PREDICIENDO LA CORRUPCIÓN EN LA CONTRATACIÓN PÚBLICA COLOMBIANA

Autor:
Hernan Rodrigo Garcia Fuentes

Director:
Jacobo Alberto Campo Robledo

Trabajo presentado como requisito para optar por el título de Magíster en Economía de las Políticas Públicas

Maestría en Economía de las Políticas Públicas
Facultad de Economía
Universidad del Rosario

Bogotá, Colombia
2026

PREDICIENDO LA CORRUPCIÓN EN LA CONTRATACIÓN PÚBLICA COLOMBIANA

Hernan Rodrigo Garcia Fuentes[1]

Resumen

A pesar de los constantes esfuerzos gubernamentales, el sector público colombiano sigue considerándose uno de los más permeados por la corrupción. Esto de acuerdo con los últimos reportes de Transparencia por Colombia, los cuales muestran que esta actividad se sigue percibiendo como una de las causas de la baja eficiencia en la gestión del Estado, el apoderamiento indebido de recursos públicos, el clientelismo, entre otros, mientras que, alternativas como las auditorías, sanciones y reformas legales han demostrado ser insuficientes para detener este fenómeno. Por ello, la presente investigación propone una solución diferente: usar inteligencia artificial para anticipar la corrupción antes de que ocurra. A través del análisis de más de 489.000 contratos registrados en la plataforma digital del Estado colombiano (SECOP), se entrenaron modelos de aprendizaje automático capaces de identificar riesgos en contratos públicos que pueden terminar en multas o sanciones. Estos resultados, permiten crear un sistema de alertas tempranas que clasifica cada contrato en niveles de riesgo (Bajo, Medio, Alto o Crítico), permitiendo a las entidades públicas y organismos de control enfocar sus recursos donde más se necesitan. Con esto se busca dar respuesta a la pregunta: ¿Pueden los modelos basados en *Machine Learning* proporcionar herramientas para la prevención de corrupción en la contratación pública? Sin embargo, esta herramienta no reemplaza al auditor humano: lo hace más eficiente. En un país donde Colombia ocupó el puesto 99 de 182 en el Índice de Percepción de Corrupción[2] (IPC) 2025, esta propuesta representa un paso concreto hacia una gestión pública más inteligente.

Abstract

The Colombian public sector continues to be one of the most corrupt despite the government's attempts. Transparency International Colombia's recent findings point out that corruption continues to be seen as a cause of inefficiency in the operation of governments, theft of public resources, clientelism and other problems. In the meanwhile, audits, punishments and legal amendments as alternatives have failed to stop this behavior. Hence, this research provides an alternative method, that is, to use artificial intelligence to predict corruption before it occurs. Based on research of more than 489 thousand contracts registered in the digital platform of the Colombian government (SECOP), machine learning models were trained to identify hazards in public contracts that could generate penalties or sanctions. These results enable the construction of an early warning system that classifies each contract into risk levels (Low, Medium, High or Critical), thereby allowing public entities and oversight bodies to concentrate their efforts in the areas where they are most needed. This is a solution to the question: Can machine learning-based models give instruments for corruption prevention in the public procurement? But this gadget does not replace the human auditor, it just makes him more efficient. This suggestion is a concrete step towards a smarter public management, in a country where Colombia occupies the 99th place out of 182 in the Corruption Perceptions Index (2025).

Palabras Clave: Contratación Pública, Corrupción, Prevención, Machine Learning.
Clasificación JEL: D73, H11, H30.

[1]Facultad de Economía, Universidad del Rosario, Email: hernanr.garcia@urosario.edu.co

[2] Este indicador califica los niveles percibidos de corrupción en el sector público de cada país, de acuerdo con las opiniones de expertos y empresarios.

1 Introducción

En Colombia, desde que se implementó del estatuto anticorrupción[3], que busca prevenir y controlar la corrupción, los distintos gobiernos y organizaciones públicas y privadas han intentado crear diferentes formas de combatir este problema. Sin embargo, García, et al. (2022) señala que la corrupción en Colombia aumentó significativamente, prueba de ello son las más de 57.000 denuncias relacionadas con corrupción en el Estado presentadas entre 2010 y 2023. Ahora bien, en lo que a contratación pública se refiere, los contratos celebrados sin cumplir requisitos legales y los contratos indebidamente adjudicados representan cerca del 32% de todos los delitos registrados. Mostrando que, a pesar de leyes, auditorías y reformas institucionales, el problema persiste, dando como resultado el puesto actual que ocupa el país en el IPC para el año 2025[4].

Ahora bien, diferentes estudios muestran que la corrupción puede generar desde un bajo crecimiento económico (Fernand Desfrancois & Pastás Gutiérrez, 2022), una afectación en la participación democrática (Olsson, 2014); hasta, una disminución en la colaboración entre entidades públicas y privadas que se vean implicadas en eventos de corrupción (Colonnelli, et al., 2021). Y la afectación en la gestión por parte de los gobiernos y la eficiencia en sus regulaciones (Olken & Pande, 2012).

Teniendo en cuenta estos efectos y, partiendo desde una perspectiva del sector público[5], tanto los hacedores de políticas públicas como investigadores independientes y demás organizaciones estatales y no estatales han generado alternativas para frenar esta actividad. Las cuales van desde la implementación de auditorías (Olken, 2007), desincentivar factores culturales relacionados con la corrupción (Varvarigos, 2023) y fortalecer las condenas y sanciones hacia los participantes en estos actos (Colonnelli, et al., 2021), entre otras como, la implementación de herramientas digitales dirigidas a brindar transparencia y seguridad, especialmente frente a actos de corrupción, las cuales han servido para generar control no solo por medio de entidades públicas, sino también por parte de otros agentes.

Como una prueba de lo mencionado anteriormente, en los últimos años han aparecido estrategias que se centran en el uso de extensas fuentes de datos y la implementación de modelos de machine learning. Por ejemplo, los gobiernos de China y Noruega han desarrollado modelos basados en *IA* y *Machine Learning* para identificar actos de corrupción en licitaciones públicas (China), como también medir posibles riesgos de corrupción al momento de invertir en fondos privados dineros públicos (Noruega). En el contexto de América Latina, Brasil usa su herramienta Analisador de Licitações e Editais (ALICE) para enviar alertas automáticas y evitar riesgos de corrupción antes de firmar un contrato. Esto hace que las auditorias de la Contraloría brasileña sean más efectivas en estos casos. Asimismo, el CAF – Banco de Desarrollo de América Latina (2020) expone cuáles pueden ser los principales datos a usar por parte de los diferentes gobiernos para la lucha anticorrupción, pone como uno de los ejemplos la plataforma OCEANO desarollada por la Contraloría General de la República de Colombia. Esta plataforma usa una combinación de datos a partir del SECOP y del Registro Único Empresarial y Social (RUES) para detectar irregularidades (en tiempo real) en la contratación pública.

[3] Creada bajo la Ley 1474 de 2011

[4] Último año de elaboración y publicación.

[5] Esto, por ser uno de los sectores más afectados por esta actividad según (Transparencia por Colombia, 2022b).

En cuanto a lo académico, estas nuevas metodologías se muestran en las investigaciones realizadas por (Colonnelli, et al., 2020), donde determinan que el efecto de una ofensiva nacional contra la corrupción afecta la práctica de esta actividad entre los gobiernos locales y el sector privado en algunas ciudades de Brasil. Asimismo, Iturriaga & Sanz (2018) encuentran que los factores económicos como los bienes raíces, el crecimiento económico, el número de empresas no financieras, entre otros, utilizando redes neuronales, Gallego et al. (2022) logra predecir, mediante el uso de modelos de aprendizaje automático, las malas conductas de los alcaldes en los municipios de Colombia. Mientras que Gallego et al. (2021) logra establecer un modelo predictivo para definir cuáles contratos públicos tienen riesgo de problemas.

El presente documento aprovecha el avance sobre las técnicas de aprendizaje automático, debido a su sencilla adaptación y manejo sobre extensos conjuntos de datos. Y, de esta forma, proveer de herramientas que, al sector público, le permitan mejorar su capacidad de prevención, reacción y control frente a la corrupción. Asimismo, como forma de contribución a estos avances, el documento se enfoca en la contratación pública colombiana, esto se debe a que, de acuerdo con Transparencia por Colombia (2022), la contratación pública es uno de los mecanismos de la gestión estatal que más se ve afectada por la corrupción. Se desarrolla un modelo de aprendizaje automático utilizando las características de 489.580 contratos que se publicaron entre 2010 y 2021 en la plataforma SECOP II, junto con los datos sociales y económicos de las regiones donde se llevaron a cabo. Para ello, se elabora una revisión de los modelos de aprendizaje automático más usados dentro de la literatura actual. Estos son: el modelo lineal Lasso, elaborado por (Hastie, Tibshirani, & Wainwright, 2015), y el modelo de Random Forest, este último fue presentado por James et al. (2023). Los resultados obtenidos en el presente documento reflejan un desempeño favorable en términos de AUC, lo cual sugiere una adecuada capacidad de discriminación entre contratos con y sin riesgo. Sin embargo, los niveles de precisión obtenidos reflejan limitaciones en cuanto a la clasificación exacta de los eventos, asociadas principalmente al desbalance que se presenta en la variable resultado, lo cual es común en fenómenos de baja frecuencia como la corrupción.

A pesar de los resultados, lejos de construir una limitación de los modelos, estos hallazgos dan muestra de la complejidad existente al momento de querer medir y predecir eventos como la corrupción y abre el espacio para el desarrollo de enfoques complementarios. Por lo tanto, el documento propone interpretar estos modelos no como una herramienta de clasificación determinística, sino como una herramienta de apoyo para la priorización de riego, que permita orientar de una mejor manera los esfuerzos en cuanto a control y supervisión de los contratos.

La literatura reciente ha avanzado bastante en el uso de métodos de aprendizaje automático y *Machine Learning* para detectar riesgos de corrupción. Por ejemplo, Gallego et al. (2020) han clasificado los municipios más vulnerables a la corrupción durante la pandemia del COVID-19. Asimismo, en un contexto acorde a este trabajo, (Decarolis & Giorgiantonio, 2020) y (Gallego, et al., 2021) llevan a cabo investigaciones frente a los contratos tanto en Italia como en Colombia. De igual forma, (Mojica Muñoz, 2021) muestra, desde un componente metodológico, cómo el uso de herramientas basadas en *Machine Learning* puede determinar qué municipios de Colombia son más riesgosos, en cuanto a corrupción, al momento de realizar un contrato. Zuleta et al. (2019), basándose en la idea de que la

corrupción deja marcas o patrones visibles, crean un índice de riesgo para la contratación pública en Colombia. Este índice se basa en varias características de los procesos de contratación que analizaron.

Este documento señala que, en la metodología utilizada, hay un desbalance en la variable resultado. Esto se debe a que hay muy pocos contratos con sanciones o multas en comparación con el total de la muestra. Se utiliza el método llamado *SubBagging,* propuesto por (Moreno Pabón, 2018), que corrige los sesgos que pueden ocurrir, mejorando así la capacidad del modelo para responder a eventos raros o inusuales. En cuanto al análisis de los resultados obtenidos, al emplear la metodología implementada por (Breiman, 2001), se logra identificar cuáles fueron las principales variables utilizadas dentro de cada modelo, lo cual permitió entender de una mejor manera cómo pueden ser afectadas estas variables por la corrupción, esto teniendo en cuenta los hechos históricos presentados en el país en relación con la contratación pública.

Bajo la lógica mencionada anteriormente, la presente investigación introduce aportes tanto en términos de datos como en estrategia empírica. Esto, teniendo en cuenta que, se crea una base de datos propia, la cual es el resultado de un proceso organizado de integración, validación y organización de la información contractual a un nivel detallado, añadiendo eventos o características relacionadas con el ciclo de vida del contrato para obtener más información predictiva y superar las limitaciones que pueden existir en las bases estandarizadas existentes. De igual manera, se elabora un sistema de alertas a partir de la probabilidad estimada y transformada a una escala de 0 a 100 para facilitar su interpretación, así como una aproximación al impacto económico mediante la estimación de perdidas esperadas. Dicho lo anterior, la investigación no solo aporta desde una perspectiva metodología, sino que plantea una aplicación práctica en el contexto de la gestión pública, dando un primer paso en la incorporación de herramientas basadas en *ML* para la lucha contra la corrupción.

La presente investigación se encuentra dividida en 5 secciones, de las cuales la presente introducción es la primera. En la segunda sección se describe la situación actual desde la literatura de la corrupción en el sector público, haciendo énfasis especialmente en la contratación. Se dan también a conocer las fuentes y los datos con que se trabajó para hacer las estimaciones de los modelos. Seguidamente, la tercera sección expone la metodología implementada para abordar la pregunta planteada y la estrategia empírica utilizada para su realización, teniendo en cuenta las ventajas y desventajas de cada modelo implementado. Los resultados obtenidos a partir de las estimaciones realizadas se exponen en la sección cuarta, donde se evalúan las principales variables obtenidas desde la literatura y los hechos históricos en Colombia. Por último, la sección cinco expone las consideraciones finales y una posible agenda de investigación.

2 Contexto y Datos

La presente sección se divide en dos principales apartados; en primer lugar, se busca establecer la situación actual de la corrupción en el sector público, especialmente en la contratación pública, dentro de la literatura económica, social y jurídica tanto colombiana como internacional. Asimismo, se describe de una manera más detallada y rigurosa la estructura y alcance de la base de datos a implementar para las predicciones. Dicho lo anterior, la presente sección se dividirá en dos subsecciones complementarias. En primer lugar, el *contexto*, donde se sintetiza una revisión literaria de los efectos de la corrupción sobre la economía y el funcionamiento institucional de las entidades públicas. De igual manera se hace una reseña del Sistema Electrónico de Contratación Pública (SECOP) y su aporte en la lucha contra la corrupción en Colombia, estableciendo así el fundamento teórico e institucional de este tipo de actividad. En la segunda parte, los *datos*, donde se describe el universo contractual observado, las fuentes consultadas, el proceso de construcción y depuración de la base y las variables empleadas en las predicciones, mostrando estadísticas descriptivas, lo que refuerza el contexto desde un punto de vista cuantitativo. En esta sección se busca vincular las ideas con la evidencia práctica que justifica el uso de modelos econométricos y de aprendizaje automático para el estudio de la corrupción en la contratación pública.

