risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 14/06/2025
Aceitação/Acceptance: 22/08/2025

Modelos de Aprendizaje Automático para Clasificar el Riesgo de Corrupción en Contratación Pública: Caso Hospitales Públicos en Colombia

Heriberto Felizzola[1], Carlos Arango[1], Yilber Erazo[1], Geraldine Camacho[1]

healfelizzola@unisalle.edu.co; cararango@unisalle.edu.co; yerazo01@unisalle.edu.co; gcamacho72@unisalle.edu.co

[1] Universidad de La Salle, Facultad de Ingeniería, Bogotá D.C., Colombia.

DOI: 10.17013/risti.59.3–20

Resumen: La contratación pública es particularmente vulnerable a la corrupción, lo que plantea desafíos significativos para la gestión pública. A nivel mundial, los gobiernos han implementado iniciativas de datos abiertos con el objetivo de promover la transparencia y fortalecer la integridad en la gestión de los recursos públicos. Sin embargo, el éxito de estas iniciativas depende en gran medida de la capacidad para analizar los datos disponibles y detectar patrones que puedan señalar riesgos de corrupción. Este estudio presenta el desarrollo y la evaluación de modelos de aprendizaje automático orientados a analizar y predecir el riesgo de corrupción en procesos de contratación pública. Para ello, se utilizaron datos abiertos de contratación en hospitales públicos colombianos entre 2014 y 2019. Como variable de riesgo, se empleó el porcentaje de contratación directa, ampliamente reconocido como un indicador de posibles irregularidades. Se implementaron y compararon tres algoritmos de aprendizaje automático: Árboles de Decisión, Random Forest y Gradient Boosting. Los resultados evidenciaron un desempeño aceptable, con niveles de exactitud entre el 46% y el 59% y un área bajo la curva ROC que oscila entre el 0.56 y el 0.72.

Palabras-clave: Corrupción en Contratación Pública; Modelos de Aprendizaje Automático; Riesgo de Corrupción; Hospitales Públicos; Datos Abiertos.

Machine Learning Models for Classifying Corruption Risk in Public Procurement: The Case of Public Hospitals in Colombia

Abstract: Public procurement is particularly vulnerable to corruption, posing significant challenges to public administration. Globally, governments have implemented open data initiatives to promote transparency and strengthen resource management's integrity. However, the success of these initiatives largely depends on the ability to analyze available data and identify patterns that may indicate corruption risks. This study presents the development and evaluation of machine learning models designed to analyze and predict the risk of corruption in public procurement processes. For this purpose, open procurement data from Colombian public hospitals between 2014 and 2019 were used. The percentage of

direct contracting, widely recognized as an indicator of potential irregularities, was employed as the risk variable. Three machine learning algorithms were implemented and compared: Decision Trees, Random Forest, and Gradient Boosting. The results demonstrated acceptable performance, with accuracy levels ranging from 46% to 59% and an area under the ROC curve between 0.56 and 0.72.

Keywords: Corruption in Public Procurement; Machine Learning Models; Corruption Risk; Public Hospitals; Open Data.

1. Introducción

Un problema persistente en la contratación pública es la corrupción, cuyo impacto económico es significativo, afectando tanto la eficiencia como la transparencia en la gestión de recursos (Neupane et al., 2014; Soreide, 2002). Según datos de la OCDE, las compras públicas representan aproximadamente el 12% del PIB en los países miembros, y cerca del 57% de los sobornos en el sector público son destinados para la obtención de contratos (OCDE, 2015; OECD, 2019). Dada la relevancia económica de la contratación pública y con el objetivo de combatir la corrupción, gobiernos, organizaciones, académicos y la sociedad civil han implementado diversas estrategias para mejorar la transparencia, integridad y rendición de cuentas en la gestión pública.

En este contexto, las iniciativas de datos abiertos han ganado relevancia en los últimos años. Estas iniciativas buscan proporcionar información estructurada sobre la gestión pública mediante plataformas accesibles y fáciles de usar (Adam et al., 2020; Camargo & Pinzon, 2022). Aunque los datos abiertos son esenciales para fomentar la transparencia, no son suficientes por sí solos; su verdadero impacto depende de su aprovechamiento en procesos analíticos que permitan extraer información clave para la toma de decisiones, la formulación de políticas públicas y el monitoreo de los procesos de contratación (Ansari et al., 2022; Duguay et al., 2019; Janssen et al., 2012).

La analítica aplicada a la contratación pública puede abordar problemas críticos como el análisis, la estimación y la detección de riesgos de corrupción (World Bank, 2022). Esto no solo permite identificar riesgos y los factores asociados, sino también desarrollar herramientas basadas en ciencia de datos capaces de procesar grandes volúmenes de información para automatizar tareas de control y vigilancia (Colonnelli et al., 2020). Estas herramientas ofrecen a los organismos de control, entidades públicas y la sociedad civil la capacidad de realizar una supervisión proactiva y efectiva, fortaleciendo así la integridad en los procesos de contratación (Gallego et al., 2020; Rabuzin & Modrušan, 2019).

El aprendizaje automático se presenta como una herramienta clave en la ciencia de datos, capaz de identificar patrones en los datos y realizar predicciones sobre características específicas (Domingos et al., 2016; Fazekas & Czibik, 2021; Fazekas & Dávid-Barrett, 2015). En el análisis de riesgos de corrupción en contratación pública, las técnicas de aprendizaje automático ofrecen un enfoque prometedor para detectar de manera automática posibles irregularidades. No obstante, el desarrollo de estos modelos enfrenta desafíos importantes. En primer lugar, la corrupción es un fenómeno complejo, difícil de medir y de detectar, lo que exige el diseño de mecanismos efectivos para identificar riesgos de manera confiable (Fazekas et al., 2013, 2016). En segundo lugar, la

contratación pública es un proceso multifacético que involucra múltiples etapas, actores y factores del entorno, lo que requiere la incorporación de un enfoque holístico en los modelos predictivos (Broms et al., 2019).

