalent words in the feminine gender or in neutral language. Both contractors and public institutions assume that the person with whom they are interacting is a male. The table further shows the similarity score of the

word “estimado (Dear Sir/Mister)” with other masculine gender-based expressions. It is also evident that there is no direct percentage score with the highlighted word “estimada (Dear Miss/Mrs/Ma’am)” which would be the feminine counterpart of “estimado”.

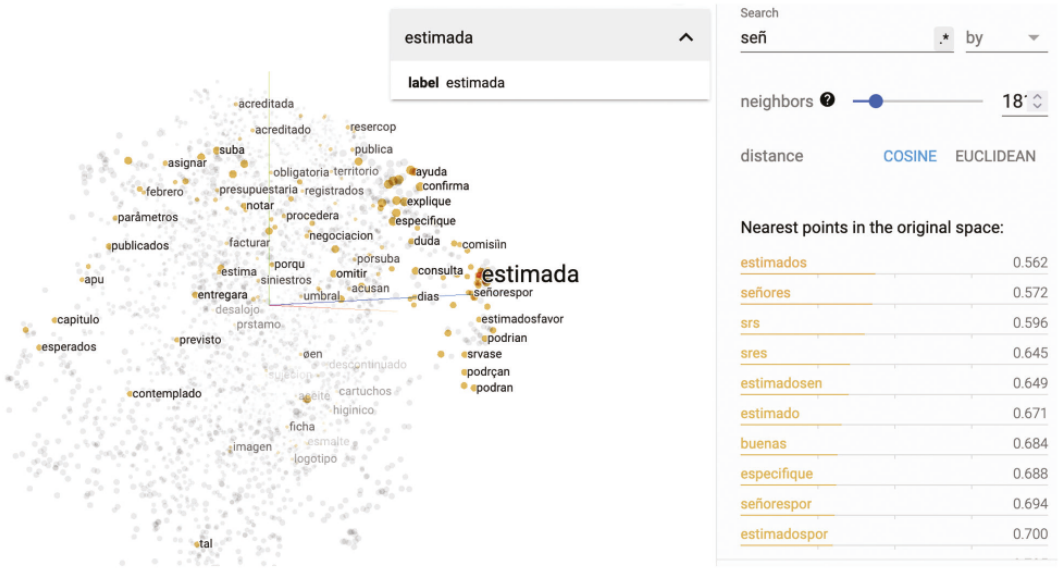


Figure 8: Word visualization in answers

Table 4 shows an example of the phrases used to detect bias in the responses made by the contracting institutions, including their evaluation, where 3 cases are observed: gender bias, favoritism (corruption), and no bias.

4.2 Sentiment Analysis

This section presents the sentiment analysis of questions and answers made by suppliers and contracting entities.

On the left side of Fig. 9, the sentiment evaluation scales are plotted at an interval of 0.15 (X-axis), while the number of questions is plotted on a Y-axis. In the questions asked by providers to contracting entities, we found a higher percentage of positive opinions. On the right side of Fig. 9, the classification of the sentiment analysis is also shown, in the responses of public institutions, highlighting only the positive and negative ones (discarding the neutral ones). There is evidence of a greater number of negative responses (105,547) compared to 73,665 positive responses, this being an indication of a possible bias. The negative sentiment was also evaluated in the responses of public institutions, highlighting that the majority are in the range of 0.1 to 0.41, these responses being “partially negative”, however, it is important to highlight the number of responses with a high percentage of negativism, which are those with scores greater than 0.41 to 0.875.

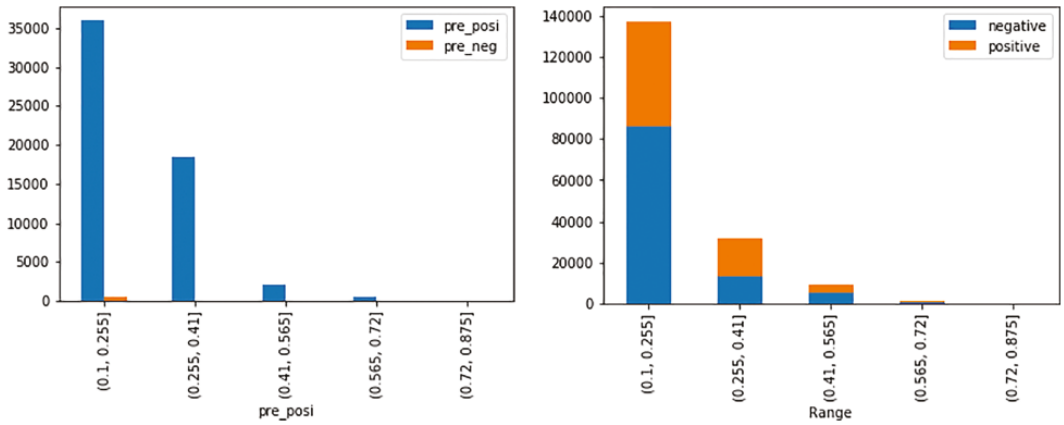


Figure 9: Distribution of sentiment in questions and responses

The result of the classification of the sentiment analysis in the questions asked by the providers is summarized in Fig. 10, which indicates that only 16% of the providers send negative questions to the contracting institution, while 39% are positive. In the case of the responses made by the contracting public entities, 32% of the institutions issue negative responses to the providers, while 45% are neutral and 23% are positive, so it can be inferred that practically 1 out of 3 responses from the contracting entities are negative.

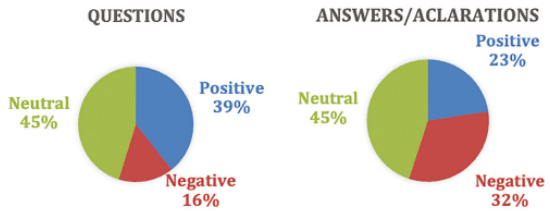


Figure 10: Distribution of sentiment analysis in questions and answers

With this, it can be seen that there is a trend of between 30% and 35% of public procurement processes, where it is intended to favor a certain contractor, either through favoritism in questions and answers or with the modification of qualification parameters.

5 Discussion

The importance of this work is based on trying to achieve equality and sustainability in public procurement processes, applying the model proposed to detect bias and favoritism in public procurement. This will contribute to achieving the goal of sustainable public procurement, which finds staunch support in the Johannesburg Implementation Plan (PIJ), approved by all governments at the United Nations World Summit for Sustainable Development.

In the evaluation of the questions and answers of each process, it can be observed in the experiments that, far from a treatment based on equality, negative responses are detected by the contracting entities towards certain suppliers. This could mean that in 32% of cases there is bias and favoritism on the part of

the state entity and this is attested with previous work by other authors [7], wherein in 35% of the cases, there was a modification of the qualification parameters of the winner of the offer, with the aim of favoring a certain contractor.

Regarding public procurement policies, the inclusion of women in the entire process can affect the quality of life of this group and their economic autonomy [36]. In addition, there is evidence that women tolerate acts of corruption less than men according to Guerra et al. [37]. Following initiatives like the UK [38] that includes women in public processes in the European construction and industry. Therefore, our work aims to demonstrate the existing bias towards male suppliers, in the responses of the contracting entities, using a Word2vec-based gender bias detection model, which reaches an accuracy of 90%.

In the evaluation of the model of bias between contractors (favoritism), an accuracy of 88% was obtained. These results support the model proposed by the authors in their previous work [11], validated by using NLP to detect corruption. This agrees with the work of Modrusan et al. [6], which makes use of the BoW model for the payment of tax advances, obtaining an accuracy of 71%, but using only 15,000 records, a much smaller quantity than the one used in the present work. Our model makes use of word embeddings just like in Skorková [14], however, it also uses Word2vect as an alternative to BERT.

6 Conclusion

With the implemented experiments, the researchers demonstrated that the indications extracted from the questions and answers to the participants in a Spanish corpus public procurement allow biased purchases to be detected. Both gender bias and favoritism towards certain suppliers on the part of the contracting public entities are detected. Also, to complement the model, sentiment analysis of the questions and answers provided by the contracting entities was conducted. The model shows an accuracy of 88% for bias detection and 90% for vendors favoritism. For this reason, the implemented algorithms produced significant improvement in terms of accuracy.

The model shows that one-third of the procurement processes carried out by the state have indications of corruption and bias, which represents millions of dollars in damages suffered by the state, discrimination against suppliers (woman bias), and potential legal cases against those involved. This shows the potential of text analysis in the detection of corruption and bias since the documents attached to procurement processes contain large amounts of information that has not been exploited.

Therefore, the work presented contributes to public contracting being directed to acquiring sustainable, innovative, environmentally friendly design goods, with the criteria of inclusion, equality, and social equity contributing directly to achieving sustainable development goals.

7 Future Work

As future work, it is intended to evaluate the model, with other machine learning algorithms like Lexicalized Dependency Paths (LDPs), for usage with active learning methods [39] with noise removal techniques [40], decomposition methods a reduction technique [41,42], or feature selection techniques [43] and finally other techniques for sentiment analysis, in addition to the review of the current methods that are being applied in other countries for the detection of bias and favoritism in public purchases. The proposed method has been empirically verified with purchases from Ecuador and it would be ideal to analyze whether it can be replicated to other Spanish-speaking countries, with the purpose of improving their skills in different dialects and local terminologies.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Y.T.-B. and V.F.L.B.; methodology, Y.T.-B., L.C.Z and V.F.L.B.; formal analysis, Y.T.-B.; investigation, Y.T.-B.; writing—original draft preparation, Y.T.-B. and L.C.Z; writing—review and editing, Y.T.-B. and V.F.L.B. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials: In this paper the dataset is available in https://drive.google.com/file/d/1lW7hp5nVP09WHSYHBJQXoxi0h6Ov49sY/view?usp=sharing,

Corpus de questions: https://www.dropbox.com/s/tlfa5jfuhvqxtyc/preguntas.txt?dl=0.

Corpus de answers: https://www.dropbox.com/s/ktmuj07mbw2jqew/respuesta.txt?dl=0.

Source code. https://colab.research.google.com/drive/1-WB6yywjyXKpWevbxwpukGtpxqa8JRse?usp=sharing.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Thi Nguyen and N. T. Bui, "Government expenditure and economic growth: Does the role of corruption control matter?," *Heiyon*, vol. 8, no. 10, pp. 127–147, 2022.

[2] E. Kehler, J. Paciello and J. Pane, "Anomaly detection in public procurements using the open contracting data standard," in *Proc. 2020 Seventh Int. Conf. on eDemocracy & eGovernment (ICEDEG)*, Buenos Aires, Argentina, pp. 127–135, 2019.

[3] Servicio Nacional de Contratación Pública, "Rendición de cuentas," SERCOP, Quito, Ecuador, 2018.

[4] M. S. Rad and A. Shahbahrami, "Detecting high risk taxpayers using data mining techniques," in *Proc. 2016 2nd Int. Conf. of Signal Processing and Intelligent Systems (ICSPIS)*, Tehran, Iran, pp.14–15, 2017.

[5] D. M. E. D. M. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University-Engineering Sciences*, vol. 30, no. 4, pp. 330–338, 2018.

[6] N. Modrusan, K. Rabuzin and L. Mršić, "Improving public sector efficiency using advanced text mining in the procurement process," in *Proc. of the 9th Int. Conf. on Data Science, Technology and Applications (DATA 2020)*, Paris, France, pp. 200–206, 2020.

[7] Y. Torres-Berru and V. F. L. Batista, "Data mining to identify anomalies in public procurement rating parameters," *Electronics*, vol. 10, no. 22, pp. 1–15, 2021.

[8] E. Ortiz-Prado, R. Fernandez-Naranjo, Y. Torres-Berru, R. Lowe and I. Torres, "Exceptional prices of medical and other supplies during the COVID-19 pandemic in Ecuador," *The American Journal of Tropical Medicine and Hygiene*, vol. 105, no. 1, pp. 81–87, 2021.

[9] Y. Torres Berru, V. F. López Batista, P. Torres-Carrión and M. G. Jimenez, "Artificial intelligence techniques to detect and prevent corruption in procurement: A systematic literature review," *Communications in Computer and Information Science*, vol. 1194, pp. 254–268, 2020.

[10] Y. Torres-Berru, V. F. L. Batista and P. Torres-Carrión, "Data mining to detect and prevent corruption in contracts: Systematic mapping review," *RISTI-Revista Iberica de Sistemas e Tecnologias de Informacao*, vol. 2020, no. E29, pp. 13–26, 2020.

[11] Y. Torres-Berru and V. F. López Batista, "Data and text mining for the detection of fraud in public contracts: A case study of Ecuador's official public procurement system," in *Doctoral Symp. on Information and Communication Technologies-DSICT. Lecture Notes in Electrical Engineering*, vol. 846, pp. 116–127, 2022.

[12] Y. Bramoullé and S. Goyal, "Favoritism," *Journal of Development Economics*, vol. 122, no. 1, pp. 16–27, 2016.

