UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Revista Científica

http://revistas.udistrital.edu.co/ojs/index.php/revcie/index

CIDC

Artículo de investigación científica y tecnológica

Modelo para definir índices de corrupción en convocatorias de contratación en Colombia basado en Big Data y procesamiento del lenguaje natural

Model to Define Corruption Indices in Contracting Announcements in Colombia Based on Big Data and Natural Language Processing

Modelo para definir índices de corrupção na contratação de processos de licitação na colômbia baseado em big data e processamento de linguagem natural

Julio-César Luna-Ortega[1]
Carlos-Alberto Cobos-Lozada[2]
Martha-Eliana Mendoza-Becerra[3]

Recibido: julio de 2022 **Aceptado:** octubre de 2022

Para citar este artículo: Luna-Ortega, J. C., Cobos-Lozada, C. A. y Mendoza-Becerra, M. E. (2023). Modelo para definir índices de corrupción en convocatorias de contratación en Colombia basado en Big Data y procesamiento del lenguaje natural. *Revista Científica*, *46*(1), 77-92. https://doi.org/10.14483/23448350.19640

Resumen

El presente trabajo de investigación propone un modelo macro que permite detectar diferentes probables delitos o anomalías relacionadas con corrupción en los procesos de contratación pública en Colombia. Para ello, el modelo propuesto consta de cinco componentes principales: 1) servicios especializados que buscan identificar situaciones específicas de probable corrupción (se propusieron tres servicios: detección de similitud entre propuestas técnicas, detección de manipulación de ofertas y detección de carteles); 2) servicios transversales que sustentan la transformación del modelo en una herramienta de software; 3) servicios adicionales que abordan situaciones generales de probable corrupción, en específico el servicio de alerta ciudadana; 4) relaciones explícitas entre servicios; y 5) salida global del modelo. En la experimentación práctica, dos de los servicios planteados en esta investigación fueron puestos a prueba en diversos escenarios. Con los resultados arrojados por algunas de las métricas clásicas del área, se determinó la calidad de la predicción obtenida por los servicios.

Palabras clave: contratación; corrupción; índices; modelo; servicios.

1. MSc. Universidad del Cauca (Popayán-Cauca, Colombia). jcluna@unicauca.edu.co.
2. Ph. D. Universidad del Cauca (Popayán-Cauca, Colombia). ccobos@unicauca.edu.co.
3. Ph. D. Universidad del Cauca (Popayán-Cauca, Colombia). mmendoza@unicauca.edu.co.

Abstract

This research work proposes a macro-model that allows detecting different probable crimes or anomalies related to corruption in public procurement processes in Colombia. To this effect, the proposed model consists of five main components: 1) specialized services that seek to identify specific situations of probable corruption (three services were proposed: the detection of similarity between technical proposals, the detection of offer manipulation, and the detection of cartels); 2) transversal services that support the transformation of the model into a software tool; 3) additional services where general situations of probable corruption are addressed, specifically the citizen alert service; 4) the explicit relationships between services; and 5) the global output of the model. In the practical experimentation stage, two of the services proposed in this research were put to the test in various scenarios. With the results obtained by some of the classical metrics of the area, the quality of the prediction obtained by the services was determined.

Keywords: contracting; corruption; indices; model; services.

Resumo

Este trabalho de pesquisa propõe um modelo macro que permite detectar diferentes crimes prováveis ou anomalias relacionadas à corrupção em processos de contratação pública na Colômbia, para isso o modelo proposto consiste em 5 componentes principais, 1) serviços especializados que buscam identificar situações específicas de provável corrupção, especificamente foram propostos três serviços dedicados a um crime específico cada, a detecção de semelhança entre propostas técnicas, a detecção de manipulação de ofertas e a detecção de cartéis, 2) os serviços transversais que suportam a transformação do modelo em ferramenta de software, 3) serviços adicionais onde são abordadas situações gerais de provável corrupção, nomeadamente o serviço de alerta ao cidadão, 4) as relações explícitas entre os serviços e 5) o output global do modelo. Na experimentação prática, dois dos serviços propostos nesta pesquisa foram colocados à prova em diversos cenários, com os resultados obtidos por algumas das métricas clássicas da área, o nível de satisfação da qualidade da previsão obtida pelos serviços foi determinado.

Palavras-chaves: contratação, corrupção, índices, modelo, serviços

Introducción

En Colombia la percepción de la corrupción se ha elevado en los últimos años, según lo revela el *Índice de Percepción de la Corrupción (IPC)* de la agencia Transparencia Internacional en el año 2021 (Transparencia por Colombia, 2022). El creciente descubrimiento de casos de corrupción en procesos de contratación pública que se ha dado en los últimos años se ha convertido en una oportunidad para la implementación de sistemas que analicen los datos y permitan detectar y prevenir nuevos casos. A nivel mundial, el promedio de las pérdidas sufridas por las entidades víctimas de corrupción es aproximadamente de USD$250.000 por caso. Siendo la contratación pública uno de los procesos más permeados por actos corruptos. El aumento de los casos de corrupción ha hecho que diversas entidades estén aplicando variadas estrategias anticorrupción, desafortunadamente la mayoría de estas son técnicas manuales y de enfoque social, que buscan prevenir, más que identificar probables casos de corrupción. Si bien las técnicas manuales pueden ser útiles como estrategias anticorrupción, se hace necesario desarrollar estrategias automáticas y predictivas (basadas en estadística, inteligencia artificial y otras técnicas) para reducir y en un futuro eliminar este flagelo de la sociedad.

Luna-Ortega, Cobos-Lozada y Mendoza-Becerra (2020) realizaron una revisión sistemática de la literatura (RSL) buscando establecer los conceptos fundamentales del estado de arte relacionados con soluciones que apliquen estrategias anticorrupción. Los resultados muestran que combinar técnicas estadísticas con tecnologías de la información (inteligencia artificial, big data, minería de datos, entre otros) es una línea de investigación prometedora en esta área de aplicación. Cabe destacar

que las investigaciones recopiladas en la RSL buscan resolver el problema de identificación de presuntas situaciones de corrupción en procesos de contratación haciendo énfasis en un delito específico, mayormente colusión, creación de carteles o manipulación de ofertas, esto mediante el uso, por lo general, de una o dos herramientas para ello. Sin embargo, las investigaciones recopiladas también muestran que, dada la complejidad del problema, se hace necesario contar con un conjunto (batería) de herramientas especializadas para atacar los diferentes delitos que ocurren en el proceso de contratación; dado esto, se logra entender la importancia que tiene continuar con el desarrollo de modelos que permitan solventar de manera más efectiva (eficaz y eficiente) el problema de la identificación de presuntas situaciones de corrupción en la contratación. Además, estos trabajos han sido desarrollados en otros países y no en Colombia.