Este trabajo aborda las limitaciones existentes mediante el desarrollo de modelos de aprendizaje automático que predicen el riesgo de corrupción a partir de una de las señales de alerta más comunes: la contratación directa. Los modelos propuestos incorporan, además, predictores que trascienden las características intrínsecas de los procesos de contratación, integrando indicadores socioeconómicos y de contratación pública asociados a las regiones y localidades donde se llevan a cabo los procesos. Esto permite analizar factores exógenos que influyen en los riesgos de corrupción. Como resultado, se busca desarrollar un modelo predictivo capaz de caracterizar y clasificar a las entidades hospitalarias públicas en Colombia según su nivel de riesgo de corrupción, utilizando la contratación directa como indicador principal. Este enfoque propone una metodología para evaluar la falta de competencia en procesos de contratación pública mediante datos abiertos, apoyándose en estadísticas, indicadores de contratación y modelos de aprendizaje automático.

El artículo se organiza de la siguiente manera: la Sección 2 presenta una revisión del estado del arte sobre modelos de aprendizaje automático aplicados a la predicción de riesgos de corrupción en contratación pública. La Sección 3 describe los datos utilizados, su procesamiento y la definición de las variables. La Sección 4 detalla los modelos y técnicas de aprendizaje automático implementados, los escenarios de análisis y las métricas empleadas. La Sección 5 expone los resultados y el análisis de los modelos desarrollados. Finalmente, la Sección 6 presenta las conclusiones y posibles implicaciones de los hallazgos.

2. Revisión de Literatura

2.1. Medición del Riesgo de Corrupción en Contratación Pública

Los índices de corrupción existentes, como el Índice de Percepción de la Corrupción de Transparencia Internacional, ofrecen una evaluación general útil de los niveles de corrupción. Sin embargo, carecen del nivel de detalle necesario para abordar de manera efectiva las intervenciones anticorrupción en procesos específicos de contratación pública. Para llenar esta brecha, investigadores han explorado el uso de modelos de aprendizaje automático para predecir el riesgo de corrupción a un nivel más granular, considerando factores como el monopolio del poder, la asimetría de información, la transparencia y la rendición de cuentas en los procesos de contratación. Estos modelos pueden ayudar a responsables políticos y funcionarios de contratación a identificar contratos de alto riesgo y a implementar medidas específicas para mejorar la transparencia y la integridad en la gestión pública (Gallego et al., 2021; Modrušan et al., 2021).

El desafío de medir y detectar la corrupción en la contratación pública es de vital importancia, dado su impacto en la asignación eficiente de recursos, el crecimiento económico y la confianza pública. Como señalan (Fazekas & Kocsis, 2020), existe una notable carencia de indicadores confiables y accionables, especialmente para la corrupción de alto nivel.

Diversos autores han propuesto indicadores innovadores para evaluar el riesgo de corrupción en la contratación pública. Fazekas & Kocsis (2020) desarrollaron dos enfoques principales: la detección de licitaciones únicas en mercados que deberían ser competitivos y una puntuación compuesta de "señales de alerta" denominada Índice de Riesgo de Corrupción (CRI). Decarolis & Giorgiantonio (2022) validaron nuevos indicadores basados en el diseño de licitaciones, utilizando técnicas de aprendizaje automático para evaluar su capacidad predictiva. Por su parte, Fazekas et al. (2023) calcularon indicadores individuales de riesgo de corrupción y diseñaron un Índice de Riesgo de Corrupción (IRC) compuesto para Bulgaria.

Estos indicadores compuestos, como el CRI, combinan diferentes señales de alerta para medir de manera más precisa los riesgos de corrupción en la contratación pública. En conclusión, el desarrollo de indicadores objetivos y accionables para evaluar el riesgo de corrupción ha sido un campo en constante evolución. Estas herramientas no solo fortalecen la capacidad de los responsables de políticas públicas y los investigadores para combatir la corrupción, sino que también establecen un marco más sólido para la rendición de cuentas en el gasto público. La transición desde enfoques estadísticos tradicionales hacia técnicas avanzadas de aprendizaje automático, como las propuestas de Decarolis & Giorgiantonio 2022), señala un camino prometedor para futuras investigaciones en este ámbito.

2.2. Modelos de Aprendizaje Automático para la Predicción de Riesgos de Corrupción

Investigaciones previas han demostrado que los modelos de aprendizaje automático son herramientas efectivas para identificar patrones y anomalías en los datos de contratación pública que podrían señalar posibles casos de corrupción (Modrušan et al., 2021; Nai et al., 2022). Estos modelos analizan factores como la distribución de precios en las ofertas, la frecuencia con la que ciertos proveedores reciben adjudicaciones, y el cumplimiento de los objetos y plazos contractuales. Además, pueden detectar señales de alerta, como la falta de transparencia en los procesos de licitación, adjudicaciones a proveedores con vínculos políticos y contrataciones aceleradas o con escaso escrutinio público. La implementación de estos modelos permite identificar contratos de alto riesgo y tomar medidas preventivas y de detección temprana para combatir la corrupción en la contratación pública (Aldana et al., 2022).

Un estudio basado en más de 2 millones de contratos públicos en Colombia exploró el uso de modelos de aprendizaje automático para detectar corrupción, logrando precisiones superiores al 80% en algunos casos (Gallego et al., 2020). De manera similar, investigaciones en México demostraron que los modelos de conjunto pueden clasificar eficazmente contratos corruptos y no corruptos (Aldana et al., 2022). Estas investigaciones resaltan la importancia de la interpretabilidad de los modelos, un factor clave para su aplicación en el sector público. Sin embargo, su implementación práctica enfrenta desafíos significativos, como el desbalanceo de clases, donde los contratos conformes superan ampliamente a los problemáticos, lo que afecta la precisión de los modelos predictivos. Además, alinear estos modelos con las complejas dinámicas de la toma de decisiones públicas, que incluyen objetivos en conflicto y la necesidad de intervención humana, resulta complicado. Para mantener su efectividad, los modelos

deben ser adaptativos, capaces de actualizarse y evolucionar frente a nuevas estrategias de corrupción (Gallego et al., 2021). Este enfoque flexible es esencial para anticipar y mitigar las amenazas emergentes en un entorno dinámico y en constante cambio.