[13] N. W. Rustiarini, T. Sustrino, N. Nurkholis and W. Andayani, "Why people commit public procurement fraud? The fraud diamond view," *Journal of Public Procurement*, vol. 19, no. 4, pp. 345–362, 2019.

[14] Z. Skorková, "Competency models in public sector," *Procedia-Social and Behavioral Sciences*, vol. 230, pp. 226–234, 2016.

[15] K. Pant, T. Dadu and R. Mamidi, "Towards detection of subjective bias using contextualized word embeddings," in *Proc. Companion Proc. of the Web Conf. 2020*, Taipei, Taiwan, pp. 75–76, 2020.

[16] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham and J. Handelsman, "Science faculty's subtle gender biases favor male students," *Proceedings of the National Academy of Sciences*, vol. 114, pp. 3–14, 2013.

[17] T. Bolukbasi, K. W. Chang, J. Zou, V. Saligrama and A. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," *Advances in Neural Information Processing Systems*, pp. 4356–4364, 2016.

[18] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief *et al.,* "Mitigating gender bias in natural language processing: Literature review," in *Proc. ACL, 2019 Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 1630–1640, 2019.

[19] M. Hamisu and A. Mansour, "Detecting advance fee fraud using NLP bag of word model," in *Proc. 2020 IEEE 2nd Int. Conf. on Cyberspace CYBER*, Niger, pp. 94–97, 2021.

[20] A. J. Ruiz, *Inclusión de mujeres en las contrataciones públicas: la experiencia latinoamericana*, La haya, Nederlands: Hivos, 2020.

[21] C. Caparrós-Laiz, J. A. García-Díaz and R. Valencia-García, "Evaluating extractive automatic text summarization techniques in Spanish," *Communications in Computer and Information Science*, vol. 1460, pp. 79–92, 2021.

[22] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed and M. T. Sadiq, "Automatic detection of offensive language for Urdu and Roman Urdu," *IEEE Access*, vol. 8, pp. 91213–91226, 2020.

[23] J. M. Conroy, J. D. Schlesinger and D. P. O'Leary, "Nouveau-ROUGE: A novelty metric for update summarization," *Association for Computational Linguistics*, vol. 37, no. 1, pp. 1–8, 2011.

[24] E. Reiter, "A structured review of the validity of BLEU," *Computational Linguistics*, vol. 44, no. 3, pp. 393–401, 2018.

[25] E. V. Tunyan, T. A. Cao and C. Y. Ock, "Improving subjective bias detection using bidirectional encoder representations from transformers and bidirectional long short-term memory," *International Journal of Cognitive and Language Sciences*, vol. 15, pp. 329–333, 2021.

[26] K. W. Church, "Emerging Trends: Word2Vec," *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, 2017.

[27] B. Athiwaratkun, A. G. Wilson and A. Anandkumar, "Probabilistic FastText for multi-sense word embeddings," in *Proc. 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, Melbourne, Australia, pp. 1–11, 2018.

[28] F. Millstein, "Natural language processing with python: Natural language processing using NLTK," in *Frank Millstein: North Charleston, USA*, 2020.

[29] J. Reback, W. McKinney, J. Van den Bossche, T. Augspurger, P. Cloud *et al.,* "Pandas," in *Zenodo [code]*, 2020. https://doi.org/10.5281/zenodo.4524629

[30] A. E. Takieldeen, E. M. El-kenawy, M. Hadwan and R. M. Zaki, "Dipper throated optimization algorithm for unconstrained function and feature selection," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 1465–1481, 2022.

[31] G. Veena, A. Vinayak and A. Nair, "Sentiment analysis using improved Vader and dependency parsing," in *Proc. 2021 2nd Global Conf. for Advancement in Technology (GCAT)*, Bangalore, India, pp. 1–6, 2021.

[32] R. Bakshi, N. Kaur, R. Kaur and G. Kaur, "Opinion mining and sentiment analysis," in *2016 3rd Int. Conf. on Computing for Sustainable Global Development*, India, pp. 452–455, 2016.

[33] K. Patel and P. Bhattacharyya, "Towards lower bounds on number of dimensions for Word Embeddings," in *Proc. of the 8th Int. Joint Conf. on Natural Language Processing*, Taipei, Taiwan, pp. 31–36, 2017.

[34] R. Popović, F. Lemmerich and M. Strohmaier, "Joint multiclass debiasing of word embeddings," *Lecture Notes in Computer Science*, vol. 12117, pp. 79–89, 2020.

[35] S. Sabra and V. Sabeeh, "A comparative study of N-gram and Skip-gram for clinical concepts extraction," in *Proc. Int. Conf. on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, USA, pp. 807–812, 2020.

[36] G. Pierri, M. J. Jarquin and R. De Michele, *Transparencia y género: el impacto de las compras electrónicas en el acceso a licitaciones públicas de las pymes lideradas por mujeres*, Banco Interamericano de Desarrollo, 2021.

[37] A. Guerra and T. Zhuravleva, "Do women always behave as corruption cleaners?," *Public Choice*, vol. 191, no. 1–2, pp. 173–192, 2022.

[38] T. Wright, "New development: Can 'social value' requirements on public authorities be used in procurement to increase women's participation in the UK construction industry?," *Public Money & Management*, vol. 35, no. 2, pp. 135–140, 2015.

[39] H. Sun and R. Grishman, "Lexicalized dependency paths based supervised learning for relation extraction," *Computer Systems Science and Engineering*, vol. 43, no. 3, pp. 861–870, 2022.

[40] M. T. Sadiq, X. Yu, Z. Yuan and M. Z. Aziz, "Motor imagery BCI classification based on novel two-dimensional modelling in empirical wavelet transform," *Electronic Letters*, vol. 56, pp. 1367–1369, 2020.

[41] M. T. Sadiq, X. Yu, Z. Yuan, Z. Fan, A. U. Rehman *et al.,* "Motor imagery EEG signals classification based on mode amplitude and frequency components using empirical wavelet transform," *IEEE Access*, vol. 7, pp. 127678–127692, 2019.

[42] M. T. Sadiq, X. Yu, Z. Yuan and M. Z. Aziz, "Exploiting dimensionality reduction and neural network techniques for the development of expert brain–computer interface," *Expert Systems with Applications*, vol. 164, pp. 114031, 2021.

[43] E. -S. M. El-kenawy, A. Ibrahim, N. Bailek, B. Kada, M. Hassan *et al.,* "Sunshine duration measurements and predictions in Saharan Algeria region: An improved ensemble learning approach," *Theoretical and Applied Climatology*, vol. 147, no. 3–4, pp. 1015–1031, 2022.

risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 29/11/2019
Aceitação/Acceptance: 31/01/2020

Minería de datos para detectar y prevenir la corrupción en los contratos: revisión sistemática de mapeo

Yeferson Torres-Berru[1,2], Vivian Félix López Batista[1], Pablo Torres-Carrión[3]

ymtorresb@usal.es, vivian@usal.es, pvtorres@utpl.edu.ec

[1] University of Salamanca, Plaza de la Merced, s/n,37008 Salamanca, Spain.

[2] Instituto Superior Tecnológico Loja, Av. Granada y Turunuma, , 1101608, Loja, Ecuador.

[3] Universidad Técnica Particular de Loja, San Cayetano Alto S/N, 1101608, Loja, Ecuador.

Pages: 13–26

Resumen: La corrupción según afirma la ONU está presente en sus diferentes formas y tipologías afectando directamente en la celebración de contratos tanto públicos como privados; en este contexto se realiza un mapping sistemático de investigaciones científicas (2015-2019) sobre corrupción en contratos en sus diversos formatos, aplicando minería de datos y sus técnicas. Se plantean seis preguntas de investigación a responder desde el análisis de 147 artículos obtenidos de las bases de datos WoS y Scopus. Las investigaciones se centran principalmente en la detección fraude, fraude financiero y corrupción, siendo las formas de corrupción más estudiadas el fraude (72.72%), y el sobreprecio (8,84%); las investigaciones se han realizado en Estados Unidos (16,32%), China (10,88%), Reino Unido (8,94%) y en LatinoAmerica Brasil (3,4%), con minimas contribuciones de Colombia y Paraguay.

Palabras-clave: Data mining, corrupción, mapping review.

Data Mining to detect and prevent corruption in contracts: Systematic Mapping Review

Abstract: Corruption according to the UN is present in various forms and types, affecting the realization of public and private contracts; in this context, a systematic mapping of scientific publications (2015-2019) is development, focused on contract corruption in its several forms, applying data mining and related techniques. Six research questions are presented to answer the analysis of 147 articles obtained from WoS and Scopus databases. The detection of fraud, financial fraud and corruption predominate in the researchs, exceling as forms of corruption, fraud (72.72%) and the overprice (8.84%). Researchs have been conducted in the United States (16.32%), China (10.88%) and the United Kingdom (8.94%); in Latin America emerge Brazil (3.4%) with minimum contributions from Colombia and Paraguay.

Keywords: Data mining, corruption, mapping review.

1. Introduction

Corruption is present in every country in the world, according to the (ONU , 2018, for the Transparency International Organization (2017) corruption is a "bribe, as an offer or receipt of any gift, loan, fee, reward or other advantage to or from any person as an incentive to do something that is dishonest, illegal or an abuse of trust, in the exercise of business activity". The word «corruption», derived from the Latin verb "corrumpĕre", with several negative meanings from the moral point of view: alter and disrupt, spoil, deprave, damage, rot, bribe, pervert, dodge (Martinez Fernandez 2015). Corruption is a situation in which a conflict of interest is used to satisfy a self-interest that is specified in obtaining a profit in breach of an existing legal framework (Cerillo Martinez 2015). This social behavior is present in different forms and aspects affecting different social contexts in family, institutional, private, governmental areas with negative repercussions.

In the public sector, corruption infers an abuse of an employee's power to obtain "profits," benefiting private entities, by taking advantage of a specific situation to break the law and benefit the other participant in the act (if any) and himself; "gains" for the corrupt include not only money but also material and intangible goods, which involve status and power (OECD 2005). Different forms and levels of corruption have been identified, such as bribery, embezzlement, fraud, extortion, abuse of trust, collusion and favoritism (Chan and Owusu 2017; Moran 2001; Vargas-Hernández 2009). The main mechanisms of corruption according to (Cassagne and Rivero Ysern 2007; Castro Cuenca 2017) are non-existence of contract, improper direct contracting, improper contracting, fractionation, contract modifications. As is visible, the study of corruption encompasses a large field of social and human sciences, with theoretical and scientific contributions.

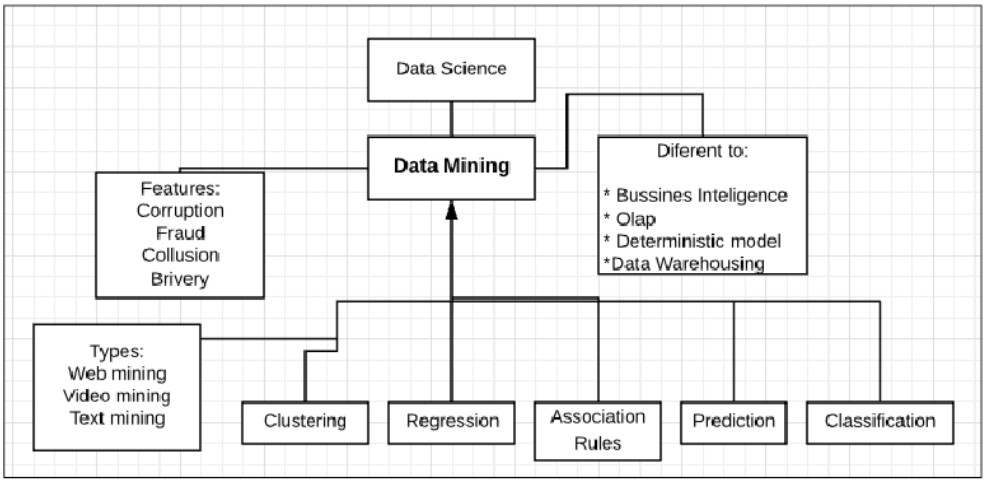


Figure 1 – Mentefacto Conceptual

As noted in (DiRienzo et al. 2007) technology has become a key ally in the fight against corruption, which shows that, in countries that use ICTS within the public administration, they get better levels in the corruption perception index. In this sense, the use of information technologies to combat corruption has shown good results in state agencies in several countries (Volosin 2015), excelling studies conducted by Goedhuys, Mohnen, and Taha (2016) demonstrating that reducing unnecessary interventions in procurement processes using technology improves transparency, because they usually cause abuse of power by public employees; (Shrivastava and Bhattacherjee 2015) consider using tools and resources to monitor the development of public employee functions at a low cost; (Lee and Lio 2016) explain how technology helps to promote transparency through the provision of information to the general public; and, (DiRienzo et al. 2007), (OECD 2007) explain how technology helps to eliminate bureaucratic procedures that generate higher costs and times. It is clear that the contribution of technology to improve anti-corruption management and control procedures, allows an improvement in the efficient administration, in search of transparency in public management processes.