2.1 Contexto

Los efectos de la corrupción en el sector público colombiano presentan diferentes efectos dentro del país. Por ejemplo, Galvis - Ciro & Hicapié - Vélez (2022) muestran que, en la distribución del gasto público, la salud y la infraestructura son los rubros más afectados en los departamentos con mayor índice de corrupción en el país. Asimismo, Ortiz Benavides (2012) sugiere que altos niveles de corrupción en Colombia reducen la cobertura de la educación secundaria y presentan un efecto sobre la mortalidad infantil. En cuanto a la contratación pública, aportando a lo mencionado anteriormente, (Martínez Cárdenas & Ramírez Mora, 2006) discuten cómo este tipo de actividades dispersa la responsabilidad y dificulta el control sobre la ejecución de los contratos. Se habla del modelo principal-agente-cliente para explicar las causas de la corrupción y otros factores que pueden contribuir a ella, como el monopolio del Estado, la baja probabilidad de ser atrapado, las sanciones débiles, los bajos salarios en el sector público, la falta de condena moral y la burocracia clientelista. Manteniendo el punto de vista jurídico sobre la contratación pública, Bastidas Vargas (2023) sugiere que establecer procedimientos claros frente a los procesos de contratación y generar un acceso abierto a la ciudadanía puede provocar menores niveles de corrupción en la contratación pública. De igual manera, argumenta que, a pesar de la existencia de la plataforma SECOP, aún se deben robustecer los mecanismos de monitoreo que respondan a las prácticas comúnmente empleadas por las redes de corrupción. Entretanto, (Transparencia por Colombia, 2022c) menciona que la contratación pública es una de las actividades del sector en las que más se presenta este tipo de práctica, donde la falta de idoneidad de contratistas, sobrecostos de adquisición en los contratos, el uso indebido de las excepcionalidades contractuales para celebrar contratos y la financiación de campañas políticas son de los principales factores que hacen de la contratación pública la segunda actividad con mayores niveles de corrupción en el país. Según datos de la Secretaría de Transparencia de la Presidencia de la República de Colombia, entre el 2010 y el 2023 se han presentado 57.582 denuncias relacionadas con corrupción. Los delitos por contratos sin

cumplimiento de requisitos legales son uno de los problemas más graves. 24,6% del porcentaje total, y la celebración indebida de contratos con un 7,4%, ocupando el segundo y el quinto lugar, respectivamente, en los delitos más frecuentes.

Ahora bien, la contratación pública colombiana está regida bajo la Ley 80 de 1993, la cual gesta la normativa y establece parámetros de esta actividad; en ella dictan los derechos y deberes tanto de la entidad contratante como del contratista, se generan los criterios de selección, modalidades y demás aspectos jurídicos, técnicos y financieros para elaborar un contrato. De igual manera, todos estos aspectos se agrupan en fases o etapas en las que intervienen diferentes actores. Asimismo, la Ley 1150 de 2007, con la cual se crea en Colombia el Sistema Electrónico de Contratación Pública (SECOP). Puede tomarse como un complemento de la Ley anteriormente mencionada, ya que es en este sistema donde se hace pública la información relacionada con los procesos contractuales de las diferentes entidades públicas del país, teniendo como objetivos principales la eficiencia, la transparencia, la participación ciudadana y el control social en cada uno de los procesos contractuales.

Tabla 1. Etapas de la Contratación Pública en Colombia

Precontractual	Contractual	Postcontractual
- **Planeación:** Se define la necesidad de la entidad y se elaboran los documentos soporte que la respaldan[6]. - **Selección:** Publicación en la plataforma SECOP, evaluación de ofertas y selección de proveedor - **Adjudicación:** Firma del contrato	- **Perfeccionamiento:** Expedición de garantías (pólizas) y registros presupuestales. - **Ejecución:** Inicio de ejecución contractual. - **Supervisión:** Vigilancia del cumplimiento técnico, financiero, administrativo y jurídico del contrato	- **Liquidación**

Fuente: Elaboración propia.

El funcionamiento de esta plataforma consiste en el diligenciamiento de ciertos campos, los cuales se ajustan según la etapa en la que se encuentre el proceso contractual. Asimismo, dentro de la plataforma se da publicación de los documentos requeridos para la participación de las empresas interesadas; cabe aclarar que, al ser una plataforma pública y de libre acceso, cualquier persona puede acceder a estas oportunidades. De esta manera, una vez publicado el proceso y los interesados presentando sus propuestas, empieza el procedimiento de selección del proveedor de acuerdo con las condiciones técnicas, financieras y jurídicas estipuladas. Seguidamente, el contrato se firma dentro de la misma plataforma para continuar con la ejecución: allí se cargan informes, pagos y demás documentos necesarios durante la vigencia del contrato y, finalmente, se realiza la liquidación una vez culminada el plazo de ejecución y cumplidas las obligaciones. SECOP permite tener una mejor trazabilidad del

[6] Estudios previos, presupuesto, riesgos y pliegos de condiciones

contrato desde su inicio hasta la terminación de este, como también un mayor control por parte de las entidades encargadas de vigilar la gestión pública y los ciudadanos interesados o veedores. La Tabla 1 ilustra las etapas del proceso contractual colombiano a partir de las leyes mencionadas anteriormente, que, para la presente investigación, se concentra en la etapa contractual, entendiendo que es la etapa donde se desarrolla el contrato, se ejecuta la necesidad de la entidad y es una de las etapas donde más actores, tanto internos como externos, intervienen, lo que genera una mayor información del comportamiento de los contratos.

De igual manera, existen otras iniciativas como el Programa Interamericano de Datos Abiertos (PIDA), estructurado por la Organización de los Estados Americanos (OEA). Las páginas web de gestión de datos, como el Portal Anticorrupción de Colombia (PACO), son lideradas por el Ministerio de las TIC y divulgan información del sector público. Estas iniciativas han buscado, mediante el uso y la divulgación de datos públicos, generar mayor control tanto de los responsables de las políticas públicas como de las entidades regulatorias y de la ciudadanía en general.

2.2 Datos

Desde la creación del SECOP[7] en el 2007, las entidades públicas colombianas, a la fecha, han registrado un poco más de 12 millones de contratos a nivel nacional, bajo distintas modalidades contractuales y con valores que van desde los cero (0) pesos hasta los miles de millones[8]. Igualmente, como las ejecuciones de los contratos se realizan en diferentes puntos del país, estos contratos han sido diseñados teniendo en cuenta las condiciones socioeconómicas de cada departamento, por lo cual entender estas características es importante para establecer las necesidades de las entidades. Con el fin de entender, desde los datos, el comportamiento de la corrupción y la contratación pública en las diferentes regiones y ciudades del país, se presentan una serie de estadísticas descriptivas de los datos utilizados para la presente investigación.

2.2.1 Variables de Corrupción

En esta investigación, se toma como variable resultado las multas y sanciones que imponen las entidades públicas contratantes a los contratos[9], por parte de las entidades públicas contratantes. Entendiendo que este tipo de contravenciones puede ser considerado como acto de corrupción, dadas las leyes colombianas como la Ley 2195 de 2022 y conceptos como el presentado por (Morgner & Chêne, 2015), donde especifican que "*la corrupción en la contratación pública puede afectar todas las etapas del proceso de contratación y adoptar diversas formas*". Ahora bien, esta es una variable dicotómica, la cual toma el valor de 1 si el contrato ha sido objeto de alguna sanción o multa, y 0 en caso contrario. La Figura 1 muestra la evolución de las sanciones contractuales en cuanto a valor y monto, en los departamentos colombianos entre el 2010 y el 2021.

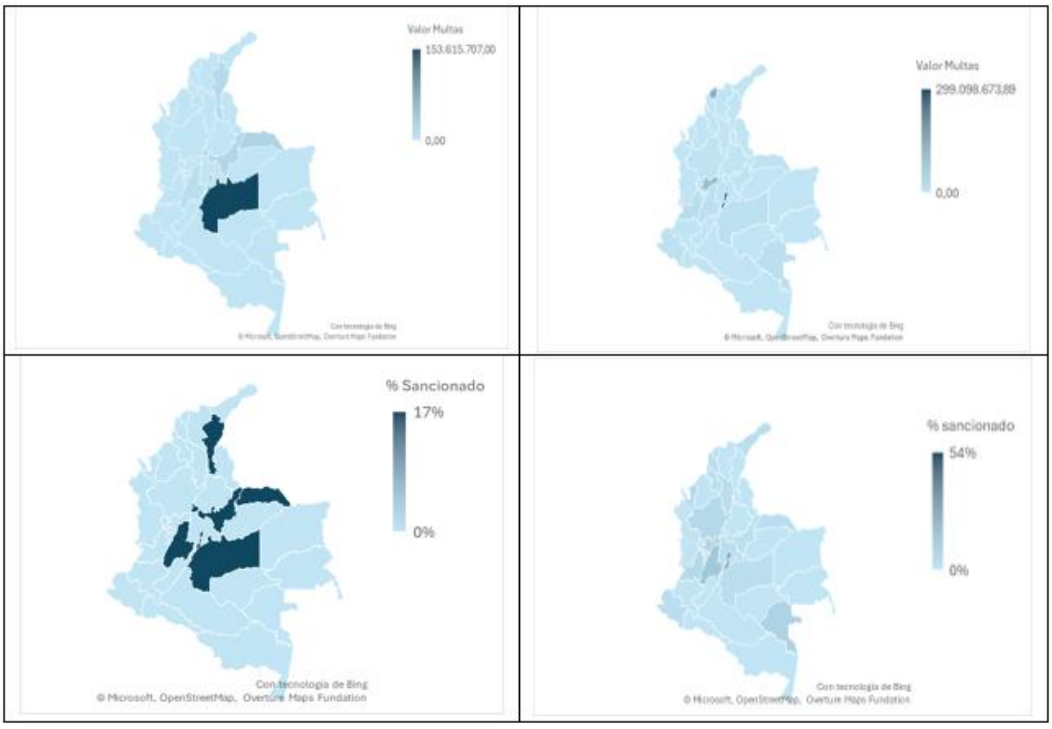
[7] A la fecha SECOP se divide en dos plataformas transaccionales SECOP I y SECOP II. Siendo que la primera registra los procesos a partir del 2011 y la segunda entra en función desde el 2015.

[8] El Acuerdo Marco de Medicamentos realizado en el 2022 presentó una adjudicación total por más de 2,2 billones de pesos.

[9] Esto, teniendo en cuenta los proveedores pueden aparecer en diferentes contratos lo que podría capturar patrones y no identificar observaciones especificas sin ser capaz de identificar otras características a partir de nuevos proveedores (Esta idea se desarrolla más adelante).

La figura mencionada muestra que la concentración, tanto en valor como en requerimientos de las multas y sanciones, pasó de estar concentrada en ciertos departamentos a una distribución más uniforme. Aunque, a pesar de esta redistribución, el aumento de esta actividad es considerable entre un año y el otro. Pasa de un máximo de multa de $153 millones, concentrando el 17% de las multas y sanciones en el 2010, a $299 millones y el 54% de concentración para el 2021. Estos valores están registrados en la ciudad de Bogotá, lo que muestra una importancia de la capital colombiana en este tipo de actividades.

Figura 1. Sanciones contractuales en departamentos colombianos (2010 y 2021).



Fuente: Elaboración propia. Con datos tomados de SECOP

2.2.2 Variables Contractuales

Para esta investigación, se tomó una muestra aleatoria de 489.580 contratos registrados entre el 2010 y el 2021 en la plataforma SECOP. Asimismo, para la selección de estos se tomó en consideración el estado de cada uno de ellos al momento de descargar la base de datos. Cada uno de los contratos seleccionados contiene un total de 20 variables que recogen las características registradas de cada uno de estos.

De acuerdo con la Gráfica 1, al analizar los datos, se logró evidenciar que la mayoría de los contratos (247.493) se celebraron en la ciudad de Bogotá, representando el 51% del total. Esta situación podría explicarse por el hecho de que la mayoría de las entidades del orden nacional tienen su sede principal en la ciudad de Bogotá y, al momento de configurar el contrato dentro de la plataforma establecen a esta ciudad como el lugar de ejecución contractual. Lo que dificulta identificar cuáles son las características que influyen en el contrato; sin embargo, para este caso se determinó tomar a la ciudad de Bogotá como lugar de ejecución.

La Gráfica 2 muestra una tendencia distinta a la mencionada antes. Al ver el valor total de los contratos en relación con el PIB del departamento para el 2021 (el último año analizado), se puede evaluar la importancia de la contratación estatal en la economía local. En este sentido, el departamento de Putumayo[10] con un 82% presenta una alta dependencia del gasto público en su economía, seguido por Bogotá con un 22%, lo que refuerza lo mencionado anteriormente en cuanto a la configuración del lugar de ejecución. y dejando a Valle del Cauca como último del mismo. Cabe resaltar que este comportamiento, según (Consejo Nacional de Política Económica y Social (CONPES), 2002) y (Galvis - Ciro & Hicapié - Vélez, 2022), da muestra de la relevancia de esta actividad no solo dentro de la gestión pública, sino también dentro de la economía local de un departamento.

2.2.3 Variables Departamentales

Teniendo en cuenta que la corrupción puede provenir de diferentes actores, para el presente trabajo, se toma en consideración 163 características de los departamentos en donde se ejecuta o se ejecutó el contrato. Estas variables se agrupan en 7 grupos compuestos de la siguiente manera: En primer lugar, se toman variables socioeconómicas; seguidamente, las relacionadas con delincuencia, otras al combate armado, recursos humanos, actividades ilegales, gestión pública y, por último, las financieras. Con estas variables se busca analizar la posibilidad de identificar factores externos que puedan influir en la contratación pública.

Tabla 2. Variables Explicativa

Sector	Fuente de Información	Grupo Variable	N° Variables
Información Contratos	SECOP I y II (Datos Abiertos)	Contractuales	20
Desarrollo Social	Departamento Administrativo Nacional de Estadística (DANE) / Global Data Lab	Socioeconómicas	13
Violencia	Ministerio de Defensa Nacional	Delincuencia	11
Conflicto Armado	Unidad de Víctimas / Agencia para la Reincorporación y la Normalización / Prosperidad Social / Ministerio de Defensa	Combate Armado	11
Capital Humano	Ministerio de Educación	Recurso Humano	16
Actividad ilegal	Ministerio de Justicia / Policía Nacional	Actividades Ilegales	4
Función Pública	Registraduría Nacional del Estado Civil (RNEC) / Departamento Nacional de Planeación	Gestión Pública	10
Comportamiento Financiero	Superintendencia Financiera de Colombia	Variables Financieras	78

Fuente: Elaboración propia.

