Universidad de los Andes
Escuela de Gobierno Alberto Lleras Camargo

No. 120
Marzo del 2025
ISSN 2215 – 7816 (En línea)

Documentos de Trabajo

Escuela de Gobierno Alberto Lleras Camargo

Sesgo de datos en aplicaciones de aprendizaje automático: un estudio de caso de un modelo no supervisado para identificar el riesgo de corrupción en la contratación pública colombiana

Kevin Steven Mojica Muñoz

Serie Documentos de Trabajo 2025
Edición No. 120
ISSN 2215-7816 (En línea)
Edición digital
Marzo 2025

Carrera 1 No. 19 -27, Bloque Aulas
Bogotá, D.C., Colombia
Teléfono: 3394949, ext. 2073
publicaciones@uniandes.edu.co
http://gobierno.uniandes.edu.co

Autores
Kevin Steven Mojica Muñoz

Directora de la Escuela de Gobierno Alberto Lleras Camargo
María Margarita, Paca Zuleta

Coordinación editorial, Escuela de Gobierno Alberto Lleras Camargo
María Alejandra Rojas Forero

Dirección de Investigaciones, Escuela de Gobierno Alberto Lleras Camargo
Diego Iván Lucumí Cuesta

Diagramación de cubierta, Escuela de Gobierno Alberto Lleras Camargo
Miguel Ángel Campos Guaqueta

Sesgo de datos en aplicaciones de aprendizaje automático: un estudio de caso de un modelo no supervisado para identificar el riesgo de corrupción en la contratación pública colombiana

Kevin Steven Mojica Muñoz[1]

Este estudio analiza el sesgo de datos en un algoritmo de aprendizaje no supervisado diseñado para identificar el riesgo de corrupción en la contratación pública de Colombia. El algoritmo empleado es un modelo de *clustering* en dos etapas utilizado para segmentar contratos electrónicos a partir de variables que indican riesgo de corrupción. El objetivo era desarrollar una herramienta de alertas tempranas de corrupción en la contratación del Programa de Alimentación Escolar (PAE), empleando datos del Sistema Electrónico para la Contratación Pública (SECOP). A pesar de que los resultados demuestran el potencial de los algoritmos de Inteligencia Artificial para la detección de riesgos de corrupción, también revelan limitaciones significativas en su implementación práctica, atribuibles a deficiencias en la disponibilidad y calidad de los datos. En particular, se identificaron sesgos de representación, de medición y de variables omitidas que afectan la confiabilidad del algoritmo. El estudio proporciona un análisis detallado de estos sesgos, evaluando su impacto en el desempeño del algoritmo, y enfatizando la importancia de reconocer y abordar los sesgos durante el desarrollo de este tipo de iniciativas. Finalmente, se presentan recomendaciones para mejorar la calidad de los datos en el SECOP, con el fin de fortalecer la fiabilidad y precisión de estos algoritmos en desarrollos futuros.

Palabras clave: sesgo de datos; aprendizaje no supervisado, Inteligencia Artificial, contratación pública, IA, justicia.

[1] MSc. Economía y MSc. Políticas Públicas. Investigador de la Universidad de los Andes, ks.mojica10@uniandes.edu.co.

Data Bias in Machine Learning Applications: A Case Study of an Unsupervised Model to Identify the Risk of Corruption in Colombian Public Procurement

Kevin Steven Mojica Muñoz[2]

This study analyzes data bias in an unsupervised learning algorithm designed to identify the risk of corruption in public procurement of Colombia. The employed algorithm is a two-stage clustering model used to segment electronic contracts based on variables indicating corruption risk. The objective was to develop an early warning tool for corruption in the Programa de Alimentación Escolar (PAE) procurement, utilizing data from the Sistema Electrónico para la Contratación Pública (SECOP). Although the results demonstrate the potential of artificial intelligence algorithms for detecting corruption risks, they also reveal significant limitations in their practical implementation, attributable to data availability and quality deficiencies. Specifically, biases of representation, measurement, and omitted variables were identified, affecting the algorithm's reliability. The study provides a detailed analysis of these biases, assessing their impact on the algorithm's performance, and emphasizes the importance of recognizing and addressing biases during the development of such initiatives. Finally, recommendations are presented to improve the quality of data in SECOP, aiming to enhance the reliability and accuracy of these algorithms in future developments.

Key Words: data bias; unsupervised learning, Artificial Intelligence, public procurement, AI, fairness.

[2] MSc. Economía y MSc. Políticas Públicas. Researcher at Universidad de los Andes, ks.mojica10@uniandes.edu.co.

Tabla de contenido

1. Introducción

La Inteligencia Artificial (IA) es el conjunto de tecnologías más importante de las últimas décadas debido a su rápido avance y las aplicaciones en casi todos los campos de la ciencia y la ingeniería (Raj, R., y Kos, A., 2023). De acuerdo con Quid via AI Index (2024) y el U.S. Bureau of Labor Statistics (2024), las inversiones globales en Inteligencia Artificial generativa pasaron de 0,889 billones de dólares en el 2019 a 22,43 billones en el 2023[3], un aumento del 2423% en tan solo cuatro años. Sin embargo, a medida que la popularidad de estos algoritmos aumenta, crecen también las preocupaciones con relación a los posibles sesgos o injusticias que pueda tener la aplicación de estas herramientas en ámbitos críticos para la sociedad.

Esta investigación explora un caso de estudio sobre sesgo de datos en algoritmos de Inteligencia Artificial en un país en desarrollo, teniendo como propósito evidenciar cómo la calidad de los datos puede afectar la utilidad y viabilidad de estas herramientas en aplicaciones de las ciencias humanas. Para lograr esto, el documento estudia la aplicación de un algoritmo de aprendizaje no supervisado utilizado para identificar riesgo de corrupción en la contratación pública de Colombia. La investigación expone los sesgos identificados en los datos que alimentan el modelo y el impacto en su uso final, de manera que los aprendizajes y buenas prácticas puedan ser aprovechados por futuras aplicaciones en ámbitos de las ciencias humanas.

A pesar de que este estudio de caso se centra en Colombia, las conclusiones de este ejercicio son de gran relevancia para todos los países en desarrollo que están actualmente explorando el uso de estas herramientas en la gestión pública. Por ejemplo, solo en América Latina y el Caribe se han implementado unos 272 sistemas con Inteligencia Artificial en el sector público, donde el 38% de estos algoritmos cumplen funciones relacionadas con análisis, monitoreo e investigación de política pública (22%), o gestión interna de procesos (16%) (Comisión Económica para América Latina y el Caribe (CEPAL), 2024).

El sesgo en aplicaciones de Inteligencia Artificial es un tema que ha empezado a tener mayor interés de estudio en los últimos años. Algunas investigaciones tempranas en el tema

[3] Considerando dólares constantes del 2021.

evidenciaron cómo estas herramientas pueden tener efectos no deseados que perpetuaban prejuicios, injusticias y discriminación. Por ejemplo, Noble (2018) descubrió que los motores de búsqueda apoyados en algoritmos de Inteligencia Artificial perpetuaban prácticas discriminatorias y contenidos sexualizados que afectaban las mujeres negras. Un resultado similar fue identificado por Buolamwini y Gebru (2018), quienes encontraron que algoritmos de reconocimiento facial basados en Inteligencia Artificial podían generar discriminación de raza y género. Mientras que Angwin y Larson (2022), encontraron que los sistemas utilizados en ámbitos judiciales para identificar el riesgo de los acusados eran discriminatorios hacia las poblaciones vulnerables y generaban recomendaciones injustas.

Si bien la mayoría de estas investigaciones iniciales identificaron sesgos e injusticias en términos de raza, los avances siguientes encontraron que estos sesgos podían ser de todo tipo y ampliarse a aplicaciones que no trabajaran con humanos. Algunos aportes como los de Nazer et al., (2023), Suresh y Guttag (2021), Mehrabi et al., (2021), Ntoutsi et al., (2020), y Hellstrom et al., (2020) han logrado generar un marco conceptual alrededor del tema que ha permitido identificar fuentes de sesgo y criterios objetivos de evaluación. Esto ha permitido el desarrollo de métodos para detectar y mitigar la existencia de estos sesgos en nuevas aplicaciones (Pagano et al., 2023).

A pesar de los avances en el tema, la mayoría de la evidencia práctica disponible en la literatura se centra en las ciencias médicas y exactas (Kumar et al., 2024; Celi et al., 2022; Challen et al., 2019; Suri et al., 2022), donde es más fácil identificar y abordar el sesgo algorítmico. En contraposición, existe una carencia de información que ejemplifique estos problemas en aplicaciones de las ciencias humanas y la administración pública. Este déficit de información es aún mayor para países en desarrollo, los cuales pueden presentar características y condiciones distintas.

Los sesgos en los algoritmos pueden generar que las decisiones que se tomen a partir de su uso no correspondan con los criterios éticos, morales y de justicia que rigen la sociedad, o que la aplicación resulte ser inútil para la tarea que dice resolver. Esto en algunos casos puede llevar incluso a la vulneración sistemática de derechos fundamentales, como el derecho a la igualdad y no discriminación. Por lo mismo, comprender las limitaciones y posibles sesgos de los algoritmos de Inteligencia Artificial en contextos de países en desarrollo es crucial para

poder implementar y aprovechar estas herramientas en las ciencias humanas y la administración pública de manera eficaz. Los países en desarrollo pueden carecer de la infraestructura tecnológica y los estándares de calidad en los datos necesarios para que replicar aplicaciones de estos algoritmos que funcionan adecuadamente en los países desarrollados. La falta de evidencia al respecto puede dar la falsa impresión de que estos problemas no existen o no son relevantes en dichas aplicaciones, aunque en realidad pueden ser determinantes para el éxito de su aplicación.

El documento se estructura en cinco secciones, siendo esta introducción la primera. La segunda sección aborda el marco conceptual que categoriza los sesgos en los datos por su origen, proporcionando un estándar para comparar los hallazgos con la literatura académica relevante. La tercera sección profundiza en el caso de estudio, describiendo el funcionamiento y características del algoritmo diseñado para identificar riesgo de corrupción en la contratación pública de Colombia. La cuarta sección expone los resultados del algoritmo y los sesgos de datos identificados en su desarrollo, mostrando el impacto que tienen en el funcionamiento del algoritmo. En concreto, se busca responder a las siguientes preguntas: ¿Cuáles son los sesgos identificados en la herramienta y su origen? ¿Cuál es el impacto de estos sesgos en la aplicación práctica del algoritmo? Finalmente, la última sección aborda las conclusiones y recomendaciones derivadas de este estudio que son relevantes para futuros proyectos similares.

2. Marco conceptual

De acuerdo con Mehrabi et al. (2021), el sesgo algorítmico puede clasificarse en tres tipos; sesgo de datos, sesgo del algoritmo, y sesgo de interacción con el usuario. El primero, como su nombre lo indica, se origina porque los datos de entrada no son adecuados para la tarea establecida o no tienen la calidad necesaria para ser utilizados. El segundo ocurre en la selección u optimización del algoritmo utilizado, por ejemplo, en el diseño de ciertas funciones de optimización, regularizaciones, o elecciones en la aplicación de los modelos. Finalmente, el tercero ocurre cuando la interfaz del algoritmo le permite anteponer al usuario creencias o comportamientos que llevan a una decisión autoseleccionada.

