INTELIGENCIA ARTIFICIAL EN EL MONITOREO DE LA CONTRATACIÓN PÚBLICA EN SALUD

Presentado para obtener el título de

MAGÍSTER EN MATEMÁTICAS APLICADAS Y CIENCIAS DE LA COMPUTACIÓN

Andrés Sebastián Salazar Mejía

Dirección:

Jorge Gallego Durán

Universidad del Rosario

Escuela de Ingeniería, Ciencia y Tecnología

Maestría en Matemáticas Aplicadas y Ciencias de la Computación

RESUMEN

El sector salud en Colombia enfrenta grandes desafíos relacionados con la adecuada gestión de los recursos en los procesos de contratación, necesarios para el funcionamiento del sistema. La contratación pública es el aspecto más susceptible al despilfarro en Colombia, relacionado con altas ineficiencias, que son más frecuentes y pueden comprometer mayores recursos. A esto se suma un aumento de la participación del Estado y la cantidad de contratos públicos debido a posibles reformas. Este trabajo propone una metodología basada en aprendizaje de máquinas para predicción temprana de ineficiencias en los contratos del sector salud, usando como medida el número de adiciones en valor (sobrecostos) reportadas en la plataforma de contratación del Estado. Además, introduce a la literatura de contratación pública el uso de modelos de regresión y la incorporación de ensambles de modelos de detección de anomalías. Entre los principales hallazgos destacan los métodos basados en árboles como los de mejor desempeño, especialmente *Random Forest*, con un RMSE de prueba menor a 0,7. El estudio también identifica el tamaño del contrato, en valor y duración, las condiciones de entrega y el presupuesto de las entidades como características valiosas para la formulación de políticas públicas; y muestra que la incorporación de medidas de detección de anomalías mejora la comprensión de las ineficiencias. Adicionalmente, se encuentra la medida de ineficiencias propuesta superior a la usada en la literatura para la priorización de la investigación de contratos por parte de las entidades de control, con un MAP_{1000} de 1. Esta metodología permitirá mejorar el cuidado de los recursos del sistema de salud colombiano por medio de una eficaz intervención por parte de las entidades de control.

ABSTRACT

The health sector in Colombia faces major challenges in managing public procurement resources, which are crucial for the system's functioning. Public procurement remains the most vulnerable aspect of resource allocation in Colombia, characterized by high inefficiencies that increasingly compromise significant financial resources. Additionally, potential reforms imply greater state participation and more public procurement processes. This research proposes a machine learning methodology for early prediction of inefficiencies in health sector contracts, utilizing the number of value additions (cost overruns) reported on the state's e-procurement platform. The study introduces to public contracting literature the use of regression models and ensembles of anomaly detection models. Among the main findings, tree-based methods emerge as top performers, with Random Forest achieving a test RMSE of less than 0.7. The research identifies several valuable characteristics for public policy formulation, including contract size (both in value and duration), delivery conditions, and institutional budgets. Furthermore, the incorporation of anomaly detection measures enhances the understanding of procurement inefficiencies. Finally, the proposed inefficiency measurement method proves superior to existing literature approaches for prioritizing contract investigations by control entities, demonstrating a MAP_{1000} of 1. This methodology aims to improve resources care in the Colombian health system by enabling effective interventions by control entities.

TABLA DE CONTENIDO

1 INTRODUCCIÓN 1

2 OBJETIVOS 3

2.1 Objetivo general 3
2.2 Objetivos específicos 4
3 PROBLEMA Y JUSTIFICACIÓN 4

4 MARCO TEÓRICO Y ESTADO DEL ARTE 9

4.1 Contratación pública, despilfarro e ineficiencias 9
4.2 Aproximaciones para la supervisión de la contratación pública 10
5 METODOLOGÍA 13

5.1 Datos 14
5.2 Modelamiento 18
6 RESULTADOS Y DISCUSIÓN 26

6.1 Modelos de predicción de ineficiencias 26
6.2 Análisis de variables explicativas 34
6.3 Predicción vs clasificación en la predicción de ineficiencias 37
7 CONCLUSIONES Y RECOMENDACIONES 40

REFERENCIAS 43

ANEXOS 48

LISTA DE TABLAS

Tabla 1. Métricas de desempeño modelos de regresión..... 28
Tabla 2. Búsqueda de mejores hiper parámetros mediante Optuna..... 30
Tabla 3. Métricas de desempeño modelos con mejores hiper parámetros (tunned). 31
Tabla 4. Métricas de desempeño en contratos directamente relacionados a la atención del paciente. 33
Tabla 5. Métricas de recomendación de contratos con ineficiencias..... 38
Tabla 6. Anexo A: Variables consolidadas..... 50
Tabla 7. Anexo B: Distribución de puntajes de atipicidad. 51
Tabla 8. Anexo C: Métricas de desempeño en regresión excluyendo atipicidad. 51

LISTA DE FIGURAS

Figura 1. Número de contratos del sector Salud y Protección social firmados mensualmente. 6
Figura 2. Distribución del conteo de adiciones en valor en los contratos. 17
Figura 3. Distribución de los puntajes de atipicidad de Local Outlier Factor (LOF) vs Isolation forest en los datos de entrenamiento..... 28
Figura 4. Densidad de probabilidad del número de adiciones observado vs predicciones..... 32
Figura 5. Importancia de variables en los mejores modelos de predicción: Random Forest, XGBoost y CatBoost..... 35
Figura 6. Gráficas de dependencia parcial de las principales variables numéricas en Random Forest..... 36
Figura 7. Densidad de probabilidad de clasificación de sobrecostos. 39

1 INTRODUCCIÓN

El sistema de salud en Colombia se financia con recursos del Presupuesto General de la Nación (PGN) y de los aportes del régimen contributivo que son recursos administrados por la Administradora de los Recursos del Sistema General de Seguridad Social en Salud (ADRES), adscrita al Ministerio de Salud y Protección Social. El ADRES se encarga de distribuir estos recursos entre las diferentes entidades prestadoras de servicios de salud públicas y privadas. Estos recursos en 2022 representaron el 5,2% del Producto Interno Bruto (PIB), de los cuales 2,8% son recursos fiscales [1]. Sin embargo, el sector tiene una particular vulnerabilidad al despilfarro, las ineficiencias y la corrupción "debido a la gran cantidad de recursos, la asimetría de la información, la gran cantidad de actores, la complejidad y fragmentación del sistema y la naturaleza globalizada de la cadena de suministro de medicamentos y dispositivos médicos" [2].

La contratación pública desempeña un papel crucial en el sector dado que permite garantizar la prestación de los servicios en salud y condiciona su calidad, por lo cual gran parte del presupuesto del sector se destina para esto. Sin embargo, la contratación es el ámbito de gestión más susceptible a la corrupción con un 38% de los casos confirmados de corrupción en Colombia entre 2021 y 2022 [3]. En este contexto, el riesgo en los recursos es aún mayor debido a que la presencia de ineficiencias, como prórrogas y sobrecostos, es más común y compromete más recursos [4]. Sumado a lo anterior, el aumento en la participación del Estado en los procesos de contratación, impulsado por la posible reforma al sistema de salud, añade más carga sobre los organismos de control para realizar investigaciones oportunas. Esta situación subraya la necesidad de contar con herramientas para detectar de manera temprana posibles riesgos en los procesos de contratación pública y priorizar los esfuerzos para las investigaciones y acciones pertinentes, especialmente en las entidades que participan directamente en la administración del sistema de salud.

De tal manera, la presente investigación tiene como objetivo proponer una metodología basada en inteligencia artificial y aprendizaje de máquinas que identifique de forma

temprana riesgos en los recursos dispuestos para la contratación pública en el sector salud. Así, el propósito de los modelos de pronósticos es predecir el riesgo de ineficiencias de los contratos, medido como el número de adiciones en valor (sobrecostos) reportados en la plataforma. De modo que, los entes de control tengan una medida de alerta para todo el sistema de contratación en salud que permita priorizar la supervisión de los contratos y tomar medidas correctivas en el tiempo oportuno, que promuevan la eficiencia y confianza en la disposición de los recursos en el sector salud.

Para el anterior objetivo, se presenta una metodología que hace uso de datos transaccionales de acceso público, y con alta frecuencia de actualización, registrados en la plataforma de compra pública SECOP II. Para el estudio se emplean contratos digitales del sector "Salud y Protección Social" firmados entre 2017 y marzo de 2024, y etiquetados como terminados, cerrados o modificados, incluyendo todos los tipos de contratos. Estos datos pasan por una etapa de limpieza e ingeniería de características resultando en un conjunto de estudio de 270.208 contratos, 36 variables explicativas y 1 variable objetivo.

Esta investigación evalúa diversos modelos de aprendizaje de máquinas para la predicción de las ineficiencias cuya salida permite tener una medida para priorizar las investigaciones e intervenciones de las entidades competentes. De igual forman, se estudia la pertinencia de la incorporación de modelos de detección de anomalías como variables explicativas, y se analiza la relación entre las variables explicativas y las predicciones para identificar características que podrían ser útiles para el diseño de políticas públicas. Adicionalmente, se compara el uso de los resultados de los modelos de predicción propuestos como medida para la priorización de la investigación de los entes de control, contra las probabilidades resultantes de los modelos de clasificación usados en la literatura.

Como resultado, se encuentra *Random Forest* como el mejor modelo de pronóstico, con un RMSE de 0,683 y un MPD de 0,341 en los datos de prueba, y un desempeño cercano en ambas métricas al evaluarlo solo en contratos directamente relacionados a la atención del paciente. Al analizar la importancia de las variables en el anterior modelo, se evidencia un

hecho común de la literatura donde contratos de mayor duración y valor son más susceptibles a sobrecostos. No obstante, se encuentran también que un menor uso de recursos propios de la entidad y dejar las condiciones de entrega del contrato a convenir se relacionan con menos sobrecostos. Además, el modelo resulta exitoso como sistema de recomendación para la priorización de contratos a investigar, con un MAP_{1000} de 1.

Las siguientes secciones de este trabajo se estructuran como se describe a continuación. La sección 2 presenta explícitamente los objetivos generales y específicos del proyecto. La sección 3 explica la problemática que da origen al trabajo y la relevancia del problema. Seguidamente, la sección 4 describe el marco teórico de la corrupción e ineficiencias en la contratación pública y las diferentes aproximaciones para su detección temprana en la literatura. Luego, la sección 5 explica las fuentes de datos utilizadas y su procesamiento, junto con una explicación detallada de la metodología de modelamiento utilizada. Posteriormente, la sección 6 presenta los resultados del proyecto y discute sobre las implicaciones del uso práctico de la metodología propuesta. Finalmente, la sección 7 presenta las conclusiones del proyecto junto con los aportes teóricos y prácticos de la metodología propuesta.

2 OBJETIVOS

2.1 Objetivo general

Proponer una metodología basada en inteligencia artificial que identifique de forma temprana riesgos en los recursos dispuestos para la contratación en el sector salud y protección social, de tal manera que apoye la labor de las veedurías, personerías ciudadanas y/o demás entidades de control de los recursos públicos.

2.2 Objetivos específicos

1. Determinar una medida de ineficiencia para los contratos públicos en el sector salud y protección social que permita dar cuenta de riesgos en el uso de los recursos monetarios durante la ejecución de los contratos.
2. Desarrollar modelos de aprendizaje supervisado para la detección temprana de ineficiencias que se puedan presentar durante la ejecución de un contrato del sector salud y protección social.
3. Incorporar técnicas de aprendizaje no supervisado para detección de anomalías en la identificación de contratos que comprometan recursos públicos en el sector salud y protección social.

3 PROBLEMA Y JUSTIFICACIÓN

Según Transparencia por Colombia [5], entre 2016 y 2020 se denunciaron en los medios de comunicación 67 casos de corrupción relacionados con el sector salud, por un costo estimado de 1,63 billones de pesos, a los que se suman otros 32 casos registrados en 2021 y 2022 [3]. Además, en el caso colombiano se encuentra que el ámbito de gestión más susceptible es la contratación pública con un 38% de los casos confirmados de corrupción entre 2021 y 2022, donde los sucesos se vieron enmarcados por falencias en la planeación, selección, ejecución y seguimiento a los procesos de contratación [3].

No obstante, esto es solo la punta del iceberg de los problemas de la contratación pública en salud. Los hechos comprobados o alegados de corrupción tienen un mayor alcance en los medios de comunicación que la falta de eficiencia en los proyectos del estado. Sin embargo, se estima que el despilfarro de los recursos es considerablemente mayor por ineficiencias en la contratación, entendidas como sobrecostos y prorrogas, que por hechos de corrupción [4]. Sumado a esto, el sector salud requiere de grandes inversiones para mejorar su cobertura y calidad, e internacionalmente se ha encontrado que los

megaproyectos en el sector tienen en promedio un 29% de sobrecostos [6]. Lo que sugiere un problema aún mayor que no es abiertamente explorado en el caso colombiano.

En Colombia, la contratación del Estado se realiza por medio del sistema electrónico para la contratación pública (SECOP II), la plataforma transaccional por la cual se crean, evalúa y adjudican los procesos de contratación pública por medio de cuentas para las entidades y los proveedores[1]. Esta obliga a las entidades a publicar todas las acciones y documentos de los procesos, así como registrar las eventualidades presentadas durante la ejecución del contrato como las adiciones en valor, sobrecostos, y adiciones en tiempo, prórrogas.

En términos de contratación, el sector salud y protección social tiene una gran participación en el total nacional. En 2023 el 18,3% de los contratos firmados SECOP II fueron del sector salud y protección social, solo por debajo del sector de servicios públicos con el 23,4% de los contratos. Adicionalmente, como se ve en la Figura 1, se presenta una tendencia creciente en los niveles de contratación desde 2020 con la crisis del Covid 19, con un crecimiento mayor a la tendencia en enero de 2023 llegando a un máximo histórico de 26.532 contratos firmados. Esto se aúna al contexto actual colombiano, donde estamos ante las consecuencias de una posible reforma al sistema de salud, donde el ADRES y las Direcciones Departamentales y Distritales de Salud pasarían a contratar directamente a nivel nacional y regional respectivamente. A su vez, se daría la transición de las Empresas Sociales del Estado (ESE) a Instituciones de Salud del Estado (ISE), creándose como entidades públicas descentralizadas [7] que crearían más carga para el sistema de contratación pública colombiano.

