Doi: 10.62544/ucomscientia.v3i1.46
UCOM Scientia,
Vol. 3, N° 1, marzo, 2025, 115-140
ISSN 3006-7278

Artículo Original/ Original Article

Optimización de hiperparámetros en algoritmos de aprendizaje no supervisado para la detección de anomalías en contrataciones públicas del Paraguay

Optimization of hyperparameters in unsupervised learning algorithms for anomaly detection in public procurement in Paraguay

Matías Fabián Sanabria
[1]Universidad Nacional de Asunción. San Lorenzo, Paraguay.
https://orcid.org/0009-0006-1424-2897
Autor corresponsal: matt31sanabria@fpuna.edu.py

Julio Manuel Paciello Coronel
[1]Universidad Nacional de Asunción. San Lorenzo, Paraguay.
https://orcid.org/0000-0003-3196-5625

Juan Ignacio Pane Fernández
[1]Universidad Nacional de Asunción. San Lorenzo, Paraguay.
https://orcid.org/0000-0003-2607-4027

Para citar este artículo:

Sanabria, M. F.; Paciello Coronel, J. y Pane Fernández, J. I. (2025). Optimización de hiperparámetros en algoritmos de aprendizaje no supervisado para la detección de anomalías en contrataciones públicas del Paraguay. *UCOM Scientia*,*3*(1), 115-140.

Fecha de recepción: 31/08/2024

Fecha de aceptación: 4/01/2025

Resumen

Este estudio aborda la optimización de hiperparámetros en algoritmos de aprendizaje no supervisado aplicados a la detección de anomalías en contrataciones públicas en Paraguay. El principal objetivo es desarrollar una herramienta capaz de identificar irregularidades en los procesos de contratación, utilizando datos abiertos proporcionados por la Dirección Nacional de Contrataciones Públicas. La metodología sigue el estándar de la industria CRISP-DM e incluye la recopilación, transformación y preparación de los datos, seguida de la aplicación de los algoritmos *Isolation Forest, Local Outlier Factor y One-Class SVM*. La optimización de los hiperparámetros se lleva a cabo mediante técnicas de *grid search y random search*, además se aborda el desbalanceo de clases en los datos utilizando la técnica de *oversampling* SMOTE. Los resultados indican que, aunque los modelos con valores altos en la métrica de *recall* detectan la mayoría de las anomalías, presentan un elevado número de falsos positivos. En contraste, para obtener modelos con altos valores de precisión, se requiere de un balanceo del conjunto de datos, disminuyendo considerablemente los falsos positivos en sacrificio de no identificar todas las anomalías. En conclusión, es deseable trabajar en un correcto etiquetado y balanceo del conjunto de datos de entrenamiento para mejorar la precisión y la utilidad práctica de los modelos.

Palabras clave: Detección de anomalías; aprendizaje de máquina; inteligencia artificial; compras públicas.

Editora responsable: Dra. Carla Fernández
Carla.fernandez@ucom.edu.py

Abstract

This study focuses on hyperparameter optimization in unsupervised learning algorithms for anomaly detection in public procurement processes in Paraguay. The main objective is to develop a tool that identifies irregularities in procurement processes using open data provided by the National Directorate of Public Procurement. The methodology follows the CRISP-DM industry standard, including data collection, transformation, and preparation, followed by the application of the algorithms Isolation Forest, Local Outlier Factor and One-Class SVM. Hyperparameter optimization is performed using grid search and random search techniques, and class imbalance is addressed using SMOTE oversampling. Results indicate that while the high recall model detects most anomalies, it produces a significant number of false positives. In contrast, to obtain models with high precision, a balancing of the data set is required, considerably reducing false positives at the cost of not identifying all anomalies. In conclusion, it is desirable to work on a correct labeling and balancing of the training data set to improve the accuracy and practical utility of the models.

Keywords: Anomaly detection; machine learning; artificial intelligence; public procurements.

1. Introducción

En el Corruption Perception Index (CPI) del 2023 (Transparency International, 2023) el Paraguay obtuvo un valor de 28/100, donde 0 indica una alta corrupción, lo que significa que los paraguayos perciben un incremento de la corrupción en el sector público del país. Quedando así, Paraguay, en la posición 136 de 180.

Las contrataciones públicas son fundamentales para transparentar, difundir y facilitar los procesos de compras de los organismos del Estado en Paraguay. La detección de anomalías en este contexto es crucial para asegurar la transparencia y eficacia de dichos procesos. Con ello, hay varios desafíos que enfrentan las entidades públicas con las contrataciones públicas, entre las que se pueden mencionar: el gran volumen de documentos generados y poco personal calificado para la revisión de los llamados a compras públicas; el direccionamiento de los pliegos de las bases y condiciones en beneficio a ciertos oferentes, la exclusión de oferentes con mejores condiciones para favorecer las ofertas menos convenientes; detectar de manera temprana y eficaz los pliegos con bases y condiciones que pueden derivar en una protesta o denuncia; así como la detección de protestas contra adjudicaciones por tener deficiencias en la evaluación de las ofertas y la asignación de las adjudicaciones.

En el contexto de las contrataciones públicas, una anomalía se refiere a cualquier desviación o irregularidad en el proceso de contratación que podría, aunque no necesariamente, indicar prácticas inadecuadas, errores o incluso corrupción. El estándar de datos de contrataciones abiertas (OCDS, por sus siglas en inglés) (Open Contracting Partnership, s.f.) permite la publicación de datos abiertos de contrataciones públicas en un formato estandarizado a nivel internacional, ayuda a incrementar la transparencia, hace posible un análisis profundo de los datos de contrataciones y facilita el uso de estos datos para una amplia gama de actores interesados. Investigaciones académicas demuestran que la mejora en la apertura y

transparencia de datos son beneficiosas para la integridad pública, la relación calidad-precio y la competencia cuando están asociadas a cambios sistemáticos que permiten a las personas usar la información.

En el contexto del OCDS, las anomalías pueden ser identificadas mediante la comparación de datos reportados contra ciertos patrones esperados o normas establecidas (las banderas rojas). Es importante su detección temprana para evitar que los fondos del estado sean malgastados, desviados o utilizados para fines que no corresponden, así como también para ir disminuyendo la corrupción.

En el artículo publicado por Open Contracting Partnership (2021), se destaca cómo Paraguay utilizó OCDS durante la pandemia para mejorar la transparencia en la rendición de cuentas. A través de la publicación en tiempo real, la DNCP permitió que la sociedad civil y los periodistas puedan monitorear de cerca las adquisiciones relacionadas con el COVID-19, lo que llevó a la reducción significativa de precios y un proceso de compra más rápido. Así, poco después de que comenzaran las contrataciones de emergencia por la pandemia y se expusieran historias sobre compras de bienes críticos a precios sobrevalorados, la DNCP emitió una regulación que requería que las instituciones no solo publicaran los precios de referencia obtenidos en el mercado, sino que también explicaran cómo llegaron a esas cifras. Además, la DNCP disponibilizó la tienda en línea para la compra de productos en un entorno similar al comercio electrónico altamente competitivo. "*Nos centramos en tener la mayor cantidad de oferentes en la tienda y los resultados fueron muy interesantes porque los precios comenzaron a caer en picada. Por ejemplo: una mascarilla comprada fuera de la tienda al comienzo de la pandemia costaba alrededor de 1 dólar (7.000 guaraníes). A medida que comenzamos a agregar más y más proveedores y la competencia se intensificó, el precio de las mascarillas cayó a 700 guaraníes*", explicó Vázquez, portavoz de la DNCP. Los organismos del estado pudieron comprar sus bienes más rápido, ya que no era necesario redactar pliegos de licitación ni recibir y evaluar ofertas, acortando los plazos de contratación de 3 semanas a 3 días.