3 Metodología

En esta sección se muestra la contribución realizada por la presente investigación, desde un punto de vista metodológico y práctico, con el fin de mejorar la comprensión existente entre corrupción y contratación pública. Este documento sugiere usar herramientas de *aprendizaje automático* para detectar características o patrones que ayuden a identificar qué contratos son más propensos a problemas de corrupción y crear un sistema de alertas basados en las

[10] Esto se debe en gran medida al contrato APP 0122015 celebrado entre la AGENCIA NACIONAL DE INFRAESTRUCTURA ANI y la Concesionario ALIADAS PARA EL PROGRESO SAS por valor de $2.969.581.000.000 COP

probabilidades que generan los modelos. En esta primera parte de la sección, se presentan los modelos utilizados para las estimaciones: el modelo lineal Lasso y el Random Forest. Ambos modelos son útiles tanto para realizar predicciones como para seleccionar características. De igual manera, una serie de autores dan muestra de ventajas para cada uno de los modelos, las cuales van desde la selección de variables (Hastie, Tibshirani, & Wainwright, 2015), la reducción de posibles sobreajustes del modelo, esto al momento de generar penalizaciones a los coeficientes estimados Hastie et al. (2009) y su fácil interpretabilidad James et al. (2013). Esto, en relación con el modelo Lasso, mientras que, en lo referido al Random Forest, presenta una reducción de la varianza al momento de combinar varios árboles de selección (Breiman, 2001), robustez ante posibles ruidos de los datos (Hastie, et al., 2009) y permite generar técnicas para el balanceo de estos (Fernandez, et al., 2018).

En segundo lugar, la sección aborda algunos posibles percances que se pueden presentar al momento de las estimaciones, como la clasificación de variables. Esto se soluciona al enseñarle a la máquina (mediante algoritmos) cuándo el contrato presenta multas o sanciones y cuándo no, teniendo en cuenta las características de las variables presentadas en cada caso.

Además, como la base tiene pocos datos positivos (contratos con multas o sanciones), esto puede causar sesgos en las estimaciones, resultando en coeficientes más bajos, lo que podría influir en su capacidad para predecir (Tibshirani, 1996). Se implementa el método *SubBagging,* el cual fue usado por (Moreno Pabón, 2018) en su estudio de la propagación de cultivos de coca en Colombia y que se explicará en los siguientes apartados como una posible solución al sesgo que se presenta.

3.1 Modelo Lasso

Este modelo es similar a una regresión OLS a la que se le incorpora un sistema de penalización, con el fin de seleccionar los mejores estimadores de acuerdo con este parámetro. Este método es mayormente utilizado en situaciones con un número reducido de casos de entrenamiento y en situaciones de datos escasos.

A continuación, se presenta la especificación del modelo:

Partiendo de querer predecir un valor real de y_i obtenemos lo siguiente:

$$y_i = \beta_0 + \sum_{j=1}^{p} X_i \beta_j \tag{1}$$

Sujeto a:

$$\min \left(y - \left(\beta_0 + \sum x_i \beta_j \right) \right)^2 + \lambda \sum |\beta_j| \tag{2}$$

De acuerdo con lo anterior, la ecuación (1) se define como el modelo Lasso, por lo cual, al querer denotar y_i, esta se encuentra sujeta a una restricción o penalización, ecuación (2). Para este caso, se determinó λ por un procedimiento de validación cruzada que, tomando el conjunto de datos de entrenamiento, se hace una partición de estos, que para este caso es a 5 partes uniformes, con el fin de entrenar 4 de estas particiones y evaluarlas o validarlas en la restante. Este ejercicio se realiza 5 veces hasta que todas las partes han sido evaluadas; cabe resaltar que este proceso no a menudo produce los mismos resultados, toda vez que al hacer

diferentes particiones esto genera una aleatorización de las bases. Ahora bien, este proceso se repite 10 veces más, probando de esta manera varios posibles valores de λ los cuales son evaluados con las métricas mencionadas anteriormente y así tomar el mejor lambda. Cabe resaltar que la presente validación se toma de (Gallego, et al., 2022) y (Gallego, et al., 2021). Al generar las estimaciones, aleatorizan los datos de entrenamiento para obtener distintos valores de λ, conservando los que arrojan mejores resultados para las predicciones y evaluaciones correspondientes.

3.2 Modelo de Bosques Aleatorios

Este modelo se basa en un método de aprendizaje supervisado[11] que usa varios árboles de decisión para hacer predicciones. Se emplea como un sistema de votación para la clasificación. Ahora bien, para la creación de los árboles de decisión Breiman (2001) establecen que estos deben ser construidos a partir de: 1. La selección aleatoria de datos de entrenamiento y 2. La selección aleatoria de un subconjunto de variables. Con esto, se busca reducir la varianza y mejorar la estabilidad del modelo. Para la aleatorización de datos, estos modelos usan la técnica *Bagging,* que fue propuesta por Breiman (1996). Esta técnica consiste en hacer varios subconjuntos aleatorios del mismo tamaño, donde una o varias variables pueden estar presentes o no en diferentes subconjuntos. Seguidamente, se entrena un modelo base en cada subconjunto con el fin de seleccionar el o los mejores predictores a partir de una votación mayoritaria entre los modelos.

Dicho lo anterior, la especificación del modelo se puede presentar de la siguiente manera:

$$\hat{y} = f(x) + \epsilon \tag{3}$$

Donde:

$$f(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{4}$$

De la ecuación (3) se establece que $\hat{y}$ es la variable objetivo, mientras que, $f(x)$ es la función de bosques aleatorios, donde, T es el número de árboles utilizados para la predicción, mientras que, h_t es la predicción del árbol t y x las variables utilizadas. Ahora bien, a pesar de que este tipo de modelos no cuenta con unas condiciones o supuestos iniciales como un modelo lineal o regresión logística, el modelo de bosques aleatorios adquiere ciertas características claves para presentar predicciones robustas. Entre ellas se encuentran: la no existencia de un sobreajuste al aumentar el número de árboles[12]. La precisión dependerá de la capacidad predictiva de cada árbol y la correlación entre ellos (a menor correlación, mejor precisión). Además, tolera de mejor manera las etiquetas erróneas, permite medir la importancia de las variables dentro de cada modelo. Estas permiten que estos tipos de modelo se puedan ajustar a las necesidades del presente documento, dado que permitirán establecer qué variables pueden afectar a la corrupción contractual.

[11] Este método consiste en el entrenamiento de un modelo a partir de ejemplos de pregunta y respuesta. Esto con el fin de que aprenda a contestar al momento de recibir nuevas preguntas.
[12] Esto se debe a la ley de grandes números.

3.3 Consideraciones de los modelos

Ahora bien, en la base original propuesta se incluyen variables relacionadas con el NIT tanto de la entidad contratante como del proveedor, para cada uno de los contratos analizados. Lo cual, partiendo de lo mencionado por James et al. (2013) y al considerar estas variables como únicas, puede provocar que los modelos, en vez de generalizar características a partir de los eventos considerados como corruptos, memoricen estos sucesos a partir de las particularidades de individuos únicos, lo que conllevaría problemas como *Leakage* o fuga de información[13].

Dicho lo anterior, para la presente investigación, se realizó la comparación entre dos bases (una teniendo en cuenta la totalidad de las variables y otra excluyendo a estas con valores únicos), esto con el fin de comparar la posibilidad de generar una memoria histórica vs. la capacidad estructural de predicción, con base en las variables que dominen a cada uno de los modelos y así establecer cuáles son las que logran anticipar de mejor manera la corrupción.

Otro aspecto que la presente investigación toma en cuenta es la consideración del uso de los años como apoyo en la predicción y su uso dentro del Training Set y el Test Set. Esto, entendiendo que, a pesar de que autores como James et al. (2023) asumen independencia entre observaciones, por lo cual, al momento de realizar una partición aleatoria entre training y test, se mantiene esta independencia. Sin embargo, con el fin de implementar un esquema de validación temporal donde se garantice que el entrenamiento se haga con contratos de años pasados y las pruebas con los años futuros, en la presente investigación tomo lo realizado por (Mojica Muñoz, 2021), en cuanto al uso de los datos de años pasados para el training test (2010-2020) y futuros (2021) para test set, en todos los modelos a analizar y así comparar el uso e implementación de estas variables. La Figura 2 muestra de una manera visual el paso a paso metodológico realizado.

3.4 Método SubBagging

Como se mencionó antes, la metodología incluye técnicas y modelos para equilibrar los datos, ya que hay poca cantidad de contratos que mostraron cambios relacionados con multas o sanciones. Por ello, se reconoce que, al presentar estas particularidades dentro del set de datos, se pueden presentar posibles problemas de clasificación y predicción de estas. Esto, teniendo en cuenta que los modelos pueden no detectar qué casos son más significativos que otros (debido a su poca frecuencia) y con ello pueden llegar a estimar casos erróneamente "falsos negativos", generando inestabilidad en los modelos.

La técnica *SubBagging*, la cual fue expuesta por Zaman y Hirose (2009), consiste en crear muestras balanceadas tomando, para este caso, el 70% de los valores positivos y una muestra ***N*** aleatoria de los negativos.

Con este método se busca validar un tamaño óptimo de ***N***, lo que podría conllevar una pérdida de los datos obtenidos, pero que garantice las mejores predicciones con una varianza mínima.

[13] Teniendo en cuenta que en el evento de ingresar una nueva entidad pública o empresa proveedora, el modelo no los reconocería dado que los nuevos valores ingresados, los cuales no son iguales a los memorizados durante el training del modelo.

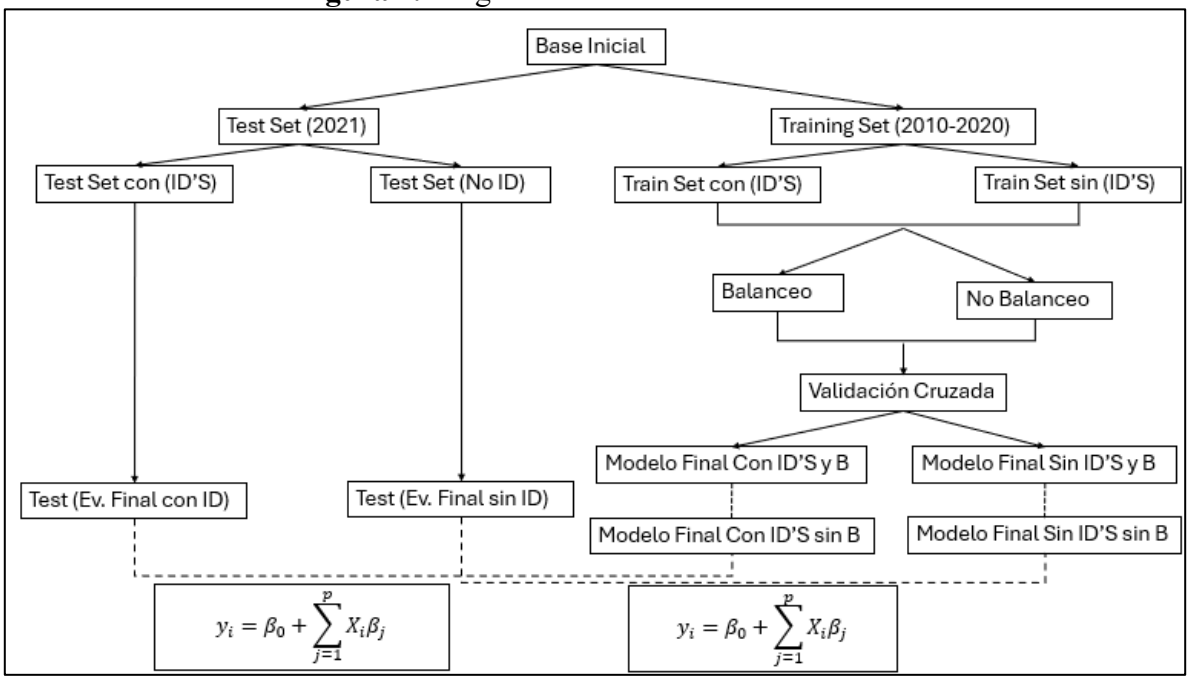
Teniendo en cuenta lo anterior, para cada uno de los modelos se tomó en consideración los siguientes parámetros, los cuales sirvieron para realizar las estimaciones y evaluar los desempeños de estos.

Tabla 2. Parámetros de los Modelos.

Modelo	Parámetro	Valor
Lasso	λ Umbral de Decisión	100 0.2
Bosques Aleatorios	N° de Arboles Umbral de Decisión*	300 0.2

*Para los modelos que presentan balanceo.
Fuente: Elaboración propia.

Figura 2. Diagrama de ensamble de modelos.



Fuente: Elaboración propia.

4 Resultados

En la sección de resultados se muestra un análisis del ejercicio empírico realizado para la presente investigación. El objetivo es evaluar, no solamente el desempeño de los modelos presentados sino también, analizar las posibles estrategias en cuanto a su uso y de esta manera profundizar en su capacidad explicativa frente a la detección de posibles actos de corrupción dentro de la contratación pública mediante la implementación de un sistema de alertas tempranas y la medición de un posible impacto económico. Ahora bien, la sección está dividida en tres aspectos relevantes. La primera, muestra la comparación entre los modelos a partir de los resultados obtenidos de las métricas estándar de clasificación, que, para el presente caso, se utiliza la curva ROC y el AUC, los cuales permiten generar y examinar la capacidad de distinguir las clases positivas y negativas de los modelos. También se presentan métricas que, según (Saito & Rehmsmeier, 2015) y (Davis & Goadrich, 2006), hacen que el

análisis de los resultados sea más sólido, ya que estas métricas funcionan mejor con datos que tienen claros desbalances.

En segundo lugar, se profundiza en la comprensión de los modelos más efectivos, analizando la importancia de las variables y relacionando los hallazgos con la literatura económica e institucional sobre corrupción. Mientras que, la tercera parte, muestra en detalle el sistema de alertas implementado para el estudio y cómo puede afectar desde un impacto económico basado en la probabilidad de estos hechos. En conjunto, este apartado busca determinar la efectividad de las herramientas basadas en *Machine Learning* tanto desde un punto metodológico como desde un punto de vista práctico.

4.1 Ajustes y desempeños de los modelos

Se toma en cuenta la curva ROC (*Receiver Operating Characteristic*) como una de las primeras comparaciones entre los modelos, basándose en lo mencionado por (Fawcett, 2006). Esta imagen muestra gráficamente la sensibilidad y especificidad del modelo para diferenciar las clases positivas (en este estudio, se refiere a los contratos con sanciones) y negativas (los que no tienen sanciones). Cabe resaltar que, al referirse a la sensibilidad, hace referencia a la capacidad para detectar las clases positivas, mientras que, la especificidad, hace referencia a la identificación correcta de las clases negativas. Con esto en mente, las siguientes especificaciones muestran el cálculo para cada uno de los casos.

$$Sensibilidad = \frac{TP}{TP + FN} \tag{5}$$

$$Especificidad = \frac{TN}{TN + FP} \tag{6}$$

Donde:

TP *(True Positive): Casos positivos bien clasificados.*

FN *(False Negatives): Casos positivos mal clasificados.*

TN *(True Negative): Casos negativos bien clasificados.*

FP *(False Positive): Casos negativos mal clasificados.*

De acuerdo con lo anterior, se puede decir que, al momento de presentar valores altos de sensibilidad y especificidad, los modelos mostrarán un mejor rendimiento en la curva ROC. De igual manera, James et al. (2013) establecen unos rangos para determinar estos rendimientos. Estos se toman mediante el AUC (*Area Under the Curve*) e indican lo siguiente: un AUC de 1.0 correspondería a un modelo perfecto, es decir, logra distinguir de manera perfecta los tipos de clases. Mientras que un AUC de 0.5 señala que es un modelo aleatorio (el caso de lanzar una moneda), lo que demostraría cierta debilidad del modelo. En cuanto a valores del AUC menores a 0.5, demostraría que el modelo es de una calidad poco fiable.

