Characterization and analysis of errors in the value and additions of public contracts in Colombia using machine learning algorithms

Ing. Heriberto Felizzola Jimenez[1], Ing. Daniela Maria Devia[1], Ing. July Marcela Martin[1]
[1]Universidad de La Salle, Colombia, healfelizzola@unisalle.edu.co, ddevia05@unisalle.edu.co, julymmartin03@unisalle.edu.co.

***Abstract*– This paper presents the implementation of a machine learning algorithm for the detection and analysis of errors in the value and cost overrun in public contracts in Colombia, in the database of the Electronic Public Procurement System - SECOP I. The research begins with the characterization of the errors that were identified in the sample and which are made when entering the information into the system. Two algorithms were implemented, one to predict contracts with errors in value and other to predict contracts with errors in cost overrun. The Random Forest classification algorithm is used, performances in accuracy of 0.91 and 0.76 are obtained, for the classification of errors in the value and cost overrun, respectively.**

***Keywords*-- Public Procurement, Data Quality, Errors, Machine Learning, Random Forest.**

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.542
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

Caracterización y análisis de errores en el valor y las adiciones de los contratos públicos en Colombia utilizando algoritmos de aprendizaje automático

Ing. Heriberto Felizzola Jimenez[1], Ing. Daniela Maria Devia[1], Ing. July Marcela Martin[1]
[1]Universidad de La Salle, Colombia, healfelizzola@unisalle.edu.co, ddevia05@unisalle.edu.co, julymmartin03@unisalle.edu.co.

Resumen *– En este trabajo se presenta la implementación de un algoritmo de aprendizaje automático para la detección y análisis de los errores en el valor y las adiciones en contratos públicos en Colombia, registrados en la base de datos del Sistema Electrónico de Compra Pública - SECOP I. La investigación parte de una caracterización de los errores que se identificaron en la muestra, los cuales se comenten al introducir la información al sistema. Se desarrollan dos modelos, uno para detectar contratos con error en el valor y otro para clasificar contratos con errores en la adición. Se utiliza el algoritmo de clasificación Random Forest con el cual se obtiene desempeños en la exactitud de 0.91 y 0.76, para la clasificación de errores en el valor y la adición, respectivamente.*

Palabras claves-- Contratación Pública, Calidad de Datos, Errores, Aprendizaje Automático, Random Forest

Abstract*– This paper presents the implementation of a machine learning algorithm for the detection and analysis of errors in the value and cost overrun in public contracts in Colombia, in the database of the Electronic Public Procurement System - SECOP I. The research begins with the characterization of the errors that were identified in the sample and which are made when entering the information into the system. Two algorithms were implemented, one to predict contracts with errors in value and other to predict contracts with errors in cost overrun. The Random Forest classification algorithm is used, performances in accuracy of 0.91 and 0.76 are obtained, for the classification of errors in the value and cost overrun, respectively.*

Keywords-- Public Procurement, Data Quality, Errors, Machine Learning, Random Forest.

I. Introducción

La contratación pública es el mecanismo que utilizan las entidades gubernamentales para la adquisición de bienes y servicios y el desarrollo de obras civiles [1]. Según datos de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) las compras públicas representan aproximadamente el 12% del producto interno bruto de un país, lo cual resalta su importancia en la dinámica económica. En Colombia toda la actividad contractual de las entidades públicas se gestiona por medio del Sistema Electrónico de Contratación Pública -SECOP-, el cual este compuesto de 3 plataformas: SECOP I, que es un sistema informativo en el que cada una de las entidades están obligadas a divulgar su actividad contractual; SECOP II, la plataforma transaccional para las compras electrónicas; y la Tienda Virtual del Estado del Colombia que soporta la administración de procesos de agregación de demanda.

Estudios realizados por Zuleta [2] y la OECD [3] sobre el SECOP, plantean la necesidad de hacer mejoras en la calidad de los registros de las compras públicas, debido a que se encuentran inconvenientes que incluyen: el registro de un mismo contratista con nombres y números de identificación diferentes; errores en los valores de contratos; errores en las fechas del contrato; objetos, procesos y modalidades de contratación mal clasificados; entre otros.