Teniendo en cuenta lo anterior, en el presente trabajo se partió de la siguiente pregunta de investigación: ¿qué servicios especializados debe incluir un modelo conceptual basado en big data y procesamiento del lenguaje natural que permita identificar presuntas situaciones de corrupción en los datos y documentos asociados a los procesos de contratación pública en Colombia? Para que este conjunto de servicios especializados pueda ser luego implementado, se buscó establecer las entradas, los procesos que se deberían realizar, los soportes tecnológicos que se podrían usar y las salidas que cada servicio debería entregar.

Modelo propuesto

El modelo propuesto en esta investigación busca ser la base para la construcción de una herramienta software que permita definir índices de riesgo de presuntas situaciones corruptas en convocatorias de contratación pública mediante la integración y comunicación de varios servicios especializados, los cuales, en su mayoría, están diseñados para abordar una situación específica donde hay probabilidad de corrupción, con base en los datos y documentos asociados a los procesos de contratación estatal en Colombia presentados por los proponentes y otros adquiridos de diversas fuentes de información. Estos servicios especializados son el corazón (*core*) del modelo y se encargan de detectar los diferentes delitos probables o anomalías relacionadas con corrupción en los procesos de contratación pública en Colombia. Los resultados arrojados por estos servicios brindan la información necesaria para determinar el índice de riesgo que presenta cada una de las convocatorias evaluadas en el modelo. Los servicios implementados e integrados en el modelo son: servicio de detección de similitud en propuestas técnicas, servicio de detección de probable manipulación de ofertas y servicios adicionales.

Servicio de detección de similitud en propuestas técnicas

El objetivo de este servicio consiste en identificar las probables similitudes que se presentan entre una nueva propuesta técnica y una fuente de información que se irá construyendo a partir de las propuestas técnicas que sean analizadas en un proceso de licitación actual o previo. El servicio se ha modelado con el objetivo de generar alertas sobre una propuesta técnica específica, señalando las partes de esta que presentan similitudes por encima de un umbral previamente definido; siendo el usuario final quien debe verificar si el reporte entregado conlleva a marcar la propuesta técnica, unas secciones de esta o unas frases como plagio. Para lo anterior se debe partir, entonces, de la existencia de la propuesta técnica, el archivo donde un proponente expone las condiciones de carácter técnico del bien o servicio a ofrecer de acuerdo con lo previamente exigido por la entidad. Esta propuesta técnica en el marco de una licitación es obligatoria, conforme al Artículo 7 del Decreto Ley 1150 de 2007. Esta propuesta técnica no podrá complementarse, adicionarse, modificarse o

mejorarse cuando se cierre la convocatoria (licitación). Esta misma es la fuente de información que el servicio de detección de similitud usará para retroalimentar la fuente de propuestas técnicas desde donde se comparará una nueva y, a su vez, es el documento al que se le realizará un análisis de similitud frente a las propuestas técnicas previamente recibidas y almacenadas.

El enfoque propuesto para la implementación del servicio sigue los principios de diseños de aplicaciones antiplagio de la literatura, basándose en las siguientes etapas: recuperación de propuestas técnicas candidatas, estrategia de comparaciones detalladas e inspección humana (Stein, Zu Eissen y Potthast, 2007). Además de ello, dada la necesidad de construir y retroalimentar la base de datos, fuente de información, el servicio realiza una primera etapa de almacenamiento sistemático de propuestas, cabe recalcar que esta etapa se debe realizar en un primer momento, con el objetivo de poblar la fuente de datos y, a su vez, ser un proceso iterativo e incremental, dado que se realizará en repetidas ocasiones, alimentando así, cada vez más, la fuente de información.

Para el almacenamiento de las propuestas técnicas, se debe partir de que el mismo se encuentra en algún formato específico, su contenido debe ser extraído ya sea para indexarlo o para analizarlo, y en este proceso, en algunas circunstancias, se requiere dividir el contenido en sus frases correspondientes. Así mismo, el archivo original de la propuesta técnica se debe almacenar en el sistema de archivos y relacionarlo en la base de datos para llevar un apropiado control de los posteriores análisis.

Una vez extraído el contenido de la propuesta técnica, el primer paso del procesamiento del archivo corresponde a la indexación, de manera que pueda ser utilizado posteriormente para el análisis de similitud, para ello se hace uso de los índices invertidos gracias a las amplias ventajas que ofrecen a la hora de realizar búsquedas complejas sobre los datos (Pibiri y Venturini, 2021), tema de gran importancia para esta investigación. A parte de esta estrategia, se hace necesario contar con una base de datos que almacene el archivo y la relacione con información externa del mismo, evitando así sobrecargar el indexador con datos que no son representativos para la posterior búsqueda. Vale mencionar que, para una correcta indexación, el índice debe ser configurado, modificando las operaciones (filtros, reducción a la raíz léxica, remoción de palabras vacías, entre otras) que se deben realizar en el momento de la indexación, esto ayuda a que el proceso de búsqueda sea más eficiente.

Hoy en día la literatura experta reporta algunas herramientas de libre uso que permiten realizar almacenamiento y búsqueda en índices invertidos. En la presente investigación se usó Elasticsearch, un motor de búsqueda basado en Lucene que a su vez es una API de código abierto dirigido a la tarea de recuperación de información. Lucene se encuentra desarrollado en Java, pero su API se encuentra disponible en una amplia variedad de lenguajes de programación, ahora bien, Elasticsearch ofrece sus funcionalidades de una manera más práctica, y ofrece además la implementación de una aplicación distribuida mediante nodos que le permite lograr una mayor disponibilidad, el cargue masivo de documentos y herramientas de análisis y de búsqueda.

Por otro lado, para la recuperación de propuestas técnicas candidatas y la tarea de comparaciones detalladas de estas, se contempla inicialmente una estrategia para tener en cuenta características del documento, convocatoria y tipo de análisis a desarrollar, luego de esto, se usó el indexador para el envío de la petición de búsqueda aplicando filtros y configuraciones específicas de la API de Elasticsearch.

Para la tarea de inspección humana de los análisis resultantes, se debe tener en cuenta, inicialmente, que los mismos toman un tiempo para su realización, esto debido a la gran cantidad de palabras o frases de la propuesta técnica que se va a analizar y la gran cantidad de propuestas técnicas con las que se debe realizar la comparación. Por

lo anterior se hace necesario implementar una estrategia de entrega asincrónica del reporte, permitiendo así que el usuario continúe trabajando en la aplicación sin la necesidad de esperar a que la tarea de análisis termine. Para esta tarea se debe contemplar, primero, el almacenamiento de la información de los análisis en un motor de base de datos no relacional, esto debido a la estructura del reporte y con el objetivo de no tener que repetir el proceso de análisis cada vez que se desee visualizar. Posteriormente, una estrategia para la entrega y visualización del reporte. También debe ser implementada una estrategia para la configuración de los análisis generados, con el objetivo de brindarle al usuario final una fácil administración de estos, como por ejemplo poder etiquetar aquellas frases similares que no deben contar a la hora de cuantificar la similitud del documento (una lista de "frases comunes" que no deben contar como plagio).