Diversos estudios han empleado la metodología CRISP-DM, ampliamente utilizada en proyectos de minería de datos y aprendizaje automático, que abarca las etapas de recopilación, preprocesamiento, modelado, evaluación e implementación (Torres-Berru et al., 2023; Torres-Berru & López Batista, 2021). Entre las actividades clave destacaron la ingeniería de características, el manejo del desbalanceo de clases y el refinamiento iterativo de modelos. Las investigaciones identificaron variables como el número de oferentes en las licitaciones, la diferencia entre la oferta ganadora y la segunda más alta, y la relación comprador-proveedor, como predictores relevantes de corrupción. Además, los modelos detectaron anomalías en los parámetros de calificación de adquisiciones públicas y revelaron un clúster denominado "Oferta Económica Nula", caracterizado por un peso insignificante de la oferta económica, lo que podría sugerir manipulaciones (Torres-Berru & López Batista, 2021). Estos hallazgos resaltan el potencial de las técnicas de minería de datos para promover la transparencia y rendición de cuentas en la contratación pública. Asimismo, subrayan la necesidad de seguir investigando para optimizar los modelos y adaptarlos a contextos específicos, ampliando su aplicabilidad y efectividad en distintos escenarios.

Otros estudios han explorado el uso de clasificadores Naïve Bayes para evaluar riesgos en adquisiciones públicas, enfocándose en factores asociados con prácticas de colusión y corrupción. Este enfoque comienza definiendo el alcance del análisis, identificando tipos de contratos o agencias gubernamentales específicas, y seleccionando factores de riesgo como el tipo de licitación, renegociaciones posteriores y vínculos entre licitadores y funcionarios públicos (Balaniuk et al., 2013). Los factores se ponderan según su importancia para identificar casos de alto riesgo. Este método ha demostrado ser práctico y eficaz, logrando alinearse con evaluaciones de auditores expertos. Además, su capacidad para integrar datos de diversas fuentes con opiniones de expertos permite realizar evaluaciones más objetivas y completas, optimizando la asignación de recursos para auditorías y promoviendo la transparencia en los procesos de contratación pública.

En un trabajo desarrollado por Sales & Carvalho (2016) se utiliza Naïve Bayes para entrenar un modelo con indicadores de riesgo como variables de entrada y clasificaciones de contratos como "Bueno" o "Malo" como objetivo. El modelo mostró efectividad al calcular la probabilidad de que un contrato pertenezca al grupo "Malo", priorizando así los contratos con mayor riesgo para auditorías. Este enfoque destacó por su precisión y capacidad para cuantificar riesgos en adquisiciones públicas.

García Rodríguez et al. (2022) concluyen que los algoritmos de aprendizaje automático, especialmente los métodos de conjunto como Extra Trees y Gradient Boosting, ofrecen resultados prometedores para detectar colusión en licitaciones de compras públicas. Estos modelos mostraron una alta precisión en la identificación de comportamientos colusorios, superando a otros algoritmos como las Máquinas de Vectores de Soporte y Random Forest. El estudio destaca la importancia de la ingeniería de variables y el preprocesamiento de datos como factores clave para optimizar el rendimiento de los modelos. Además, los autores señalan que las técnicas de aprendizaje automático son herramientas valiosas para las autoridades de competencia y las agencias de

vigilancia de compras públicas, ya que facilitan la detección y prevención de prácticas anticompetitivas, promueven la competencia justa y garantizan una gestión eficiente de los recursos públicos.

Por su parte, Domingos et al. (2016) destacan que los clasificadores bayesianos ofrecen un enfoque prometedor para medir el riesgo asociado a los contratos públicos. Su modelo, diseñado con base en diversos indicadores de riesgo, demuestra una precisión significativa al predecir la probabilidad de problemas con contratos específicos.

En conjunto, estos estudios subrayan el potencial de las técnicas de aprendizaje automático para fortalecer la transparencia y la rendición de cuentas en las adquisiciones públicas. Además, estas herramientas permiten detectar patrones asociados a la corrupción, contribuyendo a la mejora de los sistemas de compra y al uso responsable de los dineros públicos.

2.3. Factores Exógenos Asociados a los Riesgos de Corrupción en Contratación Pública

Las investigaciones han identificado una amplia gama de factores económicos y políticos que contribuyen a la prevalencia de la corrupción en la contratación pública. Por ejemplo, un estudio en Uganda destacó que los bajos salarios en el sector público, la falta de rendición de cuentas y la interferencia política en los procesos de contratación son predictores significativos de corrupción (Basheka, 2011). De manera similar, otros estudios han señalado que marcos institucionales débiles, la ausencia de transparencia y la influencia de grupos de interés poderosos agravan este problema (Soreide, 2002; Williams-Elegbe, 2018). La falta de mecanismos efectivos de supervisión y rendición de cuentas, combinada con una ejecución deficiente, crea un entorno propicio para prácticas corruptas.

Estos hallazgos revelan que la corrupción en la contratación pública no solo depende de cómo se diseñan, planifican, evalúan y ejecutan los contratos, sino también de factores externos que son determinantes para la eficiencia, integridad y transparencia de los procesos de contratación. Analizar y abordar estos elementos estructurales es esencial para mitigar los riesgos de corrupción y mejorar la gobernanza en este ámbito.

3. Metodología

3.1. Datos y Fuentes de Información

Para desarrollar los modelos predictivos, se utilizaron datos de contratación pública hospitalaria en Colombia disponibles en Sistema Electrónico de Contratación Pública (SECOP I), datos territoriales del Departamento Nacional de Planeación (Terridata) y estadísticas de salud proporcionadas por la Superintendencia de Salud (Supersalud).

Los datos extraídos de SECOP I corresponden al período 2014-2019. El conjunto inicial contiene 1.391.442 contratos y 44 variables, representando un valor total de contratos (incluyendo adiciones) de $51.474 mil millones de pesos Colombianos (COP). Cada observación correspondía a un contrato entre una entidad hospitalaria gubernamental y un proveedor de bienes o servicios. Para el análisis, se seleccionaron 15 variables clave

relacionadas con la Identificación y ubicación de los contratistas y entidades; el tipo de contrato y tipo de proceso; los valores iniciales, adiciones y total de los contratos; las fechas de firma y registro; las adiciones en tiempo y enlaces URL de los contratos.