La calidad e integridad de los datos es un aspecto clave en la transparencia de los procesos de contratación, y más aún conocer con exactitud el valor de los contratos y sus posibles adiciones permite a las partes interesadas hacer un control efectivo sobre los gastos y presupuestos [4]. Adicionalmente, con el auge del Big Data y la Analítica es posible analizar aspectos relacionados con la eficiencia, efectividad y riesgo de corrupción en la contratación, para lo cual es necesario contar con datos confiables [5].

En Colombia la plataforma SECOP, a mayo de 2022, manejaba más de 13 millones de registros [6]. En este sentido, la identificación y corrección de los errores en los datos de contratación es una tarea de gran envergadura, dado el volumen de contratación en los sistemas de compra pública. Lo anterior expone la necesidad de desarrollar y aplicar herramientas computacionales para la detección automática de errores, lo que permite tomar acciones para prevenir y mejorar la confiabilidad de los registros.

Este trabajo presenta una metodología que integra: (i) la caracterización de patrones para la detección de errores en los datos, (ii) la utilización procesos computacionales para la detección registros con errores, (iii) la implementación de técnicas de aprendizaje máquina para clasificarlos de forma automática y (iv) el análisis de las características del contrato que se relacionan con la presencia de errores. Cabe resaltar, que

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.542
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

este trabajo se enfoca primordialmente en los errores relacionados con el valor y las adiciones de los contratos.

El presente artículo se desarrollará de la siguiente manera, en la sección II se describen los patrones de los errores más comunes en el valor y las adiciones del contrato. En la sección III se describe la base de datos utilizada para analizar los errores, en la sección IV se presentan los métodos y herramientas utilizados para el desarrollo del modelo de aprendizaje automático y análisis de los resultados, en la sección V se presentan los resultados del algoritmo de Random Forest para la clasificación de los errores, y al final unas conclusiones sobre los resultados.

II. Caracterización de los errores

Gran parte de los errores en el valor y las adiciones en los contratos registrados en el SECOP I se dan al momento de su registro en el sistema. Estos errores se pueden generar por diferentes causas, entre las cuales se tiene: valores con errores en el documento del contrato o las modificaciones, confusión en el manejo de los separadores de decimales, falta de claridad en el manejo del sistema, falta de estándares para el registro de los datos, adición y omisión de valores y problemas con el manejo del orden de magnitud. A continuación, se presenta una descripción de los errores más comunes.

A. Adición o eliminación de ceros

En este tipo de error el valor registrado contiene ceros de más o viceversa. Esto genera un cambio significativo en el orden de magnitud del contrato, ya que, al agregar o retirar un cero de la cantidad se puede pasar de un millón a diez millones. En Colombia este es un error común, debido a que la moneda maneja valores de alta denominación, por lo cual es común encontrar contratos con valores de un alto orden de magnitud (por ejemplo 10^9), aumentando el riesgo de cometer errores en digitación de valores considerablemente altos.

B. Valores con errores en cifras decimales

Este error pasa una vez que no se sitúa de manera correcta el separador decimal, lo que puede generar cifras extras y por ende cambia el orden de magnitud, ejemplo: un contrato por valor de 100.975.846,93 (orden de magnitud 10^8) al utilizar mal el separador decimal puede quedar registrado por valor de 100.975.846.930 (orden de intensidad 10^{11}).

C. Valores de cero

En este caso el valor del contrato y sus adiciones se registran por un valor de cero. Este es un error que se ha encontrado en contratos que por su naturaleza deben tener un costo registrado superior a cero, ejemplo: contratos de prestación de servicio, obra, abastecimiento, entre otros. Existe otra tipología de contratos en los que no es viable decidir el costo al instante de la firma, debido a que el contratista produce cobros a medida que se desarrolla el objeto contractual, ejemplo: los contratos para la recuperación de cartera, en cuyo caso el costo que se cobra está en función de la cartera recuperada.

D. Valores de contratos que incluyen las adiciones

Este tipo de error consiste en que el valor de contrato es registrado como su valor inicial más las adiciones. En algunos casos se ha encontrado que la adición se incluye en los dos campos (valor inicial y adición).

E. Error de digitación

Este tipo de error puede tener diferentes patrones, por ejemplo: el valor se digita doblemente o dentro del valor algunas cifras se duplican (ejemplo 1760017600). Este error se puede presentar tanto en el valor del contrato como en las adiciones.