Figure 1 shows the theoretical classification, detailing data science through data mining, its techniques (Han, et al 2011), and types (Hand 2013), and areas of knowledge different from data mining, but which are part of other branches of computer science and data science. In the "mentefacto conceptual", the relationships between characteristics of corruption and data mining are also graphically represented, among the primary forms of corruption are fraud, collusion, bribery, overprice, favoritism.

As related works, in the Scopus and Wos databases, two systematic reviews of scientific literature were found that relate the fight against corruption through the use of technology. (Indrajani et al. 2016) focus on finding fraud algorithms in online transactions, finding 25 scientific articles for study; and, (Ngai et al. 2011) reviews different data mining techniques for fraud detection in different fields of the financial field from 41 scientific articles. That is why this study is proposed as original, by encompassing corruption and data mining from a broader perspective (fraud, collusion, bribery, overprice, and favoritism) than the corruption form known as fraud.

Data science plays a vital role in detecting corruption and is commonly used to find hidden information in large amounts of data (Alvarez-Jareño et al. 2019). The corruption case known as Panama papers (Woodie 2016) revealed to the public opinion fiscal and financial fraud; in Brazil, the Observatory of Public Expenditures (Controladoria-Geral da União 2015) reviewed more than 120,000 public contracts and uncovered more than 7,500 cases involving $ 104 million in financial operations of doubtful legality. These examples illustrate the importance of data science in the fight against corruption.

Based on the forewent, this paper analyzes the scientific information published from 2015 to 2019 using the methodology of (Torres-Carrion et al. 2018) which is used for the systematic literature reviews, combining it with the methodology (García-González and Ramírez-Montoya 2019) whose components are detailed in the next section. The

results are organized for presentation following the order of the research questions, with graphic explanations and referential analysis. Finally, the relevant conclusions regarding this systematic mapping are shared. The discussion section is omitted, taking into account that the objective of mapping is to describe the state of the art in terms of general areas of research on data mining and corruption.

2. Methodology

For the systematic search, the methodology of (Torres-Carrion et al. 2018) is followed, which divides the process into three phases: planning, conducting the review, and reporting the review. The third phase has been carried out following the methodology of (García-González and Ramíre\z-Montoya 2019) applied in its mapping report.

2.1. Research questions

Through the research questions, the investigation objective is established, as well as the variables to measure and answer the questions in Table 1 which pretend to inquire the number of published articles in the topic in mind, the geographical distribution, the context of the form of corruption studied and the primary line of research studied.

Question	Type of response sought
RQ1: How many studies are in the WOS and Scopus databases from 2015 to 2019?	• Number of articles in Scopus • Number of articles in WOS • Number of duplicated articles • Number of open access articles • Type of document (Review, etc)
RQ2: Who are the authors of the most cited articles?	• Most cited authors • Most cited articles
RQ3: What is the geographical distribution of the authors?	• Countries where the authors are from
RQ4: What are the journals with more publications on this line of research?	• Journals • Q1, Q2, Q3 or Q4 • Indice JCR
RQ5: In what contexts are these studies developed?	Fraud, Bribery, Collusion, Overpricing, Favoritism, Embezzlement
RQ6: What are the main topics addressed in this line of research?	

Table 1 – Research questions

2.2. Inclusion, exclusion and quality criteria

The inclusion and exclusion criteria that allowed discriminating articles that do not correspond to the areas of knowledge referred to in the study, the years of research, and the selected databases. Relevant articles that answer the raised research questions are considered. Details are presented in Table 2.

Criteria	Inclusion	Exclusion
Theoretical field	Corruption and data mining available.	Web and mobile phone applications.
Databases	Web of Science (WOS) o Scopus	Google Scholar and other index files in WoS o Scopus
Type	Article, review, editorial, conference proceedings.	Speech documents, book chapters, ESCI
Año	2015-2019	Before 2015 or later tan the article publication date.
Area of research	Computer Science, Social Science, Decision-making Science.	

Tabla 2 – Inclusion, exclusion and quality criteria

As a quality criterion, a thorough review of the articles resulting from the search and after applying the inclusion and exclusion criteria is established. Articles that do not explicitly refer to data mining applications for the prevention and detection of corruption will be separated. This phase requires the reading of each of the articles obtained and is carried out by expert researchers.

2.3. Semantic Search Structure

According to the methodology, the input for the semantic structure corresponds to the thesaurus and synonymy of the concepts obtained in the "mentefacto conceptual". The base scripts have been established, supported by the logical conjunction and disjunction operators, as well as the sequence and word relationship operators (*W/n* y *NEAR/n*). The script is organized in three levels: in the first level, data mining is approached with its characteristics and techniques; in the second level, corruption and its typology are reviewed; and in the third level the contracts and their synonymy are analyzed, as shown in Table 3.

L1	(mining W/4 (data OR video OR text OR web)) OR classificat* OR cluster* OR regression OR (association W/2 rules) OR detection OR prediction OR (sequential W/2 patterns) OR (learning W/4 (machine OR deep OR reinforced))
L2	(corruption OR bribery OR collusion OR (embezzlement OR misappropriation) OR fraud OR (abuse W/0 of W/0 discretion) OR favoritism OR nepotism)
L3	(contract OR purchase OR investment OR procurement OR acquisition OR acquirement OR tendering)

Table 3 – Semantic Search Structure

The level structure is the input to generate the final search scripts, adaptable to the databases proposed in the methodology: Web of Science and Scopus, detailed in Table 4.

Script WOS	Script Scopus
TS=(((mining near/4 (data OR video OR text OR web)) OR classificat* OR cluster* OR regression OR (association near/2 rules) OR detection OR prediction OR (sequential near/2 patterns) OR (learning near/4 (machine OR deep OR reinforced))) AND ((corruption OR bribery OR collusion OR (embezzlement OR misappropriation) OR fraud OR (abuse near/0 of near/0 discretion) OR favoritism OR nepotism)) AND (contract OR purchase OR investment OR procurement OR acquisition OR acquirement OR tendering)) Refinado por: AÑOS DE PUBLICACIÓN: (2019 OR 2018 OR 2017 OR 2016 OR 2015) AND DOMINIOS DE INVESTIGACIÓN: (SCIENCE TECHNOLOGY) AND TIPOS DE DOCUMENTOS: (ARTICLE OR REVIEW OR EDITORIAL) AND ÁREAS DE INVESTIGACIÓN: (COMPUTER SCIENCE)	TITLE-ABS-KEY (((mining W/4 (data OR video OR text OR web)) OR classificat* OR cluster* OR regression OR (association W/2 rules) OR detection OR prediction OR (sequential W/2 patterns) OR (learning W/4 (machine OR deep OR reinforced))) AND (corruption OR bribery OR collusion OR (embezzlement OR misappropriation) OR fraud OR (abuse W/0 of W/0 discretion) OR favoritism OR nepotism) AND (contract OR purchase OR investment OR procurement OR acquisition OR acquirement OR tendering)) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "DECI") OR LIMIT-TO (SUBJAREA , "SOCI")) AND (LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015))

Tabla 4 – Search Script

3. Results

After the systematic search, 219 articles are obtained from Scopus and 250 from Web Of science, applying the inclusion, exclusion and quality criteria, 153 search articles are obtained in both databases; the number of repeated articles is 6, obtaining a total of 147 articles considered for mapping. The list of articles, as well as the search scripts, can be reviewed in the following link: http://bit.ly/DMCorruption.

3.1. RQ1: How many studies are in the WOS and Scopus databases from 2015 to 2019?

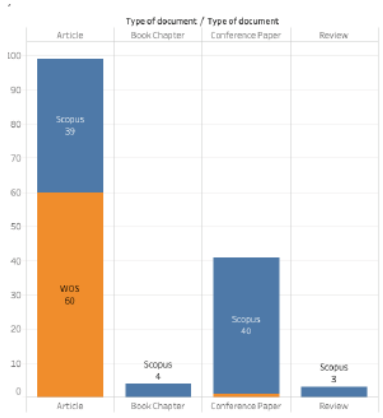


Figure 2 – Research Question 1

Of the 147 resulting articles, 61 correspond to Web of Science (WOS) equivalent to 41.49% and 86 to Scopus equivalent to 58.5%. Regarding the type of document, it is evident that the number of articles in the two databases is equivalent to 67.35%, the number of conference paper is 41, equivalent to 27.89%, of which 40 are from the Scopus database (Figure 2); the representation of book chapters is 2.72% and 2.04% revisions. Respecting the open access articles represent 16.32%.

3.2. RQ2: What are the papers with more cites?

The most cited study (Moro, Cortez, and Rita 2015) (c = 62) refers to an analysis of business intelligence literature for bank fraud using text mining, which extend to the field of study of the banking sector, to studies (7, 20, 30, 72, 84, 107) with appointments greater than 30. Types of corruption: Fraud (77.49%), overpricing (7.05%), bribery (5.05%) and favoritims (4.66%) generate greater citations in the articles. In Table 5, the articles are sorted by the number of appointments followed by the identification number.

Number of cites	Papers
>30	7, 20, 30, 72, 84, 107
21 - 30	52
16 - 20	5, 27,58, 97, 21
11 - 15	15,24,55,128,135, 144, 88,109, 147
6 - 10	2,3,43,76,112,144, 29,12, 16,63, 49,77, 83,102,123
5	39,46,56,96,131
4	35,48,69,103
3	8,37,53,59,71,98,143110,115,118,139,142
2	25,32,36,44,51,57,65,78,85,89,119,121,124, 125,132,146
1	1,6,11,14,26,33,42,45,64,67,68,70,73,80,87,95,99,106,108,111, 130,134,137
0	4,9,10,12,13,17,18,19, 22,23,28,31,34,38,40,41,47,50,54,60,61 62,66,74,75,79,81,82,86,90,91,92,93,94,100,101,104,105,113,116 117,122,126,127,129,133,136,138,140,141,145

Table 5 – Most cited articles

Analyzing the countries, in Turkey, Hungary, Egypt and South Africa, all the articles cited refer to fraud, and in the vast majority of countries they predominate widely Figure 3. Kazahistan is an exception because all quotes are from articles referring to Laundering, and countries such as Iran, Taiwan, France and Neatherlands excels Overpricing

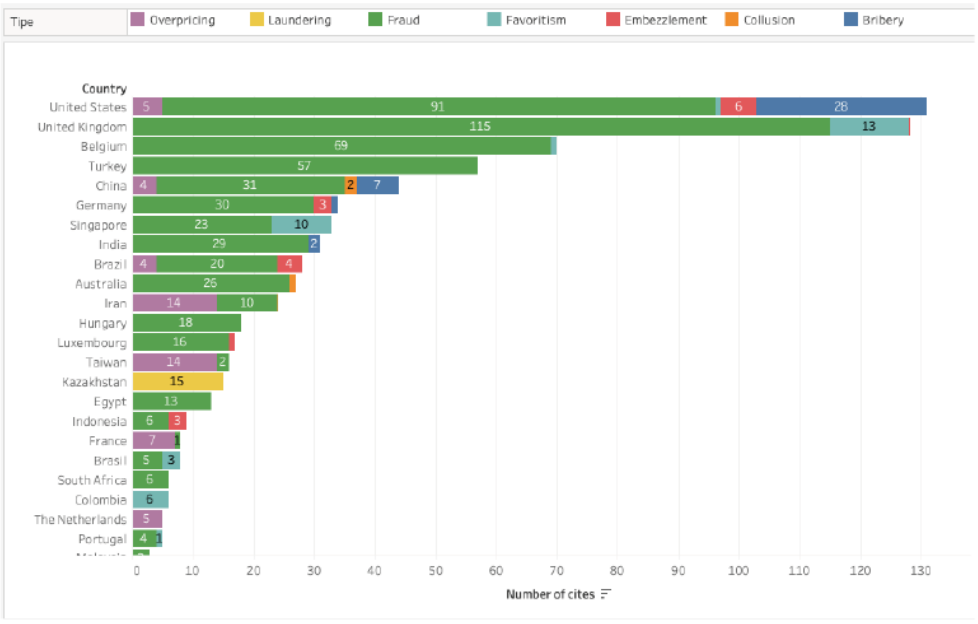


Figure 3 – Frequency of cites by country and type of corruption

3.3. RQ3. What is the geographical distribution of the authors?

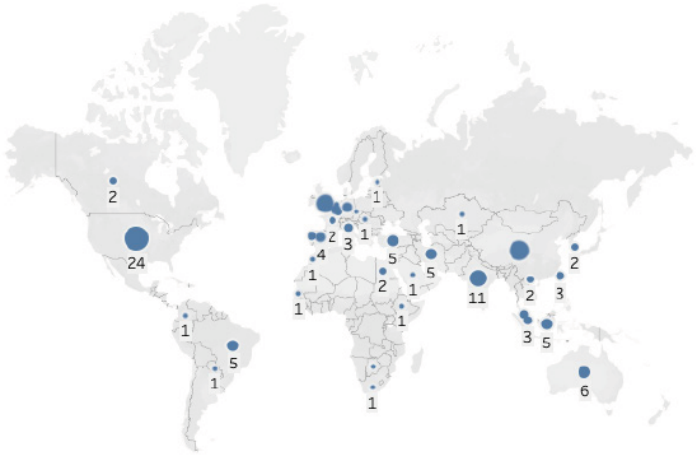


Figure 4 – Frequency of articles by country.