El primer caso es de los más críticos porque depende de factores externos al método de entrenamiento utilizado y puede ser bastante común en países con deficiencias en la calidad de los datos. Los algoritmos de aprendizaje de máquinas suelen basarse en datos de entrada existentes para aprender y generar nuevas conclusiones, ya sea mediante la predicción de atributos, o la identificación de patrones en datos no etiquetados. Si los datos de entrada no tienen la calidad necesaria o contienen sesgos su aplicación puede terminar generando injusticias o decisiones subóptimas que pueden afectar su desempeño y legitimidad.

Los sesgos de datos pueden clasificarse según su origen Mehrabi et al., (2021) hacen una clasificación a partir de lo establecido por Olteanu et al., (2016), y Suresh H., y Guttag, J., (2019), identificando las siguientes categorías:

Sesgo de medición: originado en la forma en que se seleccionan, utilizan y miden ciertas características. En este tipo de sesgo los datos no reflejan de manera precisa el constructo que dicen representar, por lo que apenas se tiene una aproximación de ese constructo en la información de entrada al algoritmo. Matemáticamente este sesgo se puede representar de la siguiente forma:

$$Y_i = \widetilde{y_i} + \varphi_i$$

Donde Y_i representa la variable disponible para la observación i, que es igual al constructo no observable que dice representar $\widetilde{y_i}$ (valor real) más un error que puede ser positivo o negativo. Si el error no es aleatorio, entonces Y, en el agregado, será diferente a $\tilde{y}$.

Sesgo de variable omitida: originado cuando una característica fundamental no se tiene disponible en los datos y no se logra integrar en el modelo de aprendizaje de máquinas. La variación no observada correspondiente a esa variable queda integrada en el error del modelo, que matemáticamente se puede representar de la siguiente forma:

$$U_i = \widehat{y_i} + \varphi_i$$

Donde U_i es el error del modelo, que se compone de un error aleatorio para cada observación φ_i y la variable omitida $\widehat{y_i}$. En este caso, la existencia de la variable omitida hace que el error del modelo no sea igual a un error aleatorio, sino que se

encuentre sesgado. En algunos modelos de aprendizaje este sesgo puede preverse en la evaluación de los algoritmos, al reportar un desempeño deficiente en las métricas de evaluación.

Sesgo de representación: originado cuando los datos no reflejan la composición real del constructo que dicen representar. En este tipo de sesgo los datos no son un reflejo confiable de la realidad porque hay subgrupos sobrerrepresentados o subrepresentados en la muestra. Al aprender de una representación sesgada de la realidad, el modelo puede identificar patrones erróneos o inexistentes, lo que a su vez conduce a predicciones o decisiones que perpetúan y amplifican las desigualdades presentes en los datos originales. Un ejemplo de este sesgo de representación es el caso de los sistemas de reconocimiento facial entrenados con bases de datos que contienen predominantemente imágenes de personas de ascendencia europea, lo que provoca que estos sistemas tengan un desempeño inferior al identificar rostros de personas de otras etnias. Este desequilibrio en la representación impide que el modelo generalice de manera adecuada en contextos diversos.

Sesgo de muestreo: similar al sesgo de representación, en este caso los datos no son un reflejo confiable de la realidad porque existen grupos sobrerrepresentados o subrepresentados en la muestra. La diferencia es que este sesgo se origina en la forma en que se realiza el muestreo en la población durante el proceso de recolección de información. Cuando la muestra no es seleccionada de manera aleatoria sobre el conjunto total de información las conclusiones derivadas de un modelo de aprendizaje de máquinas no pueden ser extrapoladas a la población general.

Sesgo de agregación: originado cuando se asumen atributos falsos sobre subgrupos de individuos a partir de observaciones en la población general. En este tipo de sesgo los datos de la población mayoritaria pueden ocultar características particulares asociadas a un subgrupo minoritario de datos. El algoritmo se entrena utilizando los datos de la población general, por lo que las conclusiones que se extraigan de su aplicación a subgrupos minoritarios con características diferentes pueden no ser adecuadas.

Sesgo de datos de corte transversal: originado cuando se agregan datos de corte transversal para diferentes periodos de tiempo buscando generar conclusiones sobre la evolución de una variable en el tiempo, en vez de utilizar datos longitudinales. Esto se debe a que las características que originan unos datos para el año inicial de análisis pueden diferir a las que aparecen en los años subsiguientes.

Sesgo de enlace: originado en análisis de redes sociales cuando los atributos de la red obtenidos a partir de las conexiones, actividades o interacciones de los usuarios difieren y tergiversan el verdadero comportamiento de los usuarios. Este es un tipo de sesgo no muy estudiado en la literatura y presente solo en casos específicos de análisis.

Los datos de calidad y confiables son imprescindibles en cualquier aplicación con algoritmos de Inteligencia Artificial. Sin embargo, en los campos de las ciencias humanas y la administración pública, es importante reconocer que estos algoritmos nunca estarán completamente libres de sesgo. Esto se debe a la naturaleza compleja de estas disciplinas y a los desafíos asociados con la recopilación y procesamiento de información.

A pesar de lo anterior, en algunos casos la presencia de sesgo puede ser más crítica y generar problemas en la aplicabilidad de los algoritmos. Esto ocurre cuando el sesgo no se alinea con los estándares éticos y morales esperados, o cuando afecta considerablemente la confiabilidad del algoritmo. En esos casos es necesario evaluar el sesgo e identificar si hay alguna forma de mitigarlo o prevenirlo, o si es mejor prescindir de su utilización hasta que existan mejores condiciones para desarrollarlo.

3. Descripción del caso analizado: aprendizaje no supervisado para detectar corrupción en la contratación pública de Colombia

El caso a estudiar es relevante por varios motivos. En primer lugar, ilustra una aplicación pionera de algoritmos de aprendizaje de máquinas para la detección de corrupción en la contratación pública de un país en desarrollo. En segundo lugar, muestra cómo los sesgos en los datos utilizados para el entrenamiento del modelo pueden afectar de forma considerable la funcionalidad de los algoritmos, pudiendo generar efectos contrarios a los deseados inicialmente. En el caso particular, el uso del algoritmo sin considerar sus limitaciones puede

llevar a asignaciones ineficientes de recursos en la gestión pública, así como generar decisiones que puedan ser injustas y contrarias al objetivo inicial del mismo.

3.1. Antecedentes y objetivos del algoritmo

En el año 2023, Kevin Steven Mojica Muñoz, investigador de la Universidad de los Andes desarrolló el proyecto "Aprendizaje no supervisado para la detección de corrupción en la contratación pública de Colombia"[4], que buscaba implementar y evaluar algoritmos de aprendizaje no supervisado para detectar posibles actos de corrupción en la contratación del país, tomando como política de estudio el Programa de Alimentación Escolar (PAE). El objetivo final del trabajo era desarrollar un esquema de alertas tempranas de riesgo de corrupción que permitiera identificar con antelación y prevenir potenciales actos de corrupción en la contratación pública de Colombia.

El estudio partía de dos precedentes importantes en la materia. El primero era la investigación de Gallego, Rivero y Martínez (2021), que utilizó aprendizaje supervisado para identificar corrupción e ineficiencias administrativas en la contratación pública de Colombia. En concreto, esta iniciativa desarrolló un algoritmo que tenía la capacidad de predecir investigaciones de corrupción, incumplimientos del contrato, o ineficiencias en la implementación utilizando los datos del Sistema Electrónico de Contratación Pública de Colombia (SECOP). Si bien los resultados indicaban un desempeño favorable de los algoritmos, las limitaciones del enfoque supervisado es que requerían de una variable objetivo a predecir que puede no estar disponible en el nivel que se requiere[5] o reflejar únicamente la corrupción que ya era previamente detectable[6].

[4] El proyecto recibió financiación parcial por parte del Centro de Estudios Manuel Ramírez.
[5] Los autores especifican que el riesgo de corrupción, aunque se estima para contratos, realmente se establece a nivel de ofertante. Esto se debe a que los datos utilizados indicaban si un ofertante había tenido casos de malversación de recursos en algún momento del tiempo, marcando cualquier contrato que haya ganado ese ofertante como riesgoso.
[6]Los autores especifican que existía un problema de etiquetado selectivo, que ocurre cuando la variable objetivo del aprendizaje supervisado es el resultado de una decisión humana. En el caso concreto, las investigaciones de corrupción eran resultado de la decisión de un auditor de iniciar una investigación, por lo que podían incorporar únicamente la corrupción que ya era detectable por la entidad, dejando entramados de corrupción ocultos fuera de las capacidades de detección del algoritmo.

El segundo precedente es el de Castiblanco (2018), que utilizó un enfoque de análisis no supervisado y supervisado para detectar irregularidades en la contratación pública de Bogotá y Antioquia utilizando los datos de la primera fase del SECOP, demostrando que este tipo de herramientas puede tener potencial para detectar riesgo de corrupción en la contratación. Este estudio, aunque es el primero en su tipo en proponer un enfoque no supervisado, solo lo hace de manera exploratoria, decantándose en últimas por un enfoque tradicional de aprendizaje supervisado, con las falencias que conlleva en términos de disponibilidad y fiabilidad de la variable objetivo. Esto se debe a que la metodología propuesta en el análisis no supervisado, *clustering* multivariado, no permitía tener contratos con características comparables, lo cual podía llevar a comparaciones erróneas que no reflejaran realmente irregularidades en la contratación.

A partir de estos precedentes, la investigación buscaba implementar y evaluar algoritmos de aprendizaje no supervisado para detectar riesgo de corrupción en la contratación pública del país, considerando los datos del SECOP en su segunda generación. La ventaja de utilizar este enfoque es que no requería de una variable objetivo para su desarrollo, lo que significa una ventaja frente a lo desarrollado previamente por Gallego, Rivero y Martínez (2021) y Castiblanco (2018). El objetivo de política era pasar de un esquema reactivo a un esquema preventivo de la corrupción, en el cual se pudieran focalizar los esfuerzos de investigación por parte de los organismos de control, la sociedad civil, y los medios de comunicación.

Como política a estudiar, y para facilitar la comparación con respecto a la investigación de Castiblanco (2018), se seleccionó la contratación inherente al Programa de alimentación Escolar (PAE) de Colombia. Este es un programa de gobierno establecido en el año 2011 por el Ministerio de Educación para garantizar la alimentación escolar de estudiantes en todo el territorio nacional (Ley 1450 del 2011). El PAE tiene la ventaja de que se trata de un programa que ofrece un servicio público relativamente estándar, lo cual facilita el análisis de presuntas irregularidades en el tema, y que ha sido frecuentemente señalado como un programa de gobierno con altos niveles de corrupción. De acuerdo con datos del Ministerio de Educación (2024), al año 2023 el servicio llegaba a 5 917 988 estudiantes en todo el territorio nacional,

siendo la principal estrategia de acceso y permanencia de los niños y adolescentes en el sistema educativo del país.

La ejecución de la política se da de manera descentralizada a través de los entes territoriales certificados en la "Certificación de Habilitación para Contratar el Programa de Alimentación Escolar (PAE)", expedida por el Ministerio de Educación. Para los municipios no certificados, la ejecución del programa se da en cabeza de la administración departamental (gobernaciones). Las entidades territoriales, en el marco de sus funciones legales, realizan la contratación incorporando los criterios generales del Ministerio de Educación a las particularidades de cada territorio. El seguimiento a la ejecución del contrato se da directamente por las entidades territoriales o por medio de contratos de interventoría.