III. Datos

Para realizar una caracterización y análisis de los errores se tomó una muestra de 1656 contratos de la base de datos SECOP I, entre 2014 y 2020. En total se seleccionaron 15 atributos de los contratos entre los cuales se tiene: clasificación de la entidad según el nivel y orden territorial, fecha de firma del contrato, régimen de contratación, tipo de proceso de contratación utilizado, tipo de contrato, clasificación del bien o servicio, valor del contrato, valor de la adición, plazo de ejecución del contrato, departamento de la entidad, entre otros. La selección de estos atributos nos permite en primer lugar realizar una caracterización de los errores, con miras a brindar recomendaciones a las agencias públicas para mejorar la calidad e integridad de los datos en función de las características del contrato, y en segundo lugar identificar patrones en los atributos de los contratos que los hacen más propensos a los errores de registro.

Adicionalmente, se crearon una serie de variables cuya finalidad es detectar la presencia patrones en los valores y adiciones que puedan indicar algún tipo de error o inconsistencia, estos son:

Número de ceros en el valor: entre la mayor la cantidad de ceros mayor riesgo de error. Algunas observaciones preliminares indican que en cifras altas (orden de magnitud 10^9) y gran cantidad de ceros al final, la posibilidad de cometer errores es alta.

Relación entre la adición y el valor del contrato: la ley establece que un contrato no puede tener una adición superior al 50% del valor inicial. Una relación entre la adición y valor del contrato superior al 50% indica una inconsistencia y por ende un posible error. En otros casos se encontró que si la relación considerablemente baja, esto es menos del 0.01%, también hay posibilidad de error.

Diferencia entre la cuantía del proceso (valor inicial) y el valor del contrato (valor al momento de la firma): cuando la diferencia es considerablemente alta (más del 100%) existe gran posibilidad de error.

Relación entre la cuantía del proceso y el valor de contrato: al igual que ocurre con las adiciones, si la relación entre la cuantía del proceso y el valor del contrato es alta la probabilidad de error también lo es.

Por último, se tienen dos atributos para el valor y la adición, donde 1 indica que el contrato presenta error 0 lo contrario. Estos atributos binarios serán las variables para predecir con los modelos de aprendizaje automático. En esta muestra se encontraron 547 contratos con errores en el valor y 99 contratos con errores en la adición, con lo cual se tienen tasas de error del 33% para el valor y 24% para la adición (ver Figura 1 y Figura 2).

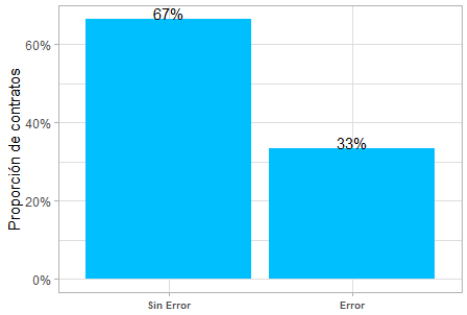

Figura 1. Proporción de contratos con y sin error en cuantía.

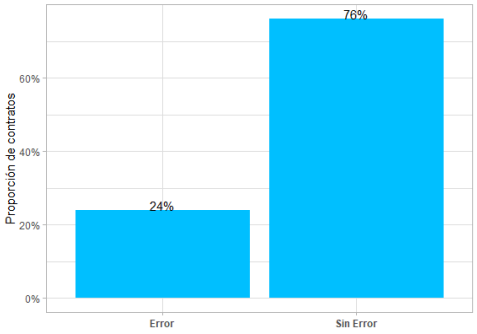

Figura 2. Proporción de contratos con y sin error en adiciones.

En la TABLA 1. Se presenta un resumen estadístico de las principales variables numéricas.

TABLA 1.
RESUMEN ESTADÍSTICO DE LAS VARIABLES NUMÉRICAS

Nombre de la variable	Media	Desviación estándar
Cuantía del proceso	$ 15.603.445.867	$ 39.411.554.311
Plazo de ejecución del contrato	102 días	767 días
Adiciones en tiempo	35.3 días	122 días
Valor del contrato	$ 15.601'980.833	$ 69.585'741.248
Valor de adiciones	$ 1.170'426.827	$ 17.512'766.006

Podemos observar que la media de la cuantía del proceso y valor del contrato se encuentran por encima de los 15 mil millones, que la media de las adiciones asciende por encima de los mil millones. Para el plazo de ejecución del contrato tenemos una media de 102 días con una desviación estándar de 767 días y para el tiempo adicional en días de los contratos tenemos una media de 35.5 con una desviación de 122.