La posición del gobierno respecto a la inviabilidad de las Entidades Promotoras de Salud (EPS), con la intervención de entidades como Nueva EPS, Sanitas, Famisanar; y la decisión de retirarse del sistema de salud de EPS Sura y EPS Compensar, con más de 7 millones de

[1] La contratación está sujeta a las disposiciones de la Ley 80 de 1993 por la cual se expide el Estatuto General de Contratación. Algunas entidades pueden estar exentas de este por competir en el mercado, por lo cual registran sus contratos bajo la modalidad de Régimen Especial.

afiliados entre las dos, crea una incertidumbre en el manejo del sistema de salud y la posibilidad de poner un mayor compromiso administrativo en el gobierno. Respecto a esta administración, la Administradora de los Recursos del Sistema General de Seguridad Social en Salud (ADRES) administra las fuentes de financiación y realiza los giros a los prestadores y proveedores del sistema de salud en Colombia. En caso de aprobarse la reforma a la salud[2], esta además pasaría a ser el pagador único del sistema de salud, así como la entidad encargada del seguimiento de los gastos y transferencias del sistema. Esto se traduce en una mayor participación en la contratación de los bienes y servicios relacionados con el sector, y, por lo tanto, un aumento de los contratos publicados a nivel nacional, comprometiendo la capacidad de los organismos de control.

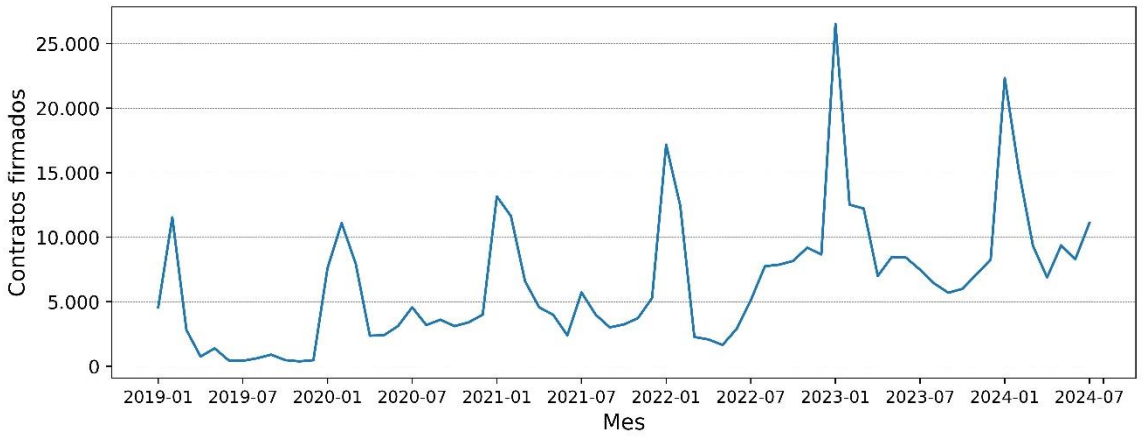


Figura 1. Número de contratos del sector Salud y Protección social firmados mensualmente.

La máxima autoridad en el control de los recursos en Colombia es la Contraloría General de la República, junto con las contraloría departamentales y territoriales, que se encarga del control fiscal y vela por el buen uso de los recursos públicos. Sin embargo, esta ejerce un control posterior y selectivo sobre las entidades, no sobre contratos, ni a manera

[2] Prevista para iniciar trámite en el Congreso en diciembre de 2024. https://caracol.com.co/2024/10/17/minsalud-se-espera-la-reforma-a-la-salud-inicie-su-tramite-en-el-congreso-en-diciembre/ Último acceso: 31 de Octubre de 2024.

preventiva. Adicionalmente, existen otras entidades de carácter especial, como la Veeduría distrital de Bogotá D.C. que se encarga del control preventivo del patrimonio del distrito, y la Superintendencia Nacional de Salud (Supersalud), que ejerce la inspección, vigilancia y control, sobre la ejecución de los recursos destinados a la salud en las entidades territoriales. Sin embargo, todas tienen en común limitaciones en tiempo, recursos y personal, por lo que la revisión detallada o auditoría de contratos por parte de expertos se realiza solo a selectos casos.

Igualmente, las estrategias implementadas para combatir las falencias del sistema suelen dirigirse a problemas, procesos o instituciones específicas, en conjunto con herramientas comúnmente normativas y prescriptivas que no capturan las complejas dinámicas del sector [2]. Así, lo anterior desencadena en un problema de vulnerabilidad del sistema de contratación pública en el sector salud ante ineficiencias, y potencial corrupción, que comprometen los recursos destinados a una atención ideal para los ciudadanos. Siendo esto especialmente relevante frente a potenciales reformas que aumentarán los desafíos para las entidades de control.

Para atacar este problema se requiere nutrir el sistema de salud con herramientas innovadoras que permitan a las agencias reguladoras y de supervisión priorizar adecuadamente sus esfuerzos en las labores de investigación e intervención. En particular, se requiere una herramienta para la supervisión temprana que permita identificar posibles ineficiencias en la contratación pública que pongan en riesgo la cobertura de nuevos bienes y servicios, o el mantenimiento de los existentes.

Para lograrlo, la inteligencia artificial ofrece alternativas que permiten la autonomía en el aprendizaje, el análisis de grandes volúmenes de datos y la imparcialidad para combatir las ineficiencias y la corrupción [8]. Así, herramientas basadas en la rama del aprendizaje de máquinas, similares a la requerida, han sido exitosas en la práctica. En Argentina, el sistema PROMETEA ha apoyado a la fiscalía en la reducción significativa en los tiempos de resolución de pliegos de contratación, procesos de requerimiento de juicios, entre otros

casos [9]. En Colombia, la Contraloría General de la República desarrolló la plataforma Océano que depura, limpia y enriquece la información contractual para identificar posibles nichos de corrupción. Sin embargo, estas herramientas son necesarias, pero no suficientes para el control de la actividad contractual del estado. Esto porque se concentran en hechos de corrupción en los cuales se puedan tomar acciones judiciales, dejando a la sombra las pérdidas por ineficiencias en la ejecución contractual. De tal forma que no dan prioridad a un control preventivo que evite las pérdidas, en veces esforzarse en recuperarlas. Más aún, en Colombia no hay propuestas que se especialicen en el sector salud, de tal manera que permitan capturar las dinámicas propias del sistema de salud, que son diferentes a otros sectores de la economía colombiana.

Finalmente, desde el punto de vista práctico, donde las entidades de control deben priorizar un pequeño grupo de casos a investigar con mayor detalle, aunque no hay proyectos estatales, hay investigaciones que han propuesto el uso de modelos binarios de clasificación para identificar los contratos de mayor interés [10, 11]. Para esto, usan la probabilidad de pertenencia al grupo de contratos con ineficiencias, de tal manera que, contratos con una mayor probabilidad deben priorizarse sobre aquellos con menor probabilidad. No obstante, esta perspectiva presenta dos problemas al usarse para recomendar los contratos a intervenir. Primero, los modelos de clasificación binarios no dan cuenta de los diferentes niveles de gravedad de las ineficiencias observadas. Y segundo, las probabilidades de tener ineficiencias no permiten priorizar entre contratos en los cuales los modelos tienen mayor certeza de la presencia de ineficiencias dado que estas probabilidades se acumulan en un único valor. Por esta razón, el presente trabajo propone el uso de modelos de regresión que ofrecen predicciones continuas sin un límite superior, que representan diversos niveles de riesgo, para sugerir a los contratos en los cuales se deben concentrar las entidades en control, como se explicará posteriormente.

4 MARCO TEÓRICO Y ESTADO DEL ARTE

4.1 Contratación pública, despilfarro e ineficiencias

La Organización para la Cooperación y el Desarrollo Económico (OCDE) define la contratación pública como el proceso de compra de bienes, servicios y obras por parte de los gobiernos y empresas estatales. Esta contratación corresponde a gran parte de los impuestos recolectados por la nación, por lo que se requiere una gestión eficiente que garantice la calidad de los servicios y salvaguarde el interés público. Así, Shiundu & Rotich [12] define como un sistema de contratación pública eficiente aquel que permite la adquisición de bienes y servicios con calidad, servicio y precios satisfactorios en un plazo de tiempo oportuno.

La OCDE divide los procesos de contratación en tres etapas: precontractual, adjudicación y ejecución. La etapa precontractual es en la cual la entidad pública define los objetivos y el producto a contratar, y diseña las condiciones del contrato. La etapa adjudicación comprende las actividades relacionadas a la publicación del proceso, la revisión de ofertas y la elección del proveedor al cual se adjudica el contrato. Por último, en la etapa de ejecución el proveedor pone en marcha las acciones para cumplir con el objetivo del contrato, y es la etapa donde se materializan los riesgos de la contratación. Durante la ejecución del contrato se evidencian las fallas que no pudieron ser tenidas en cuenta inicialmente y se traducen en ineficiencias que obligan a realizar adiciones al tiempo o al costo del bien o servicio. Por lo que la carencia de una adecuada gestión repercute en el despilfarro de los recursos al generar gastos adicionales y un costo de oportunidad para la ejecución de las labores de la administración pública [11].

Bandiera et. al. [4] distingue el despilfarro de recursos públicos entre activos y pasivos. El despilfarro activo se refiere al caso en el cual los servidores públicos reciben un beneficio directo, como sobornos, tráfico de influencia o favores políticos; que conocemos como corrupción. Por otro lado, el despilfarro pasivo es aquél en el cual los servidores públicos no reciben un beneficio directo, como prórrogas y sobrecostos que llamamos ineficiencias.

Bandiera et. al. [4] estima con datos de contratación pública en Italia que a las ineficiencias se les atribuye el 83% del despilfarro entre las dos distinciones, generando un riesgo considerablemente mayor de los recursos públicos. Sin embargo, el concentrase en atacar las ineficiencias no da rienda suelta a la corrupción, al contrario, Dal Bo & Rosi [13] documenta que mayores ineficiencias están asociadas con mayor corrupción, y viceversa.

La presencia de ineficiencias y la falta de transparencia en la contratación puede provocar mayores costos [14, 15], inadecuada asignación de los recursos [16] y barreras a la innovación [17]. Además, puede afectar la competitividad y funcionamiento de la nación en aspectos como la calidad de la educación [18], la disminución del crecimiento por desincentivos a la inversión [19], la prestación de servicios de agua y saneamiento [20], y la calidad de la prestación de los servicios de salud [21, 22].

4.2 Aproximaciones para la supervisión de la contratación pública

Las ineficiencias en la contratación son un riesgo mayor para la prestación adecuada de los servicios del Estado, por lo cual existe la necesidad de supervisar los procesos de contratación y tomar las medidas oportunas. Olken [23] afirma que un mayor monitoreo y auditoría desde las altas esferas del Estado tienen un rol significativo en la reducción de la corrupción y las ineficiencias. Esto reafirma la necesidad de tener entidades que supervisen la ejecución de los recursos públicos, en especial aquellas entidades que los administran y distribuyen.

Para la medición y detección del riesgo de corrupción e ineficiencias en la contratación pública la literatura ha abortado diferentes estrategias. Tradicionalmente, la medición se realiza mediante encuestas de percepción ciudadana que no permiten una identificación granular de las dinámicas. En respuesta a lo anterior, han surgido índices objetivos de riesgo de corrupción que se basan en la composición de señales de advertencia[3] que sirven como alertas tempranas a diferentes niveles de agregación [24, 25, 26]. En Colombia, esta

[3] En la literatura se encuentran comúnmente como "*red flags*".

línea ha sido tratada por Zuleta et.al. [27], donde usan los datos de SECOP para construir una serie de indicadores organizados en tres dimensiones: falta de competencia, falta de transparencia y anomalías en los procesos de compra.

Por otro lado, la contratación pública no ha sido ajena al uso de inteligencia artificial y aprendizaje de máquinas, pero la literatura se ha concentrado en mayor medida en la corrupción a altos niveles de agregación. Melo & Denle [28] explora la clasificación de la corrupción a nivel de país encontrando el Random Forest como el mejor modelo. A nivel municipal, Collonelli et.al. [29] y De Blasio et.al. [30] desarrollan modelos para clasificar la corrupción usando datos de auditorías e investigaciones policiales, en ambos casos destacando el desempeño de los modelos basados en árboles. A nivel de contrato, Decarolis & Giorgiantonio [31] usa diferentes modelos de aprendizaje de máquinas para identificar características relacionadas con corrupción e ineficiencias en proyectos viales en Italia.

En Colombia, Rodriguez [11] y Gallego et.al. [10] usan la presencia de adiciones en tiempo y valor reportadas para los contratos de SECOP II como medida de ineficiencia y entrenan diferentes modelos de clasificación[4], entre los que destacan modelos de *boosting* basados en árboles. Igualmente, los autores resaltan que en la práctica las agencias solo pueden investigar un pequeño grupo de los casos clasificados positivamente, por lo que deben priorizar los contratos dadas las limitaciones de recursos. En este caso, eligiendo los contratos con mayor probabilidad de tener ineficiencias o corrupción. Sin embargo, el uso de modelos de clasificación como sistema de recomendación sufre de dos limitaciones.

Primero, al asumir las ineficiencias como una medida dicotómica los modelos entrenados no permiten directamente discriminar entre contratos con ineficiencias, por lo que las recomendaciones de contratos a intervenir no dan cuenta de diferentes niveles de riesgo que un contrato puede representar para los recursos públicos. Además, usar la probabilidad

[4] Gallego et.al. (2021) realiza el proceso de clasificación con 3 conjuntos de datos diferentes como diferentes fallas en la contratación pública, entre ellas la ineficiencia de los contratos medida como la presencia de adiciones en tiempo o valor.

de contener adiciones da noción de la certeza de que un contrato tenga ineficiencias, sin importar su gravedad, por lo que usar esta medida puede llevar a priorizar contratos con un menor costo de inacción sobre otros en los que ponen en riesgo una mayor cantidad de recursos.