La preparación de los datos para los modelos incluyó una fase de depuración para garantizar su calidad y confiabilidad. En esta etapa se aplicaron filtros y se llevaron a cabo tareas de limpieza, entre las que destaca la corrección manual de registros que presentaban inconsistencias. Estas se encontraron principalmente en el valor del contrato, las adiciones y la identificación de los contratistas, y fueron subsanadas mediante la revisión y validación entre la base de datos abierta y los documentos contractuales y sus informes de cierre. En los filtros se eliminaron registros con contratos inferiores a COP $1.000.000, valores faltantes en variables clave, convenios interadministrativos, transacciones bancarias, URLs inconsistentes y entidades con menos de tres contratos. Además, se revisaron manualmente contratos con cuantías y adiciones superiores a COP $10 mil millones, detectándose errores en el 66,9% de los casos. Esto permitió corregir 90 contratos con errores en cuantías y 19 en adiciones, logrando una reducción del 32,42% en el valor total de los contratos, como se detalla en la Tabla 1, evidenciando el impacto positivo del procesamiento de datos en su calidad.

Del DNP, se emplearon datos extraídos del aplicativo Terridata. Este sistema gestiona y publica la información de indicadores agrupados en subcategorías y dimensiones para departamentos y municipios en Colombia. Para este estudio, se seleccionaron indicadores municipales y departamentales de cuatro dimensiones: Economía, Educación, Finanzas Públicas y Salud, correspondientes al período 2014-2019. Se priorizaron indicadores con menos del 40% de datos faltantes. Además, se incorporaron indicadores de la Medición del Desempeño Municipal (MDM), que evalúa el cumplimiento de metas de los planes de desarrollo, la provisión de servicios básicos, la ejecución presupuestal y la gestión administrativa y fiscal (Departamento Nacional de Planeación, 2020). De la Superintendencia de Salud (Supersalud), se utilizaron dos indicadores del Reporte de Circular Única, que proporciona información relevante sobre la gestión hospitalaria.

Atributo	*Valor Inicial (MM*)*	*Valor Final (MM*)*	*Reducción*
Valor del Contrato	$ 42.580	$ 28.719	32,55%
Valor Adiciones	$ 4.644	$ 3.194	31,22%
Valor total con Adiciones	$ 47.225	$ 31.913	32,42%

* Los datos se presentan em miles de millones de pesos colombianos.

Tabla 1 – Resumen del Ajuste en Valor de los Contratos

3.2. Variables e Indicadores de Contratación Pública

Tras la obtención de los datos de contratación, se consolidó una tabla maestra para el desarrollo de los modelos predictivos. En esta tabla, cada fila representa un hospital público, mientras que las columnas incluyen características específicas del hospital, estadísticas e indicadores de contratación, así como los indicadores proporcionados por el DNP (municipales y departamentales) y la Supersalud. Esta tabla contiene datos agregados por hospital y por año. Este nivel de granularidad de los datos se utilizó para

el cálculo de los indicadores de contratación y los indicadores tomados del DNP y la Supersalud.

Para representar las prácticas y características de la contratación en los hospitales, se emplearon indicadores basados en estudios sobre riesgos de corrupción en compras públicas realizados en Colombia (Zuleta et al., 2019), Chile (Jorquera, 2019), y México (Martínez & Torres, 2019). Cabe resaltar que todos los indicadores se normalizaron en una escala de 0-100. Los indicadores calculados son los siguientes:

Índice Herfindahl-Hirschman (HHI): Mide el nivel de concentración en la contratación pública de cada hospital, identificando posibles monopolios. Se calcula como:

$$HHI = \sum_{i=1}^{C} S_i^2 \tag{1}$$

Donde S_i es la participación porcentual (en cantidad y valor total) de cada contratista en la contratación del hospital.

Índice de diversidad de Simpson (ID): Originalmente diseñado para medir biodiversidad, en este contexto mide la diversidad de contratistas en un hospital. Su fórmula es:

$$ID = \frac{\sum_{i=1}^{C} q_i \left(q_i - 1\right)}{Q\left(Q - 1\right)} \tag{2}$$

Donde q_i es el volumen de contratación de cada contratista y Q es el volumen total de contratación de cada hospital.

Número de empresas ganadoras diferentes por cada 100 contratos: Mide el número de empresas diferentes adjudicatarias de los procesos de contratación. Dado que cada entidad tiene un volumen diferente de contratación, el índice se expresa en número de contratistas por cada 100 contratos.

Índice de concentración de las cuatro empresas con mayor número y valor de contratos - IC4k: Mide la proporción de procesos que el hospital adjudicó a los cuatro contratistas con más contratos. Un alto porcentaje, indica que la gran parte de la contratación del hospital está concentrada en pocos contratistas. Se calcula como:

$$IC4k = \sum_{c \in K} S_c \tag{3}$$

Donde S_c es la participación porcentual de cada contratista entre los cuatro principales.

Nivel de riesgo de corrupción: Para definir el nivel de riesgo de corrupción de un hospital público se utilizó el indicador de porcentaje de contratación directa de la entidad en función del número de contratos. La contratación directa es una modalidad de adquisición pública en la que una entidad estatal adjudica un contrato sin un proceso de

licitación abierta, generalmente por razones específicas como emergencias, exclusividad o conveniencia administrativa. Aunque este mecanismo está legalmente permitido en situaciones excepcionales, su uso excesivo o indebido representa un riesgo significativo de corrupción, ya que compromete principios fundamentales de la buena contratación, como la transparencia y la competencia justa. Este indicador se definió en tres niveles:

- Bajo o Moderado (≤ 25%).
- Medio (≤ 55%).
- Alto (> 55%).

En la tabla maestra se eliminaron variables con un porcentaje de datos faltantes igual o superior al 40%. Para las variables numéricas restantes, los valores faltantes fueron imputados utilizando la mediana de cada atributo. El resultado final fue una tabla maestra con 892 observaciones (hospitales) y 345 atributos, de los cuales uno corresponde a la variable objetivo (Nivel de Riesgo de Corrupción) y 344 son predictores. Entre las variables predictoras, 342 son numéricas y 2 categóricas.

El nombre de todas las variables predictoras está formado de la misma manera: primero con la categoría a la que pertenece (indba = variable calculada a partir de SECOP, ine = educación, inec = economía, inf = finanzas públicas, inmdm = medición del desempeño municipal, ins = salud, y supers = Supersalud); segundo, la codificación del nombre del indicador; después el nivel de la variable (dto = departamento, mun = municipio y hos = hospital) y por último el año al que pertenece (si se establecía por años). En la Tabla 2 se presenta el listado de variables por categoría y una descripción general.