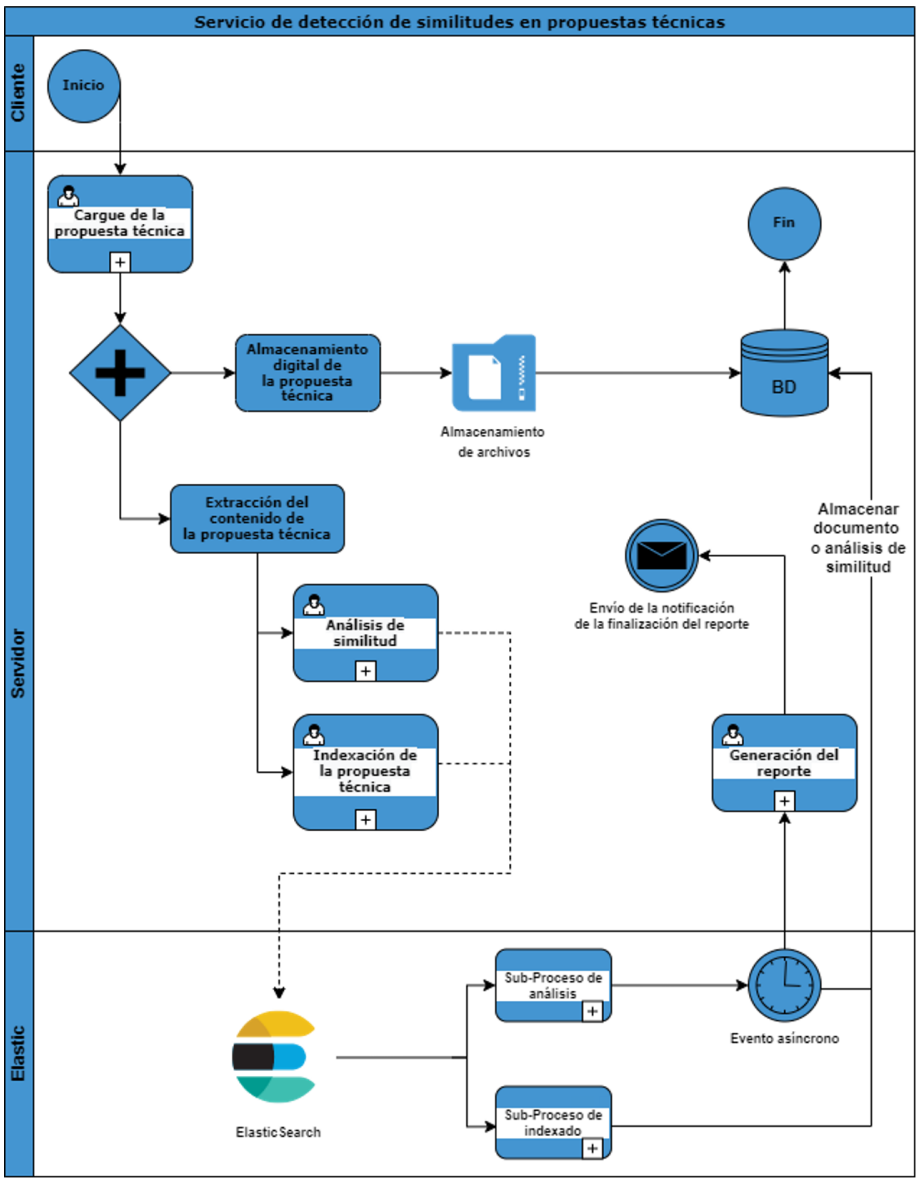


Figura 1. P.N. del módulo de detección de similitudes.

Con el objetivo de entender las principales acciones e interacciones que involucran el servicio de detección de similitudes, en la Figura 1 se presenta un diagrama de procesos de negocio que modela el funcionamiento de este. En la figura se aprecia cómo el proceso inicia con la petición por parte del usuario, partiendo del cargue de propuestas técnicas, del cual se dividen dos subprocesos, uno para el almacenamiento de archivos y otro para la extracción de datos, de donde se despliegan las dos siguientes tareas del usuario: la indexación de propuestas técnicas y el análisis de similitud, que se comunican con un servicio de indexación para el desarrollo de sus respectivas tareas. Una vez realizado el cargue de información, el proceso finaliza con la entrega, bien sea, de la referencia de la indexación creada, o del propio análisis de similitud, en ambos casos se genera una notificación del estado final de la tarea al usuario.

El objetivo de este servicio, al igual que los demás servicios del modelo propuesto, más que juzgar una propuesta, busca alertar al tomador de decisiones sobre irregularidades que se puedan presentar en los documentos. Para este servicio en específico se maneja un índice de similitud sobre el documento analizado, el cual es un valor porcentual que indica la semejanza de una propuesta técnica frente a otras almacenadas en la fuente de datos, esto basado en la similitud que exista entre las frases de las diferentes propuestas.

Servicio de detección de probable manipulación de ofertas

El delito de manipulación de ofertas se presenta cuando en un proceso de licitación, los proponentes confabulan entre sí para la elección del ganador del proceso mediante el aumento o la reducción de los precios y la calidad de los productos o servicios ofertados. Lo ideal para una entidad es obtener un equilibrio entre el costo y la calidad de los productos o servicios, para ello se recurre a un proceso de licitación competitivo, pero esto solo es efectivo si los competidores contienden con honestidad y transparencia. La manipulación de ofertas resulta particularmente dañina cuando se ven afectados los rubros y adquisiciones públicas, las cuales según la OCDE representan cerca del 15 % del PIB de los países pertenecientes a dicha organización (Abrantes-Metz, 2013). Según algunas investigaciones, en un proceso de licitación pública el delito de manipulación de ofertas se puede presentar de diferentes formas, aunque todas ellas pretenden lo mismo, a saber: impulsar un ganador previamente escogido por parte de todos o un segmento de los ofertantes, o elevar el monto de la oferta, aumentando las utilidades del proponente ganador. Las principales modalidades en las que se puede presentar este delito están enmarcadas en cuatro tipos (Ossa Bocanegra, 2014): ofertas de resguardo, supresión de ofertas, rotaciones de ofertas y asignaciones de mercado.