Segundo, la densidad de las probabilidades de tener ineficiencias resultantes de modelos de clasificación. Mejores modelos de clasificación implican una mejor separación entre las dos clases, por lo que entre mejor sea el modelo más se acumularán las probabilidades de salida en 0 y 1. Esto implica que en la búsqueda de mejores modelos se obtengan más contratos con probabilidad de 1, provocando que no se pueda priorizar entre estos contratos en los que el modelo tiene certeza.

Por otro lado, la literatura también ha explorado el uso de métodos estadísticos y de aprendizaje no supervisado de detección de anomalías para caracterizar la corrupción e ineficiencias en la contratación. Por un lado, diversos autores han usado algoritmos como *Isolation Forest*, mapas auto-organizados y análisis de componentes principales, para usar sus salidas como clasificadores o construir indicadores interpretables del riesgo en los procesos de contratación [32, 33, 34]. Por otro lado, se ha experimentado exitosamente con el uso de puntajes de atipicidad[5] generados con modelos de aprendizaje profundo y grafos directamente como rankings para la priorización de contratos para las actividades de investigación [35, 36]. Sin embargo, en la literatura de contratación pública no se ha explorado el ensamblaje de modelos de detección de anomalía como variables predictoras en modelos de aprendizaje supervisado, exitoso en otros campos [37, 38].

El presente trabajo ayuda a cerrar la brecha del despilfarro pasivo en la literatura, que se ha enfocado en el despilfarro activo, priorizando los casos de corrupción detectados en el sistema judicial. Además, el estudio acerca el uso de métodos de aprendizaje supervisado y no supervisado en la literatura de corrupción e ineficiencias en la contratación pública,

[5] Los puntajes de atipicidad son resultado de puntajes asignados por los modelos de detección de anomalías.

que han sido excluyentes en esta. Adicionalmente, la metodología propuesta para medir el riesgo de ineficiencias por medio del conteo de sobrecostos permite cubrir dos limitaciones del uso modelos de clasificación como sistemas de recomendación planteado en estudios previos. Primero, permite discriminar el nivel de riesgo entre contratos que presentan ineficiencias, y segundo, evita la acumulación del valor por el cual se ranquean los contratos en un único valor.

En el caso Colombiano, este estudio ayuda a destacar la ventajas y limitaciones de SECOP II para el estado colombiano, que se une a las plataformas digitales de contratación[6] que internacionalmente han mostrado potencial para reducir las ineficiencias y corrupción, y mejorar la transparencia [39, 40]. Por otro lado, los anteriores autores que trabajan el contexto colombiano, tanto enfocados en ineficiencias como en corrupción, no discriminan los contratos por los sectores que tratan, ignorando la heterogeneidad de las dinámicas en los servicios y productos. Este trabajo, al centrarse en el sector salud, revela que especializar las estrategias de generación de alertas mejora el desempeño de los modelos planteados y su uso como sistemas de recomendación, lo que sugiere dirigir próximos estudios de forma sectorial que permita identificar las dinámicas propias de sus contratos.

5 METODOLOGÍA

Esta investigación trabaja sobre la premisa de que la disponibilidad de nuevas fuentes de información y la mejora de la capacidad predictiva mediante el uso de aprendizaje de máquinas puede tener efectos importantes en la implementación de políticas públicas [41]. Sin embargo, debe tenerse en cuenta que para una implementación efectiva la fuente de información debe ser confiable, amplia y escalable, y se deben usar herramientas que capturen patrones complejos en los datos [10]. Con esto presente, la metodología desarrollada usa datos precisos, consistentes, de acceso público, de frecuente actualización

[6] Llamadas en la literatura en la literatura anglosajona como estrategias *e-procurement*.

y que cubren claramente la necesidad de la investigación. Así mismo, se implementan modelos de aprendizaje de máquinas para regresión y se incorporan modelos de detección de anomalías como nuevas variables explicativas.

A continuación, se detallan los datos y métodos usados para el objetivo de la investigación.

5.1 Datos

La fuente de información de este proyecto es el conjunto de información generada por la plataforma de contratación pública SECOP II[7], administrada por Colombia Compra Eficiente, entidad rectora del sistema de compra pública en Colombia. La información transaccional es publicada para el escrutinio público en el portal de datos abiertos del Estado colombiano, desde el cual se obtiene toda la información usada. Estos datos están disponibles programáticamente desde la API Socrata a la cual se accede mediante la librería sodapy del lenguaje de programación Python.

La tabla principal usada es SECOP II - Contratos Electrónicos[8] que contiene información al nivel más detallado, los contratos, y es actualizada diariamente. Para el propósito de la investigación se seleccionan los contratos firmados entre enero de 2017 y marzo de 2024 correspondientes al sector "Salud y Protección Social", y cuyo estado se encuentra en "Modificado", "terminado" o "cerrado" para utilizar contratos a los que efectivamente les pudieron haber reportado adiciones. También, se incluyen todas las tipologías de contratos, como prestación de servicios, suministros, obras, entre otros. Además, solo se usarán variables explicativas generadas antes de la etapa de ejecución del contrato, debido a que el seguimiento desincentiva las ineficiencias e irregularidades [42], por lo que debe iniciar

---

[7] También existe la versión anterior SECOP I, sin embargo, esta es una plataforma publicitaria, por lo que su uso está sujeto al juicio de las entidades para cargar la información y las actualizaciones pertinentes. Por el contrario, SECOP II es una plataforma transaccional que obliga el cargue y actualizaciones de la información para la creación de los procesos de compra.

[8] Ver https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Contratos-Electr-nicos/jbjy-vk9h/about_data Último acceso: 7 de Agosto de 2024.

lo más temprano posible[9]. Esta tabla contiene información que permite identificar a los actores implicados, el valor, las formas de financiación, los tiempos, entre otra información asociada al contrato. Seguidamente, se considera la información de las tablas SECOP II – Proveedores registrados[10] y SECOP II - Garantías[11] con las cual se adiciona a cada contrato la información del tipo de empresa, la fecha de creación de la empresa proveedora a la cual se adjudica, y un indicador de si se cuenta con una póliza en caso de que el proveedor no cumpla con sus obligaciones. Así mismo, se usa la tabla SECOP II – Ofertas por proceso[12] que contiene todas las ofertas que diferentes proveedores han hecho a los procesos de compra de cada contrato. Esta información es agrupada por proceso de compra con lo que se obtiene el número de ofertas, el valor promedio, el valor máximo y la fechas de la primera y última oferta de cada contrato.

Respecto a la variable objetivo y en consideración al primer objetivo específico relacionado con determinar una medida de ineficiencia en la contratación pública, se toma como referencia los trabajos desarrollados por Gallego et.al. [10] y Rodríguez [11]. Ambos autores emplean la presencia o no de adiciones en tiempo y dinero, medidas de prórrogas y sobrecostos, reportadas a los contratos en SECOP II como medida de ineficiencias en la contratación. Las adiciones presentadas son en sí mismas indicadores de despilfarro pasivo o ineficiencias [4] y tienen una correlación positiva con la corrupción [13, 19]. Además, estas adiciones son sucesos observados y documentados en la plataforma durante los

[9] Sin embargo, como trabajo posterior se plantea un sistema de alarmas continuo que se actualiza a medida que avanza la ejecución del contrato y se obtienen más características.
[10] Ver https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Proveedores-Registrados/qmzu-gj57/about_data Último acceso: 7 de Agosto de 2024.
[11] Ver https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Garantias/gjp9-cutm/about_data. Último acceso: 7 de Agosto de 2024.
[12] Ver https://datos.gov.co/Gastos-Gubernamentales/SECOPII-Ofertas-Por-Proceso/wi7w-2nvm/about_data Último acceso: 7 de Agosto de 2024.

procesos de ejecución, por lo que son una medida objetiva que no transmite sesgos a los modelos a desarrollar[13].

De esta manera, se usa la tabla SECOP II – Adiciones[14] que contiene el identificador de los contratos, el tipo de modificación (adición en valor, adición en tiempo, modificación general, entre otros), la fecha de la modificación y una descripción. Para el propósito de este proyecto solo se tiene en cuenta las adiciones en valor, dado que estas dan cuenta directamente de ineficiencias en los recursos monetarios del sector, como se define en los objetivos. Con esta información se realiza la agrupación por identificador del contrato y se obtiene el número de adiciones en dinero que tuvo cada uno. Si bien la mejor medida sería la magnitud relativa del valor adicionado a los contratos, esta no es publicada abiertamente y se encuentra dispersa en el registro de documentos de cada contrato, lo que impide su uso. Luego, este conteo se adiciona a la información previa mediante el identificador del contrato y en aquellas observaciones donde no se puede cruzar información se marcan el conteo como 0, indicando que fue un contrato sin adiciones. Así, la variable objetivo toma el valor de 0, contratos sin ineficiencias, en alrededor de 122 mil observaciones (cerca del 45%), y sigue la distribución presentada en la Figura 2.

---

[13] Aunque sería ideal usar medidas de corrupción o delitos como tal para atacar objetivamente el despilfarro activo y pasivo, no existe dicha medida a nivel de contrato en Colombia. En general los datos judiciales están dispersos y son de difícil acceso, así como están sujetos a un sesgo de selección al no incluir casos no detectados por los entes judiciales.

[14] Ver https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Adiciones/cb9c-h8sn/about_data Último acceso: 7 de Agosto de 2024.

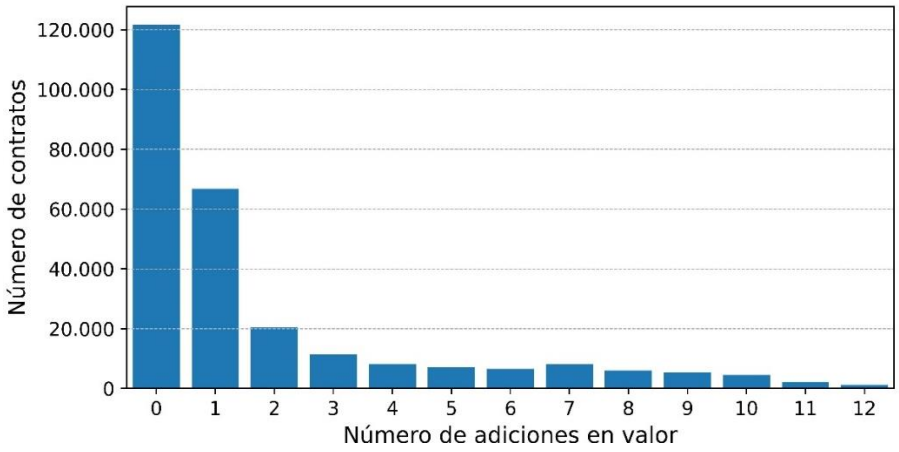


Figura 2. Distribución del conteo de adiciones en valor en los contratos.

De esta forma se obtiene un conjunto de datos inicial compuesto por 270.208 contratos y 81 variables.

A diferencia del común en la literatura, donde los autores plantean un problema de clasificación dada la generación de variables dicotómicas que indican la presencia o no de adiciones, en este trabajo se usa el conteo de adiciones como variable objetivo convirtiéndose en un problema de regresión. La adopción de esta perspectiva soluciona las dos limitaciones del uso de modelos de clasificación descritos en la sección anterior al predecir valores reales positivos que permiten comparar cualquier contrato teniendo en cuenta no solo el hecho de ser o no ineficiente, si no también cuantificar que tan ineficiente es el contrato.

Una vez se obtienen los datos anteriores se procede a la limpieza de los datos y a la ingeniería de variables. En esta etapa inicialmente se eliminan variables creadas durante la ejecución del contrato, direcciones, datos personales (nombre, documento, dirección), enlaces web, variables con un único valor (o un mismo valor en más del 99,9% de las observaciones) y variables con más del 40% de valores no definidos. Seguidamente, se modifican múltiples variables categóricas uniendo categorías de acuerdo con su significado. Por ejemplo, se agrupan los departamentos en 7 zonas, y los códigos de

producto se agrupan para tener segmentos de productos según el Código Uniforme de Productos y Servicios de las Naciones Unidas[15] (UNSPSC). De igual modo, se vuelven binarias (1 y 0) las variables que solo tienen dos categorías, se obtiene la señal anual de las fechas de firma del contrato y de inscripción del proveedor mediante el coseno, y se usan las variables de fechas para crear características del número de días entre sucesos. Por ejemplo, los días entre la firma del contrato y su inicio, *dias_inicio_firma*, y los días entre el registro del proveedor en la plataforma y la firma del contrato, *dias_proveedor_inscrito*. Finalmente, se tratan los valores nulos, principalmente en las características correspondientes a la presentación de ofertas, debido a que solo alrededor del 5% de los contratos registran ofertas formales. Para esto se imputa con 0 el número de ofertas y con -1 el valor medio y máximo de las ofertas para indicar la ausencia de ofertas.

Así se obtiene un conjunto de datos para el modelamiento conformado por 270.208 contratos y 44 variables: 7 variables identificadoras, 1 variable objetivo (Adiciones en valor) y 36 variables explicativas observadas antes de la ejecución del contrato, entre ellas, el valor, la entidad, el saldo, el tipo de contrato, el tipo de proveedor adjudicado, el segmento de producto adquirido, los días transcurridos entre la firma y el inicio del contrato, entre otras (ver listado completo en anexo A).

5.2 Modelamiento

Para predecir las ineficiencias en un contrato, definidas como el número de adiciones en valor, el primer paso es la división de los datos en 85% de entrenamiento y 15% prueba, de tal manera que se usa la mayor cantidad de datos de entrenamiento mientras se mantiene la misma distribución de la variable objetivo en las dos muestras, lo que mejora la generalización de los modelos [43]. Luego, se realiza el preprocesamiento de las variables explicativas, que incluye la codificación de las variables categóricas como numéricas para poder ser usadas en los modelos. Por su parte, para la variable *codigo_entidad* se usa *Target*

[15] Ver https://usa.databasesets.com/unspsc. Último acceso: 12 de Agosto de 2024.