El análisis de datos abiertos y la supervisión ciudadana han dado lugar a nuevas herramientas de monitoreo y ayudaron a detectar irregularidades que antes estaban ocultas entre la gran cantidad de información disponible sobre los procesos de contratación. Existen varios ejemplos de hechos de corrupción que pudieron ser identificados gracias a los datos publicados por la DNCP, por mencionar el caso de Petróleos Paraguayos (Petropar) que adquirió botellas de agua tónica a un precio unitario aproximado de 10 usd, siendo el precio de mercado entre 1 y 2 usd, lo cual desencadenó el sumario y judicialización de la cabeza de la institución y varios funcionarios. También fueron identificadas otras situaciones donde los procesos de contratación de sumas muy elevadas eran ejecutados en plazos de 24 horas o menos, y otros

casos donde en compras mayoristas del estado los precios unitarios eran iguales o más elevados que los precios minoristas del supermercado.

En la publicación de Gómez Scifo, (2023), se han analizado los expedientes procesados durante los primeros 80 días de los períodos correspondientes al 2013, 2018 y 2023 respectivamente, en los cuales se observa que en el primer periodo (2013) se procesaron 3241 expedientes, en el segundo periodo (2018) se procesaron 3274 expedientes; mientras que en el actual (2023) se procesaron 4085 expedientes, como se muestra en la Figura 1. Un notable crecimiento en la cantidad de licitaciones procesadas, lo cual lleva también a un aumento en los totales de las licitaciones, atendiendo que en el primer periodo del 2018 el monto fue de más de 600 millones de dólares, mientras que en el periodo actual el monto ya supera los 900 millones de dólares, observada en la Figura 2. Así también, analizando el mismo periodo comprendido entre los primeros 80 días del gobierno anterior y el actual, se observa un aumento en la cantidad de protestas procesadas por parte de la actual administración de la DNCP, observándose un claro aumento, ya que en el periodo anterior se finalizaron 262 protestas y en el actual han sido finalizadas más de 300, que se puede ver en la Figura 3. Estos números demuestran la importancia de contar con mecanismos y/o herramientas que permitan automatizar el control de los procesos de contratación.

Figura 1. Licitaciones procesadas en los últimos 3 períodos de gobierno

Figura 2. Expedientes de Licitación procesados en Millones de USD en los últimos 2 periodos de gobierno

Figura 3. Cantidad de Protestas finalizadas en los últimos 2 periodos de gobierno

La detección de anomalías en contrataciones públicas abordada por Niessen et al. (2020), presenta un enfoque para detectar anomalías en los procesos de contratación pública utilizando el estándar de datos abiertos de contratación (OCDS) y un modelo de aprendizaje no supervisado basado en el Isolation Forest. Donde se utilizan datos de contrataciones públicas del Paraguay, transformados a un formato adecuado para su análisis de machine learning, con

el objetivo de asignar una puntuación de anomalía a cada proceso de contratación para identificar potenciales procesos irregulares. Los resultados muestran una efectividad mayor a la del 90% en la detección de anomalías conocidas como protestas y denuncias en el proceso de contratación. Además de demostrar la viabilidad de este enfoque para la detección temprana de irregularidades en contrataciones públicas.

Además, López San Martín et al. (2024) en su trabajo aborda el problema de las protestas en los procesos de contratación pública en Paraguay, las cuales pueden causar retrasos y costos adicionales significativos. Para enfrentar este desafío, los autores proponen un modelo basado en técnicas de aprendizaje automático para predecir la probabilidad de que una licitación sea objeto de protesta. Utilizando datos históricos de la Dirección Nacional de Contrataciones Públicas (DNCP), el estudio construye un modelo predictivo que incorpora banderas rojas predefinidas como indicadores de posibles irregularidades. El dataset utilizado fue cuidadosamente balanceado mediante técnicas como SMOTE para abordar el desbalanceo de clases, el modelo fue entrenado y evaluado utilizando la herramienta H2O AutoML. Los resultados obtenidos son prometedores, con una precisión del 86.1% y una F-measure de 74.4%, lo que sugiere que el modelo puede ser una herramienta eficaz para la gestión proactiva del riesgo en los procesos de contratación pública en Paraguay.

El trabajo de Kiran et al. (2020) aborda el problema de la detección de anomalías en transferencias de red, donde identificar transferencias anómalas es crucial para asegurar el rendimiento y la confiabilidad de los sistemas. Donde los autores investigaron métodos de extracción de características no supervisados, como el Análisis de Componentes Principales (PCA), autoencoders y el algoritmo Isolation Forest, aplicados a datos de transferencias TCP. Se utilizaron dos conjuntos de datos, uno generado sintéticamente y otro basado en el 1000 Genomas workflow, ambos con anomalías introducidas artificialmente. El trabajo reveló que mientras PCA y autoencoders tienen dificultades para detectar anomalías, Isolation Forest demuestra ser efectivo al identificar patrones anómalos en los datos, destacando su potencial para la detección de anomalías.

En Zenati et al. (2018) se evaluaron y compararon el rendimiento de una amplia gama de algoritmos de detección de anomalías en diferentes escenarios, como tipos variados de anomalías y la presencia de datos ruidosos o corruptos. Donde se propone la creación de ADBench, un marco de referencia integral que evalúa 30 algoritmos de detección de anomalías en 57 conjuntos de datos, incorporando métodos no supervisados, semi supervisados y supervisados; que también incluye un análisis detallado del rendimiento de los algoritmos bajo diferentes niveles de supervisión, tipos de anomalías, y condiciones de ruido, proporcionando valiosas recomendaciones para la selección de algoritmos en aplicaciones del mundo real.

Mehta et al. (2018), plantea detectar comportamientos anómalos en los registros de conexión de red en entornos con grandes volúmenes de datos, como los generados en el CERN (European Organisation for Nuclear Research). Donde, para ello se implementa un sistema de detección de anomalías utilizando aprendizaje no supervisado, en el que se aplican técnicas como PCA (Principal Component Analysis), SVD (Singular Vector Decomposition), y algoritmos de clustering como KNN, Isolation Forest, Local Outlier Factor, y SVM de una clase. La solución propuesta permite la identificación efectiva de anomalías en tiempo real en un entorno de datos distribuidos, mejorando la capacidad de respuesta ante posibles intrusiones o fallos de seguridad. Los resultados muestran que el sistema puede detectar diferentes tipos de anomalías con una precisión aceptable, validada tanto por métricas de evaluación como por la detección de incidentes reales dentro de la infraestructura del CERN.