El objetivo del servicio propuesto es la evaluación de las propuestas económicas presentadas a una licitación abierta para encontrar probables anomalías, referentes a los costos de los servicios o productos, con ello el servicio busca generar alertas sobre la probable manipulación de las ofertas. Teniendo en cuenta que como base de todos los servicios se cuenta con el almacenamiento de documentos de propuestas anteriormente presentadas a diversas licitaciones, estas mismas pueden ser utilizadas para hacer estudios post mortem que permitan identificar diferentes anomalías en las licitaciones realizadas, con esto aprender y perfeccionar los servicios que se proponen en este trabajo, incluyendo por ejemplo la identificación de colusión basada en asignaciones de mercado o territorio.

Un proponente que desee participar dentro de un proceso de licitación abierto debe entregar, cumpliendo con el pliego de condiciones de una licitación, una propuesta económica donde se expresen con claridad las exigencias realizadas por la entidad pública. Esta propuesta económica relaciona productos o servicios con su descripción, cantidad y costo propuesto por el licitante. Este servicio parte de la existencia de este documento, la propuesta económica.

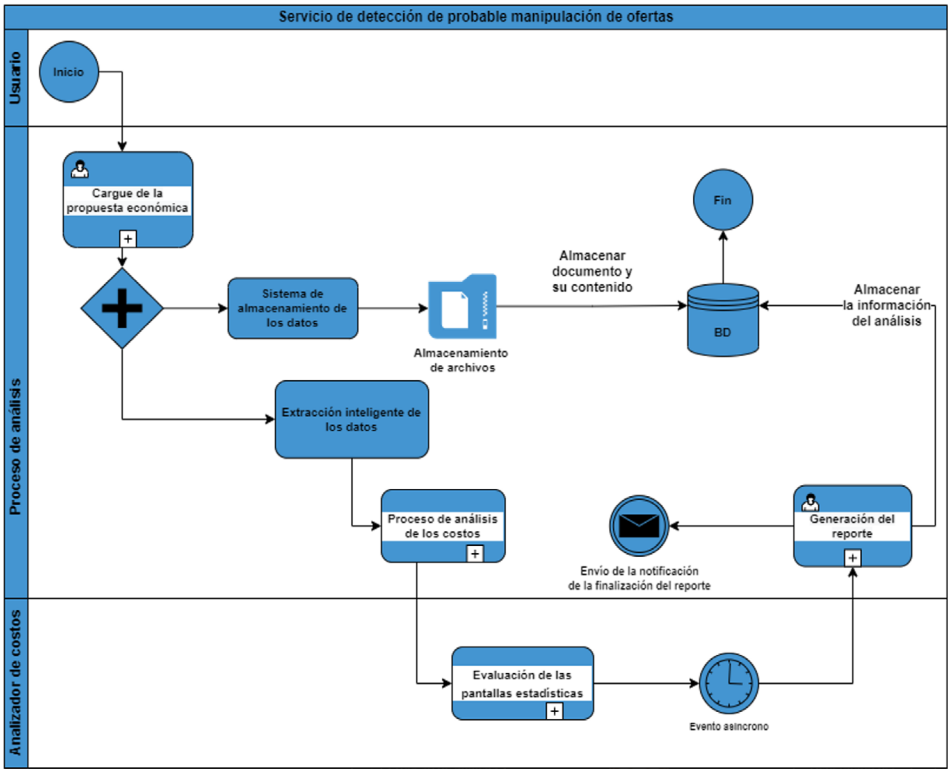


Figura 2. P.N. del servicio de detección de probable manipulación de ofertas.

La Figura 2 presenta un diagrama de procesos de negocio que modela el funcionamiento del presente servicio. En la figura se aprecia cómo el proceso inicia con una petición realizada por un usuario. Parte con el cargue de la propuesta económica de cada uno de los proponentes (tarea que se realiza previamente en un primer paso, en el momento del almacenamiento de la información que todo proponente debe entregar de acuerdo con el pliego de condiciones), luego se inicia el proceso de extracción inteligente de los datos necesarios para el análisis, este proceso además de extraer los datos, los almacena. Con estos datos, el subproceso de análisis se encarga de realizar comparaciones de las diferentes propuestas de una licitación con el objetivo de encontrar anomalías que indiquen la probabilidad de presentarse el delito de manipulación de ofertas (se explica en mayor detalle en una sección posterior). Los escenarios que debe tener en cuenta el servicio para desarrollar este análisis son:

- Existencia de ofertas globales anormales.
- Existencia de manipulación y variación de ofertas individuales que pretendan aparentar una competencia leal.
- Agrupamiento de ofertas que son similares en características o valores.

Finalmente, una vez realizado el análisis, el proceso culmina con la entrega del reporte, dentro de un motor de base de datos y la generación de una notificación del estado final de la tarea al usuario.

Subproceso de análisis de precios

Este proceso es el paso fundamental del servicio, en este se realiza el análisis y se determina la probabilidad de presentarse el delito de manipulación de ofertas dentro de una licitación, para llevar a cabo esta tarea, se puede usar la aplicación de pantallas (índices estadísticos específicos), con el

propósito de agrupar aquellas propuestas que se asemejan en términos de costos, bien sea por producto o servicio, o de manera global. A continuación, se exponen algunas pantallas utilizadas en la literatura como indicadores que sirven para señalar la probable presencia del delito de manipulación de ofertas en procesos de contratación. El conocimiento de estos indicadores, junto con su correcta aplicación, puede desincentivar la existencia de acuerdos colusorios de manipulación de ofertas, así como de presencia de carteles (Imhof, 2018).

1. **Curtosis o apuntamiento:** la curtosis ha sido utilizada para la detección del delito de manipulación de ofertas cuando las licitaciones presentan un escalado inteligente con un factor común a los costos reales, es decir, sin grandes variaciones entre la oferta más alta y la más baja. La ecuación (1) muestra el cálculo de la curtosis para una licitación t, donde n denota el número de ofertas totales de la licitación, b_{it} se refiere a la oferta i de la licitación, σ_t es la desviación estándar de las ofertas en la licitación, y μ_t es la media aritmética de las ofertas de la mencionada licitación.

$$Kurt(b_t) = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{b_{it} - \mu_t}{\sigma_t}\right)^4 - \frac{3(n-1)^3}{(n-2)(n-3)} \quad (1)$$

2. **Diferencia relativa normalizada:** contemplando que las ofertas de cobertura son creadas para perder y asegurar una ganadora previamente seleccionada, esto implica que la distancia entre las ofertas perdedoras será significativamente inferior en comparación con la oferta ganadora. Así, teniendo en cuenta que la diferencia entre las dos ofertas más bajas es importante y la diferencia entre las ofertas de cobertura es pequeña, se construye el indicador de diferencia relativa normalizada (Imhof, 2018). La ecuación (2) muestra el cálculo de este indicador, donde los valores de las ofertas son, previamente, ordenados de forma creciente y b_{it}, b_{jt} son ofertas adyacentes. Un valor mayor que 1 indica que la diferencia entre la segunda y la primera oferta más baja es mayor que la diferencia promedio entre todas las ofertas adyacentes en una licitación.