En este caso, los resultados del AUC muestran que usar métodos de balanceo no produce cambios significativos en el rendimiento y la capacidad de predicción de los modelos. Un resultado parecido ocurre al analizar los modelos con variables identificadoras IDS. Esto sugiere que, en este caso, los modelos tienen varianzas mínimas, lo que podría permitir

obtener resultados más precisos. Ahora bien, al analizar de manera individual los resultados, se evidenció que, los dos modelos generan un promedio de AUC superior al 0.90, tanto con balanceo como sin él, siendo 0.963 y 0.962, los mayores resultados presentados, los cuales son resultados arrojados por el modelo Random Forest, resaltando que, los dos modelos fueron tomados con balanceo.

Teniendo en cuenta lo anterior, como un primer acercamiento hacia el entendimiento de los resultados. Se puede concluir que, de acuerdo con los resultados del AUC, todos los modelos presentan rendimientos similares. La implementación de métodos de balanceo, en este caso, no representa una afectación significativa en lo resultados. Esto apoya lo evidenciado en (Gallego, et al., 2021). Al no utilizar estos métodos y presentar resultados semejantes, se demuestra que su uso puede ser irrelevante para las estimaciones. Además, se pueden sacrificar datos que son relevantes para entender los modelos. La Gráfica 3 ilustra los resultados de la curva ROC de los modelos con mejor comportamiento mencionados anteriormente.

Para entender mejor los resultados de los modelos, se realizó una evaluación llamada PRECISION – RECALL (PR). Esta evaluación mide qué tan bien un modelo identifica casos positivos en situaciones de desbalance de datos. De igual forma, para este también se evaluaron los modelos en los cuales se aplica el balanceo; esto, teniendo en cuenta que, al buscar mejorar la posible varianza de los datos, estos, se pueden comportar de una mejor manera con esta evaluación. Al momento de hablar de precisión: de acuerdo con (Saito & Rehmsmeier, 2015), esta se enfoca en medir la proporción de predicciones positivas que son positivas, mientras que, al hablar de recall, se enfoca en medir la proporción de casos positivos reales que el modelo logra detectar.

$$Precision = \frac{TP(\tau)}{[TP(\tau) + FP(\tau)]} \tag{7}$$

$$Recall = \frac{TP(\tau)}{[TP(\tau) + FN(\tau)]} \tag{8}$$

Donde:

τ: Umbral de predicción o decisión

De acuerdo con las fórmulas 7 y 8, en comparación con la evaluación por ROC, la evaluación PR se enfoca más en las clases positivas. Mientras que la anterior no toma en consideración el desbalance de los datos, lo cual puede generar que las predicciones se sobrevaloren (Davis & Goadrich, 2006).

Cabe mencionar que las evaluaciones de PR se encuentran sujetas a un umbral de decisión, τ el cual consiste en establecer un punto límite para convertirse en una clase positiva (1) o negativa (0) a estimación de cada modelo.

En cuanto a los resultados presentados aplicando esta evaluación. Se logró evidenciar que los modelos en su conjunto presentaron mejores valores en cuanto al Recall se refiere, lo que indica que los modelos presentan una buena capacidad para identificar casos positivos, siendo el 1.00 el mayor valor logrado en este indicador, el cual se obtiene con los modelos de

Random Forest, mostrando de esta manera que, para este indicador, si se ve afectado al implementar el balanceo en los modelos.

Contrario a lo presentado anteriormente, los valores de la precisión, al ser bajos, muestran la existencia de un alto número de falsos positivos, lo que indica que el modelo puede generar un alto número de alertas que requieren de validación aún más detallada. Sin embargo, para este caso, los valores no deben ser tomados a la ligera. En un contexto de política pública, este tipo de modelos van dirigidos más a ser una herramienta de priorización que permite orientar los recursos de control hacia los casos con mayor probabilidad de riesgo.

Conforme a lo anterior, se puede establecer que, en términos de PR, si se opta por una política más agresiva en cuanto a control, se puede considerar el modelo basado en Random Forest Balanceado, ya que, al ser sensible (Recall = 1), el modelo prioriza la no omisión de casos de riesgo, aunque esto le pueda costar un volumen alto de alertas. Mientras que, los modelos Lasso, pueden ser tomados para políticas intermedias debido al equilibrio existente en lo que se llamaría cobertura y eficiencia, permitiendo focalizar los esfuerzos sin perder la relevancia de los casos. Ahora bien, los resultados obtenidos en los modelos de Random Forest no balanceados, dan la posibilidad de generar políticas de corte más conservadoras. Al reducir el número de alertas, facilitan la gestión operativa de las entidades, aunque con un riesgo latente que es la omisión de detectar posibles casos potenciales.

Como un último método de análisis y evaluación de los resultados y, teniendo en cuenta lo recomendado por (Davis & Goadrich, 2006), se utiliza la métrica F1-SCORE, la cual puede ser denominada la media armónica de la evaluación PR, esto teniendo en cuenta que el F1-SCORE toma los resultados del PR generando una penalización a los valores extremos, o en este caso desbalanceados, y así produciendo una métrica más exacta para las predicciones realizadas.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{9}$$

De acuerdo con lo anterior, el F1-SCORE se centra en la detección correcta de los positivos (Recall) y en no equivocarse al detectarlos (Precisión). Generando así, resultados más precisos que nos permitan establecer los mejores modelos para estos casos.

Los resultados de esta métrica (Tabla 3) ilustran que, para los modelos basados en Random Forest con el método de balanceo, al presentar valores de 0.002, presentan una estrategia de detección total, pero con poca eficiencia, lo cual se convertiría en un volumen de alertas poco manejable. Asimismo, al analizar los modelos Lasso, este presenta un balance limitado entre detección y eficiencia, aunque aún presenta un alto nivel de falsos positivos, mientras que, los modelos de Random Forest no balanceados, de acuerdo con esta métrica son los que mejores indicadores, ya que mejoran el balance entre cobertura y eficiencia, reduciendo la carga operativa sin sacrificar completamente la capacidad de detección. Los resultados de este indicador, tal como los presentados en PR, reflejan la dificultad inherente de equilibrar la detección de casos de corrupción con la eficiencia operativa en contextos altamente desbalanceados. Esto refuerza la idea de que estos modelos pueden ser utilizados como herramientas de selección prioritaria en términos de control y revisión.

Como conclusión parcial de la presente sección, se puede establecer que los modelos desarrollados exhiben una capacidad de discriminación favorable de acuerdo con los niveles de AUC que estos muestran, pero bajo condiciones extremas de desbalance, su desempeño en la clasificación resulta altamente afectado. Sin embargo, estos valores también deben considerarse de acuerdo con la política que se quiera manejar, ya que, para el caso de la corrupción, donde su ocurrencia es escasa, estos modelos pueden verse como una herramienta de priorización final y no como mecanismos de clasificación determinística, permitiendo orientar de manera más eficiente los recursos de control y vigilancia.

Tabla 3. Resultados Métricas[14].

Modelo	Balanceo	AUC	Precisión	Recall	F1-Score
Lasso Sin IDS	Si	0.952	0.003	0.743	0.007
Lasso Con IDS	Si	0.961	0.006	0.655	0.012
Lasso Sin IDS	No	0.941	0.003	0.743	0.007
Lasso Con IDS	No	0.959	0.006	0.655	0.012
Random Forest Sin IDS	Si	0.963	0.001	1	0.002
Random Forest Con IDS	Si	0.962	0.001	1	0.002
Random Forest Sin IDS	No	0.959	0.006	0.560	0.012
Random Forest Con IDS	No	0.957	0.007	0.506	0.015

Fuente: Elaboración propia.

4.2 Interpretación de los modelos

Para la interpretación de los modelos analizados, se identificaron las variables más utilizadas por estos en sus estimaciones. Para ello, se parte de lo expuesto por (Velez y Kim, 2017) en cuanto a la generación de formas para mostrar e interpretar los resultados obtenidos de una manera clara, evitando sesgos y teniendo una justificación valedera, especialmente para políticas públicas. Por ello, la presente investigación se centra en determinar cuáles son las principales variables para cada modelo, entendiendo su interpretabilidad y posible efecto sobre los contratos que presentan corrupción. Resaltando que esta metodología es la misma implementada por (Gallego, et al., 2021).

Con el fin de determinar las variables más importantes, se implementa el *Mean Decrease in Gini,* el cual mide la contribución de cada variable dentro de un nodo en el modelo de bosques aleatorios. Mientras que, para el modelo Lasso, se analizan desde el coeficiente estimado en cada una de las variables. La idea, para el primer caso, es validar el nivel de "impureza"[15] que una variable puede tener sobre un nodo, lo cual ayuda a separar de mejor forma las clases durante el procesamiento del modelo. Mientras que, para los coeficientes, un valor positivo de estos puede aumentar el riesgo de corrupción y un valor negativo reducirlo.

En primer lugar, el análisis mediante el *Mean Decrease in Gini,* es implementado por (Breiman, 2001), el cual sugiere, que esta es una de las mejores medidas para la determinación de principales variables y poder analizar los resultados de los modelos. La Gráfica 4 ilustra las principales variables utilizadas por los modelos de Random Forest

[14] IDS se refiere al NIT de las empresas (públicas y privadas) usados dentro de los modelos.g
[15] La impureza Gini define qué tan mezcladas están las clases en un nodo del modelo (en este caso 0 y 1).

balanceados y no balanceados, que cuentan con variables generalizadas y no generalizadas (IDS). Ahora bien, estas variables están distribuidas de tal manera que se muestran según sus apariciones y la posición que estas ocupan, esto con el fin de presentar un top para estas. Cabe resaltar que, la importancia de cada variable se mide por valores normalizados; es decir, para la variable más importante, la cual será, la más influyente en la reducción de las impurezas o mejora de la predicción, toma un valor de 0.12, mientras que las demás se expresan en proporción a la importancia máxima.

De acuerdo con lo presentado anteriormente, se puede evidenciar que, los modelos emplean variables similares para realizar sus predicciones, recalcando que, tanto en los modelos balanceados como en los no balanceados, la variable relacionada con el departamento del contratista es tomada como primera variable en todos los modelos. Asimismo, se resalta la importancia del sector financiero en estas estimaciones.

Ahora bien, para interpretar los resultados del modelo Lasso, la presente investigación se basa en lo mencionado por Zhao, et al. (2025), el cual sugiere, que, a partir de los coeficientes estimados, se logra realizar una proyección espacial de las características, permitiendo observar la evolución de estas desde su penalización. Dicho esto, la Gráfica 5, evidencia que la mayoría de las características analizadas están asociadas a un menor riesgo, siendo en su mayoría pertenecientes a categorías contractuales y algunas características socioeconómicas de los departamentos.

Los resultados anteriormente mencionados, permiten identificar ciertos patrones que, de acuerdo con los modelos, pueden ser objeto de un estudio o control más riguroso, ya sea por su importancia acumulada, en el caso de Random Forest, o su efecto dentro de la corrupción en la contratación, en el caso del modelo Lasso. Cabe resaltar que, el uso o importancia de cada modelo podrá o será establecido de acuerdo con la necesidad de cada entidad y la misionalidad de estas.

Asimismo, al analizar las particularidades de los resultados, estos, pueden ir en concordancia con lo expuesto (Colonnelli, et al., 2021), donde encuentran que, en Brasil, proveedores (en su mayoría jurídicos), los cuales celebran contratos en zonas o departamentos cercanos a su zona de creación o casa matriz, al momento de exponer actos de corrupción en ese departamento, deciden alejarse por miedo a verse involucrados en estos actos, lo que demostraría la importancia de las variables dirigidas al lugar de residencia u origen de los proveedores y de las entidades públicas. De igual manera, Gallego et al. (2021) exponen los efectos geográficos según la relevancia de cada municipio y determinan que las ciudades con mayor número de contratos adjudicados -en este caso, según la Gráfica 1- son Bogotá, Valle del Cauca y Antioquia. Los datos, también reflejan que, en algunas ocasiones, la ciudad y/o departamentos donde se adjudica el contrato, difieren del departamento de residencia del contratista, lo que puede repercutir en la ejecución de los contratos, esto, teniendo en cuenta que posiblemente el contratista no tiene en claro las condiciones, tanto logísticas como económicas, del territorio donde se ejecuta el contrato. Prueba de ello se muestra en un caso reciente. En el 2022, la Contraloría General de la Nación adelantó un total de 144 procesos por una estimación de 42.000 millones de pesos por irregularidades en el Programa de Alimentación Escolar (PAE)[16]. Argumentando la presencia de fallas en la logística y cobertura del plan. Portafolio (2017): “El año pasado, pese al escándalo, esa corporación

[16] Para más información: (Universidad Nacional de Colombia, 2022).

logró un contrato de 999 millones para prestar servicios educativos en Cúcuta y luego obtuvo otro en Córdoba por 23.049 millones de pesos para prestar servicios de educación. Y este año ganó un contrato de 24.806 millones en Norte de Santander para administrar la educación en escuelas de zonas rurales. En todos esos negocios aparece el nombre del representante legal de esa firma, José Antonio Manrique Torres, expresidente del Cúcuta Deportivo."

Continuando con las características del contratista, (Lyra, et al., 2021) muestran que las empresas, al tener incentivos "perversos", pueden generar colusiones, carteles con el fin de manipular los precios y generar mayores beneficios. Esto, entendiendo que la corrupción no se debe mirar como un comportamiento aislado; por el contrario, se debe entender que en estas actividades participan diferentes individuos. De acuerdo con lo mencionado anteriormente, al momento de hablar sobre el valor de los contratos, los resultados sugieren una semejanza a lo evidenciado por (Gallego, et al., 2020) y (Gallego, et al., 2022). Ya que, en estas investigaciones, muestran que el valor del contrato se considera como una variable crítica al momento de querer prevenir la corrupción, este resultado se apoya de acuerdo con lo mostrado por (Organización de los Estados Americanos (OEA), Banco de Desarrollo de América Latina (CAF) y Datasketch, 2021) y (Medina Arnáiz, 2016), donde demuestran que, a mayor valor de los contratos, estos se vuelven más susceptibles a prácticas de corrupción. Lo cual, para el presente caso, aplica, dado el nivel de importancia que tiene esta variable dentro de cada uno de los modelos, sin importar la metodología implementada.