IV. MÉTODOS Y HERRAMIENTAS

A. *Algoritmo Random Forest*

Para desarrollar una herramienta que permita clasificar de forma automática los contratos con errores en el valor o la adición se utilizó el algoritmo de Random Forest, un modelo para cada tipo de error. Este es un algoritmo de aprendizaje automático que construye una cantidad determinada de árboles de decisión a partir de la selección aleatoria de un subconjunto de variables y datos [7], [8]. El conjunto de datos se dividió en dos partes, para el entrenamiento del algoritmo se seleccionó de forma aleatoria el 70% de los registros, y el 30% restantes se utilizó para la validación. Por parámetros utilizados en el Random Forest fuerton:

- Criterion = 'entropy'
- max_depth = 11
- min_samples_leaf = 5
- min_samples_split = 5
- max_features = 1.0
- n_estimators = 60

B. *Métricas de evaluación*

Las métricas utilizadas en el análisis del desempeño del algoritmo son: Exactitud (Accuracy), Precisión, Recuperación (Recall), el F1-Score y el Área Bajo la Curva ROC (AUC). Para entender las métricas de desempeño es necesario partir de la matriz de confusión y algunos conceptos que se derivan de esta (ver Figura 3), como son:

- *Verdadero Positivo (VP):* Número de casos positivos clasificados correctamente.
- *Falso positivo (FP):* Número de casos negativos clasificados como positivos.
- *Verdadero Negativo (VN):* Número de casos negativos clasificados correctamente.
- *Falso Negativo (FN):* Número de casos positivos clasificados como negativos.

		Clase Predicha	
		1: Positivo	0: Negativo
Clase Real	1: Positivo	VP	FN
	0: Negativo	FP	VN

Figura 3. Matriz de confusión y métricas

A partir de estos conceptos se definen las siguientes métricas:

Exactitud (Accuracy): Indica la proporción casos, positivos o negativos, clasificados correctamente. Esto es:

$$\text{Exactitud} = \frac{VP + VN}{VP + FP + FN + VN} \quad (1)$$

Se debe tener cuidado con el uso e interpretación de esta métrica en los casos donde las clases están desbalanceadas, ya que, su valor no representa correctamente la exactitud en la predicción de la clase minoritaria [9]. Por esta razón, se debe complementar con otras métricas como la precisión o la recuperación.

Precisión: es una medida del desempeño de la predicción de la clase positiva. Se calcula como la proporción de casos positivos clasificados correctamente sobre el total de casos clasificados como positivos. Esto es:

$$\text{Precisión} = \frac{VP}{VP + FP} \quad (2)$$

Recuperación o sensibilidad (Recall): mide la capacidad para detectar correctamente todos los casos positivos. Se calcula como la proporción de casos positivos clasificados correctamente sobre el total de casos reales positivos.

$$\text{Recuperación} = \frac{VP}{VP + FN} \quad (3)$$

F1-score: es una medida que permite hacer un balance entre la precisión y la recuperación. Se calcula como la media armónica entre estas métricas:

$$\text{F1-Score} = 2\frac{\text{Precisión} \rightarrow \text{Recuperación}}{\text{Precisión} + \text{Recuperación}} \quad (4)$$

Área bajo la curva ROC: La curva ROC es un gráfico que ilustra la relación entre sensibilidad y especificidad (capacidad para detectar las clases negativas). Se utiliza para calibrar la probabilidad a partir de la cual se predice la clase positiva, esto con el fin de obtener el mejor balance entre sensibilidad y especificidad. El área bajo esta curva, la cual puede tomar un valor máximo de 1 cuando se tiene un clasificador perfecto, se utiliza para evaluar y comparar la capacidad predictiva de los modelos y algoritmos de clasificación [10].

C. Lenguajes de programación y entorno de desarrollo

Para el desarrollo del algoritmo de Random Forest y todas las fases de entrenamiento y validación se utilizó el entorno Google Colab y el lenguaje de programación Python. Para el preprocesamiento de los datos, el entrenamiento y validación del algoritmo de Random Forest se utilizó la librería Scikit Learn [11], adicionalmente se utilizó la librería Yellowbrick [12] para la visualización de los resultados de la validación del Random Forest. Para análisis exploratorio de los datos se utilizó el lenguaje R y las librerías dplyr [13] y ggplot2 [14], las cuales permiten flujos de trabajos mucho más ágiles en tareas relacionadas con ciencia de datos.