En el artículo de Campos et al. (2016) se aborda el problema de la evaluación de algoritmos de detección de anomalías no supervisados, subrayando la falta de conjuntos de datos de referencia y métricas de evaluación adecuadas. Para abordar este problema, los autores proponen un estudio empírico exhaustivo en el que se evalúan varios algoritmos basados en k-nearest neighbors (kNN) sobre una amplia gama de datasets. La solución propuesta incluye el desarrollo de un conjunto de medidas de evaluación adaptadas y una caracterización detallada de los datasets utilizados, destacando la importancia de un correcto ajuste de los parámetros y la necesidad de benchmarks estandarizados. Los resultados muestran que los algoritmos clásicos como LOF y kNN siguen siendo robustos en diversos escenarios, aunque no existe un método universalmente superior.

Domingues et al. (2017), en su artículo menciona que identificaron los algoritmos de detección de anomalías más eficaces para diferentes aplicaciones, como la detección de fraudes y la identificación de intrusiones. Los autores realizaron un exhaustivo estudio comparativo de 14 algoritmos de detección de anomalías no supervisados, evaluándolos en 15 conjuntos de datos, tanto públicos como industriales. Compararon estos algoritmos en términos de precisión, escalabilidad y uso de memoria. Con ello se llegó a la conclusión de que, aunque no existe un algoritmo universalmente superior, el Isolation Forest mostró un rendimiento destacado en términos de detección de anomalías, escalabilidad y eficiencia en el uso de memoria, siendo recomendado para implementaciones en entornos de producción. Otros algoritmos, como Kernel Density Estimation (RKDE) y One-Class SVM, también demostraron un buen desempeño, aunque con mayores requerimientos computacionales.

Da Alesandro (2019) explora cómo identificar y gestionar anomalías en un sistema de compensación en tiempo real (RTCS), donde el gran volumen de transacciones y mensajes intercambiados genera desafíos significativos para la detección manual de irregularidades. El

estudio propone la implementación de modelos de aprendizaje automático no supervisados, específicamente SVM One-Class, Isolation Forest, y Local Outlier Factor (LOF), para automatizar el proceso de detección de anomalías. Donde se llevó a cabo una evaluación exhaustiva utilizando datos del RTCS de Cinnober en São Paulo, y se emplearon métricas como el F-score, la precisión y el coeficiente de correlación de Matthews para medir la efectividad de los modelos. Los resultados muestran que el modelo SVM One-Class superó a los otros dos en la mayoría de las métricas, destacándose como el más adecuado para la detección de anomalías en el RTCS. Sin embargo, se señala la necesidad de un ajuste cuidadoso de los hiperparámetros para alcanzar un rendimiento óptimo.

La investigación presentada en "Outlier Selection and One-Class Classification" de Janssens (2013), explora los desafíos en la selección de anomalías y clasificación de una sola clase, enfocándose en mejorar la detección de anomalías en datos complejos. El estudio se enmarca en el proyecto Poseidón, cuyo objetivo es aumentar la seguridad marítima mediante la identificación automática de anomalías. Para abordar esta tarea, se desarrolló el algoritmo "Stochastic Outlier Selection" (SOS), que utiliza un enfoque basado en afinidad para calcular probabilidades de anomalía para cada punto de datos. Este método se distingue por su capacidad para gestionar variaciones en la densidad de datos y perturbaciones, mostrando una mayor robustez en comparación con otros algoritmos. Los resultados de las pruebas empíricas, realizadas en un conjunto de 25 datasets, demuestran que SOS supera a algoritmos tradicionales en términos de rendimiento promedio y resiliencia, aunque no es universalmente el mejor para todos los conjuntos de datos.

Zhao et al. (2019) en su artículo "PyOD: A Python Toolbox for Scalable Outlier Detection" se centra en la carencia de herramientas dedicadas específicamente a la detección de anomalías en Python, a pesar de la creciente importancia de este lenguaje en el campo del aprendizaje automático. Los autores abordan esta necesidad desarrollando PyOD, una caja de herramientas de código abierto diseñada para facilitar la detección de anomalías en datos multivariados. Esta herramienta reúne más de 20 algoritmos de detección de anomalías, desde enfoques clásicos como el Local Outlier Factor (LOF) hasta modelos basados en redes neuronales, todo bajo una API unificada y bien documentada. Los resultados obtenidos mediante pruebas en múltiples aplicaciones académicas y comerciales demuestran que PyOD es una solución robusta y escalable, capaz de manejar grandes volúmenes de datos y diferentes tipos de anomalías con alta eficiencia. Además, la herramienta ha sido adoptada ampliamente por la comunidad de usuarios, mostrando un impacto significativo en la investigación y la industria.

Feurer y Hutter (2019) en el capítulo uno del libro “Automated Machine Learning: Methods, Systems, Challenges”, aborda la optimización de hiperparámetros (HPO) en modelos de

aprendizaje automático, destacando la importancia de esta práctica para mejorar el rendimiento y la eficiencia de los modelos. Se describen métodos tradicionales como la optimización por random search, grid search y optimización bayesiana, así como enfoques más avanzados como la optimización bayesiana y métodos multi-fidelidad, que permiten una evaluación más rápida y efectiva en datasets grandes. El grid search es el método de Optimización de Hiperparámetros más básico, donde el usuario especifica un conjunto finito de valores para cada hiperparámetro, y el grid search evalúa el producto cartesiano de estos conjuntos. Este método sufre la maldición de la dimensionalidad, ya que el número de evaluaciones crece exponencialmente con la dimensionalidad del espacio de configuración. Un problema adicional del grid search es que al aumentar la resolución de la discretización se incrementa sustancialmente el número necesario de evaluaciones de funciones. Otras ventajas sobre el grid search son la mayor facilidad de paralelización y la flexibilidad en la asignación de recursos. El random search realiza muestreos de configuraciones al azar hasta que se agota un determinado límite para la búsqueda. Funciona mejor que el grid search cuando algunos hiperparámetros son mucho más importantes que otros. Cuando se ejecuta con un límite fijo de B evaluaciones de funciones, el número de valores diferentes que el grid search puede permitirse evaluar para cada uno de los N hiperparámetros es sólo B1/N, mientras que el random search explorará B valores diferentes para cada uno de los N hiperparámetros. El random search es un punto de partida útil porque no hace suposiciones sobre el algoritmo de aprendizaje automático que se está optimizando y, dados los recursos suficientes, se espera que alcance un rendimiento arbitrariamente cercano al óptimo. La búsqueda aleatoria también es un método útil para inicializar el proceso de búsqueda, ya que explora todo el espacio de configuración y, por tanto, suele encontrar configuraciones con un rendimiento razonable. La optimización bayesiana es un marco de optimización de vanguardia para la optimización global de costosas funciones de caja negra, que recientemente ha ganado tracción en HPO al obtener nuevos resultados de vanguardia en el ajuste de redes neuronales profundas para la clasificación de imágenes, el reconocimiento del habla y el modelado neuronal del lenguaje, y al demostrar una amplia aplicabilidad a diferentes escenarios de problemas. La optimización bayesiana es un algoritmo iterativo con dos ingredientes clave: un modelo sustituto probabilístico y una función de adquisición para decidir qué punto evaluar a continuación. En cada iteración, el modelo sustituto se ajusta a todas las observaciones de la función objetivo realizadas hasta el momento. A continuación, la función de adquisición, que utiliza la distribución predictiva del modelo probabilístico, determina la utilidad de los distintos puntos candidatos, sopesando la exploración y la explotación. En comparación con la evaluación de la costosa función de caja negra, la función de adquisición es barata de calcular y, por tanto, puede optimizarse a fondo. Muchos desarrollos recientes en optimización bayesiana no se dirigen directamente a la HPO, pero a menudo pueden aplicarse fácilmente a la HPO, como