Algo similar pasa con la variable relacionada con la modalidad de contratación y la justificación de ella, esto debido a la importancia que esta variable adquiere dentro de la configuración de un proceso contractual en la plataforma SECOP, ya que, desde un contexto jurídico, la Ley 80 de 1993 en su artículo 5 y (Colombia Compra Eficiente, 2017) mencionan que, cada modalidad o proceso deberá estar debidamente justificado y sustentado técnicamente. Igualmente, de acuerdo con lo sustentado por (Gallego, et al. 2020), una de las principales causas de corrupción durante la pandemia del COVID-19 fue la flexibilidad que se tuvo al momento de realizar las contrataciones públicas, lo que generó inconsistencias durante las auditorías realizadas por parte de los entes de control.

Teniendo en cuenta lo anterior, en el presente apartado se buscó entender la importancia de cada una de las variables analizadas. Además, se analizó cómo estas, tanto desde la teoría como los hechos que pasaron en el país, pueden ser susceptibles o estar relacionadas con actos de corrupción dentro de la gestión pública colombiana. Entendiendo que no solamente se debe tener en cuenta aspectos como el valor del contrato, sino también aspectos tales como, el lugar de adjudicación del proceso, y algunas variables departamentales y del contratista, pueden afectar o ser susceptibles para presentarse actos de corrupción.

4.3 Sistema de Alertas e Impacto Económico

Con el fin de generar una herramienta que sea útil al momento de una toma de decisiones en un contexto de control público, la presente investigación desarrolla un sistema de alertas basado en las probabilidades estimadas por los modelos de *Machine Learning.* Este sistema consiste en transformar estas probabilidades en un índice de riesgos, el cual tiene una escala de 0 a 100, permitiendo así, clasificar los contratos en diferentes niveles de alerta (Bajo, Mediano, Alto y Crítico). Con esto, se busca facilitar la priorización en los procesos de monitoreo, auditoría o intervención.

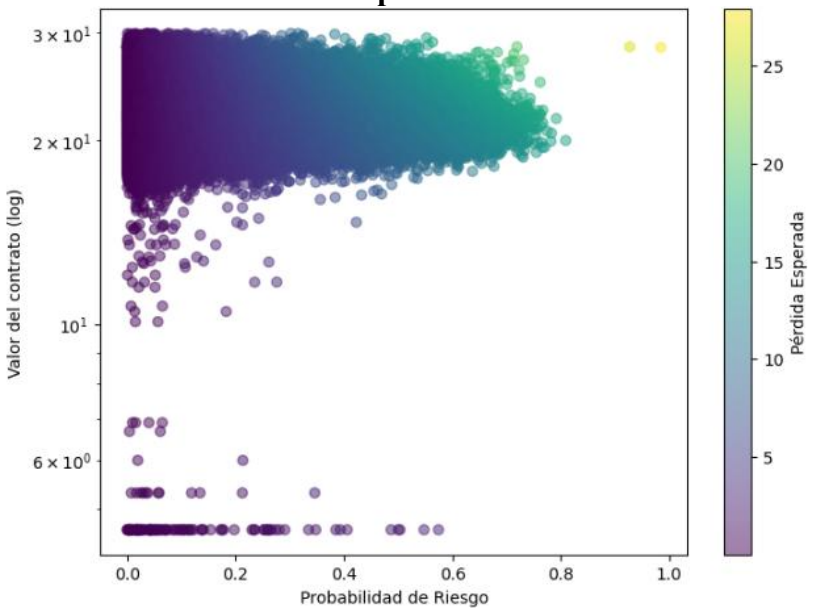
En cuanto a la metodología utilizada para este sistema, la investigación se basa en lo mencionado por (Bolton & Hand, 2002) y (OECD, 2019). En este sentido, el sistema de alertas propuesto se alinea con enfoques modernos de analítica de datos aplicados al sector público, en los cuales los modelos predictivos se utilizan como herramientas de apoyo a la decisión, permitiendo optimizar la asignación de recursos de control en escenarios caracterizados por un alto desbalance de clases y limitada información observada sobre eventos irregulares.

Ahora bien, como un complemento al indicador realizado, se establece una medida de riesgo económico basada en la probabilidad estimada de irregularidad y el valor monetario de los contratos. Se crea una variable de pérdida esperada, que resulta de la probabilidad de riesgo que el modelo asigna y el valor del contrato. Esto ayuda a estimar el impacto económico posible de cada caso. La siguiente ecuación muestra lo mencionado anteriormente.

$$\mathbb{E}[L_i] = P(Y_i = 1|\, X_i) * V_i \tag{10}$$

De acuerdo con la ecuación (10), se puede decir que la pérdida esperada $\mathbb{E}[L_i]$ o riesgo económico estará definida por el cálculo de la probabilidad estimada de riesgo $P(Y_i = 1|\, X_i)$ por el valor del contrato V_i. Este enfoque se basa en principios de la literatura sobre gestión de riesgos, especialmente en la idea de la pérdida esperada, donde el riesgo se define como la combinación de la probabilidad de que ocurra un evento negativo y la gravedad de sus efectos (Jorion, 2007). Esto, permite que el modelo sirva no solo como un clasificador de riesgo (cualitativo), sino que también, permite cuantificar este riesgo en cada contrato, la Tabla 4. Muestra los resultados obtenidos en cada uno de los modelos realizados.

Gráfico 6. Impacto Económico



Fuente: Elaboración propia.

De acuerdo con la Gráfica 6, muestra que la distribución del riesgo no está concentrada exclusivamente en contratos de baja cuantía, sino que se extiende a lo largo de distintos niveles de valor. Asimismo, la persistencia de una alta concentración de observaciones en niveles bajos de probabilidad refleja o sugieren los posibles efectos a causa del desbalance en los datos. Sin embargo, los casos de alta probabilidad y alto valor son poco frecuentes, pero representan los eventos de mayor relevancia en términos de impacto económico esperado.

Ahora bien, como un ejemplo cuantificable se tomó como ejemplo un contrato con valor de $3.001.920.236 tiene un 60% de pérdida espera o costo para la entidad, cuantificable a $1.827.845.216. Estos resultados son arrojados por el modelo Random Forest sin SubBagging, que, de acuerdo con los análisis, es el modelo más conservador que detecta riesgo en los contratos, pero no los exagera de una manera desproporcionada como puede pasar con otros modelos.

Desde una perspectiva administrativa, las entidades públicas pueden establecer criterios de priorización basados en un análisis de costo-beneficio, donde la decisión de intervenir un contrato dependerá de la relación entre la pérdida esperada estimada $\mathbb{E}[L_i]$ por los costos administrativos vinculados al proceso de recisión del contrato, los cuales, en este caso, estarán asociados a un factor que incorpora el uso de estas herramientas analíticas $V_i - C_{adm,i}$ Teniendo esto en cuenta, la ecuación 11 está expresada de la siguiente manera:

$$Score_i = \mathbb{E}[L_i] * V_i - C_{adm,i} \tag{11}$$

Cabe resaltar que, esta ecuación es una simplificación de un costo-beneficio completo, esto, teniendo en cuenta que no se toman en cuenta otros efectos como beneficios indirectos, efectos sociales, externalidades y un horizonte temporal. Sin embargo, este permite orientar la intervención hacia aquellos casos donde el beneficio esperado de control supera los costos operativos, optimizando la asignación de recursos públicos. Ahora bien, al sustituir las ecuaciones 10 y 11 por los valores que previamente se tienen del modelo, se obtiene lo siguiente:

$$\mathbb{E}[L_i] = 0.60 * \$3.001.920.26 = \$1.801.152.141^{17}$$

$$C_{adm,i} = \$143.213.460^{18}$$

[17] El valor actual difiere de los $1.827.845.216 por redondeos del modelo

[18] Este valor es estimado a 12 meses, de acuerdo con la Resolución No. SDH-000179 del 24 de noviembre de 2025 – Tabla de Honorarios Vigencia 2026 de la Secretaria Distrital de Hacienda. Para el nivel de Profesional Especializado 8 con un valor mensual de $11.934.455

$$Score_i = \$1.801.152.141 - \$143.213.460 = \$1.657.938.681$$

Teniendo en cuenta el resultado y de acuerdo con la interpretación del modelo, al comparar la pérdida esperada con los costos administrativos asociados a su revisión, se evidencia que el beneficio previsto (en este caso) de la intervención supera dichos costos, justificando su priorización dentro de los procesos de control, lo cual da muestra, desde un punto de vista práctico, del uso de estos modelos dentro de la gestión pública.

5 Conclusiones

Esta última sección se divide en tres (3) partes. Siendo la primera un análisis general de los resultados obtenidos y las conclusiones de esto. Seguidamente, se exponen las recomendaciones en políticas públicas desde el punto de vista de *Machine Learning* y los posibles actores que puedan intervenir; por último, se menciona la posible agenda de investigación de aspectos que no se abordan dentro de la presente investigación.

5.1 Conclusiones Generales

La corrupción ha sido una actividad que a lo largo del tiempo ha afectado al sector público colombiano de diversas maneras. Esto ha hecho que las distintas administraciones, tanto a nivel nacional como departamental y municipal, centren sus esfuerzos en la prevención y control de este tipo de actividades. Siendo la implementación de la plataforma SECOP uno de los esfuerzos más relevantes dentro de la lucha contra la corrupción en el sector público. La presente investigación toma como referencia de gestión pública a las actividades relacionadas con la contratación pública. Entiende que estas pueden verse como uno de los principales mecanismos para la materialización de las necesidades de una entidad. A su vez, pueden ser las iniciativas de un gobierno local o nacional. Por lo tanto, se entendería que la contratación pública es un instrumento clave para la implementación de políticas públicas.

A pesar de los constantes esfuerzos, los hechos de corrupción en la contratación pública tienden a hacerse persistentes a lo largo del tiempo, lo que da muestra de la necesidad de implementar nuevas alternativas que permitan no solo generar mayores controles, por parte de entes de control, sino también, generar mecanismos de prevención de estos actos, principalmente por parte de las entidades que realizan los contratos. Es en este punto, donde los modelos basados en *machine learning* pueden adquirir cierta relevancia, ya que, de acuerdo con lo mencionado por (Colonnelli, et al., 2020), este tipo de herramientas puede brindar alternativas de manejo sobre los actos de corrupción, debido a las ventajas que estos modelos presentan frente al uso de grandes y complejas bases de datos.

Sin embargo, se hace importante establecer los alcances de los resultados obtenidos. Esto, tiendo en cuenta que, los modelos desarrollados en esta investigación a pesar de mostrar una buena capacidad para identificar posibles riesgos no son precisos en todos los casos. Lo cual, podría resultar en la generación de alertas sobre contratos que finalmente no presenten problemas, o dejar de identificar algunos casos que si lo presenten. Ahora bien, estos percances o limitaciones no corresponden a un error del modelo, sino a la naturaleza misma del problema: *La dificultad de observar actos de corrupción,* toda vez que, estos actos son de baja frecuencia y los datos disponibles no siempre registran esta actividad de una manera

directa. Por tal motivo, los resultados de este estudio deben interpretarse como una herramienta que ayuda a priorizar la atención institucional. Es decir, permite identificar qué contratos podrían requerir mayor seguimiento o revisión, facilitando una mejor asignación de los recursos de control.

Dicho lo anterior, uno de los principales aportes de la presente investigación es la construcción de un sistema de alertas que logra traducir los riegos presentes en los procesos contractuales, no solo desde un punto de vista cualitativo, en cuanto al sistema de alertas, sino también cuantitativo con las estimaciones del posible impacto económico que puede generar cada uno de estos. De igual manera, estos resultados apoyan de una forma positiva la evidencia nacional e internacional, ya sea desde un aspecto académico como práctico, esto teniendo en cuenta lo siguiente:

Desde un contexto local, (Mojica Muñoz, 2021) al evaluar la capacidad de algoritmos de *Machine Learning* para detectar riesgos de corrupción en la administración pública municipal colombiana. Mostrando así una similitud entre ambos trabajos, dado que comparten una misma preocupación: utilizar herramientas de *inteligencia artificial* y *aprendizaje automático* para fortalecer la prevención de la corrupción en Colombia. Ahora bien, a pesar de presentar distintas metodologías, los dos concluyen que cada día es más necesario el uso de estos modelos dentro de la gestión pública. Ahora bien, la Contraloría General de la República, ha implementado un modelo predictivo que analiza anomalías en la contratación pública, con el fin de establecer vigilancia e identificar patrones dentro de las actividades contractuales, lo que demuestra la importancia de estos tipos de investigaciones y su importancia dentro de la gestión de administración pública.

Asimismo, al observar los avances internacionales, primero, el Gobierno de China está utilizando modelos de *ML* para identificar problemas en los documentos presentados en los procesos de licitación pública. De acuerdo con (MarketScreener, 2026), "*Un organismo anticorrupción en la provincia de Zhejiang detuvo en enero de 2025 a un administrador de activos estatales después de que la IA detectara posibles irregularidades en los procesos de licitación de varios proyectos públicos, según informó la cadena estatal CCTV en una serie documental sobre funcionarios corruptos emitida el mes pasado".*

En Europa, la división de gestión de inversiones del Banco Central noruego (NBIM, por sus siglas en inglés) utiliza modelos basados en IA y ML para detectar riesgos éticos y de corrupción en las empresas donde invierte. El NBIM hoy en día posee participación en más de 7000 empresas a lo largo del mundo, lo que la posiciona como uno de los fondos de inversión más destacados del mundo. Ahora bien, la Organización para la Cooperación y el Desarrollo Económico (OECD, 2025) busca promover el uso de estos métodos mediante informes y guías para su uso, con el fin de anticipar riesgos de corrupción y priorizar acciones de control. Estos ejemplos anteriormente mencionados, demuestran los avances de estos modelos en la gestión pública en diferentes partes del mundo.

Si bien, este tipo de modelos requiere un alto nivel computacional, lo que puede conllevar posibles sobrecostos para las entidades públicas que adquieran máquinas adecuadas para este fin. Sin embargo, tal como se menciona en este documento, estos sobrecostos pueden ser compensados al momento de determinar el costo-beneficio de incorporar estas tecnologías frente a los actos de corrupción y el control que estas entidades pueden generar. Es importante destacar que crear oportunidades para mejorar la plataforma de transacciones y la gestión de

contratos puede aumentar la transparencia, optimizar el análisis de datos y encontrar nuevas características, las cuales pueden ser determinantes para nuevos estudios. Esto representa un desafío tanto para los gobiernos nacionales y locales como para quienes hacen políticas públicas, tal como lo mencionan (Gallego et al. 2021). Finalmente, esta investigación debe entenderse como un punto de partida. Más que ofrecer una solución definitiva, propone una nueva forma de abordar el problema, abriendo la puerta a futuros desarrollos que permitan mejorar la precisión de estas herramientas y fortalecer su uso dentro de las entidades públicas.

