Maestría en

Ciencias de Datos y Máquinas de Aprendizaje con Mención en Inteligencia

Trabajo previo a la obtención del título de Magíster en Ciencias de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

AUTORES:

David Paúl Guerrero Rambay

Galo Ernesto Guerra Velasco

Mauricio Rodrigo Miranda Jaramillo

Ramiro Fernando Mosquera Jordán

TUTORES:

Karla Estefanía Mora Cajas

Bryan Steven Vinueza Bustamante

TEMA:

Detección de Patrones Atípicos en la Contratación Pública Ecuatoriana del Sector Salud

Quito - Ecuador

Julio – 2026

REINVENTEMOS EL FUTURO

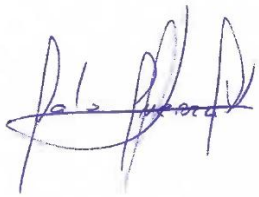
Certificación de autoría

Nosotros, David Paúl Guerrero Rambay, Galo Ernesto Guerra Velasco, Mauricio Rodrigo Miranda Jaramillo y Ramiro Fernando Mosquera Jordán, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

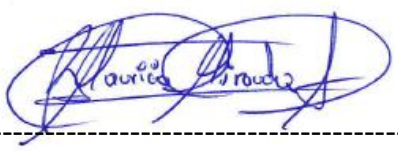
Firma

David Paúl Guerrero Rambay



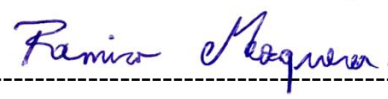
Firma

Galo Ernesto Guerra Velasco



Firma

Mauricio Rodrigo Miranda Jaramillo



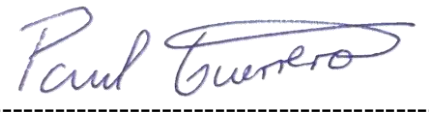
Firma

Ramiro Fernando Mosquera Jordán

Autorización de Derechos de Propiedad Intelectual

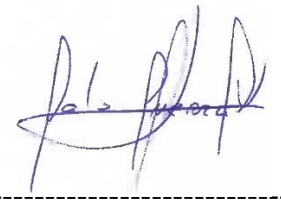
Nosotros, David Paúl Guerrero Rambay, Galo Ernesto Guerra Velasco, Mauricio Rodrigo Miranda Jaramillo y Ramiro Fernando Mosquera Jordán, en calidad de autores del trabajo de investigación titulado ***Detección de patrones atípicos en contratación pública ecuatoriana***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, JULIO 2026



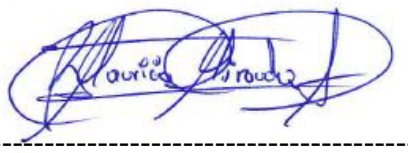
Firma

David Paúl Guerrero Rambay



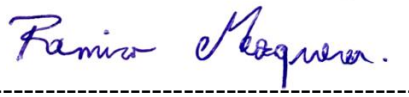
Firma

Galo Ernesto Guerra Velasco



Firma

Mauricio Rodrigo Miranda Jaramillo

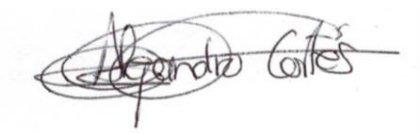


Firma

Ramiro Fernando Mosquera Jordán

Aprobación de dirección y coordinación del programa

Nosotros **Director Alejandro Cortés López de EIG y Coordinadora Karla Estefanía Mora Cajas de UIDE**, declaramos que **David Paúl Guerrero Rambay, Galo Ernesto Guerra Velasco, Mauricio Rodrigo Miranda Jaramillo y Ramiro Fernando Mosquera Jordán**, son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López
Director/a de la
Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial

Karla Estefanía Mora Cajas
Coordinador/a de la
Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial

DEDICATORIA

Este trabajo lo dedico a mi esposa Taty por su apoyo incondicional, a mis hijos Gabriel y Juan Diego por ser el motor que me impulsa todos los días; sin ustedes y su amor no lo hubiera logrado. A mis padres, Yolanda y Galo, que siempre me inculcaron la responsabilidad y me dejaron el más grande legado y apego al estudio. A Dios por iluminar mi camino.