V. Resultados

A. Predicción de errores en la cuantía del contrato

En la Tabla 2 se presentan los resultados de la matriz de confusión para la predicción de errores en la cuantía del contrato con el algoritmo Random Forest. En este caso los contratos con error son la clase positiva (1) y los contratos sin error la clase negativa (0). Se puede observar que para la cuantía del contrato fueron clasificados correctamente 136 contratos con error (VP) y 316 contratos sin error (VN). Por otro lado, fueron clasificados de forma incorrecta 21 contratos como falsos positivos y 24 como falsos negativos.

Con estos resultados el algoritmo Random Forest tiene una exactitud del 0.91, lo cual indica que en promedio que 91 de cada 100 contratos serán clasificados correctamente. Por otro lado, se tiene una precisión de 0.85, lo que significa que de cada 100 predicciones de contratos con error 85 de estas serán correctas. Además, se tiene una tasa de recuperación de 0.87, lo que significa que, de cada 100 contratos con errores, se espera que el clasificador identifique 87 de estos. Por último, el F1-Score es de 0.86 y el área bajo la curva ROC presenta un valor de 0.97 lo que muestra una buena capacidad predictiva del algoritmo en general (ver Figura 4).

Tabla 2.
Matriz de confusión para la detección de errores en cuantía.

		Clase predicha	
		1: Positivo	0: Negativo
Clase real	1: Positivo	136	21
	0: Negativo	24	316

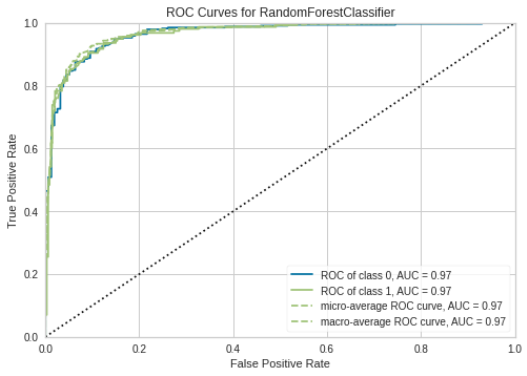

Figura 4. Curva ROC para clasificación de error en cuantía con algoritmo Random Forest.

En la Figura 5 se presenta la gráfica de importancia de las variables, esta permite identificar cuáles son las variables que brindan más información para predecir los contratos con

errores. En el Random Forest las tres variables más importantes son el tipo de consulta (patrón de error), la cuantía de proceso y la cuantía de contrato.

En la Figura 6 se muestra una gráfica de barra para representar la proporción de contratos con y sin error por tipo de consulta. En esta podemos observar que la mayor proporción de contratos con error se tiene cuando los contratos tienen gran cantidad de ceros (tipo de consulta ceros adicionales y cuando se presenta una gran diferencia entre la cuantía del proceso y el valor del contrato).

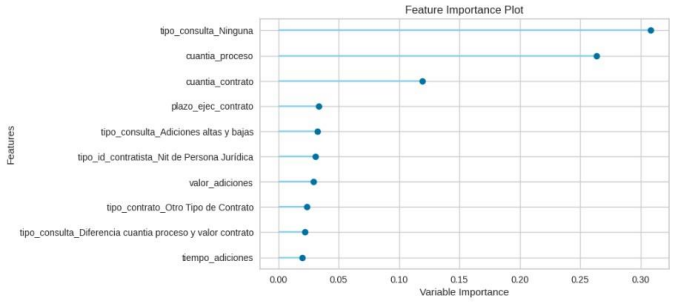

Figura 5. Importancia de las variables para la detección de errores en cuantía con Random Forest

En la Figura 7 se presenta un gráfico donde se puede observar que la cuantía del proceso es una de las variables que aporta información relevante para la detección de errores en cuantía. Se puede ver por el tamaño de la caja de error (roja) la cual es más pequeña que la caja sin error (azul) y los datos que conforman la caja de error son extremos (muy grandes o pequeños) entonces podemos concluir que los contratos con cuantías muy bajas o muy por encima de la mediana son más propensos a presentar errores.