Si bien los resultados de implementación de los algoritmos son específicos al programa estudiado, su aplicación muestra la viabilidad de su uso en otros contextos de política pública. Esto permite extraer aprendizajes sobre los desafíos y oportunidades que estas herramientas presentan en la identificación de corrupción dentro de programas ejecutados mediante contratación pública.

Para implementar el algoritmo se establecieron un conjunto de supuestos:

1. Las entidades que están legalmente obligadas a publicar sus contratos en SECOP II efectivamente lo hacen en la totalidad de su contratación[7]. Esto implica que los datos en SECOP II contienen el universo de contratos elegibles para las diferentes aplicaciones que se desarrollen con Inteligencia Artificial en el conjunto de entidades legalmente obligadas.
2. Las entidades digitan la información de los contratos de manera rigurosa para la mayor parte de su contratación y no manipulan los datos, lo que implica que la información en las bases de datos del SECOP coincide con la información real de la contratación y la documentación del proceso.
3. Las bases de datos del SECOP II se encuentran estructuradas y son de libre acceso.

[7] Esta obligación se encuentra amparada en el Decreto 1082 del 2015 y la Ley 1150 del 2007.

4. Las bases de datos del SECOP II contienen información suficiente para entrenar modelos de aprendizaje de máquinas con alto grado de confiabilidad.

Estos supuestos no se cumplieron a cabalidad en el caso de estudio desarrollado, lo cual fue parcialmente responsable de los sesgos identificados. Esto se profundiza en las secciones posteriores.

3.2. Metodología de investigación

La investigación siguió tres fases consecutivas de trabajo: i) Preparación de la muestra, ii) entrenamiento de algoritmos, y iii) evaluación de desempeño. En la primera fase, el objetivo era establecer una muestra óptima para la implementación de los algoritmos, esto implica la compilación y estandarización de los datos, el tratamiento de los valores faltantes, y el filtrado de las variables que se utilizaron en el análisis. Esta es la fase que demandó una mayor cantidad de esfuerzo y tiempo porque implicaba la unificación y estructuración de las bases a partir de los datos del SECOP.

La segunda fase de investigación suponía la aplicación de los algoritmos de aprendizaje de máquinas en la submuestra de trabajo. El algoritmo a utilizar era un *clustering* en dos etapas con modelo base de *kmeans* y con criterio de distancia euclidiana. El *clustering* multivariable con modelo base *kmeans* es un algoritmo de aprendizaje no supervisado que tiene por objetivo segmentar la información existente en grupos de datos que comparten características en común para un conjunto de variables determinadas, donde cada grupo presenta una distribución similar en las variables de interés que se diferencia al de los demás grupos (Kanungo et al., 2002).

El algoritmo funciona iterando sucesivas rondas en las que asigna cada observación a uno de los K grupos basado en la distancia entre el punto de datos (vector de datos en las variables de interés) y el centroide (la media) de cada grupo. En cada iteración se recalcula el centroide de cada grupo, entendiéndose este como la media en las variables de todos los datos asignados al grupo. Este procedimiento se repite hasta que no hay cambios significativos en la asignación las observaciones en los grupos. A medida que aumenta el número de grupos,

en este caso el parámetro ‘k’ en el modelo *kmeans*, cada grupo se vuelve más específico para retratar una combinación en la distribución de las variables.

La primera etapa del *clustering* en dos etapas busca segmentar la muestra en conjuntos de contratos que sean similares entre sí, debido a que cualquier comparación que indique posibles irregularidades relacionadas a la corrupción debe basarse en contratos que sean comparables. Esta primera etapa define un conjunto de grupos donde cada contrato comparte características similares con sus pares en el mismo grupo. Las variables que se utilizaron en la primera etapa fueron la duración del contrato y el valor del contrato. Si bien es cierto que otras características pueden generar diferencias entre los contratos que no estén directamente relacionadas con la corrupción, esta agrupación de la primera etapa se basa en la utilización exclusiva de variables esenciales y comunes para garantizar la comparabilidad. Esto permite conformar grupos con un tamaño mínimo adecuado para el análisis posterior, evitando que la inclusión de variables adicionales genere una segmentación excesivamente fragmentada o introduzca ruido y sesgos derivados de información irrelevante.

El procesamiento de los datos en la primera etapa permite implementar el segundo *clustering*, que tiene por objetivo estimar el nivel de riesgo de corrupción a través de un *clustering* multivariado en cada una de las submuestras de la primera etapa. En este caso, se busca identificar cuáles son los contratos que en el tema analizado resultan atípicos y deberían analizarse de manera individual porque tienen características asociadas a la corrupción. Por lo tanto, se identifica como anomalía al *cluster* de contratos que en la segunda etapa presenta una distribución de las variables concordante con criterios de irregularidades relacionadas a la corrupción. Este subconjunto es entonces la alerta temprana de riesgo de corrupción que se extrae del uso del algoritmo de aprendizaje no supervisado.

Para esta segunda etapa se establecieron por defecto cinco conjuntos de agrupamiento (k=5) y un conjunto de variables tales que cada una representa un factor de riesgo de corrupción en la escala 0 – 1, donde para todas un valor más cercano a 1 representa mayor riesgo de corrupción. Una vez desarrollado el agrupamiento, los cinco subgrupos resultantes se clasifican en las categorías de riesgo de corrupción considerando la media global en el centroide de cada subconjunto, teniendo que el grupo con los valores más bajos se clasifica en

riesgo ‘Muy bajo’, el siguiente ‘Bajo’, el siguiente ‘Medio’, el siguiente ‘Alto’, hasta el último que se clasifica en la categoría ‘Muy alto’.

De esta forma se logra obtener una categorización del riesgo de corrupción que funciona como alerta temprana, al seleccionar el subconjunto de contratos activos que representan mayor riesgo. También funciona como un mecanismo para investigar contratos ya finalizados, al permitir identificar el subconjunto de contratos ya terminados que presentan la mayor cantidad de combinaciones en factores de riesgo de corrupción.

Las variables que se utilizaron en esta segunda etapa como factores de riesgo de corrupción se basaron en el marco conceptual establecido por Zuleta et al. (2018). Los factores considerados son: si el contrato tuvo un proceso competitivo, la concentración de la contratación en la entidad - contratista, si se trata de una modalidad de contratación excepcional, si el contrato tuvo adiciones o modificaciones, y si el proceso de contratación tuvo alguna garantía rechazada durante el proceso contractual. Adicionalmente se integra una medida externa de riesgo en la contratación de la entidad, que es un indicador desarrollado por Instituto Nacional Anticorrupción (INAC) que evalúa los procesos en la contratación de las entidades.

Estas variables por sí mismas no representan un riesgo de corrupción porque es natural que en algunos casos existan modalidades excepcionales, pocos oferentes o adiciones. Sin embargo, cuando más de uno de estos componentes se encuentran de manera simultánea, indican que el contrato presenta características que lo hacen más vulnerable a un acto de corrupción, aunque no implica que efectivamente exista uno. Por lo tanto, los contratos serán más riesgosos cuando la mayor cantidad de componentes estén activos. En el caso extremo, el contrato con mayor riesgo posible sería uno en el cual se dio una modalidad de contratación excepcional, con un solo oferente, con una alta concentración de la contratación, con modificaciones o adiciones, con garantías rechazadas y por una entidad con altos niveles de riesgo contractual según el INAC.

Debido a que se trata de una metodología de *clustering* en dos etapas, los grupos que se conformen en la primera etapa debían cumplir con un requisito de observaciones mínimas. Esto se explica porque en la segunda etapa el algoritmo *kmeans* requiere un mínimo de 10

observaciones por variable para conformar los subgrupos de riesgo. Por esta razón, al tratarse de seis variables, se requieren al menos 60 observaciones por cada grupo conformado en la primera etapa. Para lograr esto, se implementó una aproximación iterativa en la conformación de los grupos de la primera etapa. En la primera iteración, el total de observaciones se dividió en un conjunto de grupos k, tal que cada grupo tuviera aproximadamente 60 observaciones.

Esto resultó en un subconjunto de grupos que cumplían el requisito y un subconjunto de grupos que no cumplían el requisito. La segunda iteración se hace sobre el subconjunto de grupos que no cumplen el requisito de observaciones mínimas, de manera que la nueva iteración genere un nuevo conjunto de subgrupos donde se cumple o no la condición. Esto se hace de manera sucesiva hasta que se llega a que todos los grupos cumplen la condición o no hay suficientes observaciones para continuar. El enfoque presentado garantiza que exista un mínimo de observaciones en cada grupo, lo cual permite implementar el algoritmo en la segunda etapa. Para esta primera etapa también se evaluó el uso de una metodología de *clustering* jerárquico con criterio Ward, en reemplazo de la aproximación no jerarquizada *kmeans* que se planteó, aunque los resultados no indicaron una mejora en el uso de este enfoque.

La última fase es la evaluación de los algoritmos. Esta fase incluye realizar una evaluación de los resultados de la estimación en dos frentes: 1) la validación de los datos de origen, y 2) la evaluación de las alertas identificadas. La primera parte implica evaluar la calidad de los datos que alimentan el algoritmo de aprendizaje de máquinas, es decir, revisar que efectivamente los valores que se expresan para cada variable en la base de datos sean efectivamente los datos reales del contrato. Esto se hace revisando la documentación respectiva del subconjunto de contratos que reporte las alertas más altas. Si efectivamente se encuentra que la información de las bases de datos corresponde a la información documental del contrato, entonces se puede confiar en que las alertas realmente reflejan contratos atípicos en las variables seleccionadas. Se utilizaron dos medidas de evaluación, la proporción de contratos con inconsistencias en la información y el promedio de inconsistencias en el subconjunto de contratos con inconsistencias.

La segunda evaluación integra una revisión de la calidad del agrupamiento. Para esto se hace uso del coeficiente de Silhouette, que es una medida estándar desarrollada por

Rousseeuw (1987) que identifica qué tan bien están agrupados los objetos dentro de un clúster en comparación con otros clústeres. El coeficiente se calcula para cada observación en la muestra de la siguiente manera:

$$s_i = \frac{(b_i - a_i)}{(\max(a_i, b_i))}$$

Donde 'a' es igual al promedio de la distancia entre una observación y las demás observaciones en el mismo clúster al que pertenece y 'b' representa la menor distancia entre la observación y alguna observación de otro clúster. En general un agrupamiento se considera mejor cuando 'a' tiende a cero y 'b' tiene a infinito. Es decir, cuando hay poca distancia entre las observaciones del mismo clúster y una gran distancia entre las observaciones de diferente clúster. Por esta razón, cuando 'a' tiende a cero y se cumple que 'b' > 'a' el valor del coeficiente se aproxima a 1, indicando un agrupamiento perfecto. Por el contrario, cuando 'b' tiende a cero y se cumple que 'b' < 'a', entonces el coeficiente tiende a -1, un agrupamiento completamente imperfecto. Cuando 'b' y 'a' son iguales, el coeficiente tiende a cero, lo que indica que el agrupamiento es arbitrario, es decir que la observación pudo asignarse a uno u otro clúster.