Galo Ernesto Guerra Velasco

Dedico este trabajo a Dios, quien me ha dado fortaleza en momentos de dificultad y angustia; y ha puesto alegría en mi corazón. También dedico este trabajo a mi madre, quien me ha apoyado durante mi vida, incentivándome al estudio desde que era pequeño utilizando recompensas cuando tenía buenas calificaciones, motivándome a estar siempre entre los mejores.

Ramiro Fernando Mosquera Jordán

Dedico este trabajo a mis padres, gracias a su apoyo constante y comprensión durante este proceso académico, fueron fundamentales para alcanzar esta meta. A Dios por brindarme fortaleza y sabiduría. A mis compañeros quienes compartieron conocimientos y experiencias que enriquecieron mi crecimiento profesional.

Mauricio Rodrigo Miranda Jaramillo

Dedico este trabajo a mi esposo, por su amor y apoyo incondicional en cada etapa de este camino. A mi familia materna, por ser fuente constante de fortaleza y motivación. A mis compañeros Galo, Mauricio y Ramiro, por su compromiso, dedicación y por hacer de este proyecto un éxito.

David Paúl Guerrero Rambay

AGRADECIMIENTOS

Un agradecimiento eterno a mis profesores de colegio y universidad, en especial a mis tutores de matemáticas Graciela (Chelita) Espín, Tomás Andrade y Oswaldo Guerra por señalarme el camino para escoger mi profesión. A la UIDE y sus docentes por compartir sus conocimientos en esta Maestría.

Galo Ernesto Guerra Velasco

Doy gracias a Dios por esta oportunidad de superarme en la vida y adquirir nuevos conocimientos. También quiero agradecer a la UIDE, porque como funcionarios nos da la oportunidad de capacitarnos, tanto mediante becas y descuentos, así como con el apoyo de las jefaturas para nuestra capacitación.

Ramiro Fernando Mosquera Jordán

Agradezco profundamente a Dios por permitirme alcanzar esta meta. A mis padres, por su apoyo incondicional y motivación constante. Este logro es el resultado del esfuerzo, la dedicación y el apoyo de quienes estuvieron presentes en este importante camino.

Mauricio Rodrigo Miranda Jaramillo

Agradezco a la Universidad Internacional del Ecuador y a los docentes que he conocido durante la Maestría, cuyo acompañamiento y enseñanzas hicieron posible llevar este trabajo.

David Paúl Guerrero Rambay

RESUMEN

En Ecuador, al igual que en otros países, la contratación pública concentra un volumen significativo de recursos y enfrenta constantes riesgos de irregularidades y falta de transparencia en varios sectores, especialmente en el de la salud. Debido a que no existen etiquetas oficiales que permitan clasificar los contratos como regulares o irregulares, la presente investigación se ha orientado a la detección de patrones atípicos en los procesos contractuales mediante técnicas de aprendizaje automático no supervisado. Para el entrenamiento de los modelos se segmentó la información del periodo 2020–2025, enfocándose en bienes y servicios del sector salud y utilizando los datos abiertos disponibles en la Plataforma de Información Abierta de Contratación Pública del Servicio Nacional de Contratación Pública (SERCOP). A partir de variables como montos, fechas de publicación y firma, tipo de contrato, proveedor, entidad contratante y datos tributarios asociados al RUC del proveedor, se entrenaron modelos de detección de anomalías sobre más de un millón de contratos del sector salud, sin recurrir a técnicas de aumento de datos para evitar introducir sesgos artificiales. Se identificaron subconjuntos de procesos con comportamientos atípicos, generando insumos cuantitativos para la priorización de auditorías y el fortalecimiento de la transparencia en los procesos de contratación pública. Los resultados obtenidos evidencian que las técnicas de aprendizaje automático no supervisado permiten detectar patrones atípicos en procesos de contratación pública, incluso en escenarios donde no existen etiquetas previas para la clasificación de irregularidades.

Palabras Claves: contratación pública, sector salud, aprendizaje automático no supervisado, detección de anomalías, transparencia, Ecuador.

ABSTRACT

In Ecuador, as in many other countries, public procurement involves a significant volume of public resources and faces persistent risks related to irregularities and lack of transparency, particularly in the health sector. Since there are no official labels available to classify contracts as regular or irregular, this research focuses on the detection of atypical patterns in procurement processes through unsupervised machine learning techniques.