Grupo de Variables en Indicadores	***Indicadores***
2 variables categóricas de la base de datos de SECOP	Orden Entidad Departamento de la Entidad
79 estadísticas de total, porcentajes y de medidas de tendencia central, que indican patrones de comportamiento de los hospitales.	• Cantidad total de contratos (general y por años). • Valor inicial de los contratos • Valor (Cuantía) total de los contratos (general, mediana, media, desviación estándar y por años) • Valor (Cuantía) total con adiciones (general y por años) • Porcentaje de las adiciones sobre el valor inicial del contrato (general y por años) • Porcentaje de los contratos con adiciones en tiempo (general y por años) • Porcentaje de contratación directa en número. • Porcentaje de contratos por tipo de contrato. • Cantidad de contratos con alguna modificación. • Porcentaje de la cantidad y el valor de contratos por mínima cuantía por hospital y departamento. • Cantidad de contratos con adiciones • Cantidad de contratos a fin de año (por años) • Porcentaje de contratos del total de contratos del año (por años) • Cantidad de contratistas por entidad (por años) • Diferencia entre las fechas de cargue y firma (por años)

Grupo de Variables en Indicadores	*Indicadores*
18 indicadores de concentración de contratos e identificación de riesgos de corrupción.	• Índice Herfindahl-Hirschman para el número o valor de contratos de la entidad estatal (IHH) • El índice de diversidad de Simpson por número o valor de contratos (ID) • Índice de concentración de las cuatro empresas • con mayor número y valor de adjudicaciones. (IC4K) • Cantidad de contratos por contratista por entidad (por años) • Valor de los contratos por contratista por entidad (por años) • Repitencia del mayor contratista en cantidad y proporción • Número de empresas ganadoras diferentes por cada 100 contratos • Índice de Riesgo de Corrupción
243 indicadores de desempeño municipal del aplicativo Territata del DNP.	• Indicadores de la dimensión de Educación • Indicadores de la dimensión de Economía • Indicadores de la dimensión de Finanzas Públicas • Indicadores de la dimensión de Salud • Indicadores de Medición del Desempeño Municipal
2 Indicadores de SuperSalud	• Indicadores de Reporte de Circular Única

Tabla 2 – Listado de Variables e Indicadores de la Tabla Maestra

3.3. Modelos y Escenarios

3.3.1. Modelos de Machine Learning

Para predecir el nivel de riesgo de las entidades por contratación directa, se emplearon tres modelos de clasificación multiclase: Árboles de Decisión, Random Forest y Gradient Boosting (Extreme Gradient Boosting). Estos modelos se entrenaron utilizando la libreria tidymodels e el lenguaje estadístico R. El desarrollo de todos los modelos incluyo las siguientes tareas de preprocesamiento: división de datos en conjuntos de entrenamiento (75%) y evaluación (25%), normalización de variables numéricas, creación de la categoría "otros" para los departamentos con menos del 5% del total de datos, generación de variables dummy para las variables categóricas y balanceo de clases.

Dado que la variable objetivo, nivel de riesgo de corrupción, presentaba una distribución desigual (540 observaciones en "bajo o moderado", 83 en "medio" y 269 en "alto"), se aplicó balanceo de clases mediante el método SMOTE. Este procedimiento genera nuevas instancias sintéticas para las clases minoritarias utilizando los vecinos más cercanos de cada caso, mejorando así la representatividad de las clases en el conjunto de entrenamiento.

Se utilizó el método Boruta, basado en el algoritmo Random Forest, como estrategia para seleccionar el mejor conjunto de variables. Este enfoque genera en cada iteración variables "sombra" a partir de los predictores originales, calcula su importancia relativa y las compara con las variables sintéticas. Las variables con menor importancia que las sombras son rechazadas. Este proceso continúa hasta que las variables son clasificadas como importantes, no importantes o tentativas, con un límite de 500 iteraciones para este estudio debido a su alto costo computacional. Para el análisis se consideraron las variables importantes y tentativas.

3.3.2. Escenarios

Se definieron cuatro escenarios para evaluar el desempeño de los modelos en diferentes configuraciones de la tabla maestra:

- *Escenario 1. Solo variables de SECOP (indba):* Considera exclusivamente las variables internas relacionadas con contratación pública para determinar si son suficientes para predecir el nivel de riesgo.
- *Escenario 2. Variables de todas las fuentes:* Incluye indicadores de SECOP, Terridata, MDM y Supersalud, con el objetivo de evaluar la influencia de factores externos como los indicadores departamentales y municipales.
- *Escenario 3. Selección de variables en todas las fuentes:* Aplica técnicas de selección de variables a los datos combinados para eliminar aquellas que no aportan información relevante al modelo.
- *Escenario 4. Selección de variables sin MDM ni Supersalud:* Similar al escenario 3, pero excluyendo variables de estas fuentes, ya que no fueron seleccionadas como relevantes durante el proceso de reducción de dimensionalidad.

4. Resultados

4.1. Comparación de Modelos y Escenarios

Los resultados del desempeño de los modelos en los diferentes escenarios se presentan en la Tabla 3. Para la evaluación, se utilizaron las métricas de exactitud y el área bajo la curva ROC (ROC AUC). La exactitud se define como el porcentaje de predicciones que el modelo clasificó correctamente sobre el total de casos; se calcula dividiendo el número de predicciones correctas entre el número total de predicciones. Por su parte, el ROC AUC mide la capacidad del modelo para discriminar correctamente entre las clases; en este contexto, el nivel de riesgo de corrupción de la entidad (bajo, medio y alto). Esta métrica se obtiene de la curva ROC, que representa la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos a través de diferentes umbrales de decisión. Un valor de AUC cercano a 1.0 indica una capacidad de discriminación perfecta, mientras que un valor de 0.5 sugiere un desempeño equivalente al azar.

Los resultados muestran que el modelo Random Forest y el Gradient Boosting presentan el mejor desempeño en términos de exactitud, mientras que el modelo Gradient Boosting sobresale en la métrica del área bajo la curva ROC (ROC AUC), obteniendo consistentemente los valores más altos en todos los escenarios.