$$Diferencia\ relativa\ Normalizada = \frac{b_{2t} - b_{1t}}{\frac{(\sum_{i=1,j=i}^{n-1} b_{jt} - b_{it})}{n-1}} \quad (2)$$

Similar al servicio de detección de similitudes, el objetivo de este servicio es proporcionar al usuario final del modelo un índice de probabilidad, basado en los indicadores presentados en la sección anterior, que le indique una forma de medir la existencia del delito de manipulación de ofertas, pero esta vez no por cada propuesta económica presentada sino de forma global y relacionada con el proceso de licitación. Es preciso comentar que además de obtener este índice global se debe calcular la probabilidad de existencia de este delito para cada una de las propuestas con el objetivo de buscar identificar las que están implicadas o no en el probable delito. Para suplir esta necesidad se propone la implementación del algoritmo de procesamiento del lenguaje natural LexRank (Erkan y Radev, 2004), este algoritmo trata los objetos a estudiar como nodos y los enlaces representan las relaciones entre ellos. Los autores del algoritmo plantean la hipótesis de que si un nodo es muy similar a muchos otros, este se puede considerar como el más central. En un proceso de contratación se esperaría que las ofertas económicas presentadas por los ofertantes sean, en un grado considerable, similares (centrales), ofertas que se alejen considerablemente de este centro pueden, en algunos casos y soportando con la información de las pantallas estadísticas mencionadas previamente, generar alerta de probable participación en el delito de manipulación de ofertas. Con esta hipótesis, en esta investigación se plantea el uso de este algoritmo para otorgarle un índice a cada oferta económica con respecto a su probabilidad de pertenencia en el delito de manipulación de ofertas.

El algoritmo es planteado en la Figura 3. La primera tarea realizada por este es la generación de la matriz de similitud de cosenos para cada una de las ofertas (pasos 1 al 11), seguidamente se aplica un umbral con el propósito de eliminar aquellas relaciones débiles entre las ofertas (nodos del grafo), es decir, aquellos vértices con similitud de coseno que no supera un valor dado, así mismo el algoritmo calcula el grado de centralidad de cada nodo contando el número de propuestas que tienen relación de similitud con otra propuesta después de haber sido aplicado el umbral. Es importante recalcar que un valor de umbral demasiado bajo puede tener en cuenta en forma errónea similitudes débiles, mientras que un valor demasiado grande puede eliminar muchas de las relaciones de similitud en el grupo de ofertas.

Luego de esto se realiza una normalización de cada fila de la matriz de similitud (con ello cada fila se puede ver como un vector de cambios de estado con probabilidades que sumarán 1). Esto se logra dividiendo cada elemento de la matriz por el grado de la fila en la que se encuentra (pasos 12 al 16). Se continua con la tarea de garantizar que la matriz estocástica sea irreducible y aperiódica, aplicando un factor de amortiguamiento a cada celda de la matriz (pasos 17 al 21). Finalmente, se aplica PowerMethod usando una tolerancia de error específico, por ejemplo 0,0001, este método usa la propiedad de convergencia de las cadenas de Markov para calcular la distribución estacionaria de un arreglo. En cada iteración, el vector propio se actualiza multiplicándolo con la transpuesta de la matriz estocástica y gracias

Entrada: Arreglo de propuestas económicas s de tamaño n; Umbral t; Valor amortiguamiento $dampingFactor$
Salida: Arreglo L con los scores definidos por el algoritmo para cada propuesta económica

```
1:  cosineMatriz[n][n] = calcular matriz de similitud cosenos
2:  Para i = 1 hasta n incremento en 1 hacer
3:      suma = 0
4:      Para j = 1 hasta n incremento en 1 hacer
5:          Si cosineMatriz[i][j] > t entonces
6:              cosineMatriz[i][j] = 1
7:              suma++
8           Si no  cosineMatriz[i][j] = 0
9:          Fin si
10:     Fin para
11: Fin para
12: Para i = 1 hasta n incremento en 1 hacer
13:     Para j = 1 hasta n incremento en 1 hacer
14:         cosineMatriz[i][j] = cosineMatriz[i][j] / suma
15:     Fin para
16: Fin para
17: Para i = 1 hasta n incremento en 1 hacer
18:     Para j = 1 hasta n incremento en 1 hacer
19:         cosineMatriz[i][j] = (dampingFactor / n) + (1 − dampingFactor) *
    cosineMatriz[i][j]
20:     Fin para
21: Fin para
22: L = calcular el PowerMethod a cosineMatriz
23: Retornar L
```

Figura 3. Pseudocódigo algoritmo LexRank adaptado de Erkan y Radev, (2004).

a que el arreglo fijado anteriormente es irreducible y aperiódico se garantiza que el método no se quede estancado en un bucle. Con ello, al final el algoritmo retornará un vector con un valor de centralidad para cada oferta, valor que para esta investigación será utilizado como índice de probabilidad de participación en el delito de manipulación de ofertas.

Servicios adicionales propuestos en el modelo

En la investigación se propone además una serie de servicios que complementan al modelo para solventar la tarea de identificación de índices de corrupción, servicios que debido a las limitantes del documento no son especificados detalladamente, pero sí son mencionados a continuación.

Servicio especializado de detección del delito de carteles

Servicio que hace uso de los documentos recolectados en el proceso de cargue de una licitación, documentos referentes a información personal o empresarial del proponente, por ejemplo, el registro único de proponentes (RUP), para así poder crear una topología de red que junto a una estrategia de análisis de imágenes permita la detección de patrones de comportamiento con el objetivo de alertar sobre la probable formación de carteles dentro de un proceso de licitación.

Por otro lado, se cuenta con servicios adicionales que pueden, sin necesidad de identificar un delito específico, ayudar a la detección de situaciones anómalas en un proceso licitatorio. Estos servicios se comunican y correlacionan entre sí para producir una salida final del modelo con respecto a la definición de índices de riesgo de corrupción en un proceso de contratación. Este componente del modelo, entonces, aloja aquellos servicios adicionales que tienen como objetivo generar una alerta temprana sobre las convocatorias existentes, sin necesariamente realizar análisis sobre la información existente, así se podrían entender estos servicios, como servicios de índole social con aplicaciones de las tecnologías de la información. Con base en una revisión de la literatura y el análisis del impacto que el servicio de alerta ciudadana ha tenido en propuestas previas, una adaptación de este fue incorporado en el modelo (Owusu, Chan y Shan, 2019). En este modelo el servicio plantea una recolección de información referente a quejas de situaciones anómalas dentro de un proceso de licitación, por medio de un servicio de denuncia ciudadana accesible al público en general. El servicio permite, además de recolectar las denuncias, hacerles su respectivo seguimiento, gestión y socialización de la investigación realizada por medio de reportes. Este servicio, si bien no se enfoca en un delito como tal, puede permitir la detección de probables situaciones corruptas que los anteriores servicios podrían pasar por alto, dado que estos se enfocan detalladamente en un servicio específico.