Encoding[16], que codifica las categorías (entidades) dependiendo del promedio del número de adiciones en cada entidad, lo que evita aumentar la dimensionalidad de los datos debido a su alta cardinalidad [44]. Por otro lado, las demás variables categóricas se transforman mediante *One-Hot Encoding*[17] debido a que tienen una baja cardinalidad, con hasta 8 categorías por variables. Las variables monetarias se escalaron aplicando el logaritmo seguido de estandarización, transformación estándar en econometría, y las demás variables numéricas se estandarizaron, para evitar que diferentes magnitudes de las variables afecten el entrenamiento de los modelos. Seguidamente, se usan modelos no supervisados de detección de anomalías cuyas salidas son incluidas en el conjunto de variables explicativas, detallado en la sección 5.2.1.

Se prueban diversos modelos de aprendizaje supervisado para regresión, descritos en la sección 5.2.2, y se realiza el ajuste de hiper parámetros según las métricas de desempeño seleccionadas, profundizado en la sección 5.2.3. Además, se evalúan los modelos discriminando los contratos directamente relacionados a la atención del paciente. Esta distinción debido a que dentro del sector salud se celebran diversidad de contratos que permiten el funcionamiento del sistema, por ejemplo, los servicios administrativos, financieros y de investigación. Estos contratos se seleccionaron mediante el código UNSPSC[18], segmentos 41, 42, 51, 72, 85 y 86[19].

A pesar de que los modelos planteados se enfocan en la efectiva predicción de ineficiencias, y no en la interpretabilidad, el entendimiento de las variables que influyen en las salidas de los modelos aporta a las investigaciones de las entidades encargadas del control de los

---

[16] El Target Enconding fue implementado en la librería sklearn de Python. Para más información consulte https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.TargetEncoder.html.

[17] El One-hot Encoding fue implementado en la librería sklearn de Python. Par más información consulte https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

[18] *United Nations Standard Products and Services Code*.

[19] Estos segmentos del UNSPSC corresponden a: Equipos de laboratorio y medición, observación y prueba; equipamiento, suplementos y accesorios médicos; medicamentos y productos farmacéuticos; construcciones y mantenimiento de inmuebles del sector; servicios de cuidado de la salud; y entrenamiento del personal de salud.

recursos públicos, y al diseño de políticas públicas destinadas a prevenir las ineficiencias. Para este propósito, se usa el método *shapley additive explanations* (SHAP) [45] que se basa en teoría de juegos para obtener la contribución de cada variable en la salida de los modelos. Lo que permite obtener interpretaciones de la relación entre las variables explicativas y las predicciones. En particular, se usa el método TreeSHAP, una variación optimizada para modelos basados en árboles [46]. También, se estiman las dependencias parciales para analizar el efecto marginal de una variable en la predicción del número de adiciones en valor, lo que permite observar la forma de la relación entre estas, por ejemplo, lineal, exponencial, etc [46].

Finalmente, se entrenan modelos de clasificación binaria para contrastar los resultados de la metodología propuesta con los modelos de clasificación comúnmente usados en la literatura. A continuación, se explica en mayor detalle algunos de los pasos anteriores.

5.2.1 Modelos de detección de anomalías

La detección de anomalías es el proceso de identificar patrones y componentes que se desvían del comportamiento usual o esperado de un conjunto de datos. Proceso que aporta a solucionar problemas en campos diversos como la detección de fraude, préstamos bancarios, monitoreo de condiciones médicas, ciber ataques, daños industriales, entre otros [47]. De esta manera, la contratación pública no ha sido ajena al uso de métodos de detección de anomalías para identificar contratos con riesgos de corrupción, dado que se asume que la mayor parte de la contratación se realiza de buena fe. Garzón et.al. [48] explora la combinación del modelo *isolation forest* de detección de anomalías con un índice de riesgo de corrupción para la ciudad de Quebec, Canadá. Por su parte, Niessen et.al. [32] usa *isolation forest* directamente para clasificar los contratos como normales, anómalos y muy anómalos definiendo umbrales de la salida del modelo.

En esta investigación se propone combinar la detección de anomalías para crear nuevos regresores de los modelos de predicción ineficiencias en los contratos públicos, metodología con resultados exitosos en casos como la detección de intrusos en

ciberseguridad [49] y en la detección de tierras mineralizadas [50]. Particularmente, se usan dos modelos de detección de anomalías *Isolation Forest* [51] y *Local Outlier Factor* [52], entrenados con todas las variables explicativas disponibles.

Isolation forest es un modelo basado en árboles donde se asume que los datos anómalos están aislados de los demás. Así que construye arboles binarios que crecen hasta que se aíslan las observaciones, de tal manera que las observaciones en hojas más cercanas a la raíz del árbol tienen una atipicidad más alta que aquellos más alejados debido que fueron más fáciles de aislar. Aparte, *Local Outlier Factor* (LOF) se basa en el principio de agrupamiento, este define para cada punto una vecindad de k vecinos y calcula la densidad local como una medida inversamente proporcional a la distancia promedio entre el punto y sus vecinos. Por último, el algoritmo compara la densidad local de un punto con sus vecinos, de tal manera que los puntos que tienen una menor densidad que sus vecinos tienen una atipicidad mayor que aquellos con mayor densidad.

La metodología propuesta usa los modelos de detección de anomalías para calcular la atipicidad de cada registro teniendo en cuenta todas las variables explicativas, de tal manera que se obtendrán dos medidas, según los dos algoritmos propuestos, de cómo difieren los contratos del comportamiento usual. Luego, estás medidas llamadas puntajes de atipicidad, se incorporan como variables explicativas adicionales a los modelos de predicción. La inclusión de estas nuevas variables permite tener características que ayuden a los modelos de pronóstico dando cuenta de contratos especiales cuyo comportamiento difiere significativamente de los demás.

5.2.2 Modelos de predicción

La principal meta de los modelos propuestos es la predicción del número de adiciones en valor que ha tenido un contrato, por lo que el problema se debe plantear como un problema supervisado de regresión, pero con la anotación de que la variable objetivo es un conteo de valores enteros no negativos. Para la elección de los modelos se debe tener en cuenta la disyuntiva entre la capacidad predictiva y la interpretabilidad de los modelos de

aprendizaje de máquinas [53]. En esta investigación se prioriza la capacidad predictiva sobre la complejidad o dificultad de interpretación, por lo que se comparan un grupo de modelos lineales generalizados diseñados para conteos [54], modelos basados en árboles [55], y otros modelos populares para regresión: *Lasso, ElasticNet* [53], Gradiente descendiente estocástico y Perceptrón multicapa [56].

En primer lugar, los modelos lineales generalizados (GLM) extienden la regresión lineal tradicional agregando la flexibilidad de asumir variedades de distribuciones para la variable dependiente y modificar la función de enlace [57]. En este caso, se utilizan dos GLM que permiten tratar la variable dependiente como un conteo: la regresión Poisson que asume esta distribución en el número de adiciones en valor, con igual media y varianza [54]; y la regresión Binomial Negativa, en la que se asume una relación no lineal entre media y varianza [58].

Por su parte, los modelos basados en árboles son una clase de modelos que se basan en la construcción de árboles de decisión[20] para particionar consecutivamente el espacio de variables, que al no asumir una forma funcional permiten relaciones complejas entre las variables dependientes e independientes [55]. En este proyecto se implementan variedad de algoritmos de esta clase truncando la salida de las predicciones en 0 para evitar números negativos. El popular *Random Fores*t funciona como un ensamblaje de múltiples árboles de decisión en el que se promedian las salidas de cada árbol [59]. También se usan diferentes modelos que hacen uso del Boosting, que usa el principio de ensamblaje del *Random Forest*, pero en vez de entrenar los árboles de forma independiente, se enfoca en un aprendizaje secuencial modificando los pesos de las observaciones al construir un nuevo árbol basado en los errores de los árboles anteriores [59]. Entre este tipo de modelos se

[20] Un árbol de decisión es un modelo "débil" con la estructura de un árbol binario donde los nodos representan decisiones respecto a las variables explicativas y las hojas representan las salidas del modelo dadas las decisiones tomadas en los nodos superiores.

usan *Gradient Boosting Regresor* (GBR) [60], *Histogram-Based Gradient Boosting* (HBGB) [61], *AdaBoost* [62], *XGBoost* [63], *CatBoost* [64] y *LightGBM* [65].

Adicionalmente, se comparan los resultados de la metodología propuesta pronosticando los conteos de adiciones como medida de ineficiencia, con el uso de una salida binaria y modelos de clasificación en la literatura. Para esto se crea una variable binaria que toma el valor de 0 si el conteo de adiciones es 0 y 1 en cualquier otro caso. Luego, esta variable, que toma el valor de 1 en el 55% de contratos, es usada para entrenar modelos de clasificación cuya probabilidad de tener adiciones se usa como métrica de recomendación de contratos a intervenir. Estos modelos de aprendizaje de máquinas de contraste se escogen según su uso en la literatura de ineficiencias en la contratación pública e incluyen diferentes familias de modelos. Los modelos entrenados son: *Naive Bayes*, K vecinos más cercanos, regresión logística, *Gradient Boosting Clasisfier* (GBC), *XGBoost* y *Random Forest*.

5.2.3 Métricas de desempeño y ajuste de hiper parámetros

El desempeño de los modelos de regresión se mide por medio de un número que cuantifica cuán cerca están las predicciones del modelo con los datos observados, en este caso, el número de adiciones en valor que tiene un contrato. Para esté propósito se usan dos métricas de error: la raíz del error cuadrático medio (RMSE por sus siglas en inglés) y la desviación Poisson media (MPD por sus siglas en inglés). La RMSE es una métrica basada en la distancia Euclídea comúnmente usada en métodos de regresión dado que penaliza en mayor medida los errores de predicción grandes sobre los pequeños al elevarlos al cuadrado. Por su parte, la MPD se incluye debido a que esta métrica basada en desviaciones asigna diferentes pesos a las distancias entre el valor real y el predicho, dependiendo de la magnitud de los valores [66]. Siendo esto contrario a la distancia Euclídea que es indiferente a la magnitud de los valores, lo que da cuenta de la distribución del número de adiciones que se acumula en 0 y 1.

Sea $\hat{y}_i$ la predicción del modelo, y_i el valor actual de la variable dependiente y *n* el número de observaciones, el RMSE se define como la raíz del promedio de la diferencia entre ambos valores al cuadrado.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Así, RMSE está en la misma magnitud que la variable dependiente y entre más cercano a 0 sea su valor, mejor es el desempeño del modelo. Por su parte, MPD es una métrica que asume que la salida de la regresión tiene una distribución Poisson, parte de la familia de distribuciones *Tweedie*, y es comúnmente usada para conteos que incluyen el 0 [67]. Esta función de costo se define como

$$MPD = \frac{1}{n}\sum_{i=1}^{n} 2\left(y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + \hat{y}_i - y_i\right)$$

De tal manera que entre menor sea el valor del MPD, mejor es la capacidad el modelo para predecir los conteos[21].

Asimismo, los modelos de aprendizaje de máquinas se configuran por medio de hiper parámetros que pueden variar su desempeño y complejidad. Los hiper parámetros son variables propias de cada modelo que permiten configurar su entrenamiento y se eligen de forma experimental. Por esta razón, el ajuste de hiper parámetros es un paso infaltable en el uso de modelos de aprendizaje de máquinas, tradicionalmente realizado por medio de la prueba de diferentes combinaciones en una grilla de opciones. En esta investigación se usa el software Optuna[22], herramienta que crea dinámicamente el espacio de búsqueda de hiper parámetros y evita al usuario definir de manera explícita toda la estrategia de optimización

[21] La MPD se implementó mediante la librería sklearn, la cual no tiene en cuenta valores predichos de 0, así que los valores $\hat{y}_i = 0$ se convierten a 0,00001 para la construcción de la métrica.
[22] Más información en https://optuna.org/.

[68]. Optuna recibe los rangos de los hiper parámetros a estudiar y selecciona diferentes muestras aleatorias de los datos de entrenamiento y combinaciones de hiper parámetros con los cuales evalúa diferentes modelos. Luego, obtiene las pérdidas de validación de estos modelos y utiliza optimización bayesiana para refinar los valores con el objetivo de minimizar el error de pronóstico. Finalmente, itera este proceso para identificar regiones en el espacio de los hiper parámetros que reduzcan el error.

En este trabajo se usa Optuna para encontrar la mejor combinación de hiper parámetros que minimicen el RMSE para los tres (3) modelos con mejor desempeño con la configuración por defecto. Esta limitación a tres (3) modelos se da debido a los largos tiempos de ejecución de la búsqueda de hiper parámetros y la diferencia en el error de prueba entre estos y los demás. Como se encontrará en los resultados, estos tres modelos son: Random Forest, XGBoost y CatBoost, los cueles tienen los menores errores de predicción. Para la búsqueda de la mejor configuración se usa el 10% de los datos de entrenamiento como validación y 50 iteraciones, según la documentación de la librería, y se proporciona a Optuna los rangos a evaluar para hiper parámetros que modifican la complejidad de los tres modelos. Como el número de árboles, la profundidad de cada árbol, la proporción de observaciones en cada árbol, la tasa de aprendizaje, entre otros; definidos según las recomendaciones en la documentación de cada modelo y el código fuente de Optuna (el listado completo lo encontrará en la Tabla 2 de la sección 6.1).

Con relación a los modelos de clasificación planteados como contraste, las métricas usadas se basan en las coincidencias entre las categorías clasificadas y reales, en este caso presentar y no presentar ineficiencias. Se usan dos métricas propias de clasificación: *accuracy*, un valor entre 0 y 1 que mide la proporción de las observaciones que fueron correctamente clasificadas en general; y *recall*[23], un valor entre 0 y 1 que indica que

[23] Para detalle de estas métricas consulte https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics .

proporción de los contratos con ineficiencias fueron clasificados de esta manera por los modelos.

En la práctica las entidades requieren seleccionar adecuadamente un pequeño grupo de contratos a intervenir que deben priorizar. Desde la perspectiva de un modelo de clasificación se espera que estas elijan aquellos contratos con la mayor probabilidad de tener ineficiencias, como plantean anteriores autores. Sin embargo, esta investigación propone usar la predicción del número de adiciones como medida para recomendar los contratos a intervenir, de tal manera que contratos con una mayor predicción del número de adiciones se priorizan sobre los de menor predicción. Por tanto, desde el punto de vista de los usuarios es de interés saber si los contratos sugeridos son de interés.