5.2 Recomendaciones de Política Pública

Los resultados obtenidos en el presente documento no solo pretenden aportar al debate desde un contexto académico, sino que también, buscan ser apoyo dentro de la formulación de políticas públicas; para ello, se proponen una serie que van de la mano con los resultados y el contexto del documento.

Como primera medida, de acuerdo con Zuleta et al. (2019), se propone la creación de un sistema nacional de alertas tempranas. Es similar a la presentada en Brasil con la herramienta Analisador de Licitações e Editais (ALICE). Esta herramienta se integra de manera directa con la plataforma SECOP, utilizando modelos de *Machine Learning.* Así, genera alertas en tiempo real a partir de la configuración del(los) contrato(s), teniendo en cuenta la probabilidad de riesgo que dan los modelos.

La experiencia colombiana demuestra que las reformas normativas en contratación pública, aunque necesarias, no han sido suficientes para contener dinámicas estructurales de corrupción. Por lo tanto, la nueva serie de reformas debe incluir herramientas de análisis avanzado, inteligencia artificial y sistemas de decisión automática que ayuden a mejorar la capacidad del Estado para prevenir y predecir. Dicho esto, una propuesta va encaminada hacia una reforma estructural de la gestión de datos por parte de las diferentes entidades, tanto de orden nacional como departamental. El fin es fortalecer la interoperabilidad entre entidades que permita establecer protocolos para la estandarización, actualización y procesamiento de datos. Estos protocolos deben dar como resultado insumos que posibiliten la generación de modelos más robustos con altos niveles de desempeño.

Ahora bien, para lograr lo anteriormente mencionado, las entidades deben proveer de una mayor capacidad tanto física como tecnológica a los diferentes grupos y unidades encargadas de la analítica de datos; si bien, esto puede conllevar posibles costos adicionales por parte de las diferentes entidades, en el caso de que estos costos sean menores a la pérdida esperada con corrupción, las entidades optarán por aumentar esta capacidad y minimizar los riesgos.

Sin embargo, a pesar de que, en Colombia, el uso de sistemas de decisión automatizada dentro del sector público ha venido creciendo significativamente, especialmente en entidades del orden nacional y organismos de control (Gutiérrez & Muñoz-Cadena, 2023), la implementación y el uso de estas herramientas debe ir ligada a un marco de transparencia y uso estratégico de estas herramientas, (OECD, 2022) y, aunque dentro (CONPES, 2025) se reconoce la necesidad de impulsar el uso y adopción de sistemas de IA en entidades públicas, promoviendo simultáneamente principios de gobernanza, ética, mitigación de riesgos y supervisión institucional de una manera muy específica, se recomienda que, en futuros documentos, haya lineamientos dirigidos al control fiscal y la contratación pública.

Por último, cabe resaltar que Colombia durante los últimos años ha avanzado en la formulación de políticas ligadas al uso de inteligencias artificiales dentro de la política

pública. Sin embargo, existe una brecha importante frente a la creación de herramientas y aplicaciones que permitan una mejor gestión en cuanto a la contratación pública se refiere. Por lo cual, la presente investigación contribuye a cerrar dicha brecha mediante la construcción de un modelo predictivo de riesgo de corrupción contractual basado en Machine Learning.

5.3 Agenda de Investigación

Como una agenda de investigación. En primer lugar, como una idea principal, se buscará identificar posibles modelos predictivos con el fin de poder mejorar la precisión, reducir los falsos positivos y aumentar la interpretabilidad de los modelos, con el fin de dar respuesta a la siguiente pregunta: ¿Puede la IA anticipar riesgos de corrupción en la contratación pública en su etapa precontractual?

Otros aspectos cruciales que la presente investigación menciona, pero no aborda de manera directa, pero que incita a investigar, son los otros delitos que se pueden presentar dentro de la contratación pública como el clientelismo, el favorecimiento de contratos, colusiones, entre otros. Ahora bien, esto plantea una serie de posibles preguntas como: ¿Las conexiones políticas afectan la adjudicación del contrato? ¿Las campañas políticas predicen contratación? ¿Existen redes regionales de captura estatal?

Por último, desde un punto de vista de gestión pública y a partir de las recomendaciones de política sugeridas en la investigación, se plantea la interrogante de ¿cómo transformar la gestión pública mediante IA?

Referencias

Bastidas Vargas, D. (2023). *Redes de corrupción en la contratación pública en Colombia.* Ciudad de Mexico: [Tesis de maestría, Centro de Investigación y Docencia Económicas].

Bolton, R. J., & Hand, D. J. (2002). Statistical Fraud Detection: A Review. *Statistical Science.*, 235-255.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

CAF – Banco de Desarrollo de América Latina. (2020). *Tres preguntas sobre el uso de los datos para luchar contra la corrupción (Policy Brief 9, Transparencia e integridad pública). Dirección de Innovación Digital del Estado.* CAF. Retrieved from https://scioteca.caf.com/handle/123456789/1544.

Colombia Compra Eficiente. (2017). *Guía para la Contratación Directa sin Oferta.* Bogota: Colombia Compra Eficiente.

Colonnelli, E., Prem, M., & Teso, E. (2020). Patronage and selection in public secto organizations. *American Economic Review 110 (10)*, 3071–99.

Consejo Nacional de Política Económica y Social (CONPES). (2002). *Una Política de Estado para la Eficiencia y la Transparencia en la Contratación Pública.* Bogota: CONPES.

Consejo Nacional de Política Económica y Social (CONPES). (2025). *La hoja de ruta de Colombia en Inteligencia Artificial para los retos actuales y la transformación futura.* Bogota: CONPES.

Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. *Department of Computer Sciences and Department of Biostatistics and Medical Informatics*, Pittsburgh, 25-29 June 2006, 233-240. https://doi.org/10.1145/1143844.1143874.

Decarolis, F., & Giorgiantonio, C. (2020). *Corruption red flags in public procurement: new evidence from Italian calls for tenders.* Bank of Italy, Economic Research and International Relations Area.: Questioni di Economia e Finanza (Occasional Papers) 544.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 861-874.

Fernand Desfrancois, P. G., & Pastás Gutiérrez, E. R. (2022). Corrupción y crecimiento económico en América Latina y el Caribe. *Revista Economica del Caribe*, 32-49.

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. (2018). SMOTE for Learning from Imbalanced Data: Progress and. *Journal of Artificial Intelligence Research*, 863-905.

Gallego, J., Prem, M., & Vargas, J. (2020). *Corruption in the times of pandemia.* Documentos de Trabajo 18178, Universidad del Rosario.

Gallego, J., Prem, M., & Vargas, J. (2022). *Predicting Politicians Misconduct: Evidence From Colombia.* Documentos de Trabajo 20504, Universidad del Rosario.

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

Galvis - Ciro, J., & Hicapié - Vélez, G. (2022). The Effects of Corruption on Government Spending in the States of Colombia. *Apuntes del CENES*, 227-262 https://doi.org/10.19053/01203053.v41.n73.2022.13555.

Garcia, J., Morón, J., Valbuena, G., Fernandez, E., & Leguizamon, J. (2022). *La corrupción en Colombia: un análisis integral.* Documentos de trabajo sobre Economía Regional y Urbana 307, Banco de la Republica de Colombia.

Gutiérrez, J. D., & Muñoz-Cadena, S. (2023). Adopción de sistemas de decisión automatizada en el sector público: Cartografía de 113 sistemas en Colombia. *GIGAPP Estudios Working Papers*, 365-395.

Hastie, T., Friedman, J. H., & Tibshirani, R. (2009). *The Elements of Statistical Learning.*

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations.*

Iturriaga, F., & Sanz, I. (2018). Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces. *Social Indicators Research*, 975-998.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.*

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning.*

Jorion, P. (2007). *Value at Risk: The New Benchmark for Managing Financial.* New York: McGraw-Hill.

Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training sets: One-Sided Selection. *Proceedings of the 14th International Conference on Machine Learning*, 179-186.

Lyra, M., Curado, A., Damasio, B., Bacao, F., & Pinheiro, F. (2021). Characterization of the Firm-Firm Public Procurement Co-Bidding Network from the State of Ceará (Brazil) Municipalities. *Physics and Society*, 6, 77 (2021). https://doi.org/10.1007/s41109-021-00418-y.

MarketScreener. (10 de 02 de 2026). *China recurre a la IA para detectar la corrupción en las licitaciones públicas*. Obtenido de MarketScreener: https://shorturl.at/zHvJ6

Martínez Cárdenas, E. E., & Ramírez Mora, J. M. (2006). La corrupción en la Contratación Estatal Colombiana - Una aproximación desde el neoinstitucionalismo. *Reflexión Política*, 148–162. https://doi.org/10.29375/01240781.622.

Medina Arnáiz, T. (2016). La corrupción en la contratación pública: un burdo fraude al interés general. *Papeles de relaciones ecosociales y cambio global*.

Mojica Muñoz, K. (2021). Inteligencia artificial para detectar corrupción en la administración pública municipal de Colombia. *Universidad de los Andes, Facultad de Economía, CEDE*.

Moreno Pabón, J. (2018). El Efecto Globo: identificación de regiones propensas a la producción de coca. *Documentos CEDE*.

Morgner, M., & Chêne, M. (2015). *Trasnparency International*. Obtenido de Topic Guide: Public Procurement: https://shorturl.at/pmC0T

OECD. (2019). *Fraud and corruption risk assessments in public procurement.* OECD publications.

OECD. (2022). *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe.* OECD Publishing.

OECD. (2025). *Gobernar con la inteligencia artificial: Panorama actual y hoja de ruta en las funciones centrales de gobierno.* Paris: OCDE PUBLISHING.

Olken, B. (2007). Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy*.

Olken, B., & Pande, R. (2012). Corruption in Developing Countries. *Annual Review of Economics*.

Olsson, S. (2014). Corruption and Political Participation: A Multilevel Analysis. *QOG WP*.

Organización de los Estados Americanos (OEA), Banco de Desarrollo de América Latina (CAF), y Datasketch. (2021). *Guía para la identificación de riesgos de corrupción en contratación pública, utilizando la ciencia de datos.*

Ortiz Benavides, E. (2012). Efectos de la corrupción sobre la calidad de la salud y educación en Colombia 2004-2010. *Tendencias (Revista de la facultad de ciencias económicas y administrativas)*.

Portafolio. (03 de 12 de 2017). *Los contratistas que están involucrados en el desfalco del PAE.* Obtenido de PORTAFOLIO: https://www.portafolio.co/economia/los-contratistas-que-estan-involucrados-en-el-desfalco-del-pae-512237

Prem, M., Colonnelli, E., Lagaras, S., Ponticelli, J., & Tsoutsoura, M. (2021). *Revealing Corruption: Firm and Worker Level Evidence from Brazil.* Documentos de Trabajo 18673, Universidad del Rosario.

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *Plos One*.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society*, 267-288.

Transparencia por Colombia. (2022a). *Balance del monitoreo a la contratación pública COVID-19.* Bogota: Transparencia por Colombia.

Transparencia por Colombia. (2022b). *índice de percepción de la corrupción 2021.* Bogota: Transparencia por Colombia.

Transparencia por Colombia. (2022c). *RECOMENDACIONES EN MATERIA DE TRANSPARENCIA Y LUCHA CONTRA LA CORRUPCIÓN PARA LA CONSTRUCCIÓN DEL PLAN NACIONAL DE DESARROLLO 2022-2026.* Bogota: Transparencia por Colombia.

Universidad Nacional de Colombia. (8 de 04 de 2022). *Revista UNAL*. Obtenido de Programa de Alimentación Escolar ¿que lo hace tan vulnerable a la corrupción?: https://periodico.unal.edu.co/articulos/programa-de-alimentacion-escolar-que-lo-hace-tan-vulnerable-a-la-corrupcion/#:~:text=En%20enero%20de%202022%20la,en%20calidad%2C%20log%C3%ADstica%20y%20cobertura

Varvarigos, D. (2023). Cultural persistence in corruption, economic growth, and the environment. *Journal of Economic Dynamics and Control*.

Velez, D., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *Cornell Unicersity*.

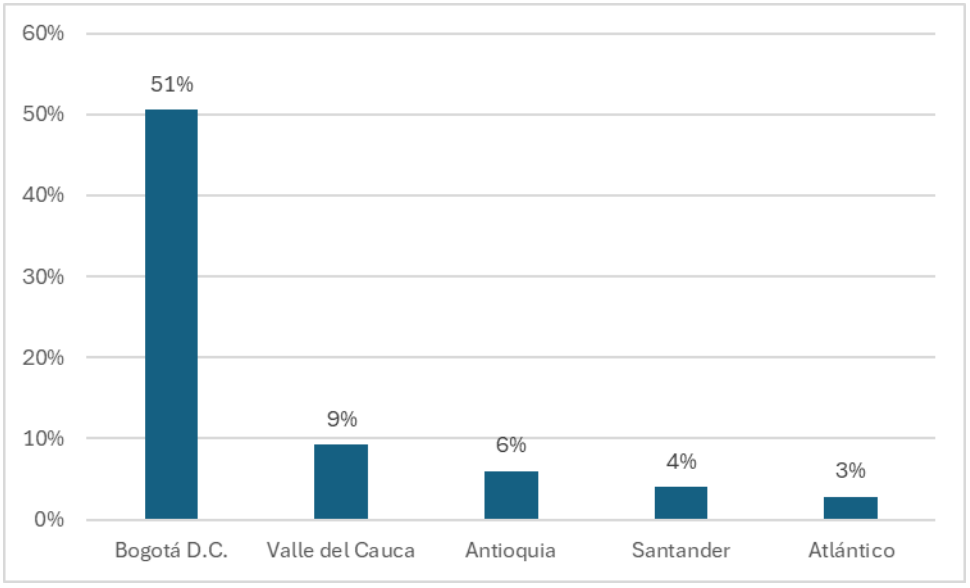
Zaman, F., & Hirose, H. (2009). Effect of subsampling rate on subbagging and related. *Pattern Recognition and Machine Intelligence*, 44-49.

Zhao, Y., Zhao, Y., Liao, H., Pan, S., & Zheng, Y. (2025). Interpreting LASSO regression model by feature space matching analysis for spatio-temporal correlation based wind power forecasting. *Applied Energy*.

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.* Fedesarrollo]/ Banco Interamericano de Desarrollo (BID).