Además de los servicios especializados y adicionales, se cuenta con servicios transversales que coadyuvan con el objetivo general del modelo, brindado características y funcionalidades vitales para la transformación del modelo en una herramienta software. Así, estos servicios facilitan el funcionamiento del modelo, ofreciendo funcionalidades que rompen el esquema independiente de los servicios especializados, logrando que estas funcionalidades abarquen y estén presentes en toda la estructura del modelo, sirviendo de soporte a cualquiera de los servicios del modelo. Los servicios transversales implementados e integrados en el modelo son:

Administración, autenticación y autorización de usuarios: servicio encargado de control, gestión de acceso, administración y autorización de usuarios y permisos de estos sobre componentes y servicios del modelo.

Administración de los servicios: servicio encargado de gestionar variables, atributos y características administrativas de los diferentes servicios del modelo.

En resumen, el modelo consta de cinco componentes principales: servicios especializados que buscan identificar situaciones específicas de probable corrupción, servicios transversales que coadyuvan a la transformación del modelo en una herramienta software, servicios adicionales donde se abordan situaciones generales de probable corrupción, relaciones explícitas entre servicios y salida global del modelo. En la Figura 4 se muestra un diagrama donde se aprecia a un alto nivel los principales componentes del modelo. Este diagrama sigue una arquitectura orientada a eventos y delega las funcionalidades a través del concepto de microservicio.

Implementación de los servicios

Para llevar a cabo la implementación de los servicios, se optó por una arquitectura de nube, con el despliegue mediante el uso de aplicaciones web, dadas las ventajas que esta trae: fácil desligue en diferentes tipos de sistemas operativos, interfaz amigable y fácil de usar, más fácil de diseñar en forma escalable, de fácil actualización y centralización de la seguridad (Attaran y Woods, 2019). Y como característica principal, la opción de escalabilidad de los recursos a utilizar, dado que en un principio el servicio (y la aplicación que acogerá el modelo planteado) no necesitará una gama amplia de recursos computacionales por la poca demanda inicial, pero luego de un tiempo, con el aumento del uso del mismo por diferentes entidades, se estima que será necesario ampliar los recursos, característica que un servicio de Cloud Computing permite de manera fácil sin afectar el rendimiento de la aplicación.

Los servicios planteados en el presente modelo, tal y como se ha mencionado anteriormente, están dirigidos a entidades gubernamentales que ofrezcan públicamente procesos de licitación. A su vez, se tendrá acceso a dichos servicios y la información de los mismos mediante usuarios previamente

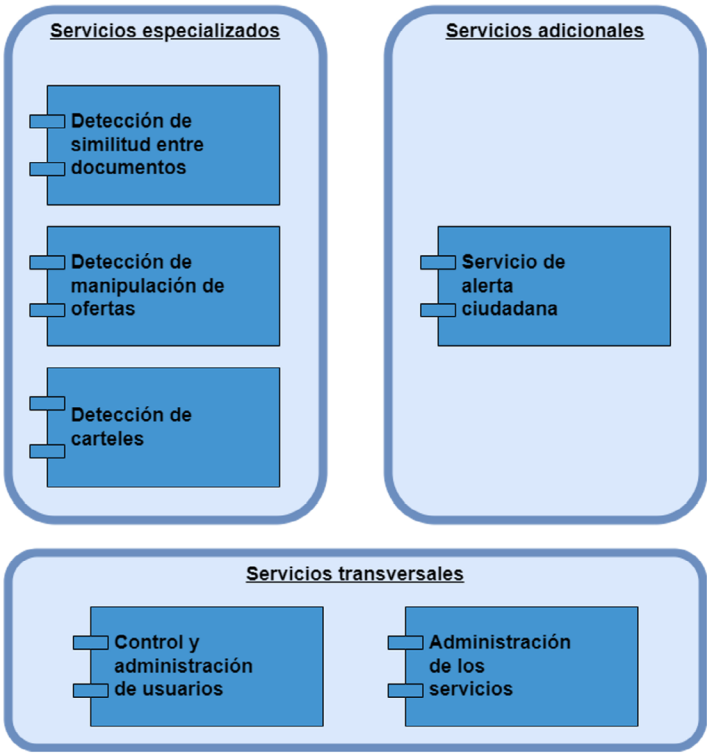


Figura 4. Diagrama macro de componentes del modelo.

registrados en la plataforma. Es por esto que la arquitectura debe permitir la conexión de múltiples usuarios de diferentes entidades a una única instancia del servicio que maneje la información de todas las entidades con la debida separación de datos y seguridad. Esta característica se conoce como arquitectura *multi-tenant* (multi-propietario) y se presenta como una de las más utilizadas para este propósito (Weber *et al.*, 2019). Esta arquitectura hace referencia a uno de los principios de la ingeniería de software donde solo se ejecuta una única instancia en un servidor y múltiples clientes aceden a ella, así un solo desarrollo software satisface las necesidades de múltiples clientes, garantizando la separación de la información sensible de cada uno. Para el almacenamiento de los datos en esta arquitectura multi-propietario se usa una única base de datos con esquemas compartidos, donde cada tabla del modelo relacional maneja la identificación del cliente que es dueño de cada registro (o fila de datos) almacenado.

Teniendo en cuenta que la aplicación desarrollada debe permitir la escalabilidad en su desarrollo debido a la futura incrustación de nuevos servicios, es necesario plantear una arquitectura que permita la comunicación entre los servicios garantizando la separación de sus funcionalidades y evitando la repetición de funciones o procedimientos (código). Planteado lo anterior, la arquitectura escogida para el desarrollo del software es una arquitectura basada en microservicios y *microfrontends*. Con los microservicios se garantiza una fácil escalabilidad de la aplicación global (Waseem, Liang y Shahin, 2020), permitiendo además la realización de ajustes de manera independiente sin afectar mayormente al todo. Así, con los microservicios se tiene un conjunto de pequeños servicios, ejecutados de forma independiente y autónoma. Con el objetivo de suplir esto mismo del lado del cliente o *frontend* y extendido de la idea de los microservicios, se implementa una arquitectura de *microfrontends* (Peltonen, Mezzalira y Taibi, 2021), siguiendo la tendencia actual del desarrollo de software de las aplicaciones *single page app*. El objetivo principal de los *microfrontends* es tratar una aplicación web como un conjunto de características independientes, cada una con su propia lógica de negocio, que es justamente lo ofrecido, del lado del servidor, por los microservicios. Con esto, el modelo implementado usa estas dos arquitecturas específicas permitiendo la comunicación cliente-servidor mediante el ofrecimiento de APIs desde el lado del servidor (Waseem *et al.*, 2020). El software implementado para el desarrollo, tanto de este servicio como del modelo se realizó siguiendo una metodología y usando patrones de diseño recomendados en la literatura (Gkantouna *et al.*, 2020), así como lenguajes de programación idóneos tanto para el *backend* como el *frontend*. Los lenguajes escogidos para cada capa del servicio cumplen con las siguientes características mínimas:

Del lado del *backend* se cuenta con un *framework* que facilitó el desarrollo de control y administración de usuarios. Así mismo, con mecanismos que facilitaron el desarrollo, el despliegue y la publicación de APIs para su posterior uso. Finalmente, cuenta con una amplia gama de librerías que ahorran el desarrollo de funcionalidades. El lenguaje escogido para esta capa fue Python en su versión 3.9.0.