nuevas funciones de adquisición, nuevos modelos y núcleos, y nuevos esquemas de paralelización.

Además Komer et al. (2019) presenta Hyperopt-Sklearn, una herramienta de AutoML basada en la popular biblioteca scikit-learn, diseñada para la optimización automática de hiperparámetros. El capítulo describe cómo Hyperopt-Sklearn emplea métodos de búsqueda secuencial y modelos probabilísticos para explorar de manera eficiente el espacio de hiperparámetros, mejorando así el rendimiento de los modelos. También se discuten técnicas para manejar el desbalanceo de datasets, crucial para asegurar que los modelos generados sean robustos y generalicen bien, especialmente en presencia de clases minoritarias.

Los trabajos revisados en la literatura proporcionan una base sólida y diversa de metodologías para la detección de anomalías, donde se destaca la aplicación de modelos de aprendizaje automáticos no supervisados, así como la optimización de hiperparámetros para mejorar la precisión. Estos enfoques son muy relevantes para el desarrollo de una herramienta que permita identificar anomalías en las contrataciones públicas, al incorporar metodologías robustas que han demostrado ser efectivas en otros contextos. Además, se subraya la importancia del manejo adecuado del balanceo de los conjuntos de datos y el uso de herramientas de optimización automática.

Entre los algoritmos no supervisados de detección de anomalías que se utilizaron se encuentran el Isolation Forest, Local Outlier Factor, OC-SVM; teniendo en cuenta que son los algoritmos más utilizados para la detección de fraudes, detección de intrusión en redes de computadoras, análisis de riesgo crediticio, entre otros; que se encuentran en la literatura. Para la optimización de los hiperparámetros se utilizaron las técnicas de grid search y la búsqueda aleatoria. Además, ya que el conjunto de datos utilizado tiene una clase minoritaria muy pequeña con respecto a la otra, se aplicó la técnica de oversampling (SMOTE, para ser más específicos) para balancear cuidadosamente el conjunto de datos.

El objetivo central de la investigación es la elaboración de un marco de trabajo para la optimización de los hiperparámetros de algoritmos no supervisados para la detección de anomalías en contrataciones públicas.

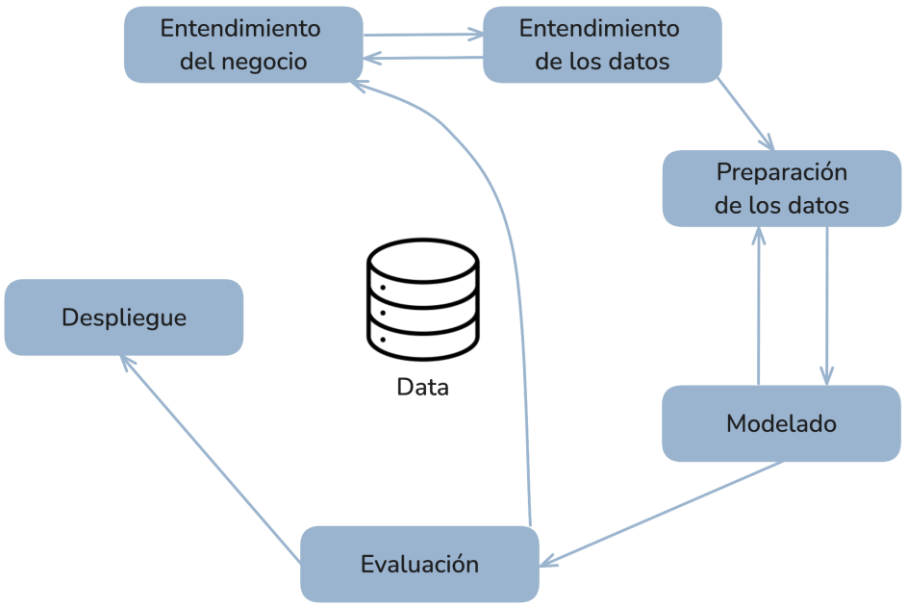
La principal contribución del trabajo es la de brindar una herramienta que permita automatizar la detección de anomalías en los procesos de contratación utilizando algoritmos de aprendizaje automático no supervisados incluyendo la optimización de sus hiperparámetros.

El presente trabajo se divide en las siguientes secciones: materiales y métodos, resultados obtenidos, discusión y por último, las conclusiones.

2. Materiales y métodos

Para llevar a cabo este trabajo se ha utilizado como referencia el estándar de la industria *Cross Industry Standard Process for Data Mining* (CRISP-DM, por sus siglas en inglés) (Chapman et al., 2000); que incluye un modelo y una guía, estructurados en seis fases: la comprensión del negocio, comprensión de los datos, la preparación de los datos, el modelado, la evaluación y el despliegue, como se muestra en la siguiente figura 4.

Figura 4. Fases de la metodología CRISP-DM



El entendimiento del negocio, es la fase inicial, donde se enfoca en entender los objetivos del proyecto y sus requerimientos desde la perspectiva del negocio, para luego este conocimiento convertirlo en una definición de problema de minería de datos y diseñar preliminarmente un plan para alcanzar los objetivos.

La fase de entendimiento de los datos inicia con una recolección inicial de datos y continúa con actividades para familiarizarse con los datos, para identificar problemas de la calidad de los datos, para descubrir las primeras percepciones de los datos o detectar subconjuntos interesantes para formar hipótesis sobre información oculta.

La fase de preparación de los datos cubre todas las actividades para construir el conjunto de datos final (datos que serán introducidos en las herramientas de modelado) a partir de los datos brutos iniciales. Es probable que las tareas de preparación de datos se realicen varias veces y no en un orden preestablecido. Las tareas incluyen selección de tablas, registros y atributos, así como la transformación y limpieza de los datos para las herramientas de modelado.

En la fase del modelado, varias técnicas de modelado son seleccionadas y aplicadas, así también sus parámetros son calibrados para valores óptimos. Típicamente, hay varias técnicas para el mismo tipo de problema de minería de datos. Por lo tanto, volver atrás a la fase de preparación de datos es a menudo necesario.