For the geographical distribution of the authors, the first author of the publication was considered; the leading countries with articles on data mining and corruption are United

States (16,32%), China (10,88%) and United Kingdom (8,94%) as shown in Figure 4. In Latin America, Brazil stands out (3,4%), with minimum contributions from Colombia and Paraguay. Therefore, there is an extensive research context in the Latin American social context.

3.4. RQ4: What are the journals with more publications on this line of research?

All journals with at least two scientific articles have been considered. To facilitate their reading, they have been organized by quartile according to Scopus, also considering revisions and proceedings at the emerging level (N / A without quartile location). The journal Expert Systems with Applications (CiteScore: 6.36, IF: 4.577, SNIP: 2.696, SJR: 1.190) houses the largest number of articles (7/147); as detailed in their web portal, this journal has partnered with Heliyon, an open access journal from Elsevier publishing quality peer reviewed research across all disciplines, and to publish papers dealing with the design, development, testing, implementation, and/or management of expert and intelligent systems. Other important journals are Decision Support Journal (CiteScore: 5,97, IF: 3.847, SNIP: 2.448, SJR: 1.536) and International Journal of Accounting Information Systems (CiteScore: 2.24, IF: 1.548, SNIP: 1.096, SJR: 0.478), both from Elsevier, and with a scope related to intelligent systems. As reviewed in these journals, they are an appropriate platform for the future publication of the proposed research results.

Quartil	Journal	Number of articles
N/A	ACM International Conference Proceeding Series	5
	International Journal of Advanced Computer Science and Applications	4
	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	3
	Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	2
	Lecture Notes in Business Information Processing	2
Q1	Expert Systems With Applications	7
	Decision Support Systems	4
	World Development	3
	Neurocomputing	2
	Journal of Development Economics	2
	IEEE Access	2
Q2	Journal of Management Information Systems	3
	Journal of Financial Crime	3
	Crime, Law and Social Change	3
	Computers & Security	2
Q3	International Journal of Accounting Information Systems	5
	Advances in Intelligent Systems and Computing	2

Table 6 – Main Journals from number of articles

3.5. RQ5: In what contexts are these studies developed?

This question is answered according to the type of corruption that is predicted or detected. According to (Ngai et al. 2011) these can be: Fraud. Favoritism, Overpricing, Collusion, Embezzlement, Bribery. In the reviewed works these can be observed that the main type of corruption studied is fraud (72.72%). These results will allow us to establish strategies to consider in the methodology of the study, focusing on the areas with less study and that require a greater contribution of science, as well as the great experience in terms of methods, algorithms, models, etc. used to study fraud, with great investigation coverage.

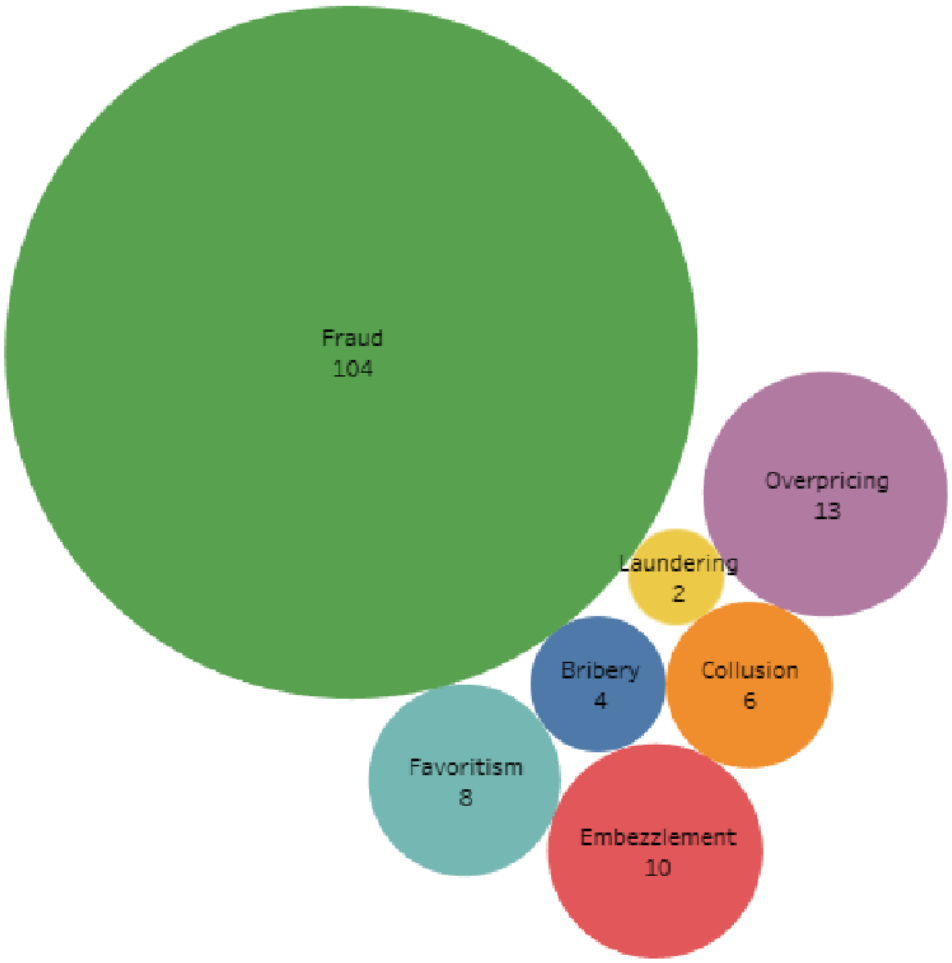


Figure 5 – Frequency of articles by research context

3.6. RQ6: What are the main topics addressed in this line of research?

To solve this question, all the keywords located in the 147 articles were summed up, adding a total of 561. In Figure 6 an analysis is carried out that identifies Fraud Detection, Financial Fraud, Data Mining and Corruption as the main ones lines of investigation. You can see some subtypes of fraud, which is the main topic of study, as discussed in the previous question.

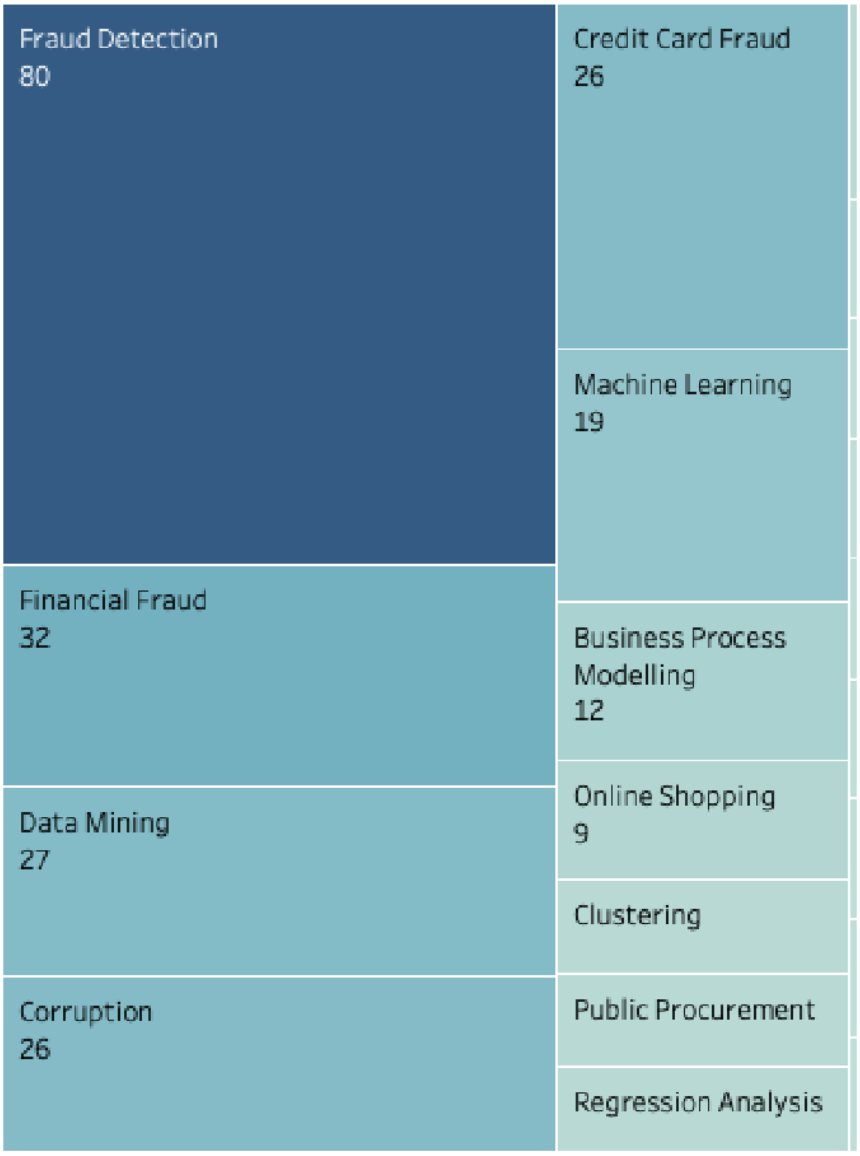


Figure 6 – Main topics of study.

Conclusions

- The articles that refer to fraud as a type of corruption, are the most cited, becoming 100% of citations in several countries in the southern European and African area. Fraud as a form of corruption in terms of citations reaches 77.49%, and other types of corruption, such as overpricing (7.05%), bribery (5.05%) and favoritims (4.66%). The interest of the scientific community in the last five years has focused on Fraud.
- Los países que más publican, no corresponden a los más citdos, así, the United States (16.32%), China (10.88%) and the United Kingdom (8.94%) are the countries with the largest publications on data mining and corruption; estos datos contrastan con los países con mayor cantidad de artículos citados, en donde China representa 7,70% y aparece Bélgica entre los tres países mas citados (9,32%); in Latin America, Brazil publishes 3.4%, and minimum contributions from Colombia and Paraguay. It is considered a research niche on the area in Latin America.
- Fraud as a typology of corruption, is also the one of greatest interest in terms of publication (72.72%), which is directly related to the citation of articles. In the field of technologies Data Mining (4.82%), machine learning (3.39%) and Clustering (1.25%) are emerging tools in this field of science. The other types of corruption and their mechanisms, supplemented with data mining, represent an important field for future research.

References

Alvarez-Jareño, José A, Elena Badal-Valero, and Jose M Pavia. 2019. *Aplicación de Métodos Estadísticos, Económicos y de Aprendizaje Automático Para La Detección de La Corrupción*. https://www.nexos.com.mx/?p=24569 (August 12, 2019).

Cassagne, Juan Carlos., and Enrique. Rivero Ysern. 2007. *La Contratación Pública*. Hammurabi. https://dialnet.unirioja.es/servlet/articulo?codigo=5400799 (July 2, 2019).

Castro Cuenca, Carlos Guillermo. 2017. *La Corrupción Pública y Privada: Causas, Efectos y Mecanismos Para Combatirla - Google Play*. ed. Editorial Universidad del Rosario. https://play.google.com/books/reader?id=2KMyDwAAQBAJ&hl=es&pg=GBS.PT178.w.5.1.35 (July 2, 2019).

Cerillo Martinez, Agustií. 2015. "El Principio de Integridad En La Contratación Pública (Dúo)."

Chan, Albert P C, and Emmanuel Kingsford Owusu. 2017. "Corruption Forms in the Construction Industry : Literature Review." 143(Johnston 1996): 1–12.

Controladoria-Geral da União. 2015. "Observatório Da Despesa Pública — Controladoria-Geral Da União." http://www.cgu.gov.br/assuntos/informacoes-estrategicas/observatorio-da-despesa-publica (August 12, 2019).

DiRienzo, Cassandra E., Jayoti Das, Kathryn T. Cort, and John Burbridge Jr. 2007. "Corruption and the Role of Information." *Journal of International Business Studies* 38: 320–32. https://www.jstor.org/stable/4540422 (August 4, 2019).

García-González, Abel, and María-Soledad Ramírez-Montoya. 2019. "Systematic Mapping of Scientific Production on Open Innovation (2015–2018): Opportunities for Sustainable Training Environments." *Sustainability* 11(6): 1–15.

Goedhuys, Micheline, Pierre Mohnen, and Tamer Taha. 2016. "Corruption, Innovation and Firm Growth: Firm-Level Evidence from Egypt and Tunisia." *Eurasian Business Review* 6(3): 299–322. http://link.springer.com/10.1007/s40821-016-0062-4 (August 4, 2019).

Han, Jiawei., Micheline. Kamber, and Jian. Pei. 2011. *Data Mining : Concepts and Techniques*. Elsevier Science.