For model training, data from the 2020–2025 period were segmented, focusing on goods and services related to the health sector and using open data available through the Open Public Procurement Information Platform managed by the National Public Procurement Service (SERCOP). Based on variables such as contract amounts, publication and signing dates, contract type, supplier, contracting entity, and tax information associated with the supplier's tax identification number (RUC), anomaly detection models were trained using more than one million health-sector contracts. Data augmentation techniques were deliberately excluded to avoid introducing artificial biases.

The analysis identified subsets of procurement processes exhibiting atypical behavior, generating quantitative inputs for audit prioritization and supporting efforts to strengthen transparency in public procurement. The results demonstrate that unsupervised machine learning techniques can effectively detect atypical patterns in public procurement processes, even in scenarios where no prior labels exist for the classification of irregularities.

Keywords: public procurement, health sector, unsupervised machine learning, anomaly detection, transparency, Ecuador.

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

CAPITULO 1:

1. INTRODUCCIÓN

La presente investigación está orientada a la detección de patrones atípicos en los procesos de contratación pública del sector salud en Ecuador mediante la aplicación de técnicas de aprendizaje automático no supervisado. En el caso del Ecuador, el Presupuesto General del Estado contempla el gasto o la inversión en los diferentes sectores a través de las instituciones públicas, las mismas que reciben por parte del Estado los recursos financieros necesarios para contratar los servicios o productos que les permitirán ejecutar sus proyectos, y las compras públicas a nivel de cada país representan un valor de gran importancia para las finanzas estatales. Los algoritmos que se usarán para el análisis de la información ayudarán con la detección de valores atípicos permitiendo priorizar auditorías o revisiones sobre aquellos procesos con características anómalas por parte de las entidades competentes. Debido a la adjudicación irregular de contratos, se vuelve pertinente y necesaria la vigilancia, la transparencia, el control y toda acción orientada a clarificar los procesos inmersos en esta dinámica contractual.

Por otro lado, a nivel nacional se han evidenciado, a través de los medios de comunicación, diversos casos de corrupción relacionados con procesos de contratación pública, por ejemplo:

- Irregularidades en la compra de pruebas Covid-19 en Quito. (Primicias, 2022)
- Se observan 202 contratos suscritos en la pandemia por anomalías en Ecuador. (El Comercio, 2020)
- Irregularidades en 54 contratos de insumos y dispositivos médicos (Primicias, 2024, junio 26).
- Compra de bolsas para cadáveres con sobreprecio (Carmen, M, 2020)
- En 2020, Contraloría reveló irregularidades en contrataciones de emergencia. (Contraloría General del Estado, 2020, 31 de diciembre).

Lamentablemente no existen cifras oficiales centralizadas que muestren el problema de la corrupción en general y más específica respecto a contratación pública en Ecuador. Como se ha citado en los ejemplos que anteceden se encuentran artículos de medios de comunicación, así como informes de gestión de entidades de control; pero no hay un portal específico que consolide todos estos casos. Para dar un contexto del problema y como se evalúa al país desde afuera, se recurre a fuentes como Transparencia Internacional quienes calculan el índice de percepción de la corrupción constituyéndose en la clasificación global de corrupción más utilizada en el mundo. Mide el grado de corrupción percibido en el sector público de cada país, según expertos y empresarios. La puntuación para cada país es una combinación de al menos tres fuentes de datos que son recopiladas por diversas instituciones como: Banco Mundial y el Foro Económico Mundial entre otras, las mismas que se extraen de trece encuestas diferentes y evaluaciones. La escala de puntuación del CPI se mide de 0 a 100, siendo 0 equivalente a alta corrupción y 100 muy baja corrupción.

Tabla 1

Transparencia Internacional, índice de percepción de corrupción para Ecuador

Año	CPI (corruption perception index o índice de percepción de corrupción)	Ranking o clasificación
2025	**33**	**116/182**
2024	**32**	**121/180**
2023	**34**	**115/180**
2022	**36**	**101/180**
2021	**36**	**105/180**
2020	**39**	**92/180**