Para esta necesidad, se utiliza el *Mean Average Precision at K* (MAP_K)[24], popular métrica para sistemas de recomendación, que promedia la precisión de K contratos sugeridos ordenados por su probabilidad de tener sobrecostos, de tal manera que 1 es el mejor valor [69]. Con la cual se obtienen medidas que permiten compara el desempeño de la clasificación versus la predicción al momento de priorizar los contratos para su indagación por parte de las entidades responsables.

6 RESULTADOS Y DISCUSIÓN

6.1 Modelos de predicción de ineficiencias

Siguiendo los pasos definidos en la sección 5.2, se utilizaron los datos de entrenamiento para ajustar los dos modelos de detección de anomalías que pretenden ayudar a los modelos a encontrar observaciones atípicas[25]. La atipicidad en cada modelo se encuentra en

[24] El MAP se realiza para un solo individuo, por lo que en este caso es la misma *Average Precision at K*. El nombre MAP se utiliza para ser consecuente con anteriores trabajos en contratación pública.
[25] Los modelos Isolation Forest y LOF fueron implementados con la librería sklearn, por tanto, se usa –output.

diferentes magnitudes por lo que se analiza la forma en la que se distribuye esta atipicidad. Como vemos en la Figura 3, LOF acumula los puntajes de atipicidad alrededor de 1, cayendo rápidamente el número de contratos a medida que aumenta. Contrariamente, Isolation forest tiene una distribución con múltiples picos en los valores bajos de atipicidad y disminuye gradualmente el número de contratos a medida que aumenta. Estas diferencias se pueden interpretar como una postura más conservadora del LOF en la cual pocas observaciones son discriminadas como anómalas respecto a las demás, mientras el Isolation forest es más flexible y crea grupos al evaluar las anomalías en los contratos. Esta oposición en la distribución y la diferencia en las magnitudes (detalle en anexo B) llevan a elegir la inclusión de ambos puntajes de atipicidad como variables explicativas adicionales para los modelos de predicción.

La Tabla 1 muestra las métricas de desempeño para los diferentes modelos usados en los conjuntos de entrenamiento y prueba. Lo primero a notar es que los valores de RMSE y MPD son consecuentes en determinar a los modelos basados en árboles como la mejor opción entre los evaluados. También, se encuentra que los modelos estadísticos tienen peor desempeño, con una diferencia de 0,67 entre el RMSE de GBR y Poisson en los datos de prueba. Este detrimento en el desempeño de los modelos estadísticos puede deberse a que realizan predicciones de valores enteras contrario a los valores continuos de los modelos arbóreos, lo que hace que una observación incorrectamente predicha por Poisson tenga un error de al menos 1. Además, al extender modelos lineales pueden obtenerse predicciones insatisfactorias, problema que vemos en el RMSE anormalmente alto de la regresión binomial negativa. En este modelo las predicciones son en promedio menores que el número real de conteos, pero el modelo es muy susceptible a valores anormalmente altos en los predictores importantes, lo que resulta en predicciones anormalmente altas que causan un aumento atípico del RMSE.

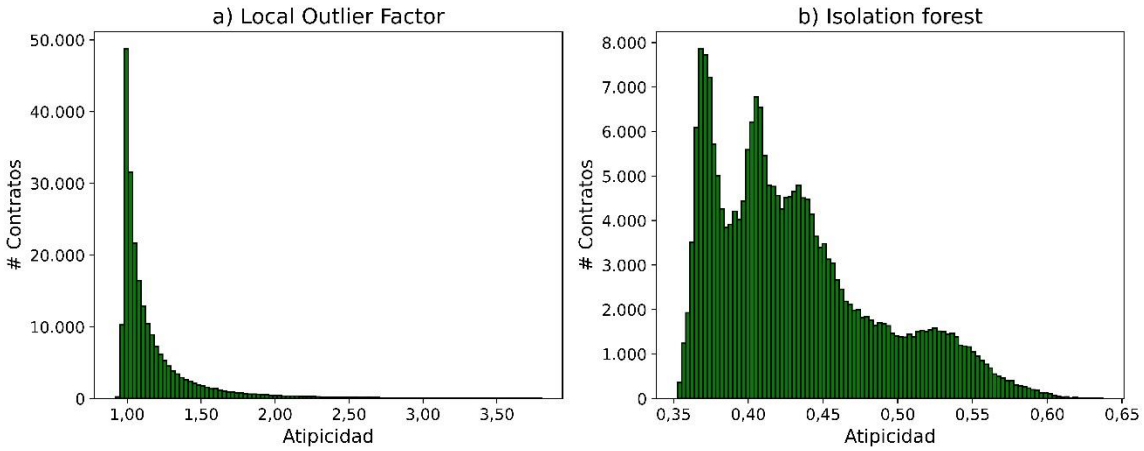


Figura 3. Distribución de los puntajes de atipicidad de Local Outlier Factor (LOF) vs Isolation forest en los datos de entrenamiento.

También se observa una alta diferencia entre las métricas en los datos de entrenamiento y prueba en Random Forest, frente a los demás modelos donde la diferencia entre el RMSE es menor a 0,1. El aumento considerable del error entre datos de entrenamiento y prueba puede ser una señal de sobre ajuste. Sin embargo, considerando únicamente los datos de prueba, Random Forest es superior a los demás modelos con una diferencia considerable.

Modelo	**RMSE Entrenamiento**	**MPD Entrenamiento**	**RMSE Prueba**	**MPD Prueba**
Random Forest	**0,244**	**0,077**	**0,683**	**0,341**
CatBoost	0,727	0,475	0,798	0,512
XGBoost	0,711	0,466	0,799	0,522
LightGBM	0,821	0,556	0,868	0,578
HBGB	0,825	0,561	0,870	0,581
MLP	0,920	0,666	0,965	0,738
GBR	1,095	0,785	1,122	0,807
Poisson	1,953	1,413	1,799	1,427
ElasticNet	1,964	1,622	1,977	1,636
Lasso	2,037	1,763	2,050	1,779
SGD	1,764	1,938	1,780	1,979
Binomial Negativo	2407,389	12,907	63,675	2,188
AdaBoost	2,160	2,621	2,181	2,640

Tabla 1. Métricas de desempeño modelos de regresión.

Adicionalmente, se consideró la exclusión de las variables de atipicidad mencionadas anteriormente como configuración de los modelos. Como resultado, se encontró que la exclusión de estas variables deteriora en pequeña medida el desempeño predictivo de los modelos (anexo C).

Posteriormente, se realizó la búsqueda de hiper parámetros de los modelos Random Forest, CatBoost y XGBoost, los mejores con los parámetros por defecto. En la Tabla 2 vemos el rango de valores evaluados para cada hiper parámetro y la mejor combinación hallada. Aquí vemos como Random Forest, que en principio es el algoritmo más simple, funciona mejor con menos árboles en comparación de los otros dos modelos que optan por aumentar la complejidad.

Modelo	**Hiper Parámetro**	**Descripción**	**Rango**	**Mejor valor**
Random Forest	n_estimators	Número de árboles	[5, 200]	72
	max_features	Número de variables a considerar en cada nodo	sqrt, log2, None	None
	max_depth	Máxima profundidad de los árboles	[6, 32]	29
	max_samples	Proporción de observaciones consideradas para cada árbol	[0,4, 1]	0,998
XGBoost	n_estimators	Número de árboles	[5, 200]	171
	colsample_bytree	Proporción de variables consideradas en cada árbol	[0,4, 1]	0,616
	max_depth	Máxima profundidad de los árboles	[6, 32]	11
	subsample	Proporción de observaciones consideradas para cada árbol	[0,4, 1]	0,731
	learning_rate	Tasa de aprendizaje de entrenamiento	[0,01, 0,5]	0,074

	min_child_weight	Mínimo de observaciones en los nodos hoja	[0, 10]	7
	gamma	Regularización del costo de agregar hojas adicionales	[1, 10]	1,383
CatBoost	iterations	Número de árboles	[50, 300]	287
	colsample_bylevel	Proporción de variables a considerar en cada nodo	[0,1, 1]	0,246
	depth	Máxima profundidad de los árboles	[4, 12]	12
	subsample	Proporción de observaciones consideradas para cada árbol	[0,4, 1]	0,532
	learning_rate	Tasa de aprendizaje de entrenamiento	[0,01, 0,5]	0,23
	l2_leaf_reg	Coeficiente de regularización L2	[1e-8, log(10)]	0,0008
	random_strength	Cantidad de aleatoriedad para elegir la división en cada nodo.	[0, 10]	3,994
	bootstrap_type	Método para definir los pesos de cada árbol	Bayesian, Bernoulli	Bernoulli

Tabla 2. Búsqueda de mejores hiper parámetros mediante Optuna.

En la Tabla 3 vemos cómo la búsqueda de hiper parámetros tiene un notable efecto en la disminución del error de CatBoost y XGBoost, especialmente en este último donde el MPD cae en 0,125, pero sin llegar a superar el Random Forest. Sin embargo, esta búsqueda no modifica el error en este último modelo con respecto a los hiper parámetros por defecto, incluso aumenta el error de entrenamiento. Esto se debe a que Random Forest puede ser más sensible a la estructura de correlación de sus datos que a sus hiper parámetros [70, 71], haciéndolo menos receptivo a la selección de hiper parámetros que otros modelos en ciertos conjuntos de datos.

Modelo	RMSE Entrenamiento	MPD Entrenamiento	RMSE Prueba	MPD Prueba
Random Forest	**0,244**	**0,077**	**0,683**	**0,341**
Random Forest Tunned	0,259	0,088	0,685	**0,341**
XGBoost Tunned	0,529	0,302	0,714	0,399
CatBoost Tunned	0,51	0,289	0,757	0,483
CatBoost	0,727	0,475	0,798	0,512
XGBoost	0,711	0,466	0,799	0,522

Tabla 3. Métricas de desempeño modelos con mejores hiper parámetros (tunned).

Continuando, la Figura 4 muestra la distribución de las predicciones en los datos de entrenamiento de los 3 mejores modelos, Random Forest con parámetros por defecto y XGBoost y CatBoost con ajuste de hiper parámetros, frente a los valores observados. Dado que los datos observados tienen valores enteros y las predicciones de los modelos son continuas se usa la estimación de la densidad de Kernel para comparar las distribuciones en la misma escala de densidad de probabilidad. En la Figura 4, se puede notar cómo los modelos no suelen predecir valores de 0, donde está la mayor densidad del número de adiciones. En cambio, tienen esta mayor densidad alrededor de 0,2 para XGBoost y CatBoost, y alrededor de 0,1 para Random Forest. Aquí también vemos cómo la distribución de las predicciones de Random Forest se ajustan ligeramente mejor a los valores observados entre 0 y 1 adiciones, y los tres modelos tienen una distribución similar desde 1 adición en adelante.

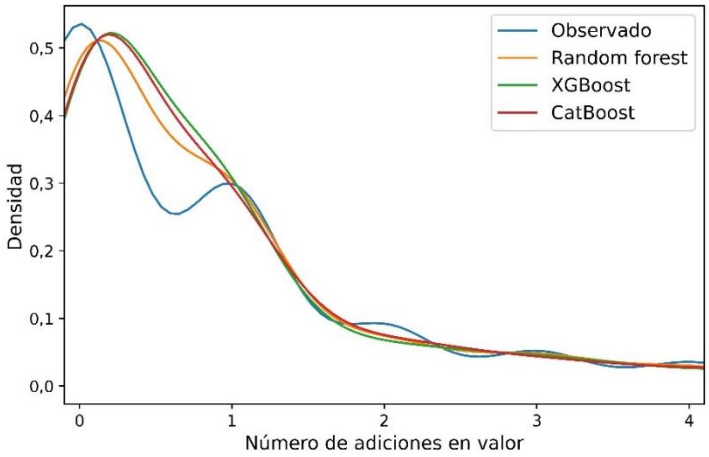


Figura 4. Densidad de probabilidad del número de adiciones observado vs predicciones.

Dentro del sector salud se celebran en gran parte contratos que, aunque no están directamente asociados a la prestación del servicio, permiten el funcionamiento del sistema, como los servicios administrativos, investigación, transporte y servicios generales, entre otros. Por esta razón, también se evaluó el desempeño de los modelos entrenados en un subconjunto de contratos cuyo objeto está destinado directamente a la atención del paciente. Para este filtrado, se eligieron 6 segmentos UNSPSC que corresponden al 28,5% de los contratos. Estos segmentos son: equipos de laboratorio y medición, observación y prueba; equipamiento, suplementos y accesorios médicos; medicamentos y productos farmacéuticos; construcciones y mantenimiento de inmuebles del sector; servicios de cuidado de la salud; y entrenamiento del personal de salud. La Tabla 4 muestra que, al evaluar las predicciones en estos contratos, el RMSE de prueba aumenta notoriamente en Random Forest, compartiendo valores similares en los tres modelos. En contraste, el MPD es menor que las métricas anteriores en los tres modelos. Esto puede deberse a que los contratos directamente relacionados a la atención del paciente tienen alrededor de un 15% más de observaciones con 0 adiciones, por lo que, sumado a los hallazgos de la Figura 4, facilita el ajuste de los datos a una distribución Poisson y disminuye el MPD.

Modelo	RMSE Entrenamiento	MPD Entrenamiento	RMSE Prueba	MPD Prueba
Random Forest	**0,254**	**0,07**	0,755	**0,323**
XGBoost Tunned	0,412	0,237	**0,754**	0,356
CatBoost Tunned	0,428	0,234	0,768	0,448

Tabla 4. Métricas de desempeño en contratos directamente relacionados a la atención del paciente.