Hand, David J. 2013. "Data MiningBased in Part on the Article 'Data Mining' by David Hand, Which Appeared in the *Encyclopedia of Environmetrics* ." In *Encyclopedia of Environmetrics*, Chichester, UK: John Wiley & Sons, Ltd. http://doi.wiley.com/10.1002/9780470057339.vad002.pub2 (August 12, 2019).

Indrajani, H Prabowo, and Meyliana. 2016. "Learning Fraud Detection from Big Data in Online Banking Transactions: A Systematic Literature Review." *Journal of Telecommunication, Electronic and Computer Engineering* 8(3): 127–31. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84984816013&partnerID=40&md5=5f1e56822f7965a82a91a5a3671635c5.

Lee, Ming-Hsuan, and Mon-Chi Lio. 2016. "The Impact of Information and Communication Technology on Public Governance and Corruption in China." *Information Development* 32(2): 127–41. http://journals.sagepub.com/doi/10.1177/0266666914529293 (September 28, 2019).

Martinez Fernandez, Jose Manuel. 2015. "TRANSPARENCIA versus CORRUPCIÓN EN LA CONTRATACIÓN PÚBLICA. MEDIDAS DE TRANSPARENCIA EN TODAS LAS FASES DE LA CONTRATACIÓN PÚBLICA COMO ANTÍDOTO CONTRA LA CORRUPCIÓN." : 0–518.

Moran, J. 2001. "Democratic Transitions and Forms of Corruption." *Crime, Law and Social Change* 36(4): 379–93.

Moro, Sérgio, Paulo Cortez, and Paulo Rita. 2015. "Business Intelligence in Banking: A Literature Analysis from 2002 to 2013 Using Text Mining and Latent Dirichlet Allocation." *Expert Systems with Applications* 42(3): 1314–24.

Ngai, E. W.T. et al. 2011. "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature." *Decision Support Systems* 50(3): 559–69. http://dx.doi.org/10.1016/j.dss.2010.08.006.

OECD. 2005. *Public Sector Integrity Management Framework*.

———. 2007. *Access for Tax Authorities to Information Gathered by Anti-Money Laundering Authorities*. https://www.oecd.org/ctp/exchange-of-tax-information/2389989.pdf (August 4, 2019).

Shrivastava, Utkarsh, and Anol Bhattacherjee. 2015. "ICT as a Corruption Deterrent." In *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development - ICTD '15*, New York, New York, USA: ACM Press, 1–5. http://dl.acm.org/citation.cfm?doid=2737856.2737864 (August 4, 2019).

Torres-Carrion, Pablo Vicente, Carina Soledad Gonzalez-Gonzalez, Silvana Aciar, and Germania Rodriguez-Morales. 2018. "Methodology for Systematic Literature Review Applied to Engineering and Education." *IEEE Global Engineering Education Conference, EDUCON* 2018-April: 1364–73.

Transparencia Internacional. 2017. "El Índice de Percepción de La Corrupción Muestra Un Estancamiento de La Lucha Contra La Corrupción En La Mayoría de Los Países." *El índice de percepción de la corrupción muestra un estancamiento de la lucha contra la corrupción en la mayoría de los países*.

Vargas-Hernández, José G. 2009. "The Multiple Faces of Corruption: Typology, Forms and Levels." *SSRN Electronic Journal*: 43.

Volosin, Natalia A. 2015. *Datos Abiertos, Corrupción y Compras Públicas*. https://idatosabiertos.org/wp-content/uploads/2015/10/5.-Corrupcion-y-compras-publicas-Volosin1.pdf (August 4, 2019).

Woodie, Alex. 2016. "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/ (August 12, 2019).

Artificial Intelligence Techniques to Detect and Prevent Corruption in Procurement: A Systematic Literature Review

Yeferson Torres Berru[1,2(✉)], Vivian Félix López Batista[2(✉)], Pablo Torres-Carrión[3(✉)], and Maria Gabriela Jimenez[2]

[1] University of Salamanca, Plaza de la Merced, s/n, 37008 Salamanca, Spain
ymtorresb@usal.es
[2] Instituto Superior Tecnológico Loja,
Av. Granada y Turunuma, 1101608 Loja, Ecuador
vivian@usal.es
[3] Universidad Técnica Particular de Loja,
San Cayetano Alto S/N, 1101608 Loja, Ecuador
pvtorres@utpl.edu.ec

Abstract. Transparency International estimates that the costs of corruption in public procurement reach between 20 and 25% of the contract value, sometimes reaching 40–50%. In this study, we analyzed differentness kinds of corruption like (bribery, collusion embezzlement, misappropriation, fraud, abuse of discretion, favoritism, nepotism), and six types of Artificial Intelligence techniques (classification, regression, clustering, prediction, outlier detection, and visualization). The methodology proposed by Torres-Carrion was used, and four research questions were raised, which allow knowing the types of research carried out, the characteristics of the organizations in which the investigations are carried out, the technological tools, and data mining methodologies and techniques. The search was done in the Scopus and Web of Science databases, getting 102 articles published between 2015 and 2019. The primary data mining techniques used are logistic models, neural networks, Bayesian networks, supported vector machines, and decision trees.

AQ1

Keywords: Artificial Intelligence · Corruption · Data mining · Procurement · Systematic literature review

1 Introduction

Corruption expenditures are equivalent to 5% of global GDP, according to the G-20 [1], being the third most lucrative "industry" of all those in the world. Transparency International estimates that the costs of corruption in public contracts average 20–25% of the contract value, and can reach 40–50% in some cases [2]; public procurement accounted for 32.5% of government expenditure. The highest risk of corruption in hiring occurs during the planning stage, potential frauds in the procurement system take very diverse forms starting from bribery, collusion embezzlement, misappropriation, fraud, abuse of discretion, favoritism, nepotism [3].

M. Botto-Tobar et al. (Eds.): ICAT 2019, CCIS 1194, pp. 1–15, 2020.
https://doi.org/10.1007/978-3-030-42520-3_21

Data mining combines Artificial Intelligence techniques (classification, regression, clustering, prediction, outlier detection, and visualization) with statistical analysis techniques (clustering, dimensional analysis, etc.), which allow analyzing information such as data, text, images, audio. In automatic learning, algorithms can be classified as supervised, unsupervised, and reinforcement [4]. These fields of knowledge, discussed for this systematic review of the literature (SLR), are detailed in Fig. 1, which allows establishing the theoretical knowledge base on which the review has its basis to contribute to the knowledge about data mining research and artificial intelligence for the detection and prevention of anomalies in contracts.

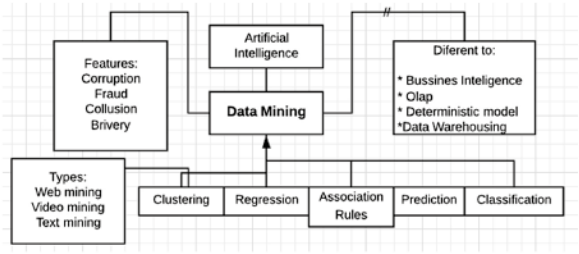


Fig. 1. Theoretical representation of data mining in contracts

1.1 Related Systematic Review of Literature

After the SLR, we proceed to look for previous literature review works related to the theoretical constructs: AI, data mining, and corruption. In the results obtained (see Table 1), the studies on the different types of fraud, at the private and governmental levels, are discussed, the different techniques of automatic learning and data mining are also considered in the process of fraud detection and prevention. All research is more than five years old and does not answer the research questions proposed in this study.

Table 1. Related reviews

Article	Analysis	#Papers
Fraud detection system: a survey	Survey about fraud prevention systems five areas of electronic frauds (e-fraud): credit card, telecommunication, health care insurance, automobile insurance and online auction in papers between 1994 and 2014	70
A survey of machine-learning and nature-inspired based credit card fraud detection techniques [5]	The survey revealed that various machine learning-based and nature-inspired algorithms had been used to handle credit card fraud detection between 2010 and 2014	47
The application of data mining techniques in financial fraud detection: a classification framework and an academic review of the literature [6]	Review data mining techniques in financial fraud, the articles were published between 1997 and 2008 analyzing fraud in bank, insurance and securities categories, the six data mining techniques (classification, regression, clustering, prediction, outlier detection, and visualization) and the main artificial intelligence methods logistic models, neural networks, the Bayesian belief network, and decision trees	49

Section two details the applied methodology, research questions, inclusion criteria, and semantic structure of the search.; in section three, research questions are answered based on the 102 articles found; finally, in section four, conclusions and future work are presented.

2 Method

2.1 Method for SLR

We used the method for a systematic review of the literature by Torres-Carrión [7] adapted from Kitchenham [8], which divides the process into three phases: planning, conducting the review, and reporting the review whit the PRISMA method [9]. As part of planning the search process, several general and specific inclusion and exclusion criteria were defined.

2.2 Research Questions

RQ1: What methods are being applied to investigate corruption in public procurement contracts?
RQ2: What are the characteristics of the organizations in which the research has been carried out?
RQ3: What technological tools are being used to investigate the detection and prevention of corruption?
RQ4: What algorithms, methodologies, and data analysis tools are used to detect corruption?

2.3 Quality Inclusion and Exclusion Protocols

As inclusion criteria, scientific articles published in the Web of Science (WOS) and Scopus databases during the years 2015–2019 are considered to be the type of document (article, review, editorial or conference proceedings), research area (Computer Science, Social Sciences and Decision Sciences); it excludes repeated documents, short papers, posters, and book chapters. As a quality criterion, a detailed review of the articles is carried out, filtering the studies that do not analyze corruption and its forms, or that do not apply data mining and artificial intelligence.

2.4 Semantic Search Structure

From the theoretical constructs (see Fig. 1), synonyms are sought in the scientific thesaurus; operators are applied (OR, AND, W/) to optimize the search, although Table 2 only shows the search query for Scopus, WOS search replaces the operators with the correspondents to its database. The procedure is detailed in levels with the resulting number of articles, as the inclusion, exclusion, and quality criteria are applied. One hundred two articles are obtained, identified as valid and explicitly related to the problem raised, and with which this SLR is working. AQ2

Table 2. Semantic search structure

Level	Thesaurus	SCOPUS script	Scopus	WOS
L1	Data mining			
	Mining W/4 (data or video or text or web)) Classificat* Cluster* Regression Association rules Detection Prediction Sequential Patterns	(Mining W/4 (data OR video OR text OR web)) OR classificat* OR cluster* OR regression OR (association W/2 rules) OR detection OR prediction OR (sequential W/2 patterns) OR (learning W/4 (machine OR deep OR reinforced))	6,962,346	8,072,286
L2	Corruption	Corruption OR	9,653	35,781
	Bribery, collusion, embezzlement, fraud, abuse of discretion, favoritism, nepotism	Bribery OR collusion OR (embezzlement OR misappropriation) OR fraud OR (abuse W/0 of W/0 discretion) OR favoritism OR nepotism)		
L3	Contracts	(Contract OR	691	1.811
	Contract, purchase, investment, procurement, acquisition, acquirement, tendering	Purchase OR investment OR procurement OR acquisition OR acquirement OR tendering)		
L4	Review protocol			
	(Last 5 years) from 2015		351	796
	(Research Areas) Computer Science, Social Sciences, Decision Sciences		222	275
	Article or review or editorial or conference proceedings		210	255
	Quality criteria		58	54
L5 Combination of results in Scopus and WOS (repeated = 5)			**102**	

3 Results

The results are presented in conformity with the research questions established in the methodology, and their corresponding variables and indicators.

RQ1: What methods are being applied to investigate corruption in public procurement contracts?

Knowing the methodology used by other researchers gives light to the planning of new studies. 87% of investigations work with a previously structured database; it is also observed that techniques such as web Scraping (3%) are rarely used for data collection (see Table 3). Quantitative research is the most used, as well as the statistical method of correlation. As for the time of validity of the data for the study, most of them

are 1–3 years, and most of the research is an experimental type, with contributions to computer science. In this sense, the research highlights [10], which performs research with multivariate analysis and correlation, using web scraping and 4-year data.

Table 3. Research question 1.