Para la fase de evaluación, ya se ha construido el o los modelos que parecen tener una gran calidad desde el punto de vista del análisis de datos. Antes de proceder al despliegue final del modelo, es importante evaluarlo más a fondo y revisar los pasos ejecutados para construir el modelo con el fin de asegurarse de que alcanza adecuadamente los objetivos del proyecto.

La creación del modelo no suele ser el final del proyecto. Aunque el objetivo del modelo es generar conocimientos a partir de los datos, habrá que organizar y presentar los conocimientos adquiridos de forma que el cliente pueda utilizarlos. A partir de las fases del CRISP-DM definidas, pudimos identificar esas fases en la investigación realizada, quedando de la siguiente manera:

En la fase del entendimiento del negocio, el objetivo principal del trabajo es la detección de los llamados de contrataciones públicas que pueden derivar en protestas o denuncias, para el fortalecimiento de los mecanismos de control y la disminución del mal uso de los recursos del estado. Donde son necesarias herramientas que puedan ayudar a las personas encargadas de la verificación de los pliegos a detectar problemas en los mismos; como el direccionamiento de los requerimientos en las bases y condiciones, la mala evaluación de los pliegos, entre otros.

Con el uso de los datos abiertos que se encuentran en la página de la DNCP, estos datos ser utilizados para el entrenamiento de un modelo o varios modelos de Inteligencia Artificial para la detección de anomalías utilizando modelos de aprendizaje no supervisado, que puedan ayudar a los analistas encargados a detectar uno o varios subconjuntos donde se presentan las anomalías en los procesos de contrataciones públicas, de manera automática y escalable para que pueda soportar el gran volumen de datos existente.

Para la fase de entendimiento y preparación de los datos, utilizamos los datos que se encuentran disponibles en la página web de la DNCP, de donde pueden ser descargados en formato abierto de planilla electrónica CSV, o desde el portal de datos abiertos de la DNCP

donde se cuenta con varias APIs, según la etapa en la que se encuentran los llamados: planificaciones, licitaciones, adjudicaciones, contratos, protestas, entre otros.

Con la disponibilidad del acceso a los datos, estos fueron recolectados desde la API de datos abiertos en su versión 1. A los datos recolectados, fueron agregados dos campos, las etiquetas de los procesos y el número de los oferentes. Entre los posibles valores de las etiquetas se encuentran: el abastecimiento simultáneo, ad referéndum, agricultura familiar, bienestar social estratégico, contrato abierto, Fondo Nacional de Inversión Pública y Desarrollo (FONACIDE), impugnado, plurianual, producción nacional, seguridad nacional, subasta a la baja electrónica, urgencia impostergable. Posteriormente se actualizó el formato de los datos para tenerlo en su versión 1.1 del OCDS utilizando el código disponible en el repositorio de Github (McKinney, 2023), según lo menciona Niessen et al. (2020) en su trabajo.

Durante la selección de variables se tuvieron en cuenta los siguientes factores: i) si la variable estaba en uso, ii) si el contenido de la variable tenía una estructura, evitando el uso de texto libre o URLs, iii) se eliminan variables altamente correlacionadas o repetidas, iv) se eliminaron los ítems del catálogo debido a la granularidad del experimento y por último, v) se definió no utilizar las variables de los estados de las fases. Algunas variables fueron agregadas teniendo en cuenta las banderas rojas del estándar del OCDS. Otras variables fueron transformadas a variables numéricas según los tipos de datos que fueron detectados como se muestra en la siguiente tabla 1.

Tabla 1. Transformaciones por cada tipo de variable

Tipo de variable	Transformación
montos de dinero	los montos que se encontraban en guaraníes, quedaron con su valor original. Los montos de los valores que estaban en dólares americanos fueron calculados con la tasa de conversión correspondiente a la fecha de la convocatoria, contrato o modificación según su etapa.
fechas	de las variables que tienen fechas, se utilizan solamente el mes ya que en un trabajo anterior se encontró una relación importante entre el mes de la adjudicación y algunos criterios como los montos y las cantidades de procesos (Vierci Codas, 2018). Se generó una variable denominada “periodo entre fechas”, el cual es un valor numérico de la cantidad de días entre dos fechas. Otro campo agregado es el porcentaje de tiempo que pasó desde el inicio de un periodo, a otra fecha, con respecto al final del periodo.

colección de variables binarias	donde se dispuso de la colección de valores posibles en un diccionario, se creó un arreglo binario inicializado con 0; para cada valor del diccionario que forma parte en la colección de variables categóricas, se cambió el valor a 1, y finalmente se sumaron los valores decimales resultantes como se puede observar en la tabla 2, donde se muestra la transformación cuando el método de adjudicación es una "Licitación Pública Nacional".
variables categóricas	se manejaron como un arreglo binario, donde las opciones en el diccionario son las opciones categóricas del campo y el resultado es el número decimal correspondiente al campo según su posición en el arreglo binario.
variables binarias	para las variables donde los posibles valores son solamente "Si" o "No", se asignó un 1 si es un "Si" y un 0 si el valor es un "No".
variables numéricas	mantuvieron su valor original.
texto libre	se convirtieron a un valor de hash único, el cual se transforma a un valor numérico único mediante la conversión a hexadecimal. El valor numérico resultante (hash) se guarda en un diccionario con su valor original. Aquí presentamos un ejemplo utilizado en el trabajo de Niessen et al. (2020) donde el valor del campo partyContractpointEmail es uoc@ips.gov.py: partyContractpointEmail = int(md5(partyContactpointEmail).hex(), 16) = int(md5(uoc@ips.gov.py).hex(), 16) = int(554a60f64e2fe04c2f14d19c79c28893,16) = 113370576234897065005746686023629441171
variables categóricas almacenadas como texto libre	se adecuaron los scripts desarrollados en Vierci Codas (2018) para limpiar los datos mencionados y se procedió a la transformación del dato como texto libre, como se indica en la tabla 3 para el valor de campo partyAddressRegion ACUNCION, cuyo cálculo del valor hash es el siguiente: partyAddressRegion = int(md5(limpiar_datos(partyAddressRegion)).hex(),16) = int(md5(limpiar_datos(ACUNCION)).hex(), 16) = int(md5(ASUNCIN).hex(),16) = int(5438a9dc9a8652f50affbcf1a5a1c186,16)

= 111949365475252733268185372591388737926

identificadores	se obtiene el hash del identificador pero se mantiene en el diccionario el valor del nombre correspondiente. En identificadores como el RUC, se sacan los caracteres como el guión y se usa el mismo número.