La medida utilizada como criterio de evaluación es el promedio de los coeficientes de *silhouette* en el clúster. Los resultados en el indicador de riesgo de corrupción se apoyan en los coeficientes de *silhouette* de la primera etapa, así como los coeficientes de *silhouette* en la segunda etapa para cada observación, de manera que esta medida se utilice como criterio de confiabilidad en los resultados.

3.3 Datos - Sistema Electrónico de Contratación Pública de Colombia (SECOP)

La investigación utilizó las bases de datos del Sistema Electrónico para la Contratación Pública de Colombia en su segunda generación (SECOP II). Este sistema funciona como una plataforma transaccional en la que se desarrolla de manera electrónica todo el proceso de compra pública. Los datos resultantes de cada transacción se almacenan en diferentes bases de datos que se encuentran publicadas para la consulta de cualquier persona en el portal

datos.gov.co. Las bases de datos del SECOP que se utilizaron para efectos de la investigación son las siguientes:

•**SECOP II Procesos de Contratación:** compila 2,24 millones de datos sobre procesos de contratación.

•**SECOP II Contratos electrónicos:** compila 2,02 millones de datos sobre contratos electrónicos.

•**SECOP II Adiciones:** compila 2,33 millones de datos sobre adiciones contractuales.

•**SECOP II Garantías:** compila 5,01 millones de datos sobre garantías asociadas a contratos.

•**SECOP II Proponentes por procesos:** compila 1,38 millones de datos sobre proponentes por procesos contractuales.

•**SECOP II Proveedores registrados:** 1,03 millones de datos sobre proveedores.

Los conjuntos de datos se descargaron en noviembre del 2022, siendo este el último mes del cual se tienen registros. La base de datos principal es la de contratos electrónicos que contiene en cada observación un contrato único y un conjunto de variables que lo describen. A la base de contratos se integra el resto de información utilizando los identificadores únicos de contrato, de proceso de contratación y de proveedores.

Una vez se hace el pegue de las bases de datos, el trabajo se centra en filtrar la información para llegar al subconjunto de contratos relevantes para la investigación, que son los contratos ejecutorios del PAE desarrollados por una entidad del orden departamental. Esta muestra se seleccionó debido a que los departamentos, a diferencia de los municipios, tienen la obligación de realizar todo el proceso de contratación del PAE a través del SECOP, por lo que se asegura que se disponga del universo de los contratos. Esta muestra objetivo también incorpora los convenios y contratos interadministrativos entre entidades públicas, que para el caso concreto es cuando la administración departamental celebra un negocio jurídico con un municipio para la ejecución del contrato de manera descentralizada.

Para esto se aplicaron seis filtros para depurar la base de datos y seleccionar únicamente los contratos de interés para la investigación. Estos seis filtros se aplicaron de

manera consecutiva, siendo los siguientes: 1) filtro por entidad contratante, dejando únicamente aquellos relacionados a la administración territorial; 2) filtro por tema, dejando aquellos relacionados al PAE; 3) filtro por tipo de contrato, depurando aquellos relacionados a consultoría o supervisión; 4) filtro por valor del contrato, eliminando aquellos con valores minúsculos, 5) filtro por duplicados, eliminando aquellos que son registros duplicados; 6) filtro por código de categoría, dejando únicamente los contratos que prestan servicios relacionados al PAE; 7) filtro por estado del contrato, dejando únicamente aquellos que terminaron activos.

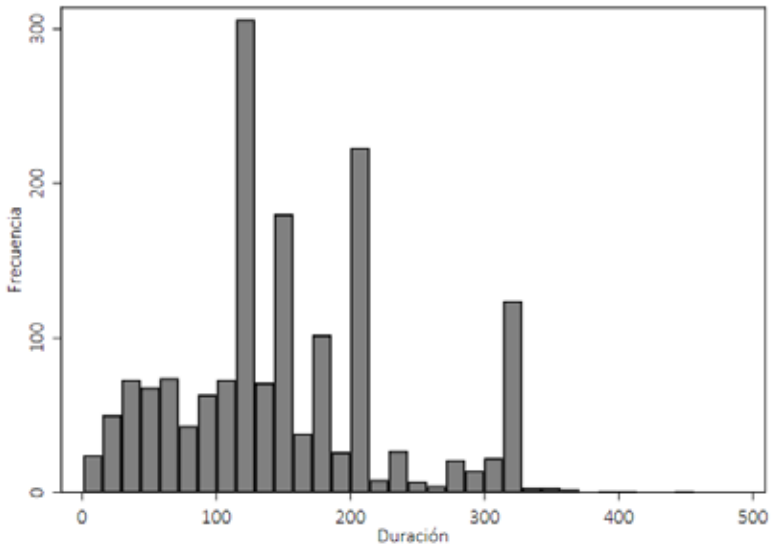
En cada uno de los filtros se hicieron pruebas de robustez del proceso, que consistían en varias revisiones de las observaciones eliminadas para identificar que no se descartaran observaciones que pudieran ser relevantes para la investigación. Esta evaluación permitía retroalimentar el proceso de filtrado, de manera que se refinara el método hasta obtener la muestra deseada. Para ahondar en los detalles de este proceso por favor remitirse al anexo 1.

Una vez filtrada completamente la base de datos del SECOP, se dispone de 2393 observaciones que contienen todos los contratos ejecutorios del PAE de entidades territoriales en el territorio nacional. El reto consistió en crear las variables que serían utilizadas en la primera y segunda etapa de acuerdo con la metodología previamente descrita.

Para el caso de la primera etapa, se utilizaron dos variables *duració*n y *valor_contrato.* Para la variable duración, se calculó la cantidad de días que hay entre la fecha de fin del contrato y la fecha de inicio de contrato. Si la fecha de inicio de contrato no se encontraba, se consideró la fecha de la firma del contrato. Si la diferencia era negativa, es decir que la fecha de firma del contrato es mayor a la fecha final del contrato, entonces la observación se eliminaba (15 observaciones eliminadas), también se eliminaban si la duración es igual a cero días (16 observaciones). De estos datos, la investigación solo consideró los contratos del orden departamental, por lo tanto, se depuraron los contratos municipales, quedando con una muestra final de 1652 observaciones. La media de la duración del contrato en esta base final es de 151,7 días, con una desviación estándar de 80,42 días. La figura 1 expuesta adelante muestra la distribución de esta variable.

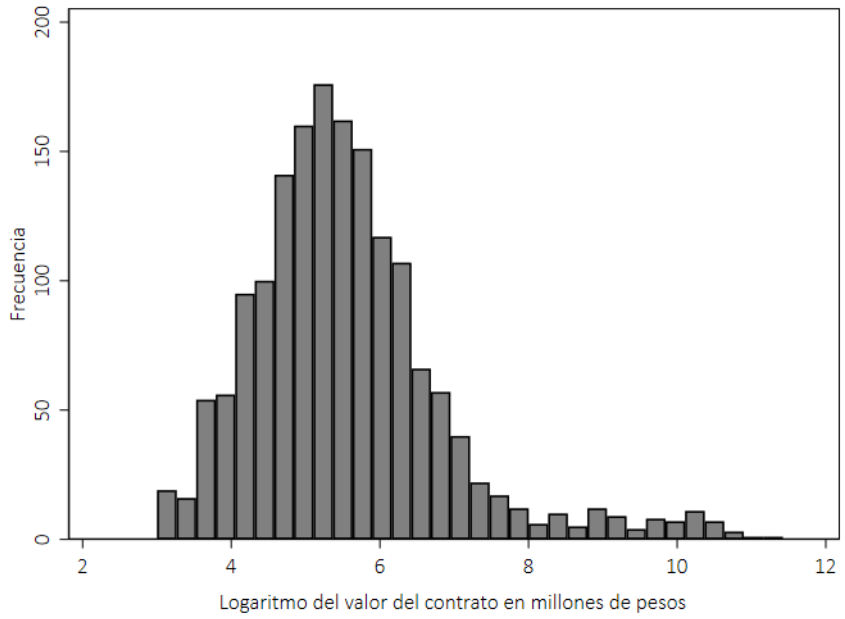
Para la variable de valor del contrato, se divide el valor reportado en 1 000 000, de manera que se puedan tener los montos de los contratos en millones de pesos colombianos. Sin embargo, existe el problema de que los datos presentan valores extremos que distan de manera considerable de la media y que pueden afectar los algoritmos de aprendizaje de máquinas. En efecto, el valor máximo es de 91 384 millones de pesos, mientras que la mediana es de 218 millones. Para solucionar esto, se hace una transformación monótona creciente en la variable de tipo logaritmo natural, que permite reducir la escala de los datos y eliminar los valores extremos. La figura 2 expuesta a continuación muestra la distribución de los datos de la variable *valor_contrato* tras la transformación. No hay valores faltantes en estas variables.

Figura 1. Duración de los contratos PAE en días



Fuente: elaboración propia.

Figura 2. Valor del contrato en escala logarítmica.



***Fuente*:** elaboración propia.

Finalmente, para que las variables puedan utilizarse en los algoritmos de aprendizaje no supervisado deben estar preferentemente estandarizadas en escala 0 a 1. Para cumplir con esto, se optó por una estandarización de tipo Min-Max, en la cual a cada valor se le resta el mínimo y se divide sobre la diferencia entre el valor máximo y el mínimo, como se aprecia en la siguiente ecuación:

$$Var_{norm} = \frac{Var - \min(Var)}{\max(Var) - \min(Var)}$$

Para la segunda etapa de la investigación se buscó identificar una variable por cada uno de los factores de riesgo de corrupción que consideraban Zuleta et al. (2018) y que se presentaron en el apartado de Marco conceptual del documento.

COMPONENTE 1. Persistencia de modalidades de contratación no competitivas:

dummy_competitivo: variable dicótoma que indica [1] si el contrato se celebró con al menos dos proponentes por proceso, [0] si se celebró con un único proponente.

Esta variable se construye utilizando los datos desagregados de la base de datos SECOP II Proponentes por proceso, que integra el número de proponentes para cada proceso de contratación.

COMPONENTE 2. concentración de contratistas:

concentracion: medido como el porcentaje del valor total de la contratación de la entidad que tiene asignado el proveedor desde que se tienen registros (valor entre 0 y 1).

Los cálculos se hacen utilizando la base de datos de SECOP II Contratos Electrónicos. Se estima para cada combinación entidad contratista el porcentaje del valor de la contratación total de la entidad desde que tiene registros en SECOP II.

COMPONENTE 3. Modalidades de contratación excepcionales:

modalidad_excepcional: Dummy que indica [1] si el contrato fue suscrito bajo una modalidad de contratación de régimen especial, [0] de lo contrario.

Para su construcción se consideró que el contrato tenía una modalidad excepcional si en la variable modalidaddecontratacion del SECOP tenía clasificación "Contratación régimen especial" o "Contratación régimen especial (con ofertas)".

COMPONENTE 4. Adiciones / Modificaciones:

dummy_adicion_modificacion: Variable dicótoma que indica [1] si el contrato tuvo alguna adición o modificación, [0] de lo contrario.

Esta información se extrae de la base de datos SECOP II Adiciones.