En particular, los mejores puntajes se obtuvieron con Gradient Boosting, destacando que, independientemente de la métrica utilizada, este modelo sería el elegido, aunque en escenarios distintos: el escenario 2 para la exactitud y el escenario 3 para ROC AUC. Dentro de cada modelo, los escenarios 2 y 4 fueron los más destacados para la exactitud, mientras que los escenarios 3 y 4 sobresalieron en ROC AUC. Es importante resaltar que el escenario 1 no logró destacarse en ninguna métrica. En promedio, el escenario con mejor rendimiento en ROC AUC fue el escenario 3, mientras que para la exactitud fue el escenario 2. Además, los valores de la curva ROC superaron el 60% en la mayoría de los casos, evidenciando el poder predictivo de los algoritmos.

Modelo	Escenario	Exactitud	ROC AUC
Árbol de decisión	1	0,4642	0,5612
	2	0,5491	0,5991
	3	0,5089	0,6706
	4	0,4196	0,6030
Random Forest	1	0,5720	0,6044
	2	0,5855	0,6291
	3	0,5540	0,6336
	4	0,5945	0,6420
Gradient Boosting	1	0,5270	0,7059
	2	0,5990	0,6622
	3	0,5765	0,7202
	4	0,4909	0,7045

Tabla 3 – Desempeño de los modelos de acuerdo con cada escenario

4.2. Análisis del Modelo con Mejor Desempeño

El modelo Gradient Boosting aplicado al escenario 3 mostró el mejor desempeño global en términos de ROC AUC, como se ilustra en la Tabla 3. Esto resalta la importancia de utilizar todas las variables disponibles en la tabla maestra, aplicando algoritmos de selección como Boruta para reducir dimensionalidad y optimizar el modelo. Los resultados indican que no es suficiente trabajar únicamente con las variables de SECOP I; es fundamental incluir indicadores externos de fuentes como Terridata y Supersalud, además de aplicar técnicas de reducción de dimensionalidad para mejorar el rendimiento del modelo.

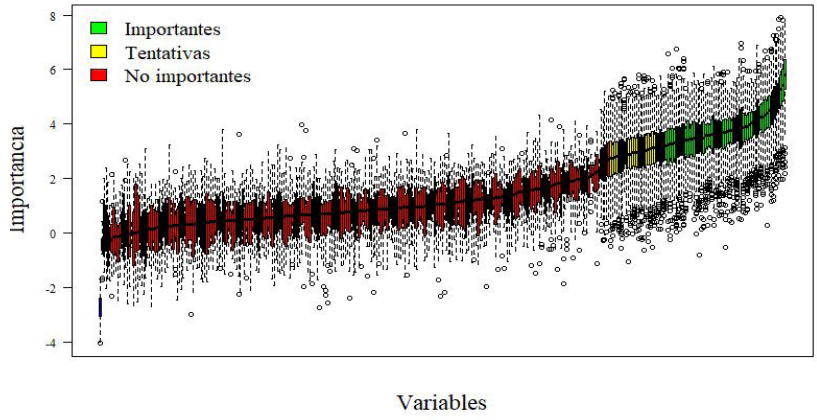

Figura 1 – Clasificación de variables según Boruta.

En la Figura 1 se muestra la clasificación de variables según el algoritmo Boruta, el cual identificó 81 de las 344 variables predictoras como importantes o tentativas. Este análisis reveló que muchas variables irrelevantes o "ruido" afectaban negativamente el rendimiento del modelo, lo que subraya la necesidad de una selección rigurosa. Es relevante señalar que ninguna variable de las fuentes MDM y Supersalud fue seleccionada como importante.

La Figura 2 presenta la curva ROC para cada categoría de la variable objetivo, Riesgo, mostrando que la categoría Medio fue la mejor clasificada. Todas las curvas ROC superan la diagonal, lo que confirma que el modelo clasifica mejor que la aleatoriedad, logrando una sensibilidad y especificidad superiores al 50% para cada clase.

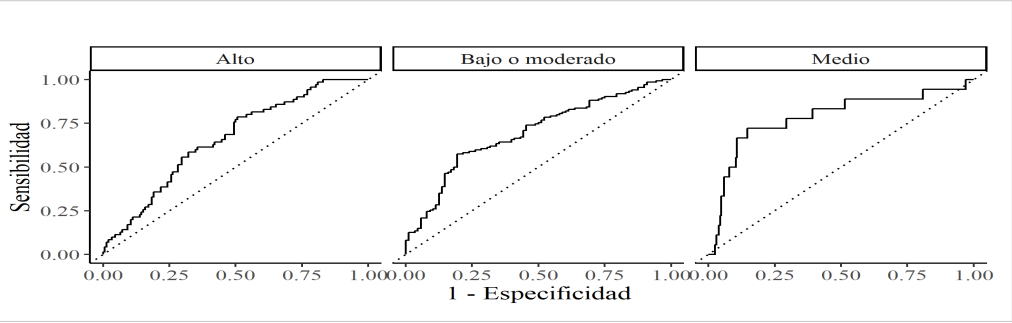

Figura 2 – Desempeño en curva ROC.

En la Figura 3 se destacan las 15 variables más importantes del modelo. Las primeras posiciones corresponden a variables internas de SECOP, seguidas por indicadores de educación, economía y salud de Terridata, excluyendo las dimensiones de finanzas públicas. Entre las variables más relevantes se incluyen los datos de pruebas de matemáticas (2015 y 2016), la cobertura de vacunación y la tasa de mortalidad neonatal. Además, el análisis muestra que las estadísticas relacionadas con la contratación al final del año y el tiempo entre la fecha de cargue y firma de los contratos son determinantes para predecir el nivel de riesgo de corrupción.

Es interesante destacar que variables aparentemente poco relacionadas, como la cobertura de vacunación o la mortalidad neonatal, resultaron altamente predictivas, mientras que algunos indicadores esperados de riesgo de corrupción fueron descartados. Este hallazgo sugiere que variables contextuales, como las relacionadas con la prestación de servicios, también influyen significativamente en los niveles de riesgo.