Niessen et al. (2020)

Tabla 2. Ejemplo de conversión de variables categóricas

Valor actual del campo	Valor binario
Acuerdo Internacional	0
Acuerdo Nacional	0
Compra directa por excepción	0
Concurso de oferta	0
Contratación directa	0
Licitación pública internacional	0
Licitación pública nacional	1
Locación de inmueble	0
Proceso de capacitación	0
Renovación de locación	0
Subasta a la baja electrónica	0

Niessen et al. (2020)

Valor binario	0	0	0	0	0	0	1	0	0	0	0
Posición del campo en la tabla	0	1	2	3	4	5	6	7	8	9	10
Valor decimal	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}

$$\text{metodoDeAdjudicación} = 0 * 2^0 + 0 * 2^1 + 0 * 2^2 + 0 * 2^3 + 0 * 2^4 + 0 * 2^5 + 1 * 2^6 + 0 * 2^7 + 0 * 2^8 + 0 * 2^9 + 0 * 2^{10} = 64$$

Tabla 3. Ejemplo de conversión de variables categóricas almacenadas como texto libre

Valor actual del campo	Valor limpio
…	…
FILA	FILADELFIA
ACUNCION	ASUNCIÓN
AUSNCION	ASUNCIÓN
EMBY	ÑEMBY
…	…

Niessen et al. (2020)

En la fase de modelado, tomamos de referencia trabajos del estado del arte para poder escoger los algoritmos de aprendizaje no supervisado para la detección de anomalías más utilizados y sus hiperparámetros. De la literatura consultada, se seleccionaron los siguientes algoritmos: Isolation Forest, Local Outlier Factor (LOF) y el One-Class SVM (OCSVM). Además, se seleccionaron los siguientes hiperparámetros por algoritmo de acuerdo a la Tabla 4.

Tabla 4. Algoritmos con sus hiperparámetros y rango de valores utilizados en los experimentos

Algoritmo	Hiperparámetros	Rango de valores
Isolation Forest	*n_estimators:* es el número de estimadores base	[10, 20, 30, …, 200]

	max_samples: es el número de ejemplos que va a extraer de X para entrenar cada estimador base.	[20, 40, 60, 80, 100]
	max_features: es el número de características a extraer de X para entrenar cada estimador base.	[23]
	contamination: es la proporción de valores atípicos en el conjunto de datos. Se utiliza al ajustar para definir el umbral en las puntuaciones de las muestras.	[0.1, 0.2, 0.3, 0.4, 0.5]
Local Outlier Factor (LOF)	*n_neighbors:* Número de vecinos a considerar para determinar el factor de localidad	[5, 10, 15, ..., 200]
	algorithm: algoritmo utilizado para calcular los vecinos más cercanos. Las opciones son: ball_tree , kd_tree, brute y auto	[‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’]
	leaf_size: tamaño de las hojas en las estructuras de árbol (BallTree o KDTree). Este parámetro puede afectar la velocidad de construcción y consulta del árbol, así como la memoria utilizada.	[10, 20, 30, …, 100]
	metric: métrica utilizada para calcular la distancia entre los puntos. El valor por defecto es "minkowski". Otras métricas posibles incluyen "euclidean", "manhattan", "chebyshev".	[‘euclidean’, ‘manhattan’, ‘chebyshev’, ‘minkowski’]

One-Class SVM (OCSVM)	*kernel:* especifica el tipo de núcleo que se utilizará en el algoritmo.	['linear', 'poly', 'rbf', 'sigmoid']
	nu: límite superior de la fracción de errores de entrenamiento y un límite inferior de la fracción de vectores de soporte.	[0.1, 0.2, …, 1]
	gamma: coeficiente de kernel para 'poly', 'rbf', 'sigmoid'	['scale', 'auto']

Para el entrenamiento de los modelos hemos utilizado inicialmente el conjunto de datos utilizados en el trabajo de Niessen et al. (2020), centrándonos en el conjunto de datos de planificación y convocatoria con entrega escrita, con los atributos ya transformados, el cual cuenta con un total de 103.748 registros, divididos en 93.373 registros de entrenamiento y 10.375 registros de validación.

Además, utilizamos el conjunto de datos de protestas y denuncias que han sido solicitados a la DNCP en el marco de la ley número 5282/2014 de libre acceso a la información pública y transparencia gubernamental (Congreso de la Nación Paraguay, 2014), el cual cuenta con 744 identificadores de procesos protestados y 494 denunciados. Es importante destacar que de estos conjuntos de datos de protestas y denuncias, para el entrenamiento fueron utilizados 440 protestas y 222 denuncias, mientras que para la validación 50 protestas y 21 denuncias, dándonos cuenta con esto que la cantidad de datos etiquetados era muy pequeña en comparación con el conjunto de registros totales.

A partir de esta problemática de desbalanceo del conjunto de datos, se han realizado pruebas con un conjunto de datos balanceado, generado a partir del conjunto de datos original, utilizando la técnica de oversampling SMOTE. Se introdujeron datos sintéticos y generando así un conjunto de datos de entrenamiento con 185.866 registros, 90.933 registros etiquetados como protestados y no protestados respectivamente. Este mismo procedimiento se aplicó para el conjunto de datos de validación, generando un total de 10.325 registros etiquetados como protestados y no protestados respectivamente. En la siguiente sección se explican los detalles de los diferentes experimentos implementados y sus resultados obtenidos.

3. Resultados

En nuestro estudio sobre la detección de posibles protestas o denuncias en llamados a licitación pública, hemos identificado dos enfoques distintos basados en los resultados obtenidos: Recall Alto (minimizar Falsos Negativos) y Precisión Alta (minimizar Falsos Positivos). Ambos enfoques de explican en detalle a continuación:

- **Enfoque de Recall Alto (Minimizar Falsos Negativos):** Este enfoque es útil cuando la prioridad es asegurarse de encontrar todo proceso de contratación que podría derivar en una protesta o denuncia. En este caso, se acepta lidiar con un mayor número de falsas alarmas (falsos positivos) a cambio de una alta capacidad para detectar todos los casos problemáticos conocidos. Este enfoque en la práctica puede llevar a mucho sobre trabajo de revisión de alertas, con el beneficio de garantizar procesar todas las posibles anomalías reales al costo de sobreprocesar procesos que posiblemente no presenten anomalías.

  Todos los modelos entrenados bajo este enfoque mostraron un recall alto y una precisión baja, de hasta 80% de recall con una precisión del 1% con el Isolation Forest optimizando los hiperparámetros con el Random Search, como se ilustran en la Tabla 5. Para el entrenamiento fue utilizado el conjunto de datos balanceado y para la validación fue utilizado el conjunto de datos original sin balancear, lo que permitió a los modelos detectar efectivamente las protestas y denuncias conocidas. Sin embargo, el desbalanceo de las clases y el posible ruido en el conjunto de datos, es decir procesos potencialmente anómalos no etiquetados correctamente, también llevaron a un gran número de falsos positivos. Esto significa que, aunque los modelos son capaces de identificar la mayoría de los casos reales, también etiquetan incorrectamente muchas instancias normales como problemáticas.

- **Enfoque de Precisión Alta (Minimizar Falsos Positivos):** Este enfoque se centra en minimizar las falsas alarmas (falsos positivos), es decir, reducir la probabilidad de predecir que un proceso de contratación terminará en una protesta o denuncia cuando en realidad no será así. Cuando el modelo predice que un proceso podría terminar en una protesta o denuncia, se busca que esta predicción sea muy precisa, a costa de no detectar todos los posibles casos anómalos. Es decir, este enfoque pretende no generar sobre trabajo de revisión de procesos, garantizando que cuando se dispara una alerta de anomalía, la misma representa una anomalía real con un alto grado de precisión.