Data collection instrument [11, 12]		*f*
Survey	[13–18]	6
Web Scraping	[13, 19]	2
Database	[10, 20–76]	57
Type of research [12]		
Qualitative	[77–79]	3
Quantitative	[10, 11, 13–102]	90
Mixt	[5, 103–110]	9
Statistics to evaluate results [12]		
Univariate	[14, 24, 26, 34, 64, 66, 80–90]	17
Multi-varied	[70, 91, 93]	6
Correlation	[10, 13, 15–23, 25, 27–33, 35–39, 41–48, 50–61, 65, 67–69, 71–77, 81, 89, 92, 95, 97–100, 111]	79
Data period		
<1 year	[26, 28, 29, 32, 40, 81]	6
1> & $\leq$ 3	[41, 43, 46, 47, 52, 53, 59, 61, 64–67, 71, 72, 75, 76, 95, 101, 102]	23
>3& $\leq$ 5	[10, 11, 13, 21, 27, 39, 49, 86, 97]	9
>5& $\leq$ 10	[92, 99]	2
>10	[20, 46, 52, 60, 82]	5
Research design [12]		
Experimental	[10, 11, 13, 16, 17, 19, 20, 22–26, 29–35, 37–76, 80, 86, 88, 91, 92, 95–97, 99, 100, 111]	70
No experimental	[5, 14, 23, 24, 36, 77–79, 82–85, 87–89, 94, 95, 98, 101, 104, 110]	23
Quasi experimental	[15, 18, 21, 81]	4
Field of science		
Computer Science	[11, 13, 17–26, 29–34, 36–40, 45–59, 62–76, 80, 81, 86–88, 90, 92, 94, 95, 99, 103, 104, 110, 111]	70
Economy	[27, 28, 35, 77–79, 82, 84, 85, 100]	10
Mathematic	[14, 34, 83, 89]	4
Statistics	[10, 16, 39, 61, 80, 91, 96]	6

RQ2: What are the characteristics of the organizations in which the research has been carried out?

50% of the works focus on the private sector, with a significant presence of experimental type works that relate the public and private sectors (17%). The main commercial activity of the organizations is the provision of services, having the banking sector as the most studied (64%) (see Table 4). In the government sector, tax fraud is prominent (33%), and public purchases account for only 6% of total investigations.

Table 4. Research question 2.

Activity sector		*f*
Public	[10, 16, 18, 24, 34, 35, 39, 42, 45, 47, 48, 50, 54, 57, 65, 66, 68, 71, 76, 77, 81, 84–87, 89, 91, 92, 97, 100]	31
Private	[11, 13, 15, 16, 18, 19, 25–29, 31–33, 36–38, 43, 46, 51–53, 58–64, 70, 72, 75, 88, 90, 91, 101, 102, 104–107, 111]	49
Mixt	[10, 11, 17, 30, 44, 49, 78–80, 83, 92, 93]	17
Commercial activity		
Services	[5, 13, 18, 21–23, 25–29, 33, 36–38, 47, 50–53, 55–57, 59, 62, 63, 70–72, 76, 81, 88, 90, 92, 104, 111]	36
Commercial	[19, 31, 32, 49, 58, 64, 65, 94]	8
Government	[10, 11, 16, 24, 35, 39, 41–43, 45, 46, 48, 54, 61, 67, 68, 86, 87, 89, 91, 100]	22
Organization		
Bank	[5, 21, 22, 25, 26, 29, 33, 36, 37, 47, 51, 55, 59, 62, 63, 70–72, 76, 79, 88, 90, 96, 104, 111]	25
Buys online	[32, 65]	2
Public buys	[10, 16, 35, 42, 43, 54, 69]	7
Taxes	[41, 67, 87, 91, 100]	5

In the mixed activity sector, the research conducted by Dhurandhar [30] proposes a bigdata-based solution for risk analysis in public and private sector procurement; focused on the public sector, in [46] a framework for the detection of crimes in contracts is presented, supported by information provided by the World Bank; in the private sector in [90] an experimental analysis of data mining tools in the banking sector is conducted.

RQ3: What technological tools are being used to investigate the detection and prevention of corruption?

Twenty articles (20/112) are selected that in their methodology propose the use or development of a technological tool to evaluate corruption (see Table 5); 89% are desktop tools; in terms of web platforms, these analyze corruption in contracts in public procedures [30, 69] and in the process of buying medicines [13].

The degree of reliability in the detection or prevention of corruption, as the case may be, has as its greatest interval between the 80%-90%; Baader et al. [17] obtain the lowest detection rate with 48.6%, applying the red flag approach with mining process to reduce the number of false positives in fraud analysis, whereas Darwish [23] obtains the best detection rate with a 98,5% in the analysis of credit card transactions in the banking environment; in the government environment, [43] work detects fake suppliers from the analysis of satellite images of the locations of companies, obtaining the best index of detection (97%). The pattern of software engineering and the computer security standard used in the tools were also analyzed, but no coincidences were found in the analysis of the work, excepting Carminati [71], which, in addition to the analysis tool, presents a mechanism for preventing the Mimicry Attack.

Table 5. Research question 3.

Platform		*f*
Desktop	[19, 23, 24, 31, 38–40, 46, 48, 52, 56, 58, 61, 64, 86, 92, 95]	17
Web	[13, 30, 69]	3
Percent of detection/prevention		
<70	[17, 24]	2
70–80	[46, 56]	2
80–90	[10, 48, 52–54, 60, 63, 64, 68]	9
91–95	[13, 32, 38, 57]	4
96–99	[23, 25, 29, 43, 65, 67]	6

RQ4: What algorithms, methodologies, and data analysis tools are used to detect corruption?

In this question (see Table 6), we evaluated factors like the data source, type, kind of mining, preprocessing methods, outliers values processing, evaluation metrics, artificial intelligence techniques, types and learning techniques, and technological tools.

Table 6. Research question 4

Data source		*f*
Re. Public	[18–24, 26, 30, 41, 43, 45, 46, 54, 61, 80, 86]	17
Re. Private	[11, 13, 25, 26, 38, 46–49, 55–57, 59, 63, 65, 67, 72, 92, 95, 111]	30
Type of data		
Data	[10, 13, 16, 17, 20–23, 25–27, 30–33, 38, 40–46, 48, 50, 53, 55, 57–61, 65, 66, 68, 70, 72, 74, 75, 78–82, 87, 90–92, 95–100, 102, 103, 105, 106, 109]	59
Text	[31, 49, 62, 64]	4
Audio	[54]	1

(*continued*)

Table 6. (*continued*)

Type of minning		
Predictive	[13, 21, 25, 46, 55–57, 59–61, 80, 88, 97]	13
Descriptive	[10, 11, 13, 16, 18–20, 22, 23, 25, 26, 29–32, 37–45, 47–49, 51–54, 63–65, 67–70, 72–75, 80, 86, 91, 92, 95, 96, 111]	51
Preprocessing of data		
Empirical assessment	[19, 25, 39, 41, 45–47, 55, 63, 75, 86]	13
Other	Cascade generalization [70] CKIP [48] Monroe and The [60]	3
Outlier values		
HBOS	[29, 45, 48, 57, 71, 80]	6
PSO	[48, 57, 59, 75]	4
Metric evaluation		
Confusion matrix	[13, 17, 18, 20, 23–25, 29, 32, 43, 49, 53, 60, 68, 80, 86, 111]	17
Curve ROC	[22, 26, 47, 54, 67, 71, 75]	7
Fraud Score	[53, 63]	3
K-fold	[57, 95]	2
Other	Matthews Correlation Coefficient [71]	
Artificial Intelligence technics		
Bayes based	[31, 38, 44, 45, 49, 52–54, 57, 60, 73, 99]	12
Neural network	[18, 19, 22, 37, 41, 47, 52, 68, 88]	9
SVM	[41, 44, 45, 48, 49, 52, 55, 87, 109]	9
Decision tree	[26, 29, 32, 51, 52, 55, 67]	7
Random Forest	[24, 25, 51, 55, 63, 67, 70]	7
Logistic- Linear regression	[14, 34, 41, 44, 45, 49, 51, 52, 67, 83, 87, 97]	12
Other	[11, 20, 36, 42, 44, 48–52, 56, 58, 59, 64, 70, 72, 86, 88]	17
Learning techniques		
Machine learning	[13, 18, 19, 23–26, 29–32, 37, 42, 46, 48–51, 53–57, 59, 64, 65, 67–70, 72–74, 76, 80, 81, 88, 91, 95, 111]	41
Deep learning	[38, 43, 47, 49]	4
Minning techniques		
Clustering	[18, 29, 30, 50, 59, 68, 70, 72, 73, 75, 76, 103]	41
Classification	[13, 18, 24–26, 29, 31, 32, 37, 38, 41, 43, 45–49, 52–54, 75, 76, 81, 95]	51
Regression	[41, 76]	2
Types of learning		
Supervised	[18–20, 25, 26, 31, 32, 37, 38, 42, 54, 55, 57, 68, 91]	15
No supervised	[18, 24, 30, 56, 59, 73]	6
Semi supervised	[29, 70, 71]	3

(*continued*)

Table 6. (*continued*)

Technological tools		
Java	[19, 64, 65, 95]	4
MatLab	[23, 41, 59, 64]	4
Python	[24, 31, 46, 65, 69, 80]	6
Weka	[44, 45, 52]	3
Others	[13, 38, 42, 50, 53, 57, 59, 61, 62, 70, 86, 92, 95]	13

Private datasets predominate in data collection sources 63% (see Table 6); this is because, as mentioned above, most work focuses on the private sector; the type of data used to generate and validate detection models in a large percentage is data (database or dataset). It should be noted that only 6% of jobs use text (documents), and only one job [54], the audio of telephone conversations are used to find fraud in public purchases.

The review focuses on the prevention and detection of corruption, with 79% of investigations of a descriptive type (detection) and 21% of a predictive type (prevention); in the techniques for the pre-processing of data, empirical assessment predominates with 81%, and amongst the methods are Cascade generalization [70] Chinese Knowledge Information Processing Group (CKIP) [48] Monroe and The [60]. In order to detect outliers in the data, the following method is used in similar percentages Particle Swarm Optimization (PSO) and the histogram-based outlier score (HBOS).

It is observed that the main techniques of artificial intelligence are those based on the theorem of Bayes, neural networks, Support Vector Machines (SVM), decision trees, Random Forest, and logistic and linear regression, reaching 76% among all of them; to a lesser extent technique such as: convolutional networks, tough set theory, graphs, natural language processing, kmeans, AdaBoost, genetic algorithms, bagging and logitBoos. The investigation with the highest number of applied techniques is [49], evaluating eleven artificial intelligence techniques to detect price manipulation in purchases.

The main evaluation metrics are: accuracy, efficiency, recalling using methods such as confusion matrix and ROC curves; in assessing of corruption, the significant contribution of the fraud score used as an evaluation indicator in some works, the confusion matrix is combined with the fraud score [53], and machine learning evaluation methods such as the Matthews Correlation Coefficient [71].

Concerning computer tools for programming, processing, and data storage Java, MatLab, Weka y Python have the highest percentage, with other tools such as R, RapidMiner, Hadoop, Spark, Neo4j, Casandra, Kafta, Visual Studio.

4 Conclusions and Future Work

Experimental quantitative research is the most widely used, as well as the statistical method of correlation. Most of the study is conducted with data over 1–3 years, with a significant contribution to computer science. The main commercial activity of the

organizations is the provision of services, specifically the banking sector. In the government sphere, tax fraud is noteworthy, with a lesser presence of public procurement processes.

Web Scraping is a rarely used technique for obtaining data on corruption studies in contracts and can be used as a basis for future work. The few jobs related to contract analysis in public procurement use datasets and are not considered documents as the initial basis for data analysis. It is also evident that the computer tools created to carry out corruption analysis in contracts in both the public and private sectors are not considered computer security standards, and the percentage of tools in the web environment is very low.

The main artificial intelligence techniques found are logistic models, neural networks, bayesian networks, and supported vector machines. As future work, they would be enhanced with mixed learning methods. Fraud Score is proposed as a specific metric to assess the risk of corruption without leaving aside the metrics used to evaluate from the confusion matrix and ROC curves, with machine learning and supervised learning as the main types of technique.