Para complementar el análisis de las variables importantes, se llevó a cabo un análisis de clúster utilizando las técnicas de PCA y UMAP, lo que permitió identificar agrupaciones de entidades según los indicadores de contratación. Como se observa en la Figura 4, las características de contratación varían significativamente entre departamentos. Departamentos como Antioquia, Valle del Cauca y Bogotá presentan comportamientos claramente diferenciados, separándose del resto del país en ambas técnicas. De manera similar, los departamentos de Santander y Boyacá tienden a formar agrupaciones

propias. Estos patrones indican que los modelos predictivos podrían beneficiarse de una segmentación por grupos de departamentos en lugar de una aplicación a nivel nacional, permitiendo incorporar las particularidades regionales y mejorar la precisión del análisis.

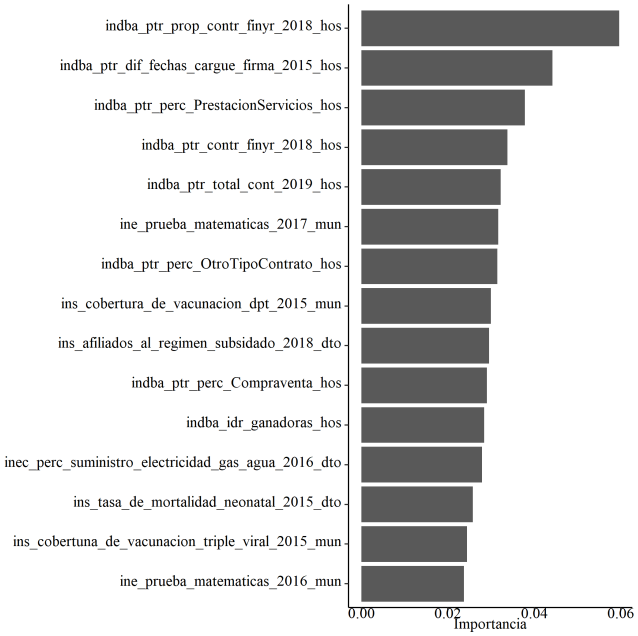

Figura 3 – Importancia de las Variables

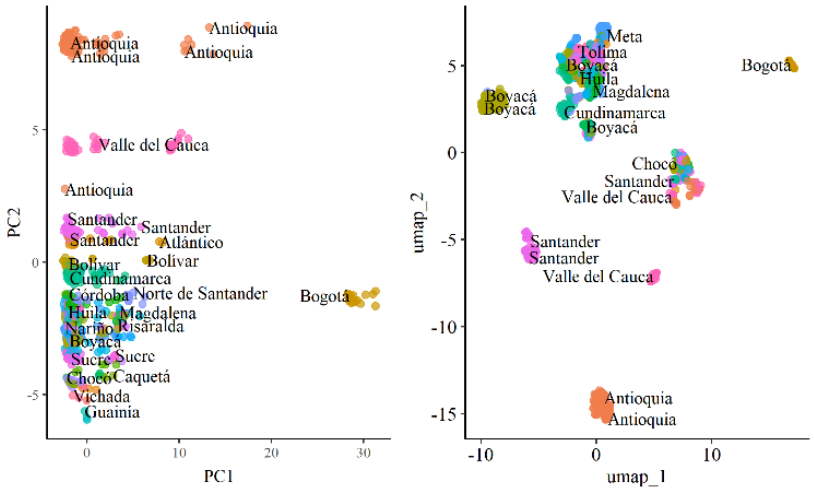

Figura 4 – Análisis de Clúster.

5. Conclusiones

Los modelos desarrollados permiten concluir que existe una relación significativa entre ciertos atributos y el nivel de riesgo de corrupción en la contratación pública. Variables como los contratos realizados al final del año, las diferencias en las fechas, el tipo de contrato y el número de contratistas, junto con características como el año y el nivel de la información, resultaron ser determinantes en la clasificación de las entidades hospitalarias. Además, se evidenciaron patrones de contratación específicos por departamento, lo que sugiere que los modelos predictivos podrían ser más efectivos al aplicarse de forma segmentada por regiones en lugar de a nivel nacional. Estos hallazgos destacan el potencial de los modelos como herramientas útiles para fomentar principios de buena contratación, apoyando la transparencia, la competencia justa y la rendición de cuentas en la contratación pública.

Las variables relacionadas con el entorno también contribuyen al desempeño de los modelos, siempre que se utilicen algoritmos de selección de variables como Boruta, ya que no solo mejoran la capacidad predictiva, sino que también reducen la complejidad del modelo mediante la disminución de su dimensionalidad. Este enfoque resulta esencial para optimizar los resultados y garantizar la eficiencia en el uso de datos heterogéneos.

En el análisis comparativo de los modelos de aprendizaje automático, el Gradient Boosting demostró ser el más efectivo, especialmente en la métrica ROC AUC, superando a los demás algoritmos evaluados. La implementación de estas técnicas de analítica representa un avance significativo en el desarrollo de herramientas destinadas a fortalecer los principios de la contratación pública eficiente y transparente.

Como trabajo futuro, sería valioso incluir información de otras fuentes tales como datos de entidades de vigilancia y control, como la Procuraduría y la Contraloría, para incorporar antecedentes sobre investigaciones relacionadas con entidades y contratistas. Asimismo, se recomienda desarrollar modelos predictivos segmentados por grupos de departamentos para evaluar su impacto en el desempeño, ya que los patrones regionales detectados en este estudio podrían mejorar la precisión y efectividad de los modelos.