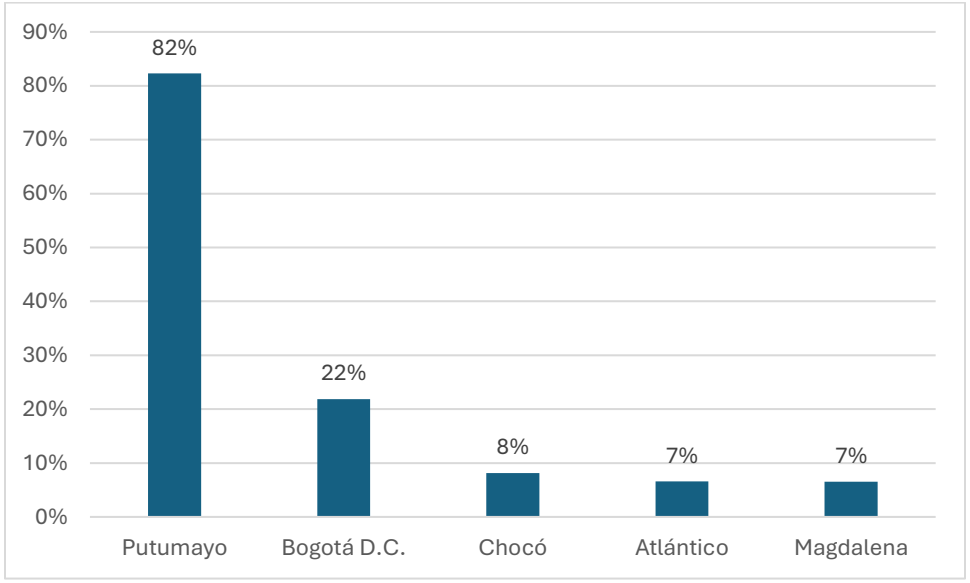
Apéndice

Gráfica 1. Departamentos con mayor número de contratos realizados (2010–2021)



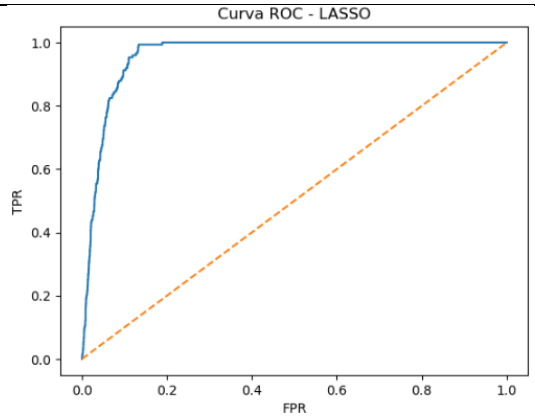
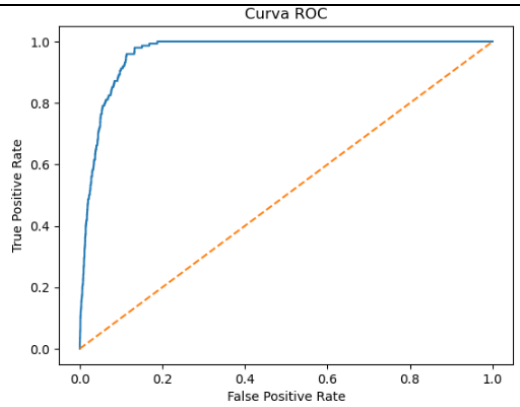
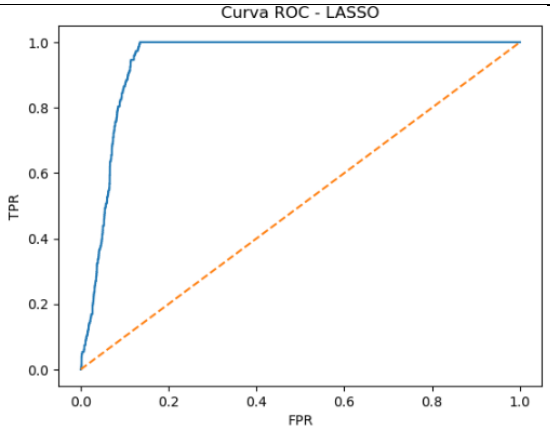
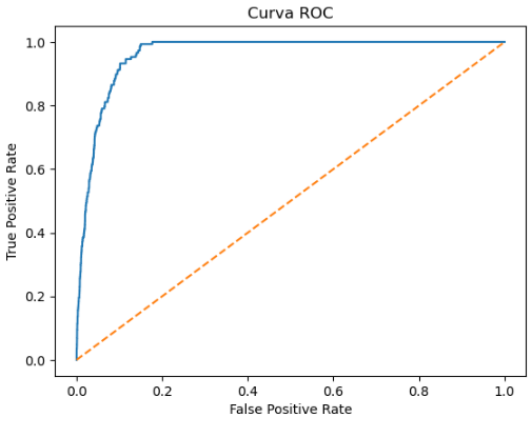
Fuente: Elaboración propia.

Gráfica 2. Departamentos con mayor valor contratado en comparación con el PIB departamental (2010–2021)



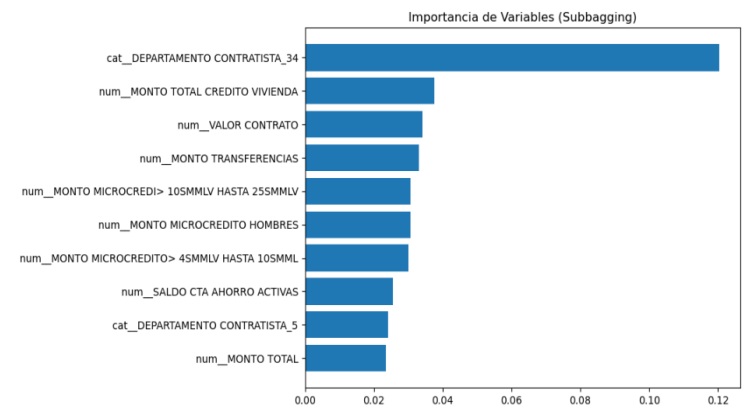
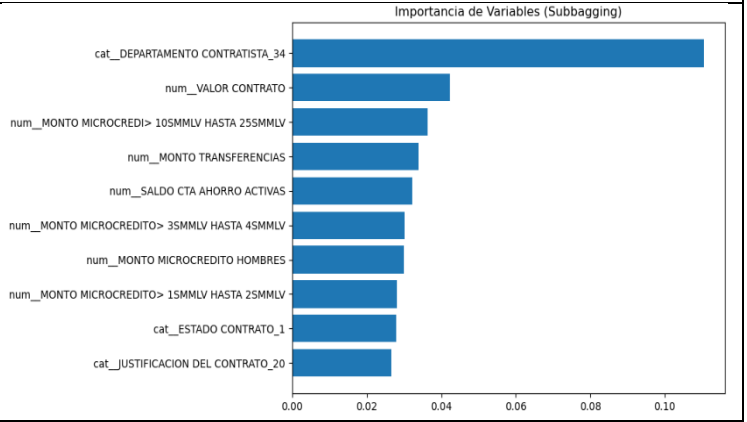
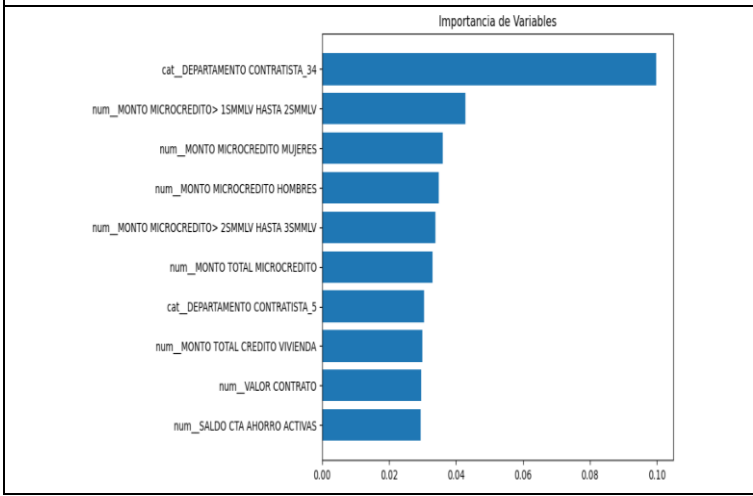
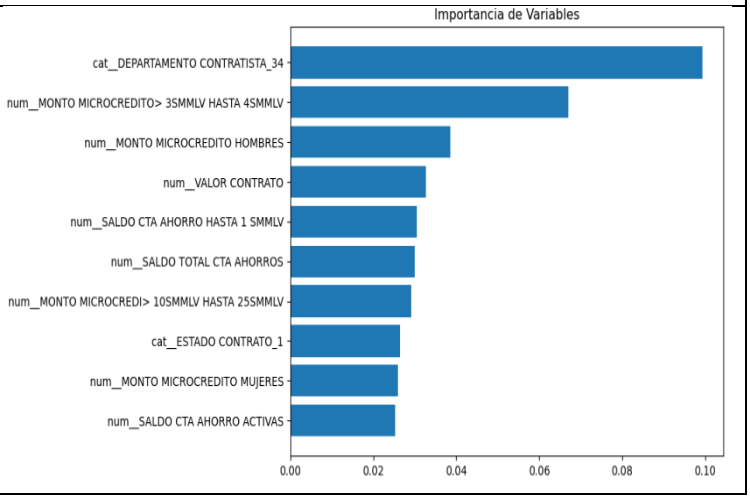
Fuente: Elaboración propia.

Gráfica 3. Curva ROC: Modelos Sin Balanceo vs. Modelos Balanceados.

Modelo Lasso Con IDS	Modelo Random Forest Con IDS
	
Modelo Lasso Sin IDS	**Modelo Random Forest Sin IDS**
	

Fuente: Elaboración propia.

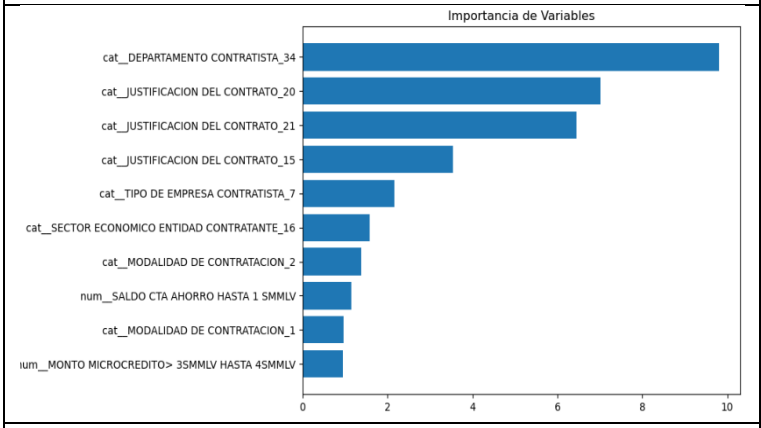
Gráfica 4. Principales Variables Random Forest

Variables RF con Balanceo Sin IDS	Variables RF con Balanceo Con IDS
	
Variables RF sin Balanceo Sin IDS	**Variables RF sin Balanceo Con IDS**
	

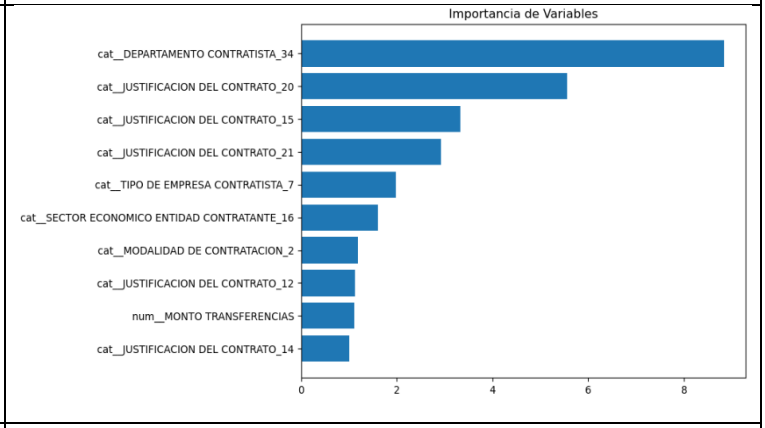
Fuente: Elaboración propia.

Gráfica 5. Principales Variables Lasso

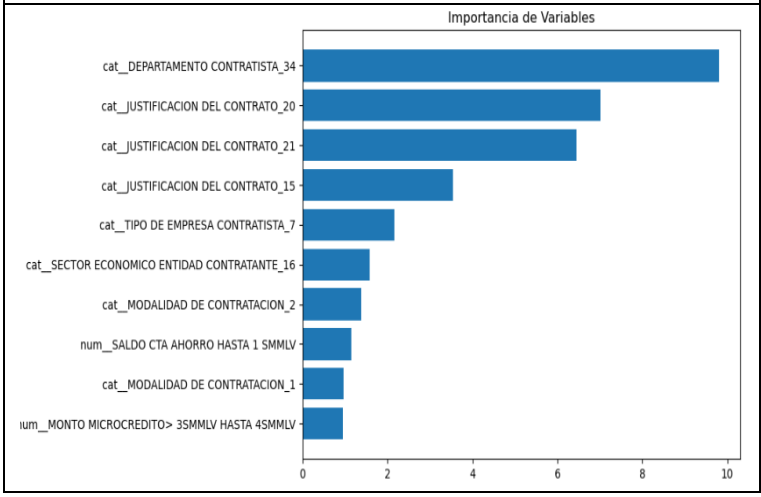
Variables RF con Balanceo Sin IDS



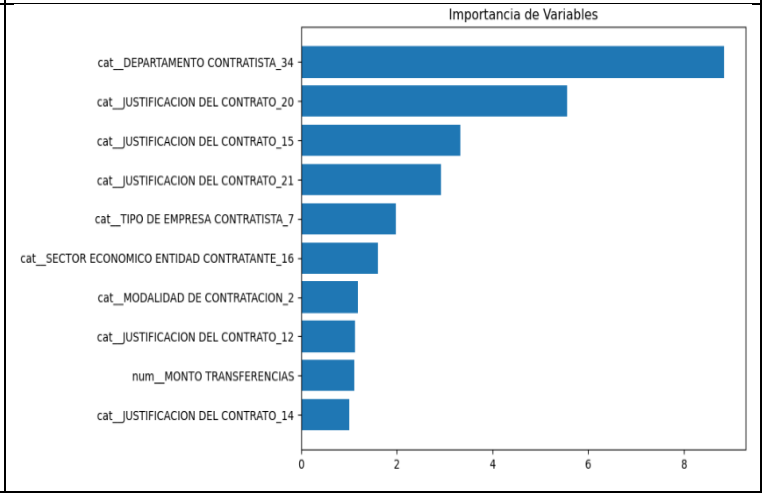
Variables RF con Balanceo Con IDS



Variables RF sin Balanceo Sin IDS



Variables RF sin Balanceo Con IDS



Fuente: Elaboración propia.