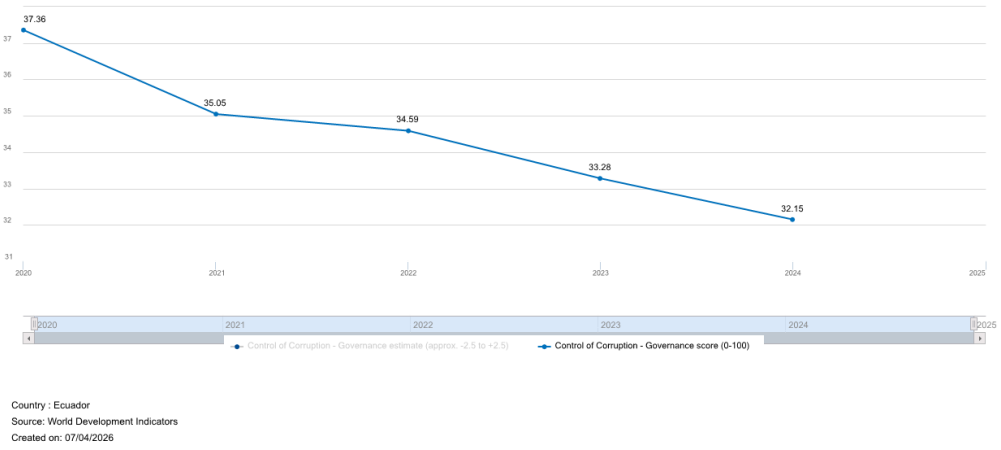
Nota. Adaptado de *Corruption Perceptions Index 2025: Ecuador*, por Transparency International, 2026, y *How CPI scores are calculated*, por Transparency International, s. f.

La clasificación de un país es relativa ya que depende del número de países incluidos en el cálculo del índice, mientras que la puntuación indica el nivel de corrupción de ese país. En la Tabla 1 se aprecia la puntuación del CPI y clasificación de Ecuador entre los años 2020 al 2025.

Otra fuente es el Banco Mundial, en donde se encuentra información sobre varios indicadores, uno de ellos es el CoC (Control of Corruption) el mismo que forma parte de los Indicadores Mundiales de Gobernanza (WGI), de igual manera es uno de los termómetros más usados a nivel global para medir el fenómeno de la corrupción. En la Figura 1 se muestra la evaluación de las percepciones sobre el grado en el que el poder público se ejerce para obtener beneficios privados. La escala del score también se mide de 0 a 100, siendo 0 el equivalente a la nula capacidad de un país para controlar la corrupción y 100 significa el mejor desempeño en el control de la corrupción. Los datos obtenidos comprenden el rango de años entre el 2020 al 2025, en donde se evidencia que Ecuador ha experimentado un retroceso paulatino en la capacidad percibida para mantener a raya la corrupción en el sector público. (World Bank, 2026).

Figura 1

Score de gobernanza, control de corrupción en Ecuador



Nota. Datos obtenidos del conjunto de datos *Worldwide Governance Indicators*, elaborado por el Banco Mundial (World Bank, 2026).

Los gobiernos de turno han intentado ejercer controles más estrictos ya sea vía leyes o mejorando la plataforma de compras públicas; sin embargo, ante el volumen que anualmente se gestionan en compras públicas los esfuerzos resultan insuficientes. En la Tabla 2 se puede dimensionar las cifras de compras públicas en millones de dólares para el rango de análisis entre los años 2020 al 2025. (Servicio Nacional de Contratación Pública [SERCOP], 2026c). Debido a su considerable magnitud y a su importante representación porcentual tanto en el Producto Interno Bruto (PIB) como en el Presupuesto General del Estado (PGE), las compras públicas constituyen un motor económico crítico para el país. No obstante, debido precisamente a esta alta concentración y flujo de recursos estatales, este sector se consolida estructuralmente como un escenario de alta vulnerabilidad frente a riesgos y actos de corrupción.

Tabla 2

Compras Públicas Generales en Ecuador

Año	Monto de compra pública general (millones de USD)
2025	**7318,7**
2024	**7996,3**
2023	**6951,9**
2022	**7698,1**
2021	**5320,5**
2020	**5072,5**

Nota. Datos obtenidos de *Cifras del Sistema Oficial de Contratación Pública del Ecuador* (sección “Compra Pública”), elaboradas por el Servicio Nacional de Contratación Pública (2026).

1.1. Definición del Proyecto

Tema Macro: Detección de patrones atípicos en los procesos de contratación pública del sector salud en Ecuador.

Título Específico del Proyecto: Detección de Patrones Atípicos en la Contratación Pública del Sector Salud en Ecuador utilizando Técnicas de Machine Learning (2020–2025).

Naturaleza del Proyecto: Debido a que no existen etiquetas que confirmen si un contrato tiene corrupción o no lo tiene en la información disponible, o si el mismo es o no un contrato irregular, el proyecto se enfoca en la búsqueda de anomalías o valores atípicos.