Del lado del *frontend* cuenta con plantillas html que permiten su fácil uso para el despliegue del servicio de forma responsiva, así como la opción de implementar una arquitectura de módulos para la independencia de los servicios. El lenguaje escogido para esta capa fue TypeScript, el cual es un superconjunto de JavaScript, y específicamente se usó el *framework* Angular en su versión 11.

Evaluación y experimentación de los servicios desarrollados

La experimentación desarrollada buscó poner a prueba la eficiencia de los dos servicios especializados propuestos en el modelo macro y que fueron desarrollados a manera de prueba de concepto, para ello y contemplando la complejidad

de obtener información de documentos que hayan sido tipificados en actos corruptos, se procedió con la creación de datos sintéticos que sirvieron como fuente de información para la evaluación de los servicios desarrollados.

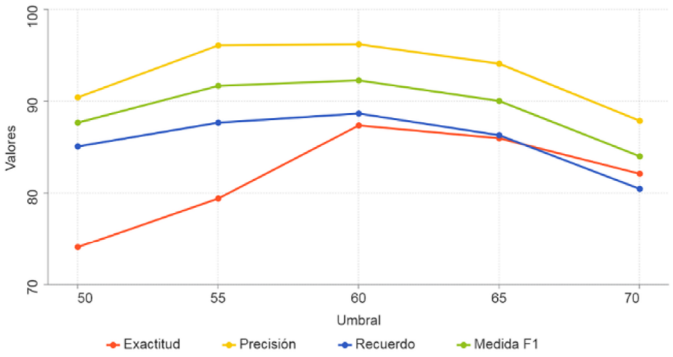
Experimentación del servicio de detección de similitudes

Para llevar a cabo la experimentación de este servicio se pusieron a prueba diferentes escenarios que se pueden presentar al momento de comparar texto, como el copiado textual (sin agregar modificaciones), el copiado parcial (copiado solo de algunas frases) y copiado inteligente (agregar modificaciones a frases con el objetivo de disimular la intención de copia). En total se realizaron pruebas en seis convocatorias, cada una de ellas con cuatro documentos de propuestas técnicas, con un total de 135 frases a evaluar. Con los resultados arrojados por la probabilidad de similitud de cada una de las frases evaluadas, se procedió a la creación de la matriz de confusión y el cálculo de métricas clásicas del área resultantes de la experimentación. El contenido de una frase es marcado como plagio, si su probabilidad de similitud es superior a un umbral dado, para la correcta escogencia de dicho umbral, se procedió a realizar diferentes pruebas variando el valor del mismo. La Figura 5 muestra, a su izquierda, la gráfica del cálculo de algunas de las métricas más representativas del área evaluadas para diferentes valores del umbral. La experimentación realizada sugiere que un valor de 60 % es el indicado.

Los resultados finales de la experimentación se observan a la derecha de la Figura 5, se aprecia que si una propuesta técnica es marcada como plagio, existe un 88,70 % de probabilidad de etiquetarla como tal. Mientras que el servicio etiqueta como partícipes del delito al 20 % de las que no son marcadas como plagio (falsos positivos). Así mismo, la figura otorga información relevante sobre métricas calculadas a partir de la matriz de confusión, destacando que el servicio obtiene una exactitud del 87,40 % y una precisión del 96,23 %, los cuales son valores considerablemente altos.

Experimentación del servicio de detección de manipulación de ofertas

Para la experimentación de este servicio se implementó una estrategia que consistió en poner a prueba diferentes escenarios que se pueden presentar en el delito de manipulación de ofertas dentro de una convocatoria. En total se realizaron pruebas en ocho convocatorias, cada una de ellas con cuatro documentos de propuestas económicas, en donde se simularon doblemente cuatro escenarios: a) todas (4) las ofertas son parte del delito, b) tres ofertas son parte del



Matriz de confusión resultados de experimentación

	Real verdadero	Real falso	Predicción general
Predice Verdadero	102 (88,70%)	4 (20%)	106
Predice Falso	13 (11,30%)	16 (80%)	29
Verdad general	115	20	Total = 135

Cálculo de métricas

Métrica	Valor (%)
Exactitud	87,40
Precisión	96,23
Recuerdo	88,70
Medida F1	92,30

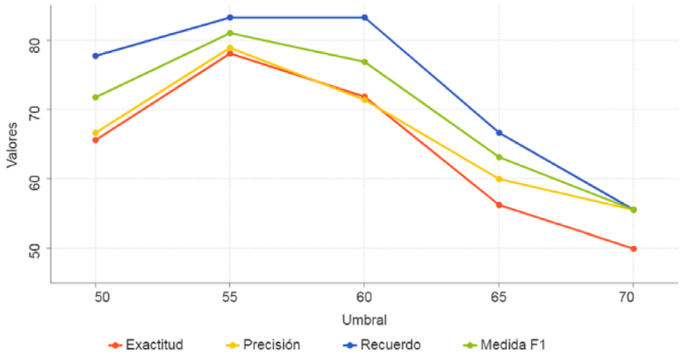
Figura 5. Gráfica de evaluación valores del umbral - Matriz de confusión y métricas sobre la experimentación del servicio de detección de similitud.

delito y d) no hay presencia intencional del delito. Similar a la experimentación del servicio de detección de similitudes, para este servicio se determina que un documento es marcada por el servicio como partícipe del delito de manipulación de ofertas si la probabilidad otorgada por el mismo supera un umbral fijado, y para la correcta selección de valor dado al umbral, se procedió a realizar diferentes pruebas variándolo. La Figura 6 muestra, a su izquierda, la gráfica del cálculo de algunas de las métricas más representativas del área evaluadas para diferentes valores del umbral. La experimentación realizada sugiere que un valor de 55 % es el indicado.