Adicionalmente se consideran dos componentes dada la disponibilidad de información adicional que puede contribuir a mejorar la precisión de los algoritmos de aprendizaje automático. Estos serían:

COMPONENTE 5. Gestión contractual:

INAC_contratacion: Indicador del Índice Nacional de Corrupción de la Secretaría de Transparencia que evalúa la gestión en la contratación de la entidad en términos de publicidad de la información y transparencia (Secretaría de Transparencia, 2021). El indicador se encuentra en una escala de 0 a 100, con una media de 68,46 y una desviación estándar de 27,16. Para efectos metodológicos, se estandarizó tipo Min-Max para asegurar los valores en el rango 0 – 1.

COMPONENTE 6. Garantías rechazadas:

garantías_rechazadas: Dummy que indica [1] si el contrato tuvo alguna garantía o modificación de garantía rechazada en el proceso de contratación, [0] de lo contrario.

4. Resultados y sesgos de datos identificados

4.1 Resultados de evaluación de los algoritmos

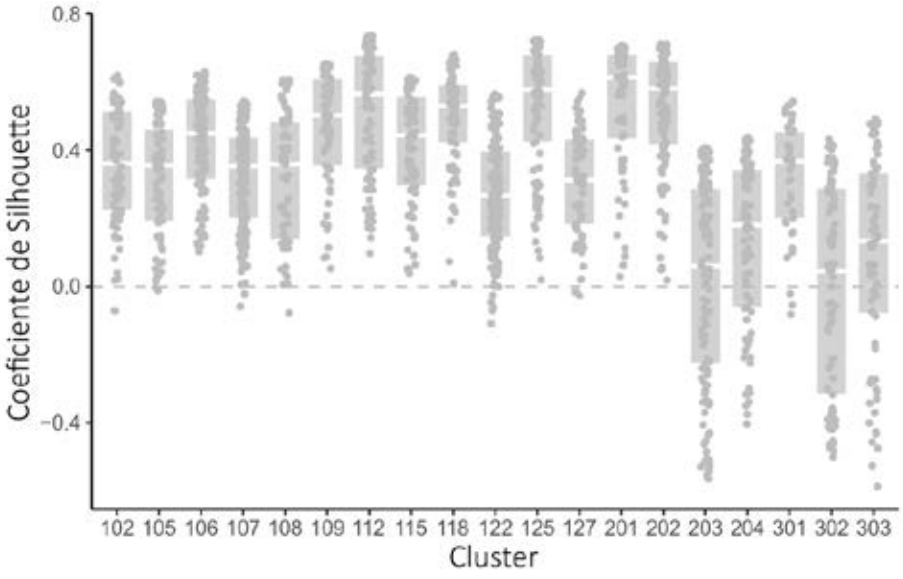
Para la primera etapa se definió un número de grupos igual a 27, que para una muestra de 1652 observaciones correspondía aproximadamente a 60 observaciones por grupo. Esto resultó en 1074 observaciones asignadas en 12 clúster que cumplían el requisito mínimo de observaciones y 578 observaciones asignadas a 15 clúster que no cumplían el requisito. La segunda iteración se hizo sobre este subconjunto de observaciones, teniendo 382 observaciones asignadas a cuatro clústeres que cumplen la condición y 196 observaciones restantes asignadas a cinco clústeres que no cumplen la condición. La última iteración resultó en 156 observaciones asignadas a dos clústeres que cumplían la condición y 40 observaciones que quedan en un clúster donde no se cumple la condición. Por la cantidad de observaciones no se pueden hacer más iteraciones.

El promedio del Coeficiente de Silhouette en el caso del clustering *kmeans* es de 0,38 para las observaciones clasificadas en la primera iteración, 0,27 para las observaciones clasificadas en la segunda iteración, y 0,09 para las observaciones clasificadas en la tercera iteración, para un resultado global de 0,32. Sin embargo, es necesario revisar a detalle cada uno de los clúster en cada iteración para determinar cuáles presentan resultados satisfactorios

y cuáles no. La figura 3 permite identificar que los clúster 203, 204, 302 y 303 (El número inicial representa la iteración y los siguientes dos números el número de clúster), si bien tienen un promedio positivo, son los que presentan un resultado de agrupación menos confiable, mientras que el resto de clúster tienen resultados de agrupación satisfactorios. Esta medida sirve como un indicador de confiabilidad en la estimación de riesgo de corrupción, puesto que si se trata de contratos más diferentes entre sí, es más probable que las diferencias en los factores de riesgo se deban a las condiciones naturales de los contratos y no a un acto de corrupción.

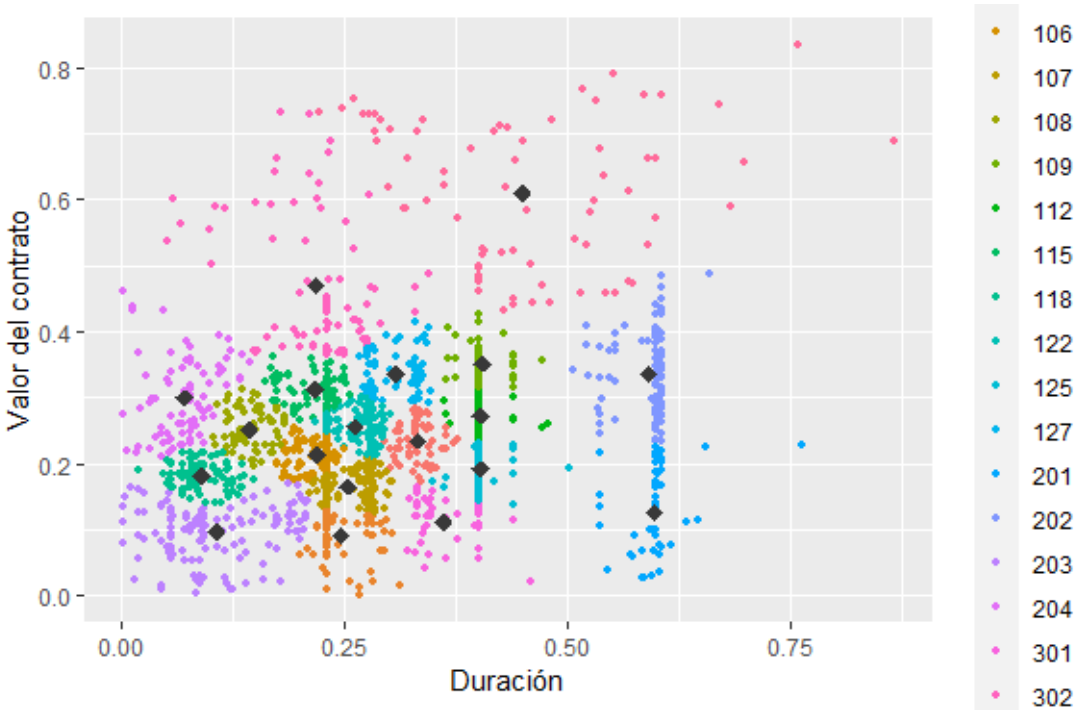
También se puede identificar que a medida que aumenta el número de iteraciones el coeficiente presenta resultados menos satisfactorios. Esto tiene sentido dado que el método que se está empleando fuerza a algunas observaciones a estar en un mismo grupo para satisfacer la condición de observaciones mínimas, aun cuando estas puedan ser diferentes entre sí. A medida que aumente el número de observaciones, y asumiendo que no existan muchos valores extremos en la muestra, este problema debería reducirse.

Figura 3. Coeficientes de *silhouette clustering* no jerarquizado (kmeans) según clúster.



Fuente: elaboración propia.

Figura 4. Resultados clústeres primera etapa



Fuente: elaboración propia.

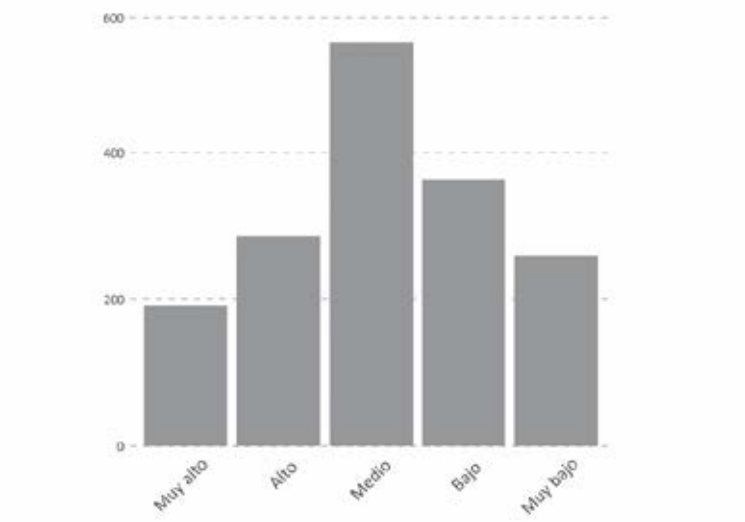
Para la segunda etapa, en cada uno de los clústeres de la primera etapa se implementó un algoritmo kmeans con cinco grupos definidos k=5. En este caso se seleccionan cinco grupos correspondientes a los diferentes niveles de riesgo de corrupción. Una vez implementado el agrupamiento, en cada subgrupo de datos de la segunda etapa se calcula el promedio de las variables de agrupamiento. Debido a que las variables van en una misma dirección y tienen una magnitud idéntica, entre más alta sea la suma de los promedios de las variables en cada subconjunto de observaciones se tiene un riesgo de corrupción más alto. A partir de este resultado, se ordenan de manera descendente los grupos, asignando la categoría 'Muy alto' al clúster de la segunda etapa con el valor de la suma más alto, 'Alto' al clúster con el valor siguiente, y así sucesivamente. En este caso no hay una restricción de observaciones mínimas en el agrupamiento, por lo cual el tamaño de los clústeres depende completamente de la distribución de las variables.

Los resultados de esta segunda etapa se muestran en las figuras 5 y 6. Del total de contratos ejecutorios del PAE, 11,6% presentan un riesgo de corrupción 'Muy Alto' y 17,31% un riesgo de corrupción 'Alto'. Esto no significa que estos contratos efectivamente tengan una

irregularidad relacionada con la corrupción, sino que presentan características que son compatibles con un mayor riesgo de corrupción en los factores considerados. Estos 192 contratos clasificados como de corrupción 'Muy Alta' serían los que los organismos de control deberían priorizar para asegurar que no existan ningún tipo de irregularidades.

El promedio de las variables en cada categoría, expuesto en la figura 6, indica también cuáles son los factores que están incidiendo más en la clasificación. En este caso se puede ver que los contratos con modalidad excepcional, el indicador INAC de contratación y las garantías rechazadas son los factores que impulsan más la clasificación en los niveles más altos de riesgo.

Figura 5. Número de observaciones por categoría de corrupción



***Fuente*:** elaboración propia.

Figura 6. Promedio de las variables de riesgos de corrupción por categoría

Clasificación de riesgo	Número de obs.	Riesgo	Dummy proponentes	Concen-tracion	Modalidad excepcional	Dummy adiciones	INAC Contratación	Garantías rechazadas
Muy alto	192	3,23	0,95	0,0038	0,74	0,56	0,78	0,19
Alto	286	2,28	1,00	0,0052	0,04	0,97	0,22	0,04
Medio	551	2,13	0,98	0,0019	0,12	0,50	0,52	0,02
Bajo	363	1,71	0,95	0,0029	0,00	0,11	0,63	0,01
Muy bajo	259	1,07	0,85	0,0034	0,00	0,00	0,19	0,03

***Fuente*:** elaboración propia.

La evaluación de esta segunda etapa también hizo uso de los Coeficientes de Silhouette para determinar la calidad del agrupamiento. La figura 7 expuesta a continuación muestra el promedio del Coeficiente de Silhouette en la segunda etapa por cada uno de los clústeres de la primera etapa. Se puede identificar que la calidad del agrupamiento es sobresaliente, teniendo apenas un valor mínimo de 0,67. Esto indica que la clasificación de los datos en la segunda etapa es consistente y puede utilizarse de manera confiable. Aunque el nivel de riesgo per sé está mediado por los resultados de agrupamiento en la primera etapa.