De esta forma se obtienen varias conclusiones. Primero, como se menciona en la literatura, los árboles son superiores en la predicción de ineficiencias, aún con diferente variable objetivo. Segundo, las diferencias en el desempeño de los modelos entre diferentes configuraciones son relativamente pequeñas. La exclusión de los puntajes de atipicidad, la búsqueda de hiper parámetros y la discriminación de contratos directamente relacionados a la atención del paciente coinciden en Random Forest como el modelo vencedor para la detección temprana de riesgos en los recursos del sistema de salud. Tercero, los modelos presentan un desempeño satisfactorio para su uso como sistema de recomendación.

Por un lado, un RMSE de 0,683 da cuenta de un error tolerable en los contratos con más adiciones, que son aquellos que se desea priorizar. Recordando que el RMSE representa la desviación estándar de los residuos, podemos construir intervalos de predicción que nos indican que aquellos contratos con predicciones mayores a 2,4 adiciones en valores son de interés para ser investigadas, dado que el valor real será mayor a 1 adición con el 95% de confianza[26]. Por su parte, un MPD de prueba cercano a 0,3 indica que el modelo no cae en la tendencia de sobreestimar las predicciones. Esto apoya el valor del RMSE indicando que los grandes errores en el pronóstico del número de adiciones no suelen suceder en los contratos con pocas adiciones observadas. Lo cual se traduce en que el modelo evita

[26] Sea el intervalo de predicción, $IC = predicción \pm (RMSE * Z - score)$, asumiendo normalidad de los errores, para el 95% de confianza $Z - score = 1{,}96$ y $predicción = 2{,}4$. $IC = 2{,}4 \pm (0{,}683 * 1{,}96) = (1{,}06, 3{,}73)$. Así, para todas las predicciones mayores a 2,4 el límite inferior del intervalo de predicción será siempre mayor a 1.

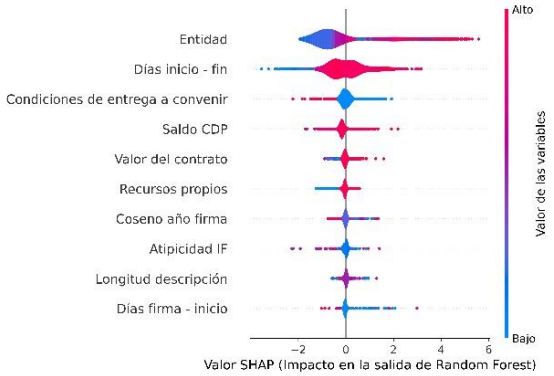
encender alarmas para contratos con pocas o ninguna adición, permitiendo una adecuada priorización para la investigación de ineficiencias.

6.2 Análisis de variables explicativas

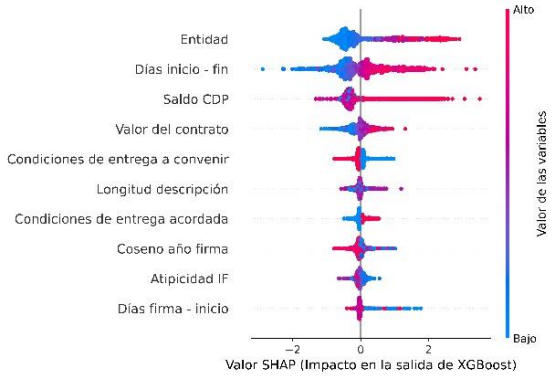
Para conocer la relevancia de las variables y sus efectos al predecir las ineficiencias, en la Figura 5 se presentan los valores shap de las 10 variables más importantes, con la mayor contribución promedio, para las predicciones de los mejores modelos de predicción. En el eje X de cada gráfica está la contribución a la predicción de ineficiencias, de tal manera que un valor positivo indica un incremento en el número de adiciones y un valor negativo una disminución. Y el color de los puntos representa el valor relativo de la variable explicativa de acuerdo con la barra de color del costado derecho.

La Figura 5 muestra la Entidad como la característica más importante para los tres modelos, de tal manera que los valores más altos contribuyen a mayores sobrecostos y viceversa. Este comportamiento se debe a la codificación de la variable entidad mediante *Target Encoding*, que codifica la variable dependiendo del promedio del número de adiciones en cada entidad [44]. En este caso, los altos valores de la variable, que tienen alta contribución en el número de adiciones, corresponden a entidades que frecuentemente tienen sobrecostos. Entre estas entidades están las Subredes Integradas de Salud de Bogotá Sur Occidente y Sur[27], con más de 30 mil contratos en la muestra y una mediana de 7 y 3 adiciones en valor por contrato respectivamente, que se encargan de la administración de la red de hospitales públicos y no están obligadas a cumplir con el Estatuto General de Contratación [11].

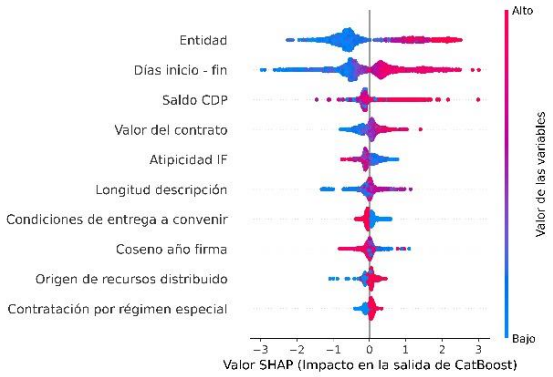
[27] Las Subredes han estado presentes en los medios por diversos escándalos como acusaciones por derroche de recursos en honorarios, falta de pago a proveedores, el retraso de más de 13 años en la entrega de la torre de urgencias del hospital de Kennedy, la orden de embargo de las cuentas de la Subred Sur en marzo de 2023, entre otros múltiples hallazgos fiscales.



(a) Random forest



(b) XGBoost



(c) CatBoost

Figura 5. Importancia de variables en los mejores modelos de predicción: Random Forest, XGBoost y CatBoost

Seguidamente, los días transcurridos entre las fechas pactadas de inicio y finalización, y el valor del contrato muestran una alta importancia y una relación directa con el número de adiciones. Además, la Figura 6 evidencia, mediante la dependencia parcial, una relación cercana a la lineal entre ambas variables transformadas y las predicciones, pero con un mayor impacto en los días transcurridos que podemos ver en el rango del eje Y. Este resultado da cuenta de que proyectos más grandes en el sector, como la construcción de obras y los servicios de gran cobertura, están expuestos a una mayor incertidumbre que deja espacio a un mayor riesgo de ineficiencias, acorde con hallazgos de la literatura en Colombia [10].

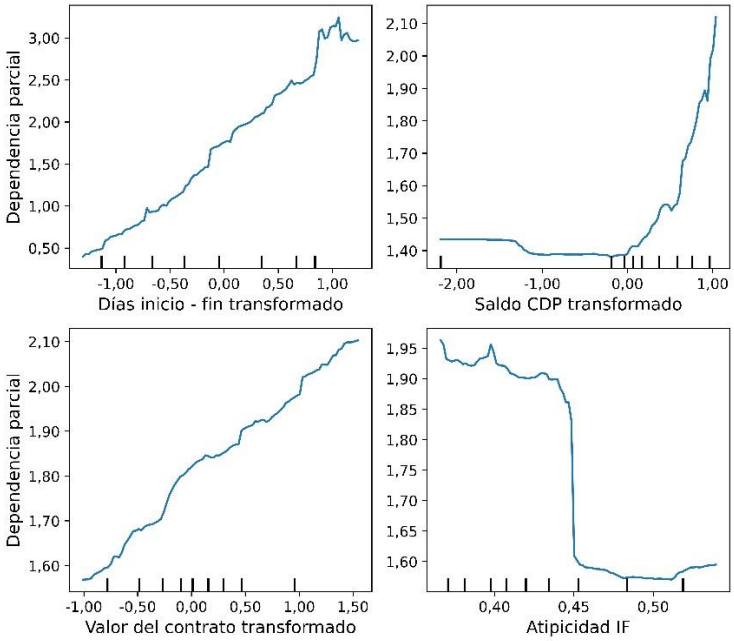


Figura 6. Gráficas de dependencia parcial de las principales variables numéricas en Random Forest.

Otra relación interesante es la condición de entrega del contrato, donde se observa que la entrega a convenir está asociado a un menor número de adiciones, es decir, pactar las condiciones en que será entregado el producto o servicio en conjunto con el proveedor durante la adjudicación previene que se presenten ineficiencias. Por su parte, el saldo del Certificado de Disponibilidad Presupuestal (CDP) de la entidad asignado al contrato se encuentra relevante para los tres modelos desatacados en la Figura 5 pero no es claro su efecto. Sin embargo, la Figura 6 confirma que tiene una baja contribución en valores por debajo de 0 en la escala transformada, correspondiente a 22.292.547 COP, pero una relación directa entre saldo y la contribución a la predicción de ineficiencias por encima de este umbral.

Por otro lado, la transformación cíclica de la fecha de firma del contrato usando el coseno (Coseno año firma) da cuenta del ciclo de gasto en las entidades públicas, que está sujeto al presupuesto aprobado por el Congreso de la República con vigencia anual. Esta resulta ser la séptima variable más relevante para Random Forest según la Figura 5, donde también

vemos que los contratos con valores altos, firmados en los últimos y primeros meses del año, están relacionados con menos adiciones mientras, los valores más bajos, contratos firmados entre mayo y agosto, contribuyen a un aumento en número de ineficiencias. Esto puede deberse a que en los meses de enero y febrero se realiza la contratación de personal y servicios acorde a la planeación de la entidad; y en noviembre y diciembre se comprometen recursos para cubrir todo el presupuesto y evitar recortes en la próxima vigencia. De esta manera, el riesgo en los recursos se presenta en los contratos firmados en fechas atípicas debido a una mala planeación o posibles requerimientos ajenos al funcionamiento de la entidad.

Respecto a las variables introducidas de los puntajes de atipicidad estimados mediante Isolation forest y LOF, el primero es una de las variables más relevantes en los tres mejores modelos, mientras la segunda está fuera del top 20 de variables más relevantes. Esto puede deberse a la distribución de los puntajes que se muestra en la Figura 3, dado que LOF es más conservador determinando menos observaciones como anómalas, opuesto a Isolation Forest, donde la partición de los nodos genera separaciones con datos más heterogéneos de la variable respuesta. En relación con la contribución de los puntajes de atipicidad de Isolation forest, la Figura 6 muestra que en Random Forest valores menores a 0,45 contribuyen de la misma forma a la predicción número de adiciones, pero en mayor medida que puntajes de atipicidad mayores a 0,45. Esto permite concluir qué la inclusión de métodos de detección de anomalías como variables exógenas, a pesar de tener poco efecto en la disminución del error, sí puede aportar al entendimiento de las decisiones del modelos y resaltar ideas para el uso de estos métodos por sí solos en la contratación pública.

6.3 Predicción vs clasificación en la predicción de ineficiencias

Atendiendo a la clasificación como sistema de recomendación, de la Tabla 5 se puede evidenciar que, en congruencia con la literatura y los resultados de los modelos de conteo anteriores, los modelos basados en árboles dan los mejores resultados. Adicionalmente, con el MAP de los modelos de clasificación vemos que los modelos cumplen con la función

de recomendar los contratos a investigar en mayor profundidad, de tal manera que de 1.000 contratos sugeridos por Random Forest o XGBoost todos son de interés. El desempeño de estos modelos es mejor que los presentados por Gallego et.al. [10] para la clasificación de prórrogas y sobrecostos, para los que menciona es suficientemente alto para el uso práctico. Estas métricas superiores pueden deberse a dos factores: primero, los contratos del sector salud tienen una mayor proporción de incidencia de adiciones en valor; y segundo, los modelos entrenados en este estudio están especializados para el sector de salud y protección social a diferencia de la diversidad de contratos que cubre Gallego et.al. [10].

Objetivo	**Modelo**	**Accuracy**	**Recall**	**MAP_{100}**	**MAP_{1000}**
Clasificación	Random forest	**0,92**	**0,92**	**1,00**	**1,00**
Clasificación	XGBoost	0,89	0,91	**1,00**	**1,00**
Clasificación	GBC	0,84	0,86	0,98	1,00
Clasificación	K vecinos	0,89	0,90	0,94	0,93
Clasificación	Regresión logística	0,78	0,81	0,88	0,84
Clasificación	Naive Bayes	0,74	0,80	0,74	0,64
Predicción	Random forest			**1,00**	**1,00**
Predicción	XGBoost			1,00	0,99
Predicción	CatBoost			1,00	1,00

Tabla 5. Métricas de recomendación de contratos con ineficiencias.

La Tabla 5 también muestra que las predicciones de Random Forest para recomendar los contratos son tan buenas como la clasificación a la luz del MAP_{100} y MAP_{1000}. Sin embargo, el uso de los modelos de regresión para recomendar los contratos a intervenir por las entidades por medio de una predicción continua muestra una solución a las limitaciones de los modelos de clasificación expuestas previamente.

Primero, al asumir las ineficiencias como una medida dicotómica los modelos entrenados no permiten directamente discriminar entre contratos con ineficiencias y puede llevar a priorizar contratos con un menor costo de inacción. Por ejemplo, se toman dos contratos para inversión en diferentes entidades y de diferente tipo[28], el primero tuvo 3 adiciones en

[28] Estos contratos puede encontrarlos en contratos electrónicos del portal datos abiertos Colombia con los ID Contrato CO1.PCCNTR.1386013 y CO1.PCCNTR.4657198.

valor y el segundo 1 adición en valor. A pesar de que el primero representó un mayor compromiso adicional de los recursos, la probabilidad de presentar adiciones con el mejor modelo de clasificación fue de 0,87, menor al 0,95 del segundo contrato, por lo que está más lejos de ser recomendado para la intervención de los entes de control.

La metodología propuesta en el presente trabajo soluciona esta limitación al realizar la recomendación mediante la predicción continua de los modelos entrenados. Al usar el conteo de adiciones en valor como variable objetivo se asignan diferentes escalas de ineficiencias que permiten discriminar entre los contratos con adiciones. Al entrenar los modelos de regresión con estos datos, las predicciones tienen en cuenta la gravedad de las ineficiencias entre los contratos, lo que hace posible dar recomendaciones que permitan a los expertos priorizar de mejor manera los contratos como se muestra en la Tabla 5. Retomando el ejemplo anterior, la predicción de adiciones para el contrato con 3 adiciones es de 1,64, mientras la predicción del contrato con 1 adición es 1, por lo que se recomendaría el primer contrato que presenta un mayor riesgo para los recursos del sistema de salud.