References

1. Báez 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) AQ3
2. Volosin, N.A.: Datos abiertos, corrupción y compras públicas (2015)
3. Padhi, S.S., Mohapatra, P.K.J.: Detection of collusion in government procurement auctions. J. Purch. Supply Manag. **17**, 207–221 (2011)
4. Martin, R.: A review of the literature of the followership since 2008: the importance of relationships and emotional intelligence. SAGE Open **5**, 2158244015608421 (2015)
5. Adewumi, A.O., Akinyelu, A.A.: A survey of machine-learning and nature-inspired based credit card fraud detection techniques. Int. J. Syst. Assur. Eng. Manag. **8**, 937–953 (2017)
6. Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y., Sun, X.: The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. Decis. Support Syst. **50**, 559–569 (2011)
7. Torres-Carrion, P.V., Gonzalez-Gonzalez, C.S., Aciar, S., Rodriguez-Morales, G.: Methodology for systematic literature review applied to engineering and education. IEEE Global Engineering Education Conference EDUCON, pp. 1364–1373, April 2018
8. Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., Linkman, S.: Systematic literature reviews in software engineering – a systematic literature review. Inf. Softw. Technol. **51**, 7–15 (2009)
9. Moher, D., et al.: PRISMA-P: preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P). Syst. Rev. 1–9 (2015) AQ4
10. Auriol, E., Straub, S., Flochel, T.: Public procurement and rent-seeking: the case of paraguay. World Dev. **77**, 395–407 (2016)
11. Lei, M., Yin, Z., Li, S., Li, H.: Detecting the collusive bidding behavior in below average bid auction. In: ICNC-FSKD 2017 - 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, pp. 1720–1727 (2018)
12. Hernández Sampieri, R., Fernández Collado, C., Baptista Lucio, M.: Metodología de la Investigación (2010)

13. Kose, I., Gokturk, M., Kilic, K.: An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Appl. Soft Comput. J. **36**, 283–299 (2015)
14. Charles Andoh, E., Ofosu-Hene, D.: Causes, effects and deterrence of insurance fraud: evidence from Ghana. J. Financ. Crime Iss. **5**, 39–44 (2016)
15. Huang, S.Y., Lin, C.C., Chiu, A.A., Yen, D.C.: Fraud detection using fraud triangle risk factors. Inf. Syst. Front. **19**, 1343–1356 (2017)
16. Seck, A.: Heterogeneous bribe payments and firms' performance in developing countries. J. African Bus. **21**, 1–20 (2019)
17. Baader, G., Krcmar, H.: Reducing false positives in fraud detection: combining the red flag approach with process mining. Int. J. Account. Inf. Syst. **31**, 1–16 (2018)
18. Choi, D., Lee, K.: An artificial intelligence approach to financial fraud detection under IoT environment: a survey and implementation. Secur. Commun. Netw. **2018** (2018)
19. Sadaoui, S., Wang, X.: A dynamic stage-based fraud monitoring framework of multiple live auctions. Appl. Intell. **46**, 197–213 (2017)
20. Yeh, C.C., Chi, D.J., Lin, T.Y., Chiu, S.H.: A hybrid detecting fraudulent financial statements model using rough set theory and support vector machines. Cybern. Syst. **47**, 261–276 (2016)
21. Ouenniche, J., Uvalle Perez, O.J., Ettouhami, A.: A new EDAS-based in-sample-out-of-sample classifier for risk-class prediction. Manag. Decis. **57**, 314–323 (2019)
22. Zakaryazad, A., Duman, E.: A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing. Neurocomputing **175**, 121–131 (2014)
23. Darwish, S.M.: An intelligent credit card fraud detection approach based on semantic fusion of two classifiers. Soft. Comput. **24**, 1243–1253 (2019)
24. Kehler, E., Paciello, J., Pane, J.: Anomaly detection in public procurements using the open contracting data standard (2019) AQ5
25. Van Vlasselaer, V., et al.: APATE: a novel approach for automated credit card transaction fraud detection using network-based extensions. Decis. Support Syst. **75**, 38–48 (2015)
26. Zareapoor, M., Shamsolmoali, P.: Application of credit card fraud detection: based on bagging ensemble classifier. Procedia Comput. Sci. **48**, 679–685 (2015)
27. Ngoc, B.H., Hai, D.B., Chinh, T.H.: Assessment of the should be effects of corruption perception index on foreign direct investment in ASEAN countries by spatial regression method. In: Anh, Ly H., Dong, L.S., Kreinovich, V., Thach, N.N. (eds.) ECONVN 2018. SCI, vol. 760, pp. 421–429. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73150-6_33
28. Burböck, B., Macek, A., Podhovnik, E., Zirgoi, C.: Asymmetric influence of corruption distance on FDI. J. Financ. Crime (2018) AQ6
29. Carminati, M., Caron, R., Maggi, F., Epifani, I., Zanero, S.: BankSealer: a decision support system for online banking fraud analysis and investigation. Comput. Secur. **53**, 175–186 (2015)
30. Dhurandhar, A., Graves, B., Ravi, R., Maniachari, G., Ettl, M.: Big data system for analyzing risky procurement entities. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1741–1750 August 2015
31. Hooi, B., et al.: BIRDNEST: Bayesian inference for ratings-fraud detection. In: 16th SIAM International Conference on Data Mining 2016, SDM 2016, pp. 495–503 (2016)
32. Snyder, P., Kanich, C.: Characterizing fraud and its ramifications in affiliate marketing networks. J. Cybersecur. **2**, 71–81 (2016)

33. Moalosi, M., Hlomani, H., Phefo, O.S.D.: Combating credit card fraud with online behavioural targeting and device fingerprinting. Int. J. Electron. Secur. Digit, Forensics (2019)
34. Anh, N.N., Minh, N.N., Tran-Nam, B.: Corruption and economic growth, with a focus on Vietnam. Crime, Law Soc. Change **45**, 307–324 (2016)
35. Ferwerda, J., Deleanu, I., Unger, B.: Corruption in public procurement: finding the right indicators. Eur. J. Crim. Policy Res. **23**, 245–267 (2017)
36. Amanze, B.C., Onukwugha, C.G.: Credit card fraud detection system in nigeria banks using adaptive data mining and intelligent agents: A review. Int. J. Sci. Technol. Res. **7**, 175–184 (2018)
37. Zanin, M., Romance, M., Moral, S., Criado, R.: Credit card fraud detection through parenclitic network analysis. Complexity **2018** (2018)
38. Randhawa, K., Loo, C.K., Seera, M., Lim, C.P., Nandi, A.K.: Credit card fraud detection using AdaBoost and majority voting. IEEE Access **6**, 14277–14284 (2018)
39. Ausloos, M., Cerqueti, R., Mir, T.A.: Data science for assessing possible tax income manipulation: the case of Italy. Chaos, Solitons Fractals **104**, 238–256 (2017)
40. Helmy, T.H., Zaki, M., Salah, T., Badran, K.: Design of a monitor for detecting money laundering and terrorist financing. J. Theor. Appl. Inf. Technol. **85**, 425–436 (2016)
41. Rahimikia, E., Mohammadi, S., Rahmani, T., Ghazanfari, M.: Detecting corporate tax evasion using a hybrid intelligent system: a case study of Iran. Int. J. Account. Inf. Syst. **25**, 1–17 (2017)
42. Van Erven, G.C.G., Carvalho, R.N., De Holanda, M.T., Ralha, C.: Graph database: a case study for detecting fraud in acquisition of Brazilian Government (Banco de Dados em Grafo: Um Estudo de Caso em Detecção de Fraudes no Governo Brasileiro). In: Iberian Conference on Information Systems and Technologies CISTI, pp. 1–6 (2017)
43. Wacker, J., Ferreira, R.P., Ladeira, M.: Detecting fake suppliers using deep image features (2018)
44. Kim, Y.J., Baik, B., Cho, S.: Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. Expert Syst. Appl. **62**, 32–43 (2016)
45. Dutta, I., Dutta, S., Raahemi, B.: Detecting financial restatements using data mining techniques. Expert Syst. Appl. **90**, 374–393 (2017)
46. Grace, E., Rai, A., Redmiles, E., Ghani, R.: Detecting fraud, corruption, and collusion in international development contracts: the design of a proof-of-concept automated system (2016)
47. Gómez, J.A., Arévalo, J., Paredes, R., Nin, J.: End-to-end neural network architecture for fraud scoring in card payments. Pattern Recognit. Lett. **105**, 175–181 (2018)
48. Chen, Y.J., Wu, C.H., Chen, Y.M., Li, H.Y., Chen, H.K.: Enhancement of fraud detection for narratives in annual reports. Int. J. Account. Inf. Syst. **26**, 32–45 (2017)
49. Wang, Q., Xu, W., Huang, X., Yang, K.: Enhancing intraday stock price manipulation detection by leveraging recurrent neural networks with ensemble learning. Neurocomputing **347**, 46–58 (2019)
50. Tan, M., Lee, W.-L.: Evaluation and improvement of procurement process with data analytics. Int. J. Adv. Comput. Sci. Appl. **6**, 70–80 (2015)
51. Correa Bahnsen, A., Aouada, D., Stojanovic, A., Ottersten, B.: Feature engineering strategies for credit card fraud detection. Expert Syst. Appl. **51**, 134–142 (2016)
52. Li, H., Wong, M.-L.: Financial fraud detection by using grammar-based multi-objective genetic programming with ensemble learning (2015)
53. Throckmorton, C.S., Mayew, W.J., Venkatachalam, M., Collins, L.M.: Financial fraud detection using vocal, linguistic and financial cues. Decis. Support Syst. **74**, 78–87 (2015)

54. Arief, H.A.A., Saptawati, G.A.P., Asnar, Y.D.W.: Fraud detection based-on data mining on Indonesian E-Procurement System (SPSE). In: Proceedings of 2016 International Conference on Data and Software Engineering, ICoDSE 2016 (2017)
55. Vimala Devi, J., Kavitha, K.S.: Fraud detection in credit card transactions by using classification algorithms. In: International Conference on Current Trends in Computer, Electrical, Electronics and Communication, CTCEEC 2017, pp. 125–131 (2018)
56. Zhou, H., Chai, H.F., Qiu, M.L.: Fraud detection within bankcard enrollment on mobile device based payment using machine learning. Front. Inf. Technol. Electron. Eng. **19**, 1537–1545 (2018)
57. Hooda, N., Bawa, S., Rana, P.S.: Fraudulent firm classification: a case study of an external audit. Appl. Artif. Intell. **32**, 48–64 (2018)
58. Fu, Y., Liu, G., Papadimitriou, S., Xiong, H., Li, X., Chen, G.: Fused latent models for assessing product return propensity in online commerce. Decis. Support Syst. **91**, 77–88 (2016)
59. Demiriz, A., Ekizoğlu, B.: Fuzzy rule-based analysis of spatio-temporal ATM usage data for fraud detection and prevention1. J. Intell. Fuzzy Syst. **31**, 805–813 (2016)
60. Chimonaki, C., Papadakis, S., Vergos, K., Shahgholian, A.: Identification of financial statement fraud in Greece by using computational intelligence techniques. In: Mehandjiev, N., Saadouni, B. (eds.) FinanceCom 2018. LNBIP, vol. 345, pp. 39–51. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-19037-8_3
61. Correa, M.A.O.S., Galindo Leal, A.: Identification of overpricing in the purchase of medication by the Federal Government of Brazil, using text mining and clustering based on ontoiogy (2018)
62. Alzaidi, A.A.: Impact of use of big data in decision making in banking sector of Saudi Arabia. Int. J. Comput. Sci. Netw. Secur. **18**, 72–80 (2018)
63. Kasa, N., Dahbura, A., Ravoori, C., Adams, S.: Improving credit card fraud detection by profiling and clustering accounts (2019)
64. Chen, Y.-J., Wu, C.-H.: On big data-based fraud detection method for financial statements of business groups (2017)
65. Weng, H., et al.: Online e-commerce fraud: a large-scale detection and analysis (2018)
66. Torres, C.F., Schütte, J., State, R.: Osiris: hunting for integer bugs in ethereum smart contracts (2018)
67. Lismont, J.: Predicting tax avoidance by means of social network analytics. Decis. Support Syst. **108**, 13–24 (2018)
68. Zhang, H., Wang, L.: Prescription fraud detection through statistic modeling (2018)
69. Martínez-Plumed, F., Casamayor, J.C., Ferri, C., Gómez, J.A., Vendrell Vidal, E.: SALER: a data science solution to detect and prevent corruption in public administration. In: Alzate, C., et al. (eds.) ECML PKDD 2018. LNCS (LNAI), vol. 11329, pp. 103–117. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-13453-2_9
70. Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.A., Caelen, O., Mazzer, Y., Bontempi, G.: SCARFF: a scalable framework for streaming credit card fraud detection with spark. Inf. Fusion. **41**, 182–194 (2018)
71. Carminati, M., Polino, M., Continella, A., Lanzi, A., Maggi, F., Zanero, S.: Security evaluation of a banking fraud analysis system. ACM Trans. Priv. Secur. **21**, 1–31 (2018)
72. Robinson, W.N., Aria, A.: Sequential fraud detection for prepaid cards using hidden Markov model divergence. Expert Syst. Appl. **91**, 235–251 (2018)
73. Ekin, T., Ieva, F., Ruggeri, F., Soyer, R.: Statistical medical fraud assessment: exposition to an emerging field. Int. Stat. Rev. **86**, 379–402 (2018)
74. Fauzan, A.C., Sarno, R., Ariyani, N.F.: Structure-based ontology matching of business process model for fraud detection. In: ICTS 2017, pp. 221–225 (2018)