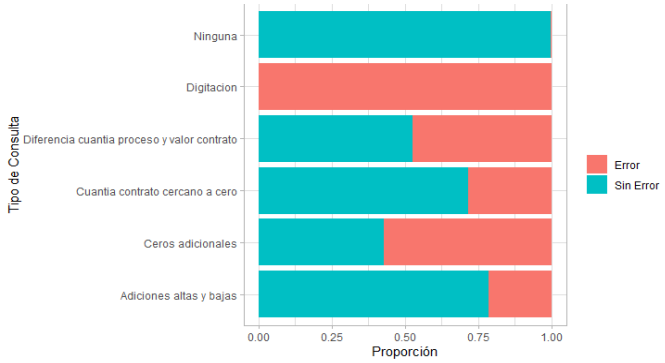

Figura 6. Proporción de error en cuantía por tipo de consulta

B. *Predicción de errores en la adición del contrato*

Para este modelo cabe resaltar que solo tomaron los contratos que tienen alguna adición (343 contratos), ya que, solo para estos tiene sentido evaluar la presencia o no de errores. En la Tabla 3 se presentan los resultados de la matriz de confusión para la predicción de errores en la adición del contrato con el algoritmo Random Forest. Se puede observar que para la adición del contrato fueron clasificados correctamente 11 contratos con error (VP) y 67 contratos sin error (VN). Por otro lado, fueron clasificados de forma incorrecta 9 contratos como falso positivo y 16 como falsos negativos.

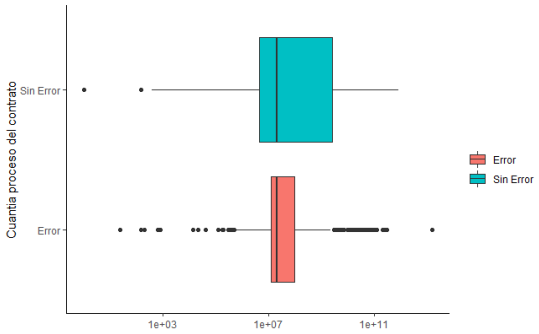

Figura 7. Cuantía del proceso para contratos con y sin error

Con estos resultados el algoritmo Random Forest tiene una exactitud de 0.76, lo cual indica que en promedio que 76 de cada 100 contratos serán clasificados correctamente, según el error en la adición. Por otro lado, se tiene una precisión de 0.55, lo que significa que de cada 100 predicciones de contratos con error en la adición 55 de estas serán correctas. Además, se tiene una tasa de recuperación de 0.41, lo que significa que, de cada 100 contratos con errores en la adición, se espera que el clasificador identifique 41 de estos.

Por último, el F1-Score es de 0.47 y El área bajo la curva ROC presenta un valor de 0.77 lo cual nos muestra buena capacidad predictiva del algoritmo en general (ver Figura 8). Con respecto a este último aspecto, la curva ROC muestra que la precisión para detectar errores en la adición se podría mejorar si se calibra el umbral de probabilidad para clasificar un contrato como positivo, con esto se podría aumentar la tasa de recuperación 0.80 sin afectar considerablemente la capacidad para detectar la clase negativa.

TABLA 3.
MATRIZ DE CONFUSIÓN PARA LA DETECCIÓN DE ERRORES EN LA ADICIÓN.

		Clase predicha	
		1: Positivo	0: Negativo
Clase real	1: Positivo	11	16
	0: Negativo	9	67

En la Figura 9 se presenta la gráfica de importancia de variables para la detección de errores en la adición. Las tres variables más importantes son el valor de la adición, la cuantía de proceso y la cuantía de contrato. A continuación, un corto análisis sobre la relación de estas variables con el error en adiciones.

En la Figura 10 se puede observar que los contratos que no presentan error tienen una menor variabilidad en el valor de la adición que los contratos que presentan error en la adición. Además, los contratos sin error tienen mayor cantidad de datos atípicos en este valor, lo cual puede ser contra intuitivo.