Figura 7. Coeficiente de Silhouette promedio en la segunda etapa por cada clúster de la primera etapa

Clúster de la primera etapa	**Coeficiente de Silhouette segunda etapa**
102	0,906
105	0,959
106	0,957
107	0,958
108	0,913
109	0,936
112	0,736

115	0,946
118	0,942
122	0,927
125	0,973
127	0,869
201	0,960
202	0,905
203	0,938
204	0,879
301	0,878
302	0,674

***Fuente*:** elaboración propia.

4.2 Identificación de sesgos en datos

Los resultados indican que los algoritmos de aprendizaje no supervisado pueden ser una herramienta poderosa para detectar riesgo de corrupción en la contratación pública del país en el futuro. Los algoritmos permiten seleccionar el conjunto de contratos que presenta anomalías en factores relacionados a la corrupción dadas las características que tienen contratos similares y evitando utilizar variables objetivo que pueden ser difíciles de conseguir dada la naturaleza del fenómeno. Esto se debe a la novedad de utilizar un enfoque no supervisado y una metodología en dos etapas que permite comparar los contratos con características similares en las variables seleccionadas. No obstante, una revisión a profundidad del proceso de investigación lleva a concluir que las alertas indicadas tienen un alcance limitado dadas las dificultades en materia de calidad y disponibilidad de la información.

En concreto, el proceso de investigación evidenció sesgos algorítmicos que son resultado de la calidad insuficiente de los datos y que ponen estrictas limitaciones de su uso en la práctica. Este documento busca mostrar cuáles eran esos sesgos en el caso estudiado, cómo pudieron ser identificados, y qué alternativas existen para superarlos, de manera que otras investigaciones similares puedan aprovechar esta experiencia en sus propios desarrollos.

A continuación, se exponen los diferentes sesgos identificados y su relación con el marco establecido por Mehrabi et al., (2021):

a. Sesgo de representación por muestra parcial de los contratos:

El primer sesgo identificado es el de representación. El algoritmo de aprendizaje no supervisado está diseñado para identificar los contratos con mayor riesgo de corrupción en las 32 gobernaciones de Colombia, una por cada departamento, identificando el subconjunto que presenta anomalías en los datos que lo hacen más susceptible a la corrupción dado el conjunto de contratos similares. Sin embargo, para que esto sea efectivo, las 32 gobernaciones de Colombia deben publicar todos los contratos relacionados al PAE a través de la plataforma SECOP, es decir que se cuente con el universo de contratos.

El marco legal del país establece que todas las gobernaciones están obligadas a publicar sus contratos a través de SECOP II, sin embargo, la evidencia indica que algunas gobernaciones no están publicando sus contratos. Esto no sería un problema si esa pérdida de información fuera aleatoria entre el conjunto de contratos y gobernaciones, es decir que no existiera un patrón sistemático en la ausencia de información, pero el proceso identificó que en efecto existe.

En el conjunto de datos filtrado es posible identificar que la distribución de contratos por departamentos no es proporcional a la población, cantidad de municipios, capacidad administrativa, o cobertura educativa de los mismos, sino que está fuertemente concentrada en algunas regiones (Figura 8). En concreto, solo Antioquia, Cauca y Boyacá concentran el 84,3% de los contratos de la muestra, dejando al resto de gobernaciones con valores mínimos de contratos relacionados al PAE. Tampoco se correlaciona con regiones donde existan más municipios habilitados para contratar directamente el PAE, porque zonas rezagadas en esta materia como los departamentos de la región de la Amazonía o Orinoquía tampoco están presentes en los datos.

Figura 8. Distribución de contratos por departamento

Departamento	Frecuencia	Porcentaje
Antioquia	671	40,62
Arauca	7	0,42
Atlántico	2	0,12
Bolívar	3	0,18

Boyacá	609	36,86
Caldas	2	0,12
Caquetá	4	0,24
Casanare	19	1,15
Cauca	114	6,9
Cesar	40	2,42
Chocó	85	5,15
Cundinamarca	7	0,42
Córdoba	7	0,42
Huila	5	0,3
La Guajira	17	1,03
Magdalena	2	0,12
Meta	2	0,12
Nariño	14	0,85
Norte de Santander	9	0,54
Putumayo	6	0,36
Quindío	3	0,18
Risaralda	5	0,3
San Andrés, Providencia y Santa Catalina	2	0,12
Santander	6	0,36
Sucre	2	0,12
Tolima	5	0,3
Valle del Cauca	4	0,24
Total	1652	100

***Fuente**:* elaboración propia.

Esto genera que las alertas de riesgo de corrupción se concentren en estos tres departamentos, no porque sean las regiones con mayores niveles de corrupción en el país, sino porque son las más transparentes. En ese sentido, utilizar los algoritmos como mecanismo de alerta temprana sin considerar esto, podría llevar a conclusiones erróneas, como que estos departamentos son los de mayor riesgo, cuando probablemente estos departamentos sean los que tienen un mayor compromiso con la transparencia y publicidad en la compra pública.

Ni siquiera es posible establecer que para esos tres departamentos las alertas son válidas, porque no se puede determinar que el total de la contratación efectivamente esté

reflejado en las bases de SECOP. La evidencia muestra que la mayoría de las gobernaciones están ignorando llevar a cabo los procesos de contratación en la plataforma sin consecuencias legales, por lo que no es posible determinar que el total de los contratos de aquellas que sí tienen la información publicada esté reflejado en los datos. Esto viola el supuesto número 1 de la investigación, "las entidades que están legalmente obligadas a publicar sus contratos en SECOP II efectivamente lo hacen en la totalidad de su contratación, lo que implica que los datos en SECOP II contienen el universo de contratos elegibles".

El problema de que este supuesto sea violado es que los funcionarios podrían deliberadamente publicar solo aquellos contratos que no tienen corrupción, o llegar al punto de no publicar ningún tipo de información de la contratación. Estos resultados fueron también sustentados con la revisión a detalle del proceso de filtrado de las bases de datos, cuyo seguimiento en cada filtro indica que no existe ningún patrón que haya podido descartar observaciones relevantes para el estudio (ver anexo 1).

Matemáticamente, el sesgo puede representarse de la siguiente manera:

Universo de Contratos (U): supongamos que el universo total de contratos relacionados al PAE en Colombia es U, con $|U|$ siendo el número total de contratos.

Muestra de Contratos (M): los contratos que efectivamente fueron publicados en la plataforma SECOP II son una muestra $M \subseteq U$, con $|M|$ siendo el número total de contratos en la muestra.

Departamentos (D): los contratos están distribuidos en 32 departamentos $D_1, D_2, \dots, D_{32}$.

Frecuencia Observada (m): la cantidad de contratos observados para el departamento D_i en la muestra M es m_i.

En teoría, cada departamento tiene características que pueden influir en la cantidad de contratos que deberían observarse w_i, como:

p_i : proporción de la población del departamento D_i respecto al total nacional.

r_i: número de municipios en el departamento D_i.

a_i: capacidad administrativa del departamento D_i, medida como un índice o un proxy (por ejemplo, el presupuesto disponible).

e_i: cobertura educativa en el departamento D_i, también medida como un índice.

s_i : proporción de municipios habilitados para contratar PAE directamente en el departamento D_i.

Entonces para cada departamento D_i la cantidad de contratos esperados w_i puede reflejarse de la siguiente forma:

$$w_i = f(p_i, r_i, a_i, e_i, s_i)$$

Donde f es una función calibrada que combina estos factores y resulta en el valor real de contratos que deberían existir para cada departamento D_i dadas sus características. Con lo anterior, la distribución esperada de contratos para cada departamento D_i en el universo U, dado que todos los contratos fueran reportados correctamente y sin sesgo, sería:

$$P(\, D_i \mid U\,) = \frac{w_i}{\sum_{j=1}^{32} w_j}$$

Sin embargo, en la muestra M, la probabilidad observada de que un contrato provenga de D_i es:

$$P(D_i \mid M) = \frac{m_i}{\mid M \mid}$$

Donde m_i es la frecuencia observada de contratos en D_i y $\mid M \mid$ es el número total de contratos en la muestra.

Ahora bien, el sesgo de representación se introduce porque $\frac{w_i}{\sum_{j=1}^{32} w_j}$, es decir la proporción de contratos esperada para el departamento D_i en el universo $\mid U \mid$ dadas sus características no es igual a $\frac{m_i}{|M|}$, la proporción de contratos observada para el departamento D_i en la muestra $|M|$. Para evidenciar el sesgo es necesario comparar la distribución observada con la distribución esperada:

$$Sesgo(D_i) = \frac{P(D_i \mid M)}{P(D_i \mid U)} = \frac{\frac{m_i}{\mid M \mid}}{\frac{w_i}{\sum_{j=1}^{32} w_j}} = \frac{\sum_{j=1}^{32} w_j * m_i}{w_i * \mid M \mid}$$

Si $Sesgo(D_i) > 1$, el departamento D_i está sobre-representado en la muestra, lo que podría llevar a un análisis incorrecto al suponer que tiene un mayor riesgo de corrupción. Esto ocurre cuando el número relativo de contratos observados en la muestra para D_i es mayor que el peso ajustado por los factores reales para ese departamento. Es decir:

$$\frac{m_i}{\mid M \mid} > \frac{w_i}{\sum_{j=1}^{32} w_j}$$

En contraposición, si $Sesgo(D_i) < 1$, el departamento D_i está sub-representado, lo que podría ocultar corrupción que realmente existe. Esto ocurre cuando el número relativo de contratos observados en la muestra para D_i es menor que el peso ajustado por los factores reales para ese departamento. Es decir:

$$\frac{m_i}{\mid M \mid} < \frac{w_i}{\sum_{j=1}^{32} w_j}$$

Con la ecuación, la estimación del sesgo puede ser estimada para cada departamento si se dispone de información sobre los factores que componen w_i. Es decir, calculando cuál debería ser la proporción de contratos real dadas las características de los departamentos. Esto en la práctica puede ser muy difícil de lograr, aunque es posible tener algunas estimaciones considerando la proporción de la contratación en ámbitos de política similares al PAE.

En el caso concreto, el sesgo pudo identificarse fácilmente porque la selección de algunos departamentos era muy evidente considerando la distribución natural de la población. Sin embargo, en otros casos puede ser más difícil de detectar, lo que demanda una revisión minuciosa de la muestra en futuros desarrollos. La revisión de expertos y comparar la distribución de las observaciones en otras muestras equivalentes también puede ser conveniente para asegurar la representatividad de la información.

b. Sesgo de medición por baja calidad en los datos del SECOP

El segundo sesgo identificado en el caso es el de medición. Este sesgo se da cuando los datos no logran representar de manera verosímil el constructo que dicen representar. Esto implica que existe una discrepancia entre los valores verdaderos de los datos y los observados en la muestra, la cual puede deberse a valores faltantes o imprecisiones en la recolección. En general, los errores de medición están presentes en una gran cantidad de aplicaciones de Inteligencia Artificial, lo que no siempre supone un problema para los propósitos de los algoritmos. El error de medición se convierte en un sesgo algorítmico cuando introduce una distorsión sistemática en los resultados del algoritmo, afectando de manera desproporcionada a ciertos grupos o situaciones. Esto ocurre cuando el error no es aleatorio, sino que sigue un patrón específico que impacta la capacidad del algoritmo para hacer predicciones precisas y justas.