Para este trabajo se utilizarán los datos abiertos que se encuentran disponibles dentro de la Plataforma de Información Abierta de Contratación Pública, desarrollada y administrada por el Servicio Nacional de Contratación Pública (SERCOP), la misma que fue construida mediante la cooperación de Open Contracting Partnership (OCP) y el Banco de Desarrollo (CAF). (SERCOP, 2026a).

En Ecuador se tiene a La Ley Orgánica del Sistema Nacional de Contratación Pública (LOSNCP), reformada en 2025 (Ley Orgánica del Sistema Nacional de Contratación Pública [LOSNCP], 2025), la misma que está compuesta por un marco normativo, principios, así como procedimientos. Todo este cuerpo legal busca regular la adquisición de bienes, obras y servicios por parte del Estado, de tal forma que los procesos de contratación sean transparentes, permitiendo la justa competencia entre los proveedores, con el objetivo de hacer uso eficiente de los recursos públicos.

Entre los datos disponibles no existe una variable objetivo que permita identificar, para cada proceso contractual, si corresponde o no a un caso de irregularidad o fraude. Esta ausencia de etiquetas imposibilita el entrenamiento de modelos de aprendizaje supervisado, ya que estos requieren un conjunto de ejemplos previamente clasificados para aprender una función de predicción que relacione las variables de entrada con una salida conocida. Debido a esta limitación, el problema se aborda mediante técnicas de aprendizaje no supervisado, cuyo objetivo no es predecir una clase previamente definida, sino descubrir estructuras ocultas y detectar

observaciones cuyo comportamiento difiere significativamente del patrón predominante en los datos.

Por lo tanto, a partir de las variables o características disponibles como: monto del contrato, fechas de publicación y firma de contrato, proveedor, la empresa contratante, el tipo de contrato, el RUC del proveedor, la fecha de inicio de operaciones en el SRI, y los datos relacionados a productos o servicios se pueden filtrar aquellos procesos relacionados al sector médico. Esto permitirá detectar anomalías en estos datos complejos y desbalanceados. Además, no se utilizará data augmentation para evitar introducir sesgos, ya que el objetivo no es construir etiquetas o indicadores que permitan medir de alguna manera el riesgo de los procesos. (Para las variables utilizadas en el modelo ver Apéndice C.5.)

Actualmente se ha trabajado sobre más de 16 GB de datos de los archivos en formato JSON, los mismos que han sido subidos a una base de datos en PostgreSQL, se tiene más de 2 millones y medio de registros de contratos desde el año 2015 hasta el 2025, y para el rango de tiempo que cubre desde el 2020 hasta el 2025 se cuenta con más de 1 millón de contratos; este volumen de información garantiza fortaleza para hacer análisis estadísticos y usarlos en modelos no supervisados.

1.2. Justificación e Importancia del Trabajo de Investigación

En Ecuador se han detectado varios problemas de corrupción durante varios años, como los mencionados en la introducción, y a pesar de los reglamentos creados por los gobiernos de turno, los recursos estatales no han sido apropiadamente invertidos o gastados en beneficio de todos los ciudadanos. El trabajo actual plantea encontrar patrones o métricas inusuales en la contratación pública, lo cual servirá para que las autoridades de control tomen las medidas que consideren necesarias. En general, a nivel mundial es una línea de investigación que tiene mucha relevancia, abordando aspectos legales, económicos y éticos. Desde el punto de vista académico se busca explotar los datos a los que se tiene acceso, y aplicar varias de las técnicas aprendidas

durante el estudio de esta Maestría en Ciencia de Datos e Inteligencia Artificial; realizar análisis con algunos modelos de machine learning (ML) y contrastar los resultados o técnicas como las descritas por Torres-Berru y López Batista (2021) en donde proponen usar conceptos de Data Mining y ML para identificar anomalías en parámetros de calificación en procesos de contratación pública, se demuestra que los modelos no supervisados permiten detectar patrones atípicos. Para el trabajo actual se obtiene la información de los datos abiertos que se encuentran disponibles dentro de la **Plataforma de Información Abierta de Contratación Pública**, la misma que fue construida mediante la cooperación de Open Contracting Partnership (OCP) y el Banco de Desarrollo (CAF), bajo el marco del Compromiso de Gobierno Abierto de SERCOP que forma parte del primer Plan de Acción de Gobierno Abierto 2019-2022. (SERCOP, 2026a).