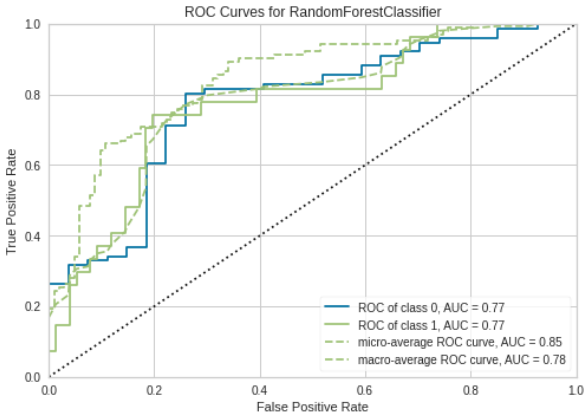

Figura 8. Curva ROC para clasificación de error en la adición con algoritmo Random Forest.

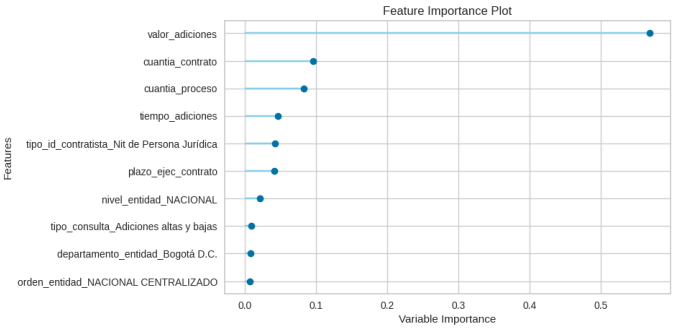

Figura 9. Importancia de las variables para la detección de contratos con error en la adición con Random Forest

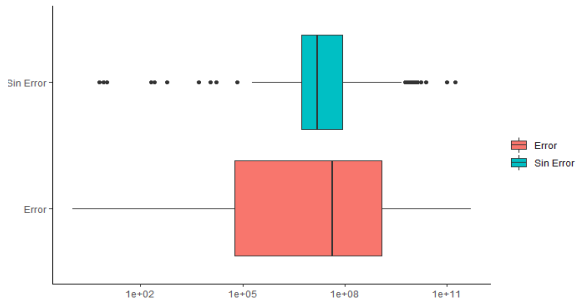

Figura 10. Error en la adición vs valor de las adiciones

Contrario con el valor de las adiciones se puede observar que los contratos sin error en la adición tienen una mayor variabilidad en la cuantía (ver Figura 11). En ambas gráficas se pueden observar leves diferencias en las medianas del valor de la adición y la cuantía del contrato, para los contratos con y sin error.

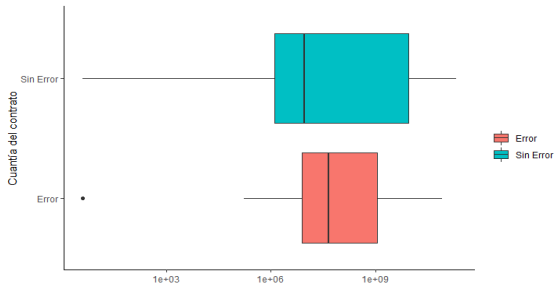

Figura 11. Error en la adición vs cuantía del contrato

C. *Comparación de resultados para los dos tipos de error*

En la Tabla 4 se presenta una comparación de las métricas del Random Forest para detectar errores en la cuantía y la adición de los contratos. Se puede observar que existe una diferencia significativa en el desempeño del Random Forest para detectar cada uno de los tipos de errores, presentando un mejor desempeño en la detección de errores en la cuantía. Esto se debe a diferentes factores como la cantidad registros, las variables incluidas y el desbalanceo de las clases. Si revisamos el AUC podemos concluir que los dos modelos tienen una buena capacidad predictiva, ya que se encuentra por arriba de 0.5, lo que indica que se pueden detectar errores en la cuantía y la adición con un desempeño superior a un modelo aleatorio.

TABLA 4.
COMPARACIÓN DE LOS RESULTADOS PARA PREDICCIÓN DE ERROR EN LA CUANTÍA Y LA ADICIÓN DEL CONTRATO.