En las bases de datos del SECOP es común que algunos datos no reflejen realmente las características de los contratos. Estos errores no son aleatorios, sino que se concentran en algunas variables. El error de medición afecta la capacidad de los algoritmos de aprendizaje no supervisado de detectar los patrones relacionados con la corrupción en la muestra porque las anomalías en los datos pueden ser el resultado de un error de los datos en vez de ser un indicador de corrupción. Tampoco es posible establecer si los datos faltantes o con errores fueron resultado de errores no premeditados de los funcionarios o fueron intencionalmente modificados para ocultar algún indicio de corrupción.

El estudio era consciente de este problema y planteó una validación de los datos de origen en el subconjunto de observaciones en la categoría de riesgo 'Muy Alta', que sería la muestra que se utilizaría para las alertas tempranas. Para esto, se tomó una muestra aleatoria correspondiente al 15% de las observaciones, donde se revisó para cada observación las variables utilizadas en la primera y segunda etapa, a excepción de las variables del Indicador INAC que provenía de información externa. Los resultados indican que, del total de observaciones analizadas, 58,62% presentaba alguna inconsistencia entre la información de la base de datos y la información documental del contrato, o no había podido verificarse dada la ausencia completa de información documental. En la mayoría de los casos con inconsistencias,

estas se encontraban en los reportes de adiciones y modificaciones al contrato, valor del contrato, y duración del contrato.

En algunos casos, la información del contrato era diferente si se revisaba la página del proceso contractual, la página del contrato electrónico, o el mismo documento, lo cual deja en evidencia las dificultades de la plataforma para garantizar la consistencia de la información. Además, en el conjunto de contratos que presentaba alguna inconsistencia en sus datos, el promedio de inconsistencias fue de 1,23, por lo que era frecuente la presencia de más de un error por contrato. Los datos que se utilizaron en esta validación pueden encontrarse en el anexo 2 del documento.

Otras variables que tienen este problema en la base de datos pero que no se utilizaron directamente en el análisis de datos son "*orden*", "*rama*", "*estadocontrato*", "*tipodecontrato*", "*codigodecategoriaprincipal*", "*estadodecontrato″* y las variables asociadas a los valores pagados, ejecutados o amortizados.

Matemáticamente, el sesgo puede representarse de la siguiente manera:

Universo de Contratos (U): supongamos que U representa al conjunto de contratos elegibles sin errores de medición, con $|\ U\ |$ siendo el número total de contratos.

Muestra de Contratos (M):): supongamos que M representa al conjunto de contratos en la muestra, con $|\ M\ |$ siendo el número total de contratos.

Ahora bien, cada contrato está representado por un vector de características $X = \{x_1, x_2, \dots, x_n\}$ donde cada x_i es una variable relevante para el algoritmo (por ejemplo, valor del contrato, duración, etc.).

Variables Verdaderas (X^*): X^* representa los valores reales de las características de los contratos en U.

Variables Observadas (X^{obs}): X^{obs} representa los valores observados de las características de los contratos en M.

Suponemos que el error de medición es sistemático, lo que significa que introduce un sesgo en la distribución de X^{obs}, entonces se tiene que: $X^* \neq X^{obs}$. Por lo que el error de medición para una característica específica x_i se define como:

$$e_i = x_i^{obs} - x_i^*$$

El algoritmo k-means agrupa los contratos en k clusters basados en la minimización de la varianza dentro de los clusters. El objetivo es encontrar los centroides $\mu_1, \mu_2, \ldots, \mu_k$, tal que minimizan la suma de distancias cuadradas de cada punto X_j^{obs} a su centroide más cercano:

$$min_{(\mu_1,\mu_2,\ldots,\mu_k)} \sum_{j=1}^{|M|} min_{(i=1,\ldots,k)} \left|\left|X_j^{obs} - \mu_i\right|\right|^2$$

Sin errores de medición, u otros sesgos, los contratos se agrupan en clusters correctos basados en X^*, y estos clusters reflejan patrones de riesgo de corrupción. Los centroides óptimos basados en X^* son ${\mu_1}^*, {\mu_2}^*, \ldots, {\mu_k}^*$. Sin embargo, cuando los datos observados X^{obs} contienen errores de medición, los centroides calculados ${\mu_1}^{obs}, {\mu_2}^{obs}, \ldots, {\mu_k}^{obs}$ pueden desviarse de los verdaderos centroides ${\mu_1}^*, {\mu_2}^*, \ldots, {\mu_k}^*$, teniendo que:

$${\mu_i}^* = {\mu_i}^{obs} + \Delta\mu_i$$

Donde $\Delta\mu_i$ es el desplazamiento del centroide debido al error de medición.

Esto puede causar que los contratos se asignen incorrectamente a los clusters, generando agrupaciones que no reflejan los verdaderos patrones de corrupción. Por ejemplo, contratos que no presentan riesgo de corrupción pueden ser agrupados en un cluster de alto riesgo debido a errores en las variables observadas. De manera análoga, contratos con riesgo de corrupción pueden no ser detectados porque las anomalías en sus características fueron suavizadas o distorsionadas por el error de medición.

La probabilidad de que un contrato c_j sea asignado incorrectamente a un cluster debido a estos errores de medición es proporcional a la probabilidad de error de medición p_{inc}, lo que se puede representar de la siguiente manera:

$$P\big(\, Asignación\ Incorrecta \mid c_j \,\big) = f(p_{inc})$$

En una forma simplificada, esta relación podría verse así:

$$P\big(\, Asignación\ Incorrecta \mid c_j \,\big) = \alpha * p_{inc}$$

Donde la constante α captura el grado de sensibilidad del algoritmo de segmentación a los errores de medición. Este valor dependería de la distribución de los datos. Por ejemplo, en un escenario donde los clústeres están bien separados y la variabilidad interna es baja, el impacto de los errores puede ser menor, lo que se traduciría en un α menor. Por el contrario, si los datos se agrupan de forma más difusa o los clústeres se solapan, la misma magnitud de error de medición puede tener un efecto mucho más pronunciado en la asignación de clústeres, lo que se reflejaría en un α más grande.

En el caso concreto, se identificó que el 58,62% de la muestra en el nivel más alto tenía errores de medición o no era posible verificarlo dada la ausencia de documentación para comprobarlo. Esto indica que la probabilidad de que existiera error de medición p_{inc} para cada observación es considerable, lo que lleva a que la asignación pueda ser incorrecta.

c. Sesgo de variable omitida por insuficiencia en la información

El sesgo de variable omitida se presenta cuando una característica fundamental no se encuentra disponible en los datos y no es posible integrarla en el entrenamiento de los algoritmos de aprendizaje de máquinas. En contextos de aprendizaje supervisado, este sesgo se manifiesta en las métricas de evaluación, donde se observan altos errores de predicción que evidencian la incapacidad del modelo para predecir con precisión la variable objetivo. Por otro lado, en algoritmos de aprendizaje no supervisado, la detección de este sesgo es más compleja, ya que no existe una variable explícita que permita una evaluación directa, lo que dificulta la identificación del impacto del sesgo en el rendimiento del modelo.

En el caso concreto, el algoritmo de *clustering* requiere que las variables incluidas representen adecuadamente todos los factores relevantes para la corrupción en la contratación pública. Si ciertas variables fundamentales que influyen en la corrupción no se incluyen en el modelo, los resultados del *clustering* podrían estar sesgados. Es decir que la clasificación de los contratos en los niveles de riesgo podría no corresponder con los riesgos reales de la

corrupción. Esto tendría un impacto mayor dependiendo de la relevancia real de las variables integradas en el análisis para los propósitos del estudio. Por ejemplo, si las variables integradas en el análisis no fueran en absoluto determinantes para detectar la corrupción, entonces la clasificación se haría de manera aleatoria en los niveles de riesgo y no habría diferencias significativas en términos de riesgo de corrupción entre los niveles. Debido a que el fenómeno de la corrupción no se comporta de manera aleatoria, entonces existiría otro conjunto de variables no observadas que sí sería determinante para la clasificación de los contratos en los niveles de riesgo.

La mejor forma de superar este sesgo en aplicaciones con aprendizaje no supervisado es asegurarse que la selección de las variables a utilizar esté fuertemente relacionada con la teoría, de forma que las variables integradas tengan una fuerte relación con el fenómeno a estudiar. En este caso, la selección de las variables fue resultado de los estudios desarrollados por Zuleta et al., (2018), que analizaban a profundidad los factores que inciden en la corrupción en la contratación de Colombia. Por lo anterior, si bien este sesgo podría afectar la aplicación del algoritmo, este impacto no sería tan significativo como los anteriores.

Matemáticamente, este sesgo puede representarse de la siguiente manera:

Supongamos que tenemos un conjunto de variables $X = \{x_1, x_2, \dots, x_n\}$ que se utilizan para realizar un clustering multivariado. Cada observación o_j está representada por un vector de características $x_j = \{x_{1_j}, x_{2_j}, \dots, x_{n_j}\}$. El objetivo del algoritmo es agrupar las observaciones en K clusters de manera que las observaciones dentro de un mismo cluster sean lo más similares posible entre sí, y lo más distintas posible de las observaciones en otros clusters.

Sin embargo, existe una variable fundamental x_{k^*} que no se incluye en X, ya sea porque no estaba disponible en los datos o porque no se consideró en la selección de variables. Esta variable omitida tiene un impacto crucial en el fenómeno de interés, es decir, en la corrupción.

La omisión de la variable x_{k^*} provoca que el algoritmo base la asignación de las observaciones a los *clusters* en una representación incompleta de la realidad. Esto puede generar una subestimación de la varianza dentro de los *clusters*, es decir que observaciones que deberían estar en *clusters* diferentes podrían ser agrupadas en el mismo *cluster* debido a

la ausencia de la variable x_{k^*}. También puede generar una sobreestimación de la varianza entre *clusters*, lo que puede llevar a una separación artificial de grupos que deberían estar juntos.

Considerando la distancia euclidiana como criterio de similitud, el objetivo del algoritmo es minimizar la suma de las distancias cuadradas dentro de los *clusters* (WSS):

$$WSS = \sum_{i=1}^{K} \sum_{x_j \in C_i} \| x_j - \mu_i \|^2 = \sum_{i=1}^{K} \sum_{x_j \in C_i} (\sum_{l=1}^{n} (x_{l_j} - \mu_{l_i})^2)$$

Donde:

K es el número de clusters y C_i es el i-ésimo cluster.

x_{l_j} representa el valor de la variable l para la observación j.

μ_{l_i} representa el valor correspondiente en el centroide μ_i para la variable l.

La suma se realiza sobre todas las variables l.