Con los datos obtenidos en la experimentación de este servicio se procedió a la creación de la matriz de confusión y el cálculo de métricas clásicas del área resultantes de la experimentación, los cuales se pueden ver a la derecha de la Figura 6. Se aprecia que, si una propuesta económica es marcada como partícipe del delito de manipulación de ofertas, existe un 83,33 % de probabilidad de etiquetarla como tal. Mientras que el servicio etiqueta como partícipes del delito al 28,60 % de las que no lo son (falsos positivos). Así mismo, la figura otorga información de métricas calculadas a partir de la matriz de confusión, destacando que el servicio obtiene una exactitud de 78,13 % y una precisión del 78,95 %, los cuales son valores considerablemente altos.

Conclusiones

El modelo planteado en esta investigación busca, de manera conjunta con los servicios planteados, detectar diferentes probables delitos o anomalías relacionadas con corrupción en los procesos de contratación pública en Colombia. Dos de los servicios propuestos en el modelo fueron implementados mediante la creación de una herramienta software que sirvió como prueba de concepto del modelo expuesto, la herramienta cuenta con las funcionalidades planteadas en el componente de servicios transversales, logrando así la transformación del modelo en una herramienta software. Debido a la complejidad de obtener información de documentos que hayan sido tipificados en actos corruptos, en esta investigación, para llevar a cabo la experimentación, se contó con un conjunto de datos sintéticos los cuales fueron puestos a prueba en los dos servicios desarrollados. La precisión de los servicios fue evaluada mediante métricas clásicas del área, arrojando valores relativamente altos. Con los datos arrojados por la experimentación se puede afirmar que los servicios propuestos e implementados en esta investigación permiten la identificación de presuntas situaciones de corrupción en los datos y documentos asociados a los procesos de contratación pública con un alto grado de precisión.



Matriz de confusión resultados de experimentación

	Real verdadero	Real falso	Predicción general
Predice Verdadero	15 (83,33%)	4 (28,60%)	19
Predice Falso	3 (16,67)	10 (71,40%)	13
Verdad general	18	14	Total = 32

Cálculo de métricas

Métrica	Valor (%)
Exactitud	78,13
Precisión	78,95
Recuerdo	83,33
Medida F1	81,08

Figura 6. Gráfica de evaluación valores del umbral - Matriz de confusión y métricas sobre la experimentación del servicio de detección de manipulación de ofertas.

Como trabajos futuros se considera importante la integración de nuevos servicios específicos dedicados a la detección de nuevos delitos, así como la implementación del servicio especializado de detección del probable delito de organización de carteles y análisis del beneficiario final y del servicio de alerta ciudadana sobre irregularidades en procesos de convocatorias de contratación pública. Esto con el objetivo de enriquecer aún más, tanto el modelo propuesto en esta investigación como la herramienta software que lo sustenta. Finalmente, el grupo de investigación espera, como trabajo futuro, poder desplegar la herramienta software desarrollada en una entidad estatal colombiana, integrándola en sus procesos de contratación, para permitir brindar una ayuda en la tarea de la prevención de la corrupción en los procesos de contratación estatal.

Contribución de autoría

Julio-César Luna-Ortega: Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Software, Validación, Escritua – Borrador original.
Carlos-Alberto Cobos-Lozada: Conceptualización, Análisis formal, Investigación, Metodología, Supervisión, Validación, Escritura – revisión y edición.
Martha-Eliana Mendoza-Becerra: Supervisión, Validación, Escritura – revisión y edición.

Agradecimientos

Los autores desean extender un mensaje de agradecimiento a la Universidad del Cauca y al clúster CreaTIC por financiar parcialmente el desarrollo de la presente investigación.

Referencias

Abrantes-Metz, R. M. (2013). Roundtable on ex officio cartel investigations and the use of screens to detect cartels. *SSRN*. https://doi.org/10.2139/SSRN.2343465

Attaran, M., Woods, J. (2019). Cloud computing technology: Improving small business performance using the Internet. *Journal of Small Business & Entrepreneurship*, *31*(6), 495-519. https://doi.org/10.1080/08276331.2018.1466850

Erkan, G., Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457-479. https://doi.org/10.1613/JAIR.1523

Gkantouna, V., Papaioannou, V., Tzimas, G., Sabic, Z. (2020). A semantic approach for domain-specific design patterns recommendations in CMS-based web development. En *International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. https://doi.org/10.1109/INISTA49547.2020.9194622

Imhof, D. (2018). *Empirical Methods for Detecting Bid-rigging Cartels* [Tesis doctoral]. Université Bourgogne Franche-Comté. https://theses.hal.science/tel-01963076/document

Luna-Ortega, J.-C., Cobos-Lozada, C.-A., Mendoza-Becerra, M.-E. (2020). Una revisión sistemática de los métodos de identificación y detección de corrupción en contratación pública. *Revista Ibérica de Sistemas e Tecnologias de Informação*, *E38*, 43-57. https://www.proquest.com/openview/f92f510dcbcd38ad0bf1425e3ad345b0/1.pdf

Ossa Bocanegra, C. E. (2014). Tratamiento de la colusión en la contratación pública: una visión del caso colombiano. *Revista de Derecho*, *42*, 233-263

Owusu, E. K., Chan, A. P. C., Shan, M. (2019). Causal factors of corruption in construction project management: An overview. *Science and Engineering Ethics*, *25*(1), 1-31. https://doi.org/10.1007/s11948-017-0002-4

Peltonen, S., Mezzalira, L., Taibi, D. (2021). Motivations, benefits, and issues for adopting micro-frontends: A multivocal literature review. *Information and Software Technology*, *136,* e106571. https://doi.org/10.1016/j.infsof.2021.106571

Pibiri, G. E., Venturini, R. (2021). Techniques for Inverted Index Compression. *ACM Computing Surveys (CSUR)*, *53*(6), 1-36. https://doi.org/10.1145/3415148

Stein, B., Zu Eissen, S. M., Potthast, M. (2007). Strategies for retrieving plagiarized documents. En *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07*, 825-826. https://doi.org/10.1145/1277741.1277928

Transparencia por Colombia. (2022). *Índice de Percepción de la Corrupción (IPC) 2021*. Transparencia por Colombia, Capítulo Transparencia Internacional. https://transparenciacolombia.org.co/2022/01/25/indice-de-percepcion-de-la-corrupcion-2021/

Waseem, M., Liang, P., Shahin, M. (2020). A Systematic mapping study on microservices architecture in DevOps. *Journal of Systems and Software*, *170*, e110798. https://doi.org/10.1016/J.JSS.2020.110798

Weber, I., Lu, Q., Tran, A. B., Deshmukh, A., Gorski, M., Strazds, M. (2019). A platform architecture for multi-tenant blockchain-based systems. En *IEEE International Conference on Software Architecture (ICSA)*, 101-110. https://doi.org/10.1109/ICSA.2019.00019