Referencias

Adam, I., Fazekas, M., & Tóth, B. (2020). Measuring the benefits of open contracting. Government Transparency Institute; http://redflags.govtransparency.eu/wp-content/uploads/2020/01/GTI-WP_OC-benefits-research_final-report_20200121.pdf

Aldana, A., Falcón-Cortés, A., & Larralde, H. (2022). A machine learning model to identify corruption in México's public procurement contracts. arXiv. https://doi.org/10.48550/arXiv.2211.01478

Ansari, B., Barati, M., & Martin, E. G. (2022). Enhancing the usability and usefulness of open government data: A comprehensive review of the state of open government data visualization research. Government Information Quarterly, 39(1), 101657. https://doi.org/10.1016/j.giq.2021.101657

Balaniuk, R., Bessiere, P., Mazer, E., & Cobbe, P. (2013). Collusion and Corruption Risk Analysis Using Naïve Bayes Classifiers. Communications in Computer and Information Science, 246, 89-100. https://doi.org/10.1007/978-3-642-42017-7_7

Basheka, B. C. (2011). Economic and political determinants of public procurement corruption in developing countries: An empirical study from uganda. Journal of Public Procurement, 11(1), 33-60. https://doi.org/10.1108/JOPP-11-01-2011-B002

Broms, R., Dahlström, C., & Fazekas, M. (2019). Political Competition and Public Procurement Outcomes. Comparative Political Studies, 52(9), 1259-1292. https://doi.org/10.1177/0010414019830723

Camargo, E. A. R., & Pinzon, M. A. R. (2022). La importancia de la seguridad de la información en el sector público en Colombia. RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação, 46, 87-99. https://doi.org/10.17013/risti.46.87-99

Colonnelli, E., Gallego, J. A., & Prem, M. (2020). What Predicts Corruption? SSRN Scholarly Paper ID 3330651. Social Science Research Network. https://doi.org/10.2139/ssrn.3330651

Decarolis, F., & Giorgiantonio, C. (2022). Corruption red flags in public procurement: New evidence from Italian calls for tenders. EPJ Data Science, 11(1), Art. 1. https://doi.org/10.1140/epjds/s13688-022-00325-x

Domingos, S. L., Carvalho, R. N., Carvalho, R. S., & Ramos, G. N. (2016). Identifying IT Purchases Anomalies in the Brazilian Government Procurement System Using Deep Learning. 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 722-727. https://doi.org/10.1109/ICMLA.2016.0129

Duguay, R., Rauter, T., & Samuels, D. (2019). The Impact of Open Data on Public Procurement. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3483868

Fazekas, M., & Czibik, Á. (2021). Measuring regional quality of government: The public spending quality index based on government contracting data. Regional Studies, 55(8), 1459-1472. https://doi.org/10.1080/00343404.2021.1902975

Fazekas, M., & Dávid-Barrett, E. (2015). Corruption Risks in UK Public Procurement and New Anti- Corruption Tools. November, 1-33.

Fazekas, M., & Kocsis, G. (2020). Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data. British Journal of Political Science, 50(1), 155-164. https://doi.org/10.1017/S0007123417000461

Fazekas, M., Poltoratskaia, V., & Tóth, B. (2023). Corruption Risks and State Capture in Bulgarian Public Procurement. World Bank. https://doi.org/10.1596/1813-9450-10444

Fazekas, M., Tóth, I. J., & King, L. P. (2013). Corruption manual for beginners.

Fazekas, M., Tóth, I. J., & King, L. P. (2016). An Objective Corruption Risk Index Using Public Procurement Data. European Journal on Criminal Policy and Research, 22(3), 369-397. https://doi.org/10.1007/s10610-016-9308-z

Gallego, J., Prem, M., & Vargas, J. F. (2021). Pandemic corruption: Insights from Latin America. En Procurement in Focus: Rules, Discretion, and Emergencies (p. 180). Centre for Economic Policy Research. https://www.govtransparency.eu/wp-content/uploads/2022/01/Procurement-in-Focus.pdf#page=110

Gallego, J., Rivero, G., & Martínez, J. (2020). Preventing rather than punishing: An early warning model of malfeasance in public procurement. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2020.06.006

García Rodríguez, M. J., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E. D., & Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. Automation in Construction, 133, 104047. https://doi.org/10.1016/j.autcon.2021.104047

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management, 29(4), 258-268. https://doi.org/10.1080/10580530.2012.716740

Jorquera, M. (2019). Compras públicas y Big Data: Investigación en Chile sobre índice de riesgo de corrupción. http://www.repository.fedesarrollo.org.co/handle/11445/3871

Martínez, A., & Torres, L. M. (2019). Compras públicas y Big Data: El caso mexicano. http://www.repository.fedesarrollo.org.co/handle/11445/3874

Modrušan, N., Rabuzin, K., & Mršic, L. (2021). Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies. International Journal of Advanced Computer Science and Applications, 12(2). https://doi.org/10.14569/IJACSA.2021.0120272

Nai, R., Sulis, E., & Meo, R. (2022). Public Procurement Fraud Detection and Artificial Intelligence Techniques: A Literature Review. The Knowledge Management for Law Workshop (KM4LAW), Bozen-Bolzano, Italy.

Neupane, A., Soar, J., Vaidya, K., & Yong, J. (2014). Willingness to adopt e-procurement to reduce corruption: Results of the PLS Path modeling. Transforming Government: People, Process and Policy, 8(4), 500-520. https://doi.org/10.1108/TG-03-2014-0007

OCDE. (2015). Informe de la OCDE sobre el Soborno Internacional. OECD Publishing. https://doi.org/10.1787/9789264226654-es

OECD. (2019). Government at a Glance 2019. OECD. https://doi.org/10.1787/8ccf5c38-en

Rabuzin, K., & Modrušan, N. (2019). Prediction of Public Procurement Corruption Indices using Machine Learning Methods: Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 333-340. https://doi.org/10.5220/0008353603330340

Sales, L. J., & Carvalho, R. N. (2016). Measuring the Risk of Public Contracts Using Bayesian Classifiers. Proceedings of the 13th UAI Bayesian Modeling Applications Workshop (BMAW 2016), 7-59.

Soreide, T. (2002). Corruption in public procurement: Causes, consequences and cures. CMI.

Torres-Berru, Y., & López Batista, V. F. (2021). Data Mining to Identify Anomalies in Public Procurement Rating Parameters. Electronics, 10(22), 2873. https://doi.org/10.3390/electronics10222873

Williams-Elegbe, S. (2018). Systemic corruption and public procurement in developing countries: Are there any solutions? Journal of Public Procurement, 18(2), 131-147. https://doi.org/10.1108/JOPP-06-2018-009

World Bank. (2022). Using Data Analytics in Public Procurement Operational Options and a Guiding FrameworkEquitable Growth, Finance & Institutions Insight. World Bank. https://doi.org/10.1596/37467

Zuleta, M. M., Ospina, S., & Caro, C. A. (2019). Índice de riesgo de corrupción en el sistema de compra pública colombiano a partir de una metodología desarrollada por el Instituto Mexicano para la Competitividad. http://www.repository.fedesarrollo.org.co/handle/11445/3872