Tabla 4. Resultados Sistema de Alertas e Impacto Económico

Probabilidad	Score	Nivel de Riesgo	Valor Contrato	Perdida Esperada	Modelo	Valor Real	Perdida Esperada Real
0,538422	53,84218	Alto	27,16412	14,62575	RF (Sin SubBagging, con IDS)	626.942.708.874,36	337.559.747.197,55
0,565132	56,51325	Alto	25,09271	14,18071	RF (Sin SubBagging, con IDS)	79.000.024.535,34	44.645.441.865,71
0,51001	51,00103	Alto	27,69463	14,12454	RF (Sin SubBagging, con IDS)	1.065.670.186.208,13	543.502.451.668,01
0,528933	52,89326	Alto	26,0611	13,78456	RF (Sin SubBagging, con IDS)	208.060.970.270,52	110.050.313.188,10
0,579065	57,90653	Alto	23,00187	13,31858	RF (Sin SubBagging, con IDS)	9.762.994.461,53	5.653.408.387,86
0,608892	60,88918	Crítico	21,82252	13,28755	RF (Sin SubBagging, con IDS)	3.001.920.236,45	1.827.845.216,61
0,52941	52,94104	Alto	25,07062	13,27265	RF (Sin SubBagging, con IDS)	77.273.893.045,29	40.909.571.717,11
0,597174	59,71739	Alto	22,13095	13,21602	RF (Sin SubBagging, con IDS)	4.086.460.380,90	2.440.327.891,50
0,57738	57,73804	Alto	22,7719	13,14879	RF (Sin SubBagging, con IDS)	7.757.298.661,23	4.478.909.101,02
0,594333	59,43328	Alto	22,04449	13,10176	RF (Sin SubBagging, con IDS)	3.747.999.244,44	2.227.559.634,95
0,563502	56,35025	Alto	27,16412	15,30705	RF (Sin SubBagging, sin IDS)	626.942.708.874,36	353.283.470.336,12

Probabilidad	Score	Nivel de Riesgo	Valor Contrato	Perdida Esperada	Modelo	Valor Real	Perdida Esperada Real
0,559834	55,98341	Alto	25,07062	14,03539	RF (Sin SubBagging, sin IDS)	77.273.893.045,29	43.260.552.639,12
0,642295	64,22947	Crítico	21,82252	14,01649	RF (Sin SubBagging, sin IDS)	3.001.920.236,45	1.928.118.358,27
0,60028	60,028	Crítico	22,77919	13,67029	RF (Sin SubBagging, sin IDS)	7.814.055.997,61	4.690.621.534,25
0,613674	61,36741	Crítico	22,04449	13,52813	RF (Sin SubBagging, sin IDS)	3.747.999.244,44	2.300.049.688,33
0,58377	58,37702	Alto	22,80676	13,31391	RF (Sin SubBagging, sin IDS)	8.032.502.818,78	4.689.134.170,52
0,585536	58,55363	Alto	22,67416	13,27655	RF (Sin SubBagging, sin IDS)	7.035.003.322,95	4.119.247.705,71
0,591212	59,12119	Alto	22,13095	13,08408	RF (Sin SubBagging, sin IDS)	4.086.460.380,90	2.415.964.414,71
0,573075	57,30745	Alto	22,77967	13,05445	RF (Sin SubBagging, sin IDS)	7.817.807.644,81	4.480.190.116,05
0,58377	58,37702	Alto	22,32918	13,03511	RF (Sin SubBagging, sin IDS)	4.982.402.292,88	2.908.576.986,51
0,761652	76,16519	Crítico	26,0611	19,84948	RF (Con SubBagging, con IDS)	208.060.970.270,52	158.470.054.128,48
0,714328	71,43278	Crítico	27,69463	19,78304	RF (Con SubBagging, con IDS)	1.065.670.186.208,13	761.238.052.773,68
0,724345	72,43446	Crítico	27,07415	19,61102	RF (Con SubBagging, con IDS)	573.000.255.751,22	415.049.870.252,12
0,71186	71,18603	Crítico	27,19716	19,36058	RF (Con SubBagging, con IDS)	648.000.302.421,52	461.285.495.281,79
0,720311	72,03109	Crítico	26,61214	19,16902	RF (Con SubBagging, con IDS)	361.000.073.903,18	260.032.324.233,28
0,708471	70,84714	Crítico	26,94125	19,0871	RF (Con SubBagging, con IDS)	501.688.376.529,00	355.431.665.807,88
0,702039	70,20394	Crítico	27,16412	19,07028	RF (Con SubBagging, con IDS)	626.942.708.874,36	440.138.232.395,45
0,71687	71,68697	Crítico	26,46238	18,97008	RF (Con SubBagging, con IDS)	310.788.980.643,63	222.795.296.554,00
0,745142	74,51421	Crítico	25,44848	18,76273	RF (Con SubBagging, con IDS)	112.754.530.911,67	84.018.136.672,58
0,699733	69,97333	Crítico	27,03532	18,91751	RF (Con SubBagging, con IDS)	551.175.440.922,94	385.675.644.803,33
0,731038	73,10381	Crítico	27,07415	19,79224	RF (Con SubBagging, sin IDS)	573.000.255.751,22	418.884.960.963,86
0,713157	71,3157	Crítico	27,69463	19,75062	RF (Con SubBagging, sin IDS)	1.065.670.186.208,13	759.990.152.985,63
0,713004	71,30036	Crítico	26,94125	19,20921	RF (Con SubBagging, sin IDS)	501.688.376.529,00	357.705.819.218,68
0,708726	70,87261	Crítico	27,00188	19,13693	RF (Con SubBagging, sin IDS)	533.047.832.972,38	377.784.858.471,18
0,682849	68,28492	Crítico	27,8853	19,04145	RF (Con SubBagging, sin IDS)	1.180.363.522.530,19	806.010.050.996,21
0,761589	76,15886	Crítico	24,88993	18,95589	RF (Con SubBagging, sin IDS)	64.499.996.080,75	49.122.487.515,14
0,722773	72,27728	Crítico	26,0611	18,83625	RF (Con SubBagging, sin IDS)	208.060.970.270,52	150.380.851.665,34
0,701303	70,13035	Crítico	26,787	18,78582	RF (Con SubBagging, sin IDS)	429.977.620.958,74	301.544.595.511,23
0,665754	66,57543	Crítico	27,69563	18,43849	RF (Con SubBagging, sin IDS)	1.066.743.856.587,95	710.188.989.498,85
0,723899	72,38992	Crítico	25,44848	18,42213	RF (Con SubBagging, sin IDS)	112.754.530.911,67	81.622.892.172,43
0,96844	96,84397	Crítico	27,07415	26,21968	Lasso Con IDS	573.000.255.751,22	554.916.367.679,71
0,967872	96,78724	Crítico	27,03532	26,16674	Lasso Con IDS	551.175.440.922,94	533.467.276.356,97
0,975947	97,59472	Crítico	26,27973	25,64763	Lasso Con IDS	258.905.249.585,28	252.677.801.617,01
0,983886	98,3886	Crítico	25,13075	24,72579	Lasso Con IDS	82.062.747.088,79	80.740.387.982,20
0,908959	90,89589	Crítico	27,19716	24,7211	Lasso Con IDS	648.000.302.421,52	589.005.706.888,77
0,937993	93,79925	Crítico	26,15514	24,53332	Lasso Con IDS	228.576.092.362,24	214.402.774.603,14

Probabilidad	Score	Nivel de Riesgo	Valor Contrato	Perdida Esperada	Modelo	Valor Real	Perdida Esperada Real
0,983563	98,35635	Crítico	24,88993	24,48083	Lasso Con IDS	64.499.996.080,75	63.439.809.645,17
0,98693	98,69301	Crítico	24,75105	24,42755	Lasso Con IDS	56.136.159.711,96	55.402.460.104,53
0,965058	96,50579	Crítico	24,82763	23,9601	Lasso Con IDS	60.604.258.509,31	58.486.624.508,48
0,979335	97,93355	Crítico	24,38595	23,88202	Lasso Con IDS	38.965.715.237,68	38.160.488.732,29

ANEXO

Tabla de Variables

Grupo	Fuente de Información	Variables	Cantidad de Variables
Actividad Ilegal	Ministerio de Justicia / Policía Nacional	Hectáreas Sembradas De Cocaína, Desmantelamiento De Laboratorios, Hectáreas Erradicadas, Incautaciones De Droga En Gramos	4
Capital Humano	Ministerio de Educación	Número Total De Docentes Universitarios ,Número Total De Docentes Universitarios En Universidad Pública ,Número Total De Docentes Universitarios En Universidad Privada ,Número De Matriculados A Educación Superior ,Numero De Admitidos A Educación Superior ,Número De Inscritos En Universidades ,Numero Empleados Administrativos ,Numero De Graduados ,Población Nini ,Puntaje Promedio En Pruebas Saber Pro Competencias Ciudadanas ,Puntaje Promedio En Pruebas Saber Pro Comunicación Escrita ,Puntaje Promedio En Pruebas Saber Pro Escritura ,Puntaje Promedio En Pruebas Saber Pro Inglés ,Puntaje Promedio En Pruebas Saber Pro Lectura Critica ,Puntaje Promedio En Pruebas Saber Pro Razonamiento Cuantitativo ,Tasa De Alfabetización	16
Característica del contrato	Colombia Compra Eficiente (Secop)		20

Grupo	Fuente de Información	Variables	Cantidad de Variables
Comportamiento Financiero	Superintendencia Financiera de Colombia	Nro. Corresponsales Propios ,Nro. Corresponsales Tercerizados, Nro. Corresponsales Activos Nro. Corresponsales Nro. Depósitos, Monto Depósitos, Nro. Giros Enviados, Monto Giros Enviados, Nro. Giros Recibidos, Monto Giros Recibidos, Nro. Pagos Nro. Retiros, Monto Pagos, Monto Retiros, Nro. Transferencias, Monto Transferencias, Nro. Total Monto, Total Nro. Cta Ahorro Hasta 1 Smmlv, Saldo Cta Ahorro Hasta 1 Smmlv, Saldo Cta Ahorro> 1 Smmlv Hasta 3 Smmlv, Nro. Cta Ahorro > 1 Smmlv Hasta 3 Smmlv, Nro. Cta Ahorro> 3 Smmlv Hasta 5 Smmlv, Saldo Cta Ahorro> 3 Smmlv Hasta 5 Smmlv, Nro. Cta Ahorro Activas, Saldo Cta Ahorro Activas Nro. Cta Ahorro Mujeres, Saldo Cta Ahorro Mujeres, Nro. Cta Ahorro Hombres, Saldo Cta Ahorro Hombres, Nro. Total Cta Ahorros, Saldo Total Cta Ahorros, Nro. Cta Ahorro Electrónicas Activas, Saldo Cta Ahorro Electrónicas Activas, Nro. Cta Ahorro Electrónicas Mujeres, Saldo Cta Ahorro Electrónicas Mujeres, Nro. Cta Ahorro Electrónicas Hombres, Saldo Cta Ahorro Electrónicas Hombres, Nro. Total Cta Ahorros Electrónicas Saldo Total Cta Ahorros Electrónicas, Nro. Crédito Consumo Mujeres, Monto Crédito Consumo Mujeres, Nro. Crédito Consumo Hombres, Monto Crédito Consumo Hombres, Nro. Total Crédito Consumo, Monto Total Crédito Consumo, Nro. Crédito Consumo Bajo Monto Mujeres ,Monto Crédito Consumo Bajo Monto Mujeres, Nro. Crédito Consumo Bajo Monto Hombres, Monto Crédito Consumo Bajo Monto Hombres, Nro. Total Crédito Cons Bajo Monto, Monto Total Crédito Consumo Bajo Monto Nro. Crédito Vivienda Mujeres, Monto Crédito Vivienda Mujeres, Nro. Crédito Vivienda Hombres, Monto Crédito Vivienda Hombres, Nro. Total Crédito Vivienda, Monto	78

Grupo	Fuente de Información	Variables	Cantidad de Variables
		Total Crédito Vivienda, Nro. Microcrédito Hasta 1 Smmlv <,Monto Microcrédito Hasta 1 Smmlv, Nro. Microcrédito > 1 Smmlv Hasta 2 Smmlv, Monto Microcrédito> 1smmlv Hasta 2smmlv, Nro. Microcrédito> 2 Smmlv Hasta 3 Smmlv, Monto Microcrédito> 2smmlv Hasta 3smmlv, Nro. Microcrédito> 3 Smmlv Hasta 4 Smmlv, Nro. Microcrédito> 4 Smmlv Hasta 10 Smmlv, Monto Microcrédito> 3smmlv Hasta 4smmlv, Monto Microcrédito> 4smmlv Hasta 10smml, Nro. Microcrédito> 10smmlv Hasta 25smmlv, Monto Microcrédito> 10smmlv Hasta 25smmlv, Nro. Microcrédito Mujeres, Monto Microcrédito Mujeres, Nro. Microcrédito Hombres Monto Microcrédito Hombres, Nro. Total Microcrédito, Monto Total Microcrédito, Nro. Prod Deposito Nivel Nacional, Monto Prod Deposito Nivel Nacional	
Conflicto Armado	Unidad de Victimas / Agencia para la Reincorporación y la Normalización / Prosperidad Social / Ministerio de Defensa	Población En Reincorporación ELN, Población En Reincorporación Farc, Población En Reincorporación Paramilitares, Población En Reincorporación Otros, Desplazados Precipita, Desplazados Número De Casos, Promedio Índice Lica, Combate Y/O Contacto Armado Farc, Combate Y/O Contacto Armado ELN, Combate Y/O Contacto Armado Paramilitar, Combate Y/O Contacto Armado Otros	11
Desarrollo Social	Departamento Administrativo Nacional de Estadística (DANE)	IDH, PIB Agricultura Y Ganadería, PIB Explotación De Minas Y Canteras, Pib Industrias Manufactureras, PIB Suministro De Electricidad, Gas Etc., PIB Construcción, PIB Comercio Al Por Mayor Y Al Por Menor, Pib Información Y Comunicaciones, Pib Actividades Financieras, PIB Actividades Inmobiliarias, PIB Actividades Profesionales, PIB Administración Pública Y Defensa, PIB Actividades Artísticas	12
	Global Data Lab		1

Grupo	Fuente de Información	Variables	Cantidad de Variables
Función Pública	Departamento Nacional de Planeación Registraduría Nacional del Estado Civil (RNEC)	Índice Eficiencia Fiscal, Índice Eficacia Total, Índice Eficiencia Total, Índice Requisitos Legales, Capacidad Administrativa, Delitos Electorales	10
Variable dependiente	Colombia Compra Eficiente (SECOP)	Multas SECOP I / II	1
Violencia	Ministerio de Defensa Nacional	Homicidios, Hurtos Personas, Violencia Intrafamiliar, Hurto A Vehículos, Secuestros, Delitos Sexuales, Delitos Medio Ambiente, Hurto A Comercios, Financiera, Residencias, Incautaciones De Armas	11