  Para poder obtener resultados prometedores con este enfoque, fue necesario aplicar el balanceo del conjunto de datos tanto al entrenamiento como a la validación, lo que resultó en modelos con una precisión cercana al 70% y un recall mayor al 80% para el Isolation Forest con el Random Search superando así a los resultados del LOF y OCSVM

en el Grid Search y en el Random Search, como se ilustran en la Tabla 6. Esto indica que los modelos entrenados bajo estas condiciones son capaces de predecir con mayor precisión los casos que efectivamente terminarán en protestas o denuncias, aunque esto significa que algunos casos verdaderamente problemáticos pueden no ser detectados. Donde, si se desea mejorar aún más la precisión, es fundamental trabajar en el correcto etiquetado de los datos de entrenamiento para reducir el impacto del ruido y el desbalanceo en los resultados.

A continuación se muestran los resultados para ambos enfoques con los distintos algoritmos utilizados (Isolation Forest, Local Outlier Factor y el One Class SVM) con las técnicas de optimización aplicadas.

Enfoque de Recall Alto

Tabla 5. Métricas de los mejores modelos de los algoritmos con Grid Search

Algoritmo	Precision	Recall	Accuracy	AUC ROC	F1 Score
Isolation Forest	0.010091	0.2	0.90159	0.552494	0.019212
Local Outlier Factor	0.005438	0.32	0.714699	0.518305	0.010695
One Class SVM	0.004615	0.06	0.933108	0.498668	0.008571

Tabla 6. Métricas de los mejores modelos de los algoritmos con Random Search

Algoritmo	Precision	Recall	Accuracy	AUC ROC	F1 Score
Isolation Forest	0.010212	0.84	0.606843	0.722857	0.020178
Local Outlier Factor	0.005438	0.32	0.714699	0.518305	0.010695
One Class SVM	0.004615	0.06	0.933108	0.498668	0.008571

Enfoque de Precisión Alta

Tabla 7. Métricas de los mejores modelos de los algoritmos con Grid Search

Algoritmo	Precision	Recall	Accuracy	AUC ROC	F1 Score
Isolation Forest	0.668357	0.191477	0.548232	0.548232	0.297674
Local Outlier Factor	0.256117	0.093269	0.411186	0.411186	0.136741
One Class SVM	0.499613	0.062567	0.499952	0.499952	0.111207

Tabla 8. Métricas de los mejores modelos de los algoritmos con Random Search

Algoritmo	Precision	Recall	Accuracy	AUC ROC	F1 Score
Isolation Forest	0.685807	0.86063	0.733172	0.733172	0.763336
Local Outlier Factor	0.256117	0.093269	0.411186	0.411186	0.136741
One Class SVM	0.499613	0.062567	0.499952	0.499952	0.111207

4. Discusión

Los resultados obtenidos con los modelos de aprendizaje no supervisado, particularmente con Isolation Forest, destacan la capacidad de estos algoritmos para identificar anomalías en los procesos de contratación pública de Paraguay. Comparado con estudios previos, como el de Niessen et al. (2020), donde también se utilizó Isolation Forest para detectar irregularidades en las contrataciones públicas, nuestros hallazgos refuerzan la viabilidad de este enfoque en diferentes contextos. Sin embargo, a diferencia de estudios anteriores, nuestra investigación incluyó un enfoque exhaustivo en la optimización de hiperparámetros y en el balanceo de los datos mediante SMOTE, lo que resultó en una mejora significativa en la precisión de los modelos, en donde los trabajos previos se centraban principalmente en atender un alto *recall* sin garantizar precisiones altas.

Un aspecto crucial que surgió durante el estudio fue la disyuntiva entre minimizar los falsos negativos o los falsos positivos. En nuestro análisis, en ambos casos resultó mejor utilizar un conjunto de datos de entrenamiento balanceado, sin embargo en validación, solamente mediante un conjunto de datos balanceado se logró un mejor rendimiento en términos de precisión, aunque con un sacrificio en el *recall*. Esto es consistente con la literatura que sugiere que la optimización del balanceo de datos es esencial para mejorar la especificidad de los modelos, especialmente en contextos donde la clase minoritaria, como las protestas en las contrataciones públicas, es crítica.

Del análisis previo podemos deducir que un desafío clave identificado es trabajar el ruido potencial presente en los datos, es decir aquellos puntos de datos erróneamente etiquetados potencialmente por desconocimiento, y también la importancia de un balanceo en las etiquetas presentes en los conjuntos de datos utilizados, lo cual afecta la capacidad del modelo para generalizar de manera efectiva. Esto refuerza de forma objetiva, con resultados numéricos, este problema que también fue observado en otros estudios de detección de anomalías, como el trabajo de Martín et al. (2024), lo que sugiere que una mejora en la calidad y cantidad de datos etiquetados podría aumentar aún más la eficacia de los modelos.

Finalmente, es importante considerar las limitaciones del estudio, como la dependencia de los datos disponibles públicamente, que pueden no reflejar completamente todas las posibles anomalías en los procesos de contratación. Futuros trabajos deberían explorar la integración de nuevas fuentes de datos y la aplicación de técnicas de aprendizaje más avanzadas, como el aprendizaje profundo (*Deep Learning*), para abordar estas limitaciones y mejorar la capacidad predictiva de los sistemas de detección de anomalías.

5. Conclusiones

Con el trabajo realizado se ha revisado el estado del arte con respecto a las técnicas de detección de anomalías utilizando algoritmos de aprendizaje de máquina no supervisado y los hiperparámetros necesarios definir para obtener mejores resultados, derivando esto en una propuesta de automatización de los mismos utilizando técnicas de optimización de hiperparámetros y múltiples algoritmos de detección de anomalías.

Debido al desbalance de clase presente en el conjunto de datos utilizado, se procedió a realizar un balance de clases utilizando la técnica de oversampling SMOTE, generando una muestra con la misma cantidad de registros con clases positivas y negativas, tanto para el conjunto de entrenamiento como para el de validación.

Han sido entrenados los algoritmos Isolation Forest, Local Outlier Factor, One Class SVM. Se han definido rangos de valores más utilizados tomados de trabajos anteriores para sus

hiperparámetros, y se ha obtenido un mejor modelo por cada algoritmo con la combinación de su espacio de búsqueda de hiperparámetros utilizando el Grid Search y el Random Search.

En los resultados obtenidos, en ambos enfoques utilizando un conjunto de datos balanceado en el entrenamiento, podemos observar que con el conjunto de datos de validación sin balancear tenemos un *recall* alto y una precisión baja en todos los modelos, lo que significa que podemos encontrar todas las anomalías etiquetadas con el dataset de protestas y denuncias, pero también presenta muchos falsos positivos, que estimamos se deben a que se tienen pocos datos etiquetados como positivos en el conjunto de datos y podría haber presencia de muchos falsos negativos (ruido en el conjunto de datos). Sin embargo al utilizar conjunto de datos de validación balanceados si pudo lograrse resultados de alta precisión con *recall* moderados, reforzando así la necesidad de un correcto etiquetado del conjunto de datos a utilizar.