1.3. Objetivos

1.3.1. Objetivo General. Utilizar algoritmos de machine learning (ML) para analizar la información de contratación pública proveniente de entidades como el SERCOP y el SRI, disponible en sus plataformas de datos abiertos, con el propósito de detectar anomalías que contribuyan a la construcción de un puntaje (scoring) de riesgo, entendido como un índice cuantitativo que representa el nivel de riesgo asociado a cada proceso contractual del sector salud en Ecuador, utilizando información correspondiente al período 2020–2025.

1.3.2. Objetivos Específicos. Los objetivos específicos son:

1. Integrar y preparar la información proveniente de las bases de datos abiertas del SERCOP y del SRI, correspondiente al período de estudio, en un repositorio unificado de datos sobre contratación pública enfocada al sector de la salud.
2. Diseñar y construir un modelo de *scoring* de riesgo, basado en técnicas de detección de anomalías aplicadas a procesos de contratación pública. Para esto se utilizarán los siguientes modelos de machine learning: Isolation Forest, HDBSCAN y Local Outlier Factor.

3. Detectar procesos contractuales que presenten comportamientos atípicos mediante el análisis conjunto de variables económicas, temporales y relacionales, derivadas del repositorio de datos integrado.
4. Evaluar y comparar el desempeño de los algoritmos de detección de anomalías mediante métricas de validación adecuadas para aprendizaje no supervisado, considerando indicadores de calidad del agrupamiento (Silhouette Score e índice de Davies–Bouldin), capacidad de separación entre observaciones normales y anómalas, y concordancia entre modelos mediante el coeficiente de Cohen (κ).
5. Formular recomendaciones para la utilización de herramientas de *machine learning* en la gestión y supervisión de la contratación pública, a partir de los resultados obtenidos y de los patrones de riesgo identificados.

1.3.3. Actividades Metodológicas. Se realizaron las siguientes actividades:

1. Descargar y organizar los archivos en formato JSON provenientes de las bases de datos abiertos del SERCOP y del SRI, correspondientes al período de estudio.
2. Procesar y transformar los archivos descargados mediante DuckDB, a partir de archivos Parquet, con el fin de optimizar el procesamiento inicial de datos, y almacenar la información resultante en una base de datos PostgreSQL, donde se realizan las operaciones de limpieza, normalización y unión necesarias para su posterior análisis.
3. Seleccionar las variables que se emplearán en la detección de anomalías.
4. Implementar y parametrizar los algoritmos Isolation Forest, HDBSCAN y Local Outlier Factor sobre el conjunto de datos preparado, generando para cada proceso de contratación un puntaje de riesgo asociado.
5. Comparar el desempeño de los algoritmos de detección de anomalías mediante métricas cuantitativas y visualizaciones, evaluando su capacidad para discriminar patrones inusuales en los datos de contratación.

6. Sistematizar los hallazgos en términos de riesgos, patrones y casos destacados, haciendo énfasis en aquellos de mayor interés potencial para la supervisión y el control de la contratación pública.

1.4. Delimitación del Proyecto

1.4.1. Límites Geográficos, de Temporalidad y Sectoriales. Se establecieron los siguientes límites:

Geográficos. Las contrataciones públicas que se gestionan en el SOCE son a nivel país. No existe restricción alguna en la Plataforma de Datos Abiertos para segmentar por provincias o regiones, en tal sentido, se cuenta con acceso a datos a nivel de todo el Ecuador.

Temporalidad Analizada. El análisis se centra en los procesos contractuales entre los años 2020 y 2025, considerando que dentro del período de la pandemia Covid-19 (años 2020 a 2023 aprox.) se registraron eventos que fueron noticia a nivel nacional e internacional en lo referente a contratos por insumos médicos.

Sectores. El análisis se circunscribe específicamente a los contratos del sector salud, para ello se usa la clasificación de productos CPC (SERCOP, 2026e), misma que se encuentra en la información de cada proceso.

1.4.2. Límites Metodológicos y de Datos. Para los límites metodológicos y de datos se definen:

Fuentes de datos: Se usarán los archivos en formato JSON que el Servicio Nacional de Contratación Pública tiene disponible en su plataforma web (SERCOP, 2026c). También se usan datasets con información de RUCs por provincia en formato CSV que el SRI publica en su sitio web (Servicio de Rentas Internas [SRI], s. f.).

Variables: Se utilizarán características extraíbles de bases de