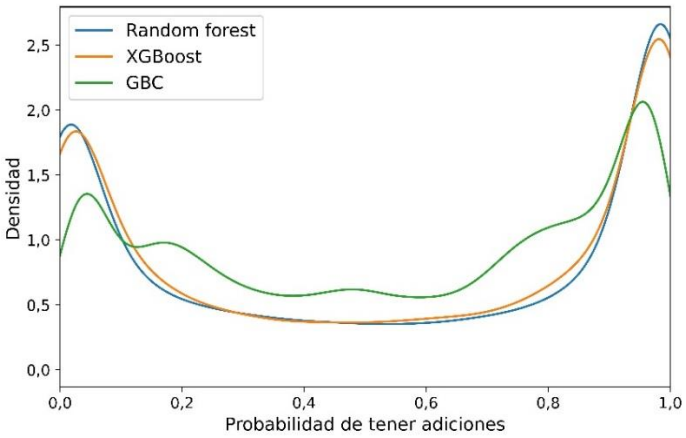


Figura 7. Densidad de probabilidad de clasificación de sobrecostos.

La segunda limitación del uso de modelos de clasificación es la acumulación de las probabilidades de clasificación en 1. La Figura 7 muestra la densidad de las probabilidades de clasificar un contrato con ineficiencias en los datos de entrenamiento, donde se observa

una distribución bimodal, con picos en 0 y 1. Especialmente, 7.171 contratos tienen probabilidad 1, aquellos que se desean investigar, esto hace que la recomendación de contratos a intervenir se convierta en una sugerencia aleatoria entre estos contratos con el mismo valor. Así, el uso de los modelos de conteos propuesto soluciona esta limitación al usar la salida continua y sin límite superior de los modelos entrenados como medida para la priorización de contratos a intervenir. Conservando así una adecuada selección de los contratos con ineficiencias y agregando una noción de orden a estos, que podemos identificar en la distribución de las predicciones en los datos de prueba de la Figura 4. Por ejemplo, se toman dos contratos con características similares[29], pero el primero tuvo 3 adiciones en valor y el segundo 12 adiciones en valor; al usar el mejor modelo de clasificación ambos obtienen una probabilidad de 1 de tener ineficiencias, evitando priorizar uno sobre el otro. Por su parte, al usar el modelo de regresión propuesto, la predicción para los anteriores contratos es de 3,07 y 11,81, respectivamente, permitiendo una priorización del contrato con la mayor predicción.

7 CONCLUSIONES Y RECOMENDACIONES

El sector salud y protección social es el segundo con mayor participación en la contratación estatal y ocupa recursos superiores al 5% del PIB, pero no está exento del despilfarro de los recursos públicos que evita una ejecución eficiente. Este despilfarro se da principalmente por ineficiencias en la contratación, sobre la corrupción [4]. Por tal razón este proyecto se centró en la predicción de ineficiencias en el sector salud usando datos transaccionales de acceso público y actualización continua. De este trabajo resultaron diversas conclusiones.

[29] Estos contratos puede encontrarlos en contratos electrónicos del portal datos abiertos Colombia con los ID Contrato CO1.PCCNTR.1299235 y CO1.PCCNTR.2750358.

Primero, los modelos basados en árboles son los mejores para la predicción del número de adiciones en valor que tendrá un contrato, especialmente Random Forest. Con estos se consigue un desempeño satisfactorio desde el punto de vista numérico, con un modelo que evita sobreestimar las predicciones y un RMSE de 0,683 que indica que en promedio las predicciones no llegan a tomar el valor de los enteros más cercanos al valor real. Además, el ensamble de modelos de detección de anomalías por medio de los puntajes de atipicidad es útil para entender la predicción de ineficiencias y su interpretación da luces para el uso de estos modelos por si solos, a pesar de encontrar poco aumento del error de predicción al excluirlos del modelo.

Segundo, a pesar de no ser modelos causales, el proyecto encuentra características relevantes para el entendimiento de las ineficiencias en el sistema de contratación del sector salud que pueden ayudar a los hacedores de política a dirigir esfuerzos en la prevención del riesgo de los recursos. Los datos confirman una asociación ampliamente detectada en la literatura entre grandes contratos, en términos de valor y duración, y la presencia de ineficiencias. Pero también se encuentran nuevas variables para la formulación de controles en la contratación, como las condiciones de entrega pactadas, el saldo del Certificado de Disponibilidad Presupuestal (CDP) asignados al contrato y los recursos propios de la entidad utilizados.

Tercero, el presente trabajo logra resultados satisfactorios para el uso de la metodología propuesta como herramienta para recomendar de forma temprana los contratos a priorizar para su investigación, aun usando solo variables previas a la ejecución del contrato. Se encuentra que, de 100 contratos sugeridos por su predicción del número de adiciones en valor, el 100% serían de interés. Sumado a esto, el uso del número de adiciones como medida de ineficiencias soluciona dos limitaciones de la clasificación binaria usada en la literatura: permite discriminar en nivel de riesgo entre contratos que presentan ineficiencias y evita la acumulación del valor por el cual se ranquean los contratos en un único valor, 1 en clasificación, que impide una adecuada priorización de los contratos.

La metodología propuesta y sus hallazgos ofrecen una herramienta innovadora para mitigar los riesgos de ineficiencias en el sistema de salud. Por un lado, la identificación de variables relevantes como alertas sugiere medidas preventivas para la creación de los contratos del sector salud desde la plataforma SECOP II o las mismas reglas de contracción pública en Colombia. Por otro lado, la supervisión de los contratos firmados debe incluirse en el funcionamiento del sistema de salud, especialmente en la actual reforma que da más facultades al Estado dentro del sistema. De esta manera, una herramienta basada en aprendizaje de máquinas como la propuesta en este trabajo debe incorporarse para priorizar los contratos a investigar para tomar de forma oportuna las medidas correctivas correspondientes, especialmente dentro del ADRES, que debería velar por los recursos que gira a las diferentes entidades del sistema de salud.

El proyecto cumple con el objetivo de identificar de forma temprana riesgos en los recursos del sector salud que permite apoyar la labor de las entidades de control. La implementación de esta metodología permitirá la mejora en el cuidado de los recursos del sistema de salud colombiano por medio de una eficaz intervención de las entidades de control para evitar ineficiencias. Asimismo, el trabajo aporta a la literatura de supervisión de la contratación pública mediante el uso satisfactorio de conteos para la medición de la gravedad de las ineficiencias y el ensamblaje de modelos de detección de anomalías como predictores de ineficiencias.

Finalmente, los desafíos y limitaciones también revelan recomendaciones para las plataformas de contratación y alertas. Por un lado, la información publicada en datos abiertos limita el uso del valor adicionado a los contratos, ideal para estudiar los sobrecostos en Colombia, información que debería ser registrada en los formularios de SECOP II. Por otro lado, se requiere de la participación de expertos investigadores de la gestión fiscal en salud que permita incluir el conocimiento experto en la definición de umbrales de desempeño minuciosos para la implementación a escala nacional. Como trabajo futuro, es de interés aprovechar la información detallada de los documentos

adjuntos como evidencias al registrar un contrato en SECOP II, y estudiar las implicaciones y retos de incorporar este tipo de herramientas en el sistema de salud colombiano.

REFERENCIAS

[1] C. Granger, J. Ramos, L. Melo y G. Silva, «Financiamiento del Sistema de Salud en Colombia: Fuentes y usos,» *Borradores de Economía,* nº 1233, 2023.

[2] K. Hussmann, «Corrupción en el sector salud,» Bergen: U4 Anti-Corruption Resource Centre, Chr. Michelsen Institute (U4 Issue 2020:16), 2020.

[3] Transparencia por Colombia, «Así se mueve la corrupción. Radiografía de los hechos de corrupción en Colombia 2016-2022,» 2024.

[4] O. Bandiera, A. Prat y T. Valletti, «Active and Passive Waste in Government Spending: Evidence from a Policy Experiment,» *American Economic Review,* vol. 99, nº 4, p. 1278–1308, 2009.

[5] Transparencia por Colombia, «Así se mueve la corrupción. Radiografía de los hechos de corrupción en Colombia 2016-2020,» 2021.

[6] B. Flyvbjerg y D. Gardner, How big things get done: the surprising factors that determine the fate of every project, from home renovations to space exploration and everything in between, New York: Currency, 2023.

[7] Brigard Urrutia, «10 claves para entender el proyecto de Reforma a la Salud,» Bogotá D.C., 2024.

[8] N. Köbis, C. Starke y I. Rahwan, «Artificial Intelligence as an Anti-Corruption Tool (AI-ACT) -- Potentials and Pitfalls for Top-down and Bottom-up Approaches,» *pre-print 2102.11567,* 2021.

[9] E. Estevez, P. Fillottrani y S. Linares Lejarraga, «PROMETEA: Transformando la administración de justicia con herramientas de inteligencia artificial,» Banco Interamericano de Desarrollo, 2020.

[10] J. Gallego, G. Rivero y J. Martínez, «Preventing rather than punishing: An early warning model of malfeasance in public procurement,» *International Journal of Forecasting,* vol. 37, nº 1, pp. 360-377, 2021.

[11] S. Rodrígues, *Predicción de ineficiencias en la contratación pública de Bogotá,* Bogotá D.C.: Tesis de maestría, 2020.

[12] D. Shiundu y G. Rotich, «Factors influencing efficiency in procurement systems among public institutions: A case of city council of Nairobi,» *International Academic Journals,* vol. 1, nº 1, pp. 79-96.

[13] E. Dal Bo y M. Rossi, «Corruption and inefficiency: Theory and evidence from electric utilities,» *Journal of Public Economics,* vol. 91, nº 5-6, pp. 939-962, 2007.

[14] A. Estache y R. Foucart, «The scope and limits of accounting and judicial courts intervention in inefficient public procurement,» *Journal of Public Economics,* vol. 157, pp. 95-106, 2018.

[15] J. Nakabayashi, «Small business set-asides in procurement auctions: An empirical analysis,» *Journal of Public Economics,* vol. 100, pp. 28-44, 2013.

[16] A. Baltrunaite, C. Giorgiantonio, S. Mocetti y T. Orlando, «Discretion and Supplier Selection in Public Procurement,» *The Journal of Law, Economics, and Organization,* vol. 37, nº 1, pp. 134-166, 2020.

[17] E. Uyarra, J. Edler, J. Garcia-Estevez, L. Georghiou y J. Yeow, «Barriers to innovation through public procurement: A supplier perspective,» *Technovation,* vol. 34, nº 10, pp. 631-645, 2014.

[18] E. Duflo, R. Hanna y S. Ryan, «Incentives Work: Getting Teachers to Come to School,» *American Economic Review,* vol. 102, nº 4, p. 1241–78, 2012.

[19] P. Mauro, «Corruption and Growth,» *The Quarterly Journal of Economics,* vol. 110, nº 3, pp. 681-712, 1995.

[20] I. Adam, M. Fazekas, A. Hernandez Sanchez, P. Horn y N. Regös, «Integrity Dividends: Procurement in the Water and Sanitation Sector in Latin America and the Caribbean,» 2022.

[21] N. Chaudhury, J. Hammer, M. Kremer, K. Muralidharan y F. H. Rogers, «Missing in Action: Teacher and Health Worker Absence in Developing Countries,» *Journal of Economic Perspectives,* vol. 20, nº 1, pp. 91-116, 2006.

[22] J. Gallego, M. Prem y J. F. Vargas, «Corruption in the times of pandemia,» *Documentos de Trabajo. Universidad del Rosario,* nº 18178, 2020.

[23] B. Olken, «Monitoring Corruption: Evidence from a Field Experiment in Indonesia,» *Journal of Political Economy,* vol. 115, nº 2, 2007.

[24] M. Fazekas, I. Tóth y L. King, «An Objective Corruption Risk Index Using Public Procurement Data.,» *European Journal on Criminal Policy and Research,* vol. 22, pp. 369-397, 2016.

[25] N. Charron, C. Dahlström, M. Fazekas y V. Lapuente, «Careers, Connections, and Corruption Risks: Investigating the Impact of Bureaucratic Meritocracy on Public Procurement Processes,» *The Journal of Politics,* vol. 79, nº 1, 2017.

[26] M. Gnaldi y S. Del Sarto, «Measuring Corruption Risk in Public Procurement over Emergency Periods,» *Social Indicators Research,* vol. 172, pp. 859-577, 2024.

[27] M. M. Zuleta, S. Ospina y C. A. Caro, «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,» Laboratorio Latinoamericano de Políticas de Probabilidad y Transparencia. Un Proyecto de Cooperación Sur-Sur ATN O/C 16465-RG, Bogotá: Fedesarrollo, 67p, 2019.

[28] M. Melo y D. Delen, «Predicting and explaining corruption across countries: A machine learning approach,» *Government Information Quarterly,* vol. 37, nº 1, 2020.

[29] E. Colonnelli, J. A. Gallego y M. Prem, «What Predicts Corruption?,» *Documento de trabajo,* 2020.

[30] G. De Blasio, A. D'Ignazio y M. Letta, «Gotham city. Predicting 'corrupted' municipalities with machine learning,» *Technological Forecasting and Social Change,* vol. 184, nº C, 2022.

[31] F. Decarolis y C. Giorgiantonio, «Corruption red flags in public procurement: new evidence from Italian calls for tenders,» *EPJ Data Science,* vol. 11, nº 16, 2022.

[32] M. E. K. Niessen, J. M. Paciello y J. I. P. Fernandez, «Anomaly Detection in Public Procurements using the Open Contracting Data Standard,» de *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*, 2020, pp. 127-134.

[33] A. Westerski, R. Kanagasabai, E. Shaham, A. Narayanan, J. Wong y M. Singh, «Explainable anomaly detection for procurement fraud identification—lessons from practical deployments,» *International Transactions in Operational Research,* vol. 28, nº 6, pp. 3276-3302, 2021.

[34] Y. Torres-Berru y V. López, «Data Mining to Identify Anomalies in Public Procurement Rating Parameters,» *Electronics,* vol. 10, nº 22, 2021.