Por otro lado, si se realiza un esfuerzo de balancear y etiquetar mejor los datos, por ejemplo estableciendo mecanismos de controles más estrictos en el sistema de contrataciones públicas, esto permitiría obtener modelos más precisos, capaces de asegurar que cuando se dice que un llamado es una anomalía estos no fallen, minimizando el sobre esfuerzo por obtener varios falsos positivos que podrían representar una sobrecarga de trabajo de revisión. Contar con más datos etiquetados claramente puede ayudar a mejorar drásticamente un modelo de detección de anomalías, para que de este modo, las personas encargadas de la revisión de los procesos de contrataciones públicas puedan hacer un uso correcto y útil de las herramientas de Inteligencia Artificial para la disminución de la corrupción y el uso indebido de los recursos del estado.

Algunos trabajos futuros derivados de esta investigación que podrían ampliar significativamente el alcance y la aplicabilidad de la herramienta de detección de anomalías, contribuyendo a mejorar la transparencia y eficiencia en las contrataciones públicas podrían incluir:

- Investigar y aplicar otros algoritmos de aprendizaje no supervisado que no fueron considerados en este estudio, como *Deep Learning* basado en autoencoders o modelos generativos adversariales (GANs) para la detección de anomalías en grandes volúmenes de datos.
- Explorar técnicas más avanzadas de optimización de hiperparámetros, como la optimización bayesiana multi-fidelidad o el uso de algoritmos genéticos, para mejorar la eficiencia y el rendimiento de los modelos.
- Extender la experimentación actual a un conjunto de datos mayor, incluyendo red flags y otros datos que puedan provenir de otros conjuntos de datos y aporten mayor información para el etiquetado correcto de los datos.

6. Declaración de financiamiento

La presente investigación se llevó a cabo con financiación propia.

7. Declaración de conflictos de intereses

Los autores declaran no tener conflictos de intereses.

8. Declaración de autores

Los autores aprueban la versión final del artículo.

9. Contribución de los autores

Autor	Contribución
Matías Fabián Sanabria	Participación importante en el desarrollo e implementación de todos los experimentos realizados durante la investigación, y redacción de los documentos científicos.
Julio Manuel Paciello Coronel	Ssupervisión y cotutoría en el diseño y desarrollo de la investigación, con énfasis en los modelos de Inteligencia Artificial; revisión y participación en la elaboración de los documentos científicos.
Juan Ignacio Pane Fernández	Supervisión y cotutoría en el diseño y desarrollo de la investigación, con énfasis en los procesos de Contrataciones Públicas y el Open Contracting Data Standard; revisión y participación en la elaboración de los documentos científicos.

10. Referencias Bibliográficas

Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenková, B., Schubert, E., Assent, I., & Houle, M. E. (2016). *On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study*. *Data Mining and Knowledge Discovery, 30*(4), 891-927. https://doi.org/10.1007/s10618-015-0444-8

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. CRISP-DM Consortium. https://www.crisp-dm.org/

Congreso de la Nación Paraguay. (2014). *Ley Nº 5282 Libre acceso ciudadano a la información pública y transparencia gubernamental*. https://www.bacn.gov.py/leyes-paraguayas/3013/ley-n-5282--libre-acceso-ciudadano-a-la-informacin-pblica-y-transparencia-gubernamental

Da Alesandro, R. (2019). *Investigation of anomalies in a RTC system using Machine Learning* (Master's thesis, Umeå University). Umeå University Publications. https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-164768

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2017). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, 74, 406-421. https://doi.org/10.1016/j.patcog.2017.09.037

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. En F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning* (pp. 3-33). Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_1

Gómez Scifo, J. D. (2023). *DNCP bate récord en gestión y control de procesos de licitación en los primeros 80 días de gobierno*. Dirección Nacional de Contrataciones Públicas. https://www.contrataciones.gov.py/dncp/dncp-bate-record-en-gestion-y-control-de-procesos-de-licitacion-en-los-primeros-80-dias-de-gobierno/

Janssens, J. H. M. (2013). *Outlier selection and one-class classification*. Wöhrmann Print Service.

Kiran, M., Wang, C., Papadimitriou, G., Mandal, A., & Deelman, E. (2020). Detecting anomalous packets in network transfers: Investigations using PCA, autoencoder and isolation forest in TCP. *Machine Learning, 109*, 1127-1143. https://doi.org/10.1007/s10994-020-05870-y

Komer, B., Bergstra, J., & Eliasmith, C. (2019). Hyperopt-sklearn. En F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning* (pp. 97-111). Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_5

López San Martín, M., Núñez Benitez, D. R., Paciello Coronel, J. M., & Pane Fernandez, J. I. (2024). *Quantifying the risk of complaints in public procurement tenders in Paraguay using machine learning*. 164-169. https://doi.org/10.54808/IMCIC2024.01.164

McKinney, J. (2023). *test_fictional_example.py* [Archivo de código fuente]. GitHub. https://github.com/open-contracting/sample-data/blob/main/tests/test_fictional_example.py

Mehta, S., Kothuri, P., & Garcia, D. L. (2018). *Anomaly detection for network connection logs* (arXiv:1812.01941). arXiv. https://doi.org/10.48550/arXiv.1812.01941

Niessen, M. E. K., Paciello, J. M., & Fernandez, J. I. P. (2020). Anomaly detection in public procurements using the open contracting data standard. 2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG), 127-134. https://doi.org/10.1109/ICEDEG48599.2020.9096674

Open Contracting Partnership. (s.f.). *¿Qué es el Estándar de Datos para las Contrataciones Abiertas (OCDS)?*. Open Contracting Data Standard. https://standard.open-contracting.org/latest/es/primer/what/

Open Contracting Partnership. (2021). *Calling for accountability: How Paraguay's open emergency procurement can help restore public trust*. Open Contracting. https://www.open-contracting.org/2021/05/03/calling-for-accountability-how-paraguays-open-emergency-procurement-can-help-restore-public-trust/

Transparency International. (2023). *Corruption Perceptions Index 2023: Paraguay.* Transparency International. https://www.transparency.org/en/cpi/2023/index/pry

Vierci Codas, M. B. (2018). *Análisis exploratorio de datos públicos categóricos usando agrupación*. https://gitlab.com/mbvierci/analisis-exploratorio-de-datos-publicos-categoricos-usando-agrupacion

Zenati, H., Romain, M., Foo, C. S., Lecouat, B., & Chandrasekhar, V. R. (2018). *Adversarially learned anomaly detection*. arXiv. https://doi.org/10.48550/arXiv.1812.02288

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PyOD: A Python toolbox for scalable outlier detection. *Journal of Machine Learning Research,20*(96),1-7. http://jmlr.org/papers/v20/19-011.html