En presencia de una variable omitida x_{k^*}. La distancia entre una observación x_j y el centroide μ_i, sin considerar esta variable ($l \neq k^*$), se calcula como:

$$\| x_j - \mu_i \|^2_{omitiendo\ x_{k^*}} = \sum_{l=1; l \neq k^*}^{n} (x_{l_j} - \mu_{l_i})^2$$

La WSS calculada sin la variable x_{k^*} es una subestimación o sobreestimación de la verdadera WSS, dependiendo de cómo influya x_{k^*} en la separación entre los clusters. La subestimación podría ocurrir si x_{k^*} aporta una variación interna importante a cada cluster, es decir que ayuda a distinguir diferencias dentro del mismo grupo. En ese caso el algoritmo agrupa en un mismo cluster observaciones que deberían estar asignadas a clusters diferentes. Por el contrario, la sobrestimación podría darse si x_{k^*} ayuda a acercar observaciones que son similares realmente. En este caso, la omisión de la variable asigna las observaciones a clusters diferentes, cuando deberían estar asignadas al mismo cluster.

Entonces es posible establecer que:

$$\sum_{i=1}^{K} \sum_{x_j \in C_i} (\sum_{l=1}^{n} (x_{l_j} - \mu_{l_i})^2) \neq \sum_{i=1}^{K} \sum_{x_j \in C_i} (\sum_{l=1; l \neq k^*}^{n} (x_{l_j} - \mu_{l_i})^2)$$

La WSS estimada no incluye la influencia de la variable x_{k^*}, por lo que la distancia calculada, y por ende la dispersión calculada, es diferente de la WSS real. La fórmula para la WSS estimada excluye el término de distancia asociado con x_{k^*}, lo que lleva a una medida incompleta de cómo las observaciones se agrupan alrededor del centroide. Esto puede resultar en una asignación incorrecta de los *clusters*, lo que puede afectar significativamente el desempeño del algoritmo de *clustering*. En otras palabras, sin la información proporcionada por x_{k^*}, el modelo puede agrupar observaciones que son realmente distintas en el mismo *cluster*, o separar observaciones similares en *clusters* diferentes. Por ejemplo, si x_{k^*}, es una variable crítica que diferencia contratos corruptos de los no corruptos, su omisión podría llevar a que ambos tipos de contratos se mezclen en el mismo *cluster*. Esto no solo distorsiona la interpretación de los resultados, sino que también reduce la efectividad del algoritmo para identificar patrones significativos en los datos.

5. Discusión final

Este estudio buscaba identificar los sesgos de datos presentes en la aplicación de un algoritmo de aprendizaje no supervisado utilizado para detectar riesgo de corrupción en la contratación pública de Colombia. La investigación encontró que, pese al potencial de los algoritmos de aprendizaje no supervisado para desempeñar tareas de detección de riesgo de corrupción, la aplicación en la tarea asignada requiere que, previamente, se solucionen problemas en la calidad y disponibilidad de los datos. En concreto, se identificó que el algoritmo presenta sesgos de representación, de medición y de variable omitida que pueden poner en riesgo la confiabilidad de los resultados y que son producto de las deficiencias en la calidad de los datos del SECOP. El estudio aquí presente detalló estos sesgos y los representó matemáticamente para ejemplificar cómo afectan la precisión y la validez del algoritmo. A través de este análisis se demostró que estos sesgos pueden distorsionar la estructura de los *clusters*, conduciendo a resultados que no reflejan adecuadamente el riesgo de corrupción real en la contratación.

Considerar este tipo de sesgos en la aplicación de los algoritmos es fundamental para garantizar una aplicación ética y responsable de los algoritmos de aprendizaje de máquinas. Confiar en modelos de IA que se entrenan con datos sesgados puede llevar a decisiones injustas y perjudiciales, especialmente en áreas tan sensibles como la contratación pública. Por ejemplo, puede indicar riesgo de corrupción en contratos que realmente no representan el mayor riesgo de corrupción, sino que hacen parte de las entidades que sí cumplen con la publicidad de la contratación. Esto no solo generaría una visión distorsionada del riesgo de corrupción en el país, sino que puede generar una asignación errónea de los recursos de investigación y prevención de la corrupción en el país.

En el caso del SECOP, algunas recomendaciones deben adelantarse para hacer que la información disponible pueda aprovecharse al máximo. En primer lugar, deben establecerse mecanismos de supervisión que hagan seguimiento a la publicación de los contratos por parte de las entidades del orden territorial, de manera que se asegure que todos los contratos sean desarrollados por medio del SECOP tal como lo establece la ley. En segundo lugar, se deben implementar mecanismos de revisión y supervisión de los datos, de manera que se asegure que la información en las bases de datos coincida a la perfección con los datos reales de la contratación. Ambos mecanismos deben asegurar sanciones aplicables que hagan exigible estos compromisos por parte de las entidades territoriales.

Futuras investigaciones podrían enfocarse en analizar la selección de contratos en el SECOP, con el objetivo de determinar la brecha entre los contratos reales de las entidades territoriales y aquellos efectivamente publicados en la plataforma. Identificar los factores que influyen en la probabilidad de que un contrato sea publicado podría revelar causas como la corrupción, la falta de capacidad administrativa, o el desconocimiento por parte de los funcionarios. Además, sería valioso realizar un estudio exhaustivo sobre la calidad de los datos en el SECOP, estimando la magnitud de los errores detectados y los factores que los ocasionan. Estos estudios no solo mejorarían la comprensión del funcionamiento del SECOP, sino que también proporcionarían información crítica para desarrollar estrategias que optimicen la herramienta en el futuro.

A pesar de las limitaciones identificadas, el caso de estudio analizado ofrece una perspectiva optimista para el desarrollo futuro de algoritmos de detección de corrupción en la

contratación pública. Si bien la calidad y disponibilidad de los datos representan un desafío definitivo para su aplicación a nivel territorial, su uso todavía es viable en instituciones que gestionen adecuadamente la contratación a través del SECOP. Es decir, aquellas que publiquen la totalidad de sus contratos en la plataforma y mantengan estándares óptimos de calidad en sus datos. En este sentido, muchas entidades del orden nacional o las grandes ciudades podrían beneficiarse de estas herramientas, logrando aprovechar sus posibilidades en la lucha contra la corrupción.

6. Reflexión sobre la posición del investigador

La transparencia metodológica es un principio rector en la investigación científica, en particular cuando se evalúan algoritmos susceptibles de influir en decisiones de gran calado para la gestión pública. En el marco del estudio "Sesgos de datos en aplicaciones de aprendizaje automático: un estudio de caso de un modelo no supervisado para identificar el riesgo de corrupción en la contratación pública colombiana", resulta relevante explicitar que el investigador principal también fue el desarrollador del algoritmo de aprendizaje no supervisado bajo análisis. Esta circunstancia no es accidental, sino que emana de la propia evolución del trabajo previo, catalizada por hallazgos inesperados y retos identificados en las fases de desarrollo e implementación inicial del algoritmo.

La investigación sobre los sesgos, en consecuencia, se deriva de una necesidad inherente de comprender en profundidad tanto las limitaciones como el potencial del algoritmo en un contexto de aplicación real. La posición dual del investigador principal, simultáneamente creador y evaluador crítico del algoritmo, le confiere una perspectiva singular y un conocimiento a detalle de los supuestos, la arquitectura interna y el funcionamiento del mismo. Esto facilitó una identificación y un análisis más exhaustivos y contextualizados de los sesgos existentes.

No se identificó que esta posición pueda generar sesgos en las conclusiones o recomendaciones del presente análisis.

7. Referencias

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics* (pp. 254-264). Auerbach Publications.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Castiblanco, L. (2018). Diseño de un índice de riesgo en la contratación pública en Colombia: caso aplicado Programa de Alimentación Escolar (PAE).

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ quality & safety*, *28*(3), 231-237.

Celi, L. A., Cellini, J., Charpignon, M. L., Dee, E. C., Dernoncourt, F., Eber, R., ... & Yao, S. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health*, *1*(3), e0000022.

Comisión Económica para América Latina y el Caribe (CEPAL). (2024). *Superar las trampas del desarrollo de América Latina y el Caribe en la era digital: el potencial transformador de las tecnologías digitales y la Inteligencia Artificial* (LC/CMSI.9/3). Santiago: Naciones Unidas. Recuperado de https://repositorio.cepal.org/server/api/core/bitstreams/f096da2a-5107-486c-94f3-871683423556/content

Gallego, J., Rivero, G., & Martínez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting*, *37*(1), 360-377.

Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in Machine Learning--What is it Good for?. *arXiv preprint arXiv:2004.00686*.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and

implementation. *IEEE transactions on pattern analysis and machine intelligence*, *24*(7), 881-892.

Kumar, A., Aelgani, V., Vohra, R., Gupta, S. K., Bhagawati, M., Paul, S., ... & Suri, J. S. (2024). Artificial intelligence bias in medical system designs: A systematic review. *Multimedia Tools and Applications*, *83*(6), 18005-18057.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1-35.

Ministerio de Educación Nacional. (2024, marzo 7). *Con el PAE ya se atienden cerca de 5.5 millones de estudiantes en todo el país*. Ministerio de Educación Nacional de Colombia. Recuperado de https://www.mineducacion.gov.co/portal/salaprensa/Comunicados/419937:Con-el-PAE-ya-se-atienden-cerca-de-5-5-millones-de-estudiantes-en-todo-el-pais

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, *2*(6), e0000278.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, *2*, 13.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... & Nascimento, E. G. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, *7*(1), 15.

Raj, R., & Kos, A. (2023). Artificial Intelligence: Evolution, Developments, Applications, and Future Scope. *Przeglad Elektrotechniczny*, *99*(2).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, *2*(8), 73.

Suri, J. S., Agarwal, S., Jena, B., Saxena, S., El-Baz, A., Agarwal, V., ... & Naidu, S. (2022). Five strategies for bias estimation in artificial intelligence-based hybrid deep learning for acute respiratory distress syndrome COVID-19 lung infected patients using AP (ai) Bias 2.0: a systematic review. *IEEE Transactions on Instrumentation and Measurement*.

Suresh, H., & Guttag, J. (2021, October). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9).

The Center for Security and Emerging Technology (2022). *Number of AI Publications by Type, 2010–21.* AI Index Report 2023.

Quid via AI Index, & U.S. Bureau of Labor Statistics. (2024). *Global investment in generative AI* [Dataset]. Our World in Data. Recuperado de https://ourworldindata.org/grapher/global-investment-in-generative-ai?tab=table

Zuleta, M. M., Saavedra, V., & Medellín, J. C. (2018). *Fortalecimiento del sistema de compra pública para reducir el riesgo de corrupción.*

Universidad de los Andes | Escuela de Gobierno Alberto Lleras Camargo

Documentos de trabajo es una publicación periódica de la Escuela de Gobierno Alberto Lleras Camargo de la Universidad de los Andes, que tiene como objetivo la difusión de investigaciones en curso relacionadas con asuntos públicos de diversa índole. Los trabajos que se incluyen en la serie se caracterizan por su interdisciplinariedad y la rigurosidad de su análisis, y pretenden fortalecer el diálogo entre la comunidad académica y los sectores encargados del diseño, la aplicación y la formulación de políticas públicas.

gobierno.uniandes.edu.co

GobiernoUAndes

Universidad de los Andes | Vigilada Mineducación
Reconocimiento como Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.