[35] S. L. Domingos, R. N. Carvalho, R. S. Carvalho y G. N. Ramos, «Identifying IT Purchases Anomalies in the Brazilian Government Procurement System Using Deep Learning,» de *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 722-727.

[36] L. Potin, R. Figueiredo, V. Labatut y C. Largeron, «Pattern Mining for Anomaly Detection in Graphs: Application to Fraud in Public Procurement,» de *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*, Springer Nature Switzerland, 2023, pp. 69-87.

[37] H. N. Akouemo y R. J. Povinelli, «Probabilistic anomaly detection in natural gas time series data,» *International Journal of Forecasting,* vol. 32, nº 3, pp. 948-956, 2016.

[38] F. Morselli, L. Bedogni, U. Mirani, M. Fantoni y S. Galasso, «Anomaly Detection and Classification in Predictive Maintenance Tasks with Zero Initial Training,» *IoT,* vol. 2, nº 4, pp. 590-609, 2021.

[39] S. Lewis-Faupel, Y. Neggers, B. A. Olken y R. Pande, «Can Electronic Procurement Improve Infrastructure Provision? Evidence from Public Works in India and Indonesia,» *American Economic Journal: Economic Policy,* vol. 8, nº 3, pp. 258-283, 2016.

[40] J. C. Bertot, P. T. Jaeger y J. M. Grimes, «Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies,» *Government Information Quarterly,* vol. 27, nº 3, pp. 264-271, 2010.

[41] J. Kleinberg, J. Ludwig, S. Mullainathan y Z. Obermeyer, «Prediction Policy Problems,» *American Economic Review,* vol. 105, nº 5, pp. 491-495, 2015.

[42] G. Jan, «Public supervision on the public procurement market in selected EU member states (selected aspects),» *Ekonomia XXI Wieku,* pp. 35-44, 2020.

[43] M. Arjovsky, Out of Distribution Generalization in Machine Learning., pre-print 2103.02667, 2021.

[44] D. Micci-Barreca, «A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems,» *SIGKDD Exploration Newsletter,* vol. 3, nº 1, pp. 27-32, 2001.

[45] S. M. Lundberg y S.-I. Lee, «A Unified Approach to Interpreting Model Predictions,» de *Advances in Neural Information Processing Systems,* Curran Associates, Inc., 2017.

[46] C. Molnar, Interpretable Machine Learning, 2 ed., 2022.

[47] A. B. Nassif, M. A. Talib, Q. Nasir y F. M. Dakalbab, «Machine Learning for Anomaly Detection: A Systematic Review,» *IEEE Access,* vol. 9, pp. 78658-78700, 2021.

[48] C. Garzon, P.-D. Louis-Alexis, Hudon y M. P-A, «Comparing and Combining the "Corruption Risk Index" and "Anomaly Detection" Methods to Uncover Suspicious Contracts in Public Project Tendering Using Open Procurement Data,» Montreal, Québec, 2021.

[49] A. Iacovazzi y S. Raza, «Ensemble of Random and Isolation Forests for Graph-Based Intrusion Detection in Containers,» de *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2022, pp. 30-37.

[50] S. Zhang, E. J. Carranza, K. Xiao, H. Wei, F. Yang, Z. Chen, N. Li y J. Xiang, «Mineral Prospectivity Mapping based on Isolation Forest and Random Forest: Implication for the Existence of Spatial Signature of Mineralization in Outliers,» *Natural Resources Research,* vol. 31, nº 4, pp. 1981-1999, 2021.

[51] F. T. Liu, K. M. Ting y Z.-H. Zhou, «Isolation-Based Anomaly Detection,» *ACM Trans. Knowl. Discov. Data,* vol. 6, nº 1, 2012.

[52] M. M. Breunig, H.-P. Kriegel, R. T. Ng y J. Sander, «LOF: identifying density-based local outliers,» *SIGMOD Record,* vol. 29, nº 2, p. 93–104, 2000.

[53] T. Hastie, R. Tibshirani y J. Friedman, The Elements of Statistical Learning, Second edition, New York, NY: Springer, 2009.

[54] O. Maxwell, B. A. Mayowa, I. U. Chined y A. E. P. , «Modelling Count Data; A Generalized Linear Model Framework,» *American Journal of Mathematics and Statistics,* vol. 8, nº 6, pp. 179-183, 2018.

[55] C. Kern, T. Klausch y a. F. Kreuter, «Tree-based Machine Learning Methods for Survey Research,» *Survey Research Methods,* vol. 13, nº 1, pp. 73-93, 2019.

[56] X. Glorot y Y. Bengio, «Understanding the difficulty of training deep feedforward neural networks,» de *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249-256.

[57] P. K. Dunn y G. K. Smyth, «Chapter 5: Generalized Linear Models: Structure,» de *Generalized Linear Models With Examples in R*, New York, NY, Springer New York, 2018, pp. 211-241.

[58] J. M. Hilbe, «8- Negative binomial regression,» de *Negative binomial regression*, Cambridge, Cambridge University Press, 2011, pp. 185-220.

[59] L. Breiman, «Random Forest,» *Machine Learning*, vol. 45, nº 1, pp. 5-32, 2001.

[60] J. H. Friedman, «Greedy function approximation: A gradient boosting machine,» *The Annals of Statistics*, vol. 29, nº 5, pp. 1189-1232, 2001.

[61] A. Guryanov, «Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees,» de *Analysis of Images, Social Networks and Texts. Lecture Notes in Computer Science(), vol 11832*, Cham, Springer, 2019, pp. 39-50.

[62] Y. Cao, Q.-G. Miao, J.-C. Liu y L. Gao, «Advance and Prospects of AdaBoost Algorithm,» *Acta Automatica Sinica*, vol. 39, nº 6, pp. 745-758, 2013.

[63] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System,» de *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 785-794.

[64] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush y A. Gulin, «CatBoost: unbiased boosting with categorical features,» *pre-print 1706.09516*, 2019.

[65] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye y T.-Y. Liu, «LightGBM: A Highly Efficient Gradient Boosting Decision Tree,» *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[66] F. Subtil, O. Boussari, M. Bastard, J.-F. Etard, R. Ecochard y C. Génolini, «An alternative classification to mixture modeling for longitudinal counts or binary measures,» *Statistical Methods in Medical Research*, vol. 26, nº 1, pp. 453-470, 2017.

[67] A. Halder, S. Mohammed, K. Chen y D. Dey, «Spatial risk estimation in Tweedie compound Poisson double generalized linear models,» *pre-print 1912.12356*, 2020.

[68] T. Akiba, S. Sano, T. Yanase, T. Ohta y M. Koyama, «Optuna: A Next-generation Hyperparameter Optimization Framework,» de *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, Association for Computing Machinery, 2019, pp. 2623-2631.

[69] A. Jadon y A. Patil, «A Comprehensive Survey of Evaluation Techniques for Recommendation Systems,» *pre-print 2312.16015*, 2024.

[70] C. J. K. Fouodo, L. L. Kronziel, I. R. König y S. Szymczak, «Effect of hyperparameters on variable selection in random forests,» *pre-print 2309.06943*, 2023.

[71] T. Trithipkaiwanpon y U. Taetragool, «Sensitivity Analysis of Random Forest Hyperparameters,» de *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2021, pp. 1163-1167.

ANEXOS

A. Tabla de variables

Listado de las variables consolidadas. Incluye nombre, tipo de variable, descripción y etapa, que indica el momento del proceso de contratación en el que es generada (Precontractual, Adjudicación y ejecución).

Nombre	Etapa	Tipo	Descripción
Variables identificadoras			
Id contrato	Adjudicación	Texto	Identificador del contrato firmado en SECOP II.
Proceso de compra	Adjudicación	Texto	Identificador del proceso de compra en SECOP II.
Nombre entidad	Precontractual	Texto	Nombre de la entidad pública solicitante.
NIT entidad	Precontractual	Texto	Número de Identificación Tributario de la entidad pública solicitante en SECOP II.
Código proveedor	Adjudicación	Texto	Identificador del proveedor en SECOP II
Documento proveedor	Adjudicación	Texto	Identificador del proveedor seleccionado.
Nombre proveedor	Adjudicación	Texto	Nombre del proveedor seleccionado.
Variables explicativas			
Código entidad	Precontractual	Categórica	Identificador de la entidad pública que publica el contrato.
Orden	Precontractual	Categórica	Orden entidad del Estado que publica el contrato (Nacional, Territorial)
Rama	Precontractual	Categórica	Rama del Poder a la que corresponde la entidad pública.
Tipo de contrato	Precontractual	Categórica	Tipo de contrato de acuerdo con su marco jurídico.

Modalidad de contratación	Precontractual	Categórica	Modalidad de contratación de acuerdo con el modelo de selección.
Condiciones de entrega	Precontractual	Categórica	Condiciones bajo las cuales se entrega el producto o servicio.
Tipo de documento proveedor	Adjudicación	Categórica	Tipo de documento del proveedor adjudicado.
Es grupo	Adjudicación	Categórica	Determina si el proveedor es un grupo de empresas u organizaciones.
Es Pyme	Adjudicación	Categórica	Determina si el proveedor es una pequeña o mediana empresa.
Obligación ambiental	Precontractual	Categórica	Determina si el contrato tiene compromisos de cumplimiento a obligaciones ambientales.
Destino gasto	Precontractual	Categórica	Destino del gasto a nivel presupuestal.
Valor del contrato	Precontractual	Numérica	Valor total del contrato.
Saldo CDP	Precontractual	Numérica	Saldo del Certificado de Disponibilidad Presupuestal (CDP) asignado al contrato.
Saldo vigencia	Precontractual	Numérica	Saldo para la vigencia (año) del CDP asignado al contrato.
Recursos PGN	Precontractual	Numérica	Valor de los recursos provenientes del Presupuesto General de la Nación.
Recursos SGP	Precontractual	Numérica	Valor de los recursos provenientes del Sistema General de Participaciones.
Recursos SGR	Precontractual	Numérica	Valor de los recursos provenientes del Sistema General de Regalías.
Recursos AGR	Precontractual	Numérica	Valor de los recursos provenientes de las alcaldías, gobernaciones y resguardos indígenas.
Recursos crédito	Precontractual	Numérica	Valor de los recursos provenientes de crédito.
Recursos propios	Precontractual	Numérica	Valor de los recursos propios de las entidades, diferentes a alcaldías, gobernaciones y resguardos indígenas.
Tipo de empresa	Adjudicación	Categórica	Tipo de empresa del proveedor adjudicado.
Asegurado	Adjudicación	Categórica	Determina si el contrato está asegurado por una póliza.
Valor medio ofertas	Precontractual	Numérica	Valor promedio de las ofertas realizadas por proveedores al contrato.
Valor máximo ofertas	Precontractual	Numérica	Valor máximo de las ofertas realizadas por proveedores al contrato.

Número de ofertas	Precontractual	Numérica	Número de ofertas realizadas por proveedores al contrato.
Zona	Precontractual	Categórica	Agrupación de departamentos de la entidad que publica el contrato.
Segmento producto	Precontractual	Categórica	Segmento de codificación UNSPSC del producto o servicio del contrato.
Proveedor colombiano	Adjudicación	Categórica	Determina si el proveedor adjudicado es persona natural o jurídica colombiana.
Origen recursos distribuido	Precontractual	Categórica	Determina si los recursos para la financiación del contrato provienen de diferentes orígenes.
Coseno año firma	Adjudicación	Numérica	Transformación cíclica de la fecha de firma del contrato usando el coseno.
Coseno año proveedor	Adjudicación	Numérica	Transformación cíclica de la fecha de registro del proveedor usando el coseno.
Días firma – inicio del contrato	Adjudicación	Numérica	Número de días entre la fecha de firma del contrato y la fecha de inicio del contrato.
Días inicio – fin del contrato	Adjudicación	Numérica	Número de días entre el inicio y la fecha de finalización de contrato.
Días proveedor inscrito	Adjudicación	Numérica	Número de días entre el registro del proveedor en SECOP II y la firma del contrato,
Días abierto a ofertas	Precontractual	Numérica	Número de días entre la publicación del contrato para recibir ofertas y el último día que recibió ofertas.
Longitud descripción	Precontractual	Numérica	Número de caracteres del texto de descripción del contrato.
Variable objetivo			
Adiciones en valor	Ejecución	Numérica	Número de adiciones en valor (sobrecostos) que tuvo el contrato.

Tabla 6. Anexo A: Variables consolidadas.

B. Estadísticas descriptivas de los puntajes de atipicidad

Estadísticas descriptivas de los puntajes de atipicidad producto del entrenamiento de Local Oulier Factor (LOF) e Isolation Forest (IF)

Modelo	Media	Std	Min	5%	10%	25%	50%	75%	90%	95%	Max
LOF	1,25	0,89	0,92	0,98	0,99	1,00	1,07	1,23	1,55	1,94	104,74
IF	0,43	0,05	0,35	0,36	0,37	0,39	0,42	0,46	0,52	0,54	0,64

Tabla 7. Anexo B: Distribución de puntajes de atipicidad.

C. Métricas de desempeño en regresión excluyendo atipicidad.

Métricas de desempeño de modelos de regresión incluyendo y no incluyendo los puntajes de atipicidad como variables explicativas.

Modelo	Incluye Atipicidad	RMSE Entrenamiento	MPD Entrenamiento	RMSE Prueba	MPD Prueba
Random Forest	No	**0,240**	**0,075**	**0,670**	**0,330**
Random Forest	Si	0,244	0,077	0,683	0,341
CatBoost	Si	0,727	0,475	0,798	0,512
CatBoost	No	0,723	0,474	0,790	0,512
XGBoost	No	0,708	0,458	0,792	0,519
XGBoost	Si	0,711	0,466	0,799	0,522
MLP	Si	0,920	0,666	0,965	0,738
MLP	No	0,924	0,669	0,963	0,749
Poisson	Si	1,953	1,413	1,799	1,427
Poisson	No	1,953	1,413	1,799	1,427

Tabla 8. Anexo C: Métricas de desempeño en regresión excluyendo atipicidad.