75. Saghehei, E., Memariani, A.: Suspicious behavior detection in debit card transactions using data mining: a comparative study using hybrid models. Inf. Resour. Manag. J. **28**, 1–14 (2015)
76. El-kaime, H., Hanoune, M., Eddaoui, A.: The data mining: a solution for credit card fraud detection in banking. In: Mizera-Pietraszko, J., Pichappan, P., Mohamed, L. (eds.) RTIS 2017. AISC, vol. 756, pp. 332–341. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-91337-7_31
77. Schlenther, B.O.: Addressing illicit financial flows in Africa: how broad is the whole of government approach supposed to be? J. Financ. Crime (2016)
78. Sadaf, R., Oláh, J., Popp, J., Máté, D.: An investigation of the influence of theworldwide governance and competitiveness on accounting fraud cases: a cross-country perspective. Sustain **10**, 1–11 (2018)
79. Wang, H., Chen, H.M.: Deterring bidder collusion: auction design complements antitrust policy. J. Compet. Law Econ. **12**, 31–68 (2016)
80. Wahid, A., Rao, A.C.S.: A distance-based outlier detection using particle swarm optimization technique. In: Fong, S., Akashe, S., Mahalle, Parikshit N. (eds.) Information and Communication Technology for Competitive Strategies. LNNS, vol. 40, pp. 633–643. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-0586-3_62
81. Coma-Puig, B., Carmona, J.: A quality control method for fraud detection on utility customers without an active contract. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 495–498 (2018)
82. Fazekas, M., Cingolani, L.: Breaking the cycle? How (not) to use political finance regulations to counter public procurement corruption. Slav. East Eur. Rev. **95**, 76–116 (2017)
83. Lehne, J., Shapiro, J.N., Vanden Eynde, O.: Building connections: political corruption and road construction in India. J. Dev. Econ. **131**, 62–78 (2018)
84. Cieślik, A., Goczek, Ł.: Control of corruption, international investment, and economic growth – Evidence from panel data. World Dev. **103**, 323–335 (2018)
85. Lourenço, I.C., Rathke, A., Santana, V., Branco, M.C.: Corruption and earnings management in developed and emerging countries. Corp. Gov. **18**, 35–51 (2018)
86. van Erven, G.C.G., Holanda, M., Carvalho, Rommel N.: Detecting evidence of fraud in the brazilian government using graph databases. In: Rocha, Á., Correia, A.M., Adeli, H., Reis, L.P., Costanzo, S. (eds.) WorldCIST 2017. AISC, vol. 570, pp. 464–473. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56538-5_47
87. Rad, M.S., Shahbahrami, A.: Detecting high risk taxpayers using data mining techniques. In: 2016 2nd International Conference of Signal Processing and Intelligent Systems ICSPIS 2016, pp. 14–15 (2017)
88. Monirzadeh, Z., Habibzadeh, M., Farajian, N.: Detection of violations in Credit Cards of Banks and financial institutions based on artificial neural network and Metaheuristic optimization algorithm. Int. J. Adv. Comput. Sci. Appl. **9**, 176–182 (2018)
89. Bramoullé, Y., Goyal, S.: Favoritism. J. Dev. Econ. **122**, 16–27 (2016)
90. Saxena, A., Sharma, N., Saxena, K., Parikh, Satyen M.: Financial data mining: appropriate selection of tools, techniques and algorithms. In: Deshpande, A.V., Unal, A., Passi, K., Singh, D., Nayak, M., Patel, B., Pathan, S. (eds.) SmartCom 2017. CCIS, vol. 876, pp. 244–251. Springer, Singapore (2018). https://doi.org/10.1007/978-981-13-1423-0_27
91. Bogdanov, D., Jõemets, M., Siim, S., Vaht, M.: How the Estonian tax and customs board evaluated a tax fraud detection system based on secure multi-party computation. In: Böhme, R., Okamoto, T. (eds.) FC 2015. LNCS, vol. 8975, pp. 227–234. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-47854-7_14

92. Kültür, Y., Çağlayan, M.U.: Hybrid approaches for detecting credit card fraud. Expert Syst. **34**, 1–13 (2017)
93. Indrajani, Prabowo, H., Meyliana: Learning fraud detection from big data in online banking transactions: a systematic literature review. J. Telecommun. Electron. Comput. Eng. **8**, 127–131 (2016)
94. Hutchings, A.: Leaving on a jet plane: the trade in fraudulently obtained airline tickets. Crime Law Soc. Change **70**, 461–487 (2018)
95. Saia, R., Boratto, L., Carta, S.: Multiple behavioral models: a divide and conquer strategy to fraud detection in financial data streams (2015)
96. Lee, P.S., Owda, M., Crockett, K.: Novel methods for resolving false positives during the detection of fraudulent activities on stock market financial discussion boards. Int. J. Adv. Comput. Sci. Appl. **9**, 1–10 (2018)
97. Fazekas, M.: Red tape, bribery and government favouritism: evidence from Europe. Crime Law Soc. Change **68**, 403–429 (2017)
98. Yaseen, M., et al.: Secure sensors data acquisition and communication protection in eHealthcare: review on the state of the art. Telemat. Inform. **35**, 702–726 (2018)
99. Jetter, M., Parmeter, C.F.: Sorting through global corruption determinants: institutions and education matter – Not culture. World Dev. **109**, 279–294 (2018)
100. Jagger, P., Shively, G.: Taxes and Bribes in Uganda. J. Dev. Stud. **51**, 66–79 (2015)
101. Williams, M.J.: The political economy of unfinished development projects: corruption, clientelism, or collective choice? Am. Polit. Sci. Rev. **114**, 705–723 (2017)
102. Kussainov, D.S.: The problems of qualification of illegal alienation of ownership of residential premises. Asian Soc. Sci. **11**, 188 (2015)
103. Ahmed, M., Mahmood, A.N., Islam, M.R.: A survey of anomaly detection techniques in financial domain. Future Gener. Comput. Syst. **55**, 278–288 (2016)
104. Moro, S., Cortez, P., Rita, P.: Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Syst. Appl. **42**, 1314–1324 (2015)
105. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., Bontempi, G.: Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE Trans. Neural Netw. Learn. Syst. **29**, 3784–3797 (2018)
106. Kumar, P., Iqbal, F.: Credit card fraud identification using machine learning approaches (2019)
107. Mahmoudi, N., Duman, E.: Detecting credit card fraud by modified fisher discriminant analysis. Expert Syst. Appl. **42**, 2510–2516 (2015)
108. Abdallah, A., Maarof, M.A., Zainal, A.: Fraud detection system: a survey. J. Netw. Comput. Appl. **68**, 90–113 (2016)
109. Rajak, I., Mathai, K.J.: Intelligent fraudulent detection system based SVM and optimized by danger theory. In: IEEE International Conference on Computer, Communication and Control IC4 2015, pp. 2–5 (2016)
110. Xu, J.J., Lu, Y., Chau, M.: P2P lending fraud detection: a big data approach. In: Chau, M., Wang, G.A., Chen, H. (eds.) PAISI 2015. LNCS, vol. 9074, pp. 71–81. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18455-5_5
111. Hajek, P., Henriques, R.: Mining corporate annual reports for intelligent detection of financial statement fraud – a comparative study of machine learning methods. Knowl.-Based Syst. **128**, 139–152 (2017)

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.

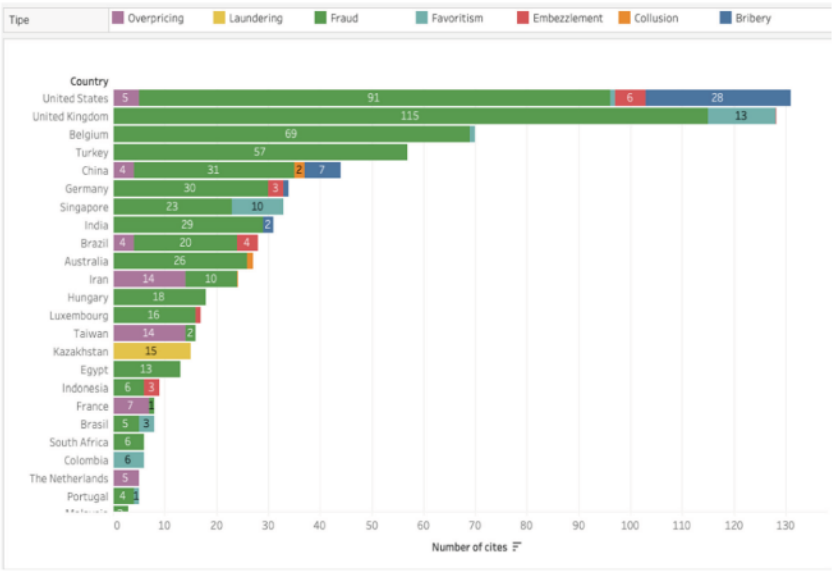


Fig. 1. Articles published by countries and type of corruption.

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.

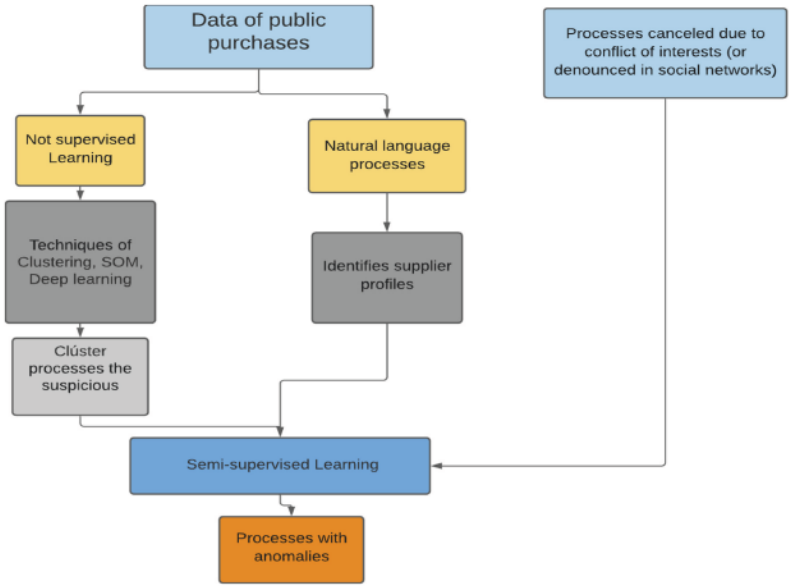


Fig. 2. Anomaly detection model

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

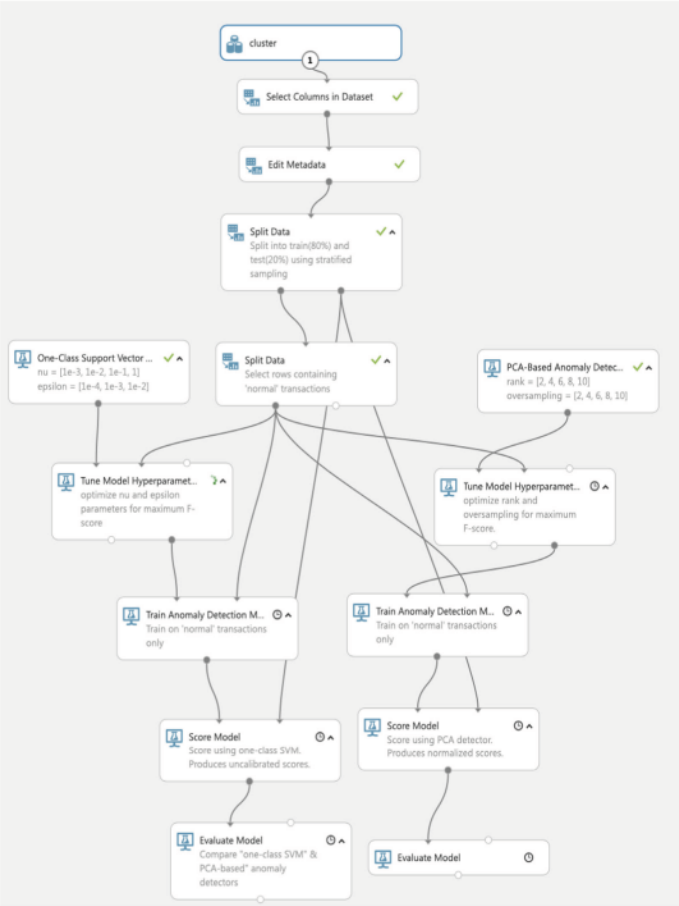


Fig. 3. Azure anomaly detection model

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.

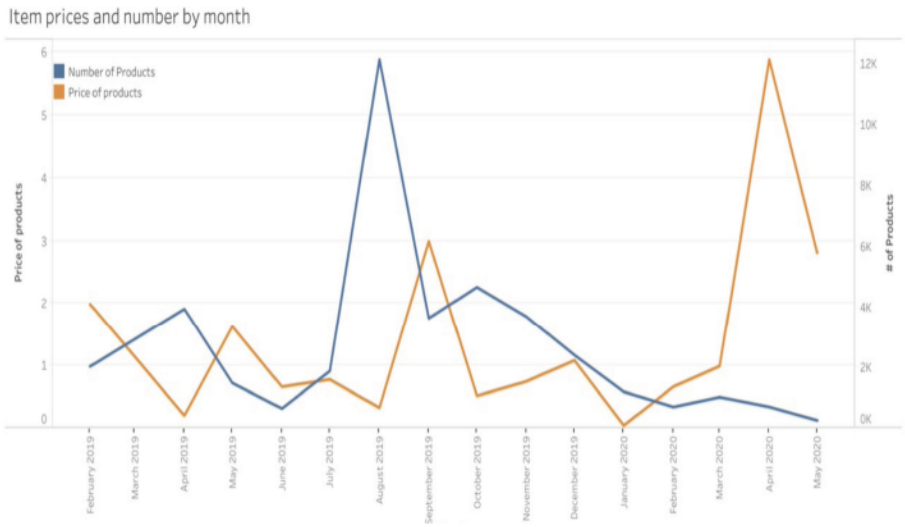


Fig. 4. COVID-19 pandemic medicine price developments.

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