	Errores en Valor	Errores en Adición
Exactitud	0.91	0.76
Precisión	0.85	0.55
Recuperación	0.87	0.41
F1-Score	0.86	0.47
AUC ROC	0.97	0.77

VI. CONCLUSIONES

A partir del modelo desarrollado se puede concluir que existe una relación entre el error en el valor del contrato y características como el tipo de consulta (patrones de error en el valor), la cuantía de proceso y la cuantía de contrato. Además, se encontró una capacidad de predicción de error en cuantía elevada, como se pudo evidenciar a partir de la métrica AUC o área bajo la curva de 0.97 por lo que se puede considerar que es un clasificador con una buena capacidad predictiva.

A partir de los resultados encontramos que el desempeño del modelo en la detección de errores en adición es menor en comparación a los resultados obtenidos para detectar errores en la cuantía, con un área bajo la curva de 0.77, que sigue considerándose aceptable. Adicionalmente, se encontraron que las variables de importancia para la detección de estos errores fueron el valor de la adición, la cuantía de proceso y la cuantía

de contrato. Además, se pudo observar que la cuantía del proceso es importante para la detección de ambos tipos de error.

Consideramos que el enfoque de aprendizaje automático es útil en la detección de problemas de calidad en los contratos presentes en el SECOP I. Este tipo de modelos puede ser la base para implementar soluciones computacionales para detectar errores y mejorar la calidad de los resultados, y partir del análisis de los resultados se pueden generar propuestas para garantizar el adecuado registro de la información y por ende la integridad de los datos.

En trabajos futuros se recomienda probar otros modelos y algoritmos de clasificación con el fin de evaluar y comparar su efectividad con respecto al algoritmo de Random Forest. Recomendamos trabajar con una muestra más grande para que el modelo puede entrenarse con mayor confiabilidad y la posibilidad de encontrar nuevos patrones, lo que mejoraría su desempeño y brindaria métricas con mayor precisión para analizar y seleccionar el modelo optimo. Por último, se recomienda incluir otras variables relacionadas con patrones de error para mejorar la capacidad predictiva y lograr propuestas que permitan mejorar la calidad de los registros desde la fuente.

REFERENCIAS

[1] OECD, «Preventing Corruption in Public Procurement». 2016. [En línea]. Disponible en: https://www.oecd.org/gov/public-procurement/publications/Corruption-Public-Procurement-Brochure.pdf

[2] M. M. Zuleta, V. Saavedra, y J. C. Medellín, «Fortalecimiento del sistema de compra pública para reducir el riesgo de corrupción», abr. 2018, Accedido: 24 de junio de 2020. [En línea]. Disponible en: http://www.repository.fedesarrollo.org.co/handle/11445/3544

[3] OECD, *Towards Efficient Public Procurement in Colombia: Making the Difference*. OECD, 2016. doi: 10.1787/9789264252103-en.

[4] C. Csáki y E. Prier, «Quality Issues of Public Procurement Open Data», en *Electronic Government and the Information Systems Perspective*, vol. 11032, A. Kő y E. Francesconi, Eds. Cham: Springer International Publishing, 2018, pp. 177-191. doi: 10.1007/978-3-319-98349-3_14.

[5] A. Soylu *et al.*, «Data Quality Barriers for Transparency in Public Procurement», *Information*, vol. 13, n.º 2, p. 99, feb. 2022, doi: 10.3390/info13020099.

[6] Colombia Compra Eficiente, «SECOP Integrado». https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-Integrado/rpmr-utcd (accedido 6 de junio de 2022).

[7] G. James, D. Witten, T. Hastie, y R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-7138-7.

[8] L. Breiman, «Random Forests», *Mach. Learn.*, vol. 45, n.º 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[9] I. H. Witten, E. Frank, y M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA: Morgan Kaufmann, 2016. doi: 10.1016/c2009-0-19715-5.

[10]T. Fawcett, «An introduction to ROC analysis», *Pattern Recognit. Lett.*, vol. 27, n.º 8, pp. 861-874, jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[11]F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in Python», *J. Mach. Learn. Res.*, vol. 12, n.º 85, pp. 2825-2830, 2011.

[12]B. Bengfort *et al.*, *Yellowbrick V0.6*. Zenodo, 2018. doi: 10.5281/ZENODO.1206264.

[13]H. Wickham, R. François, L. Henry, y K. Müller, *dplyr: A Grammar of Data Manipulation*. Rstudio, 2022.

[14]H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, 2nd ed. 2016. Cham: Springer International Publishing : Imprint: Springer, 2016. doi: 10.1007/978-3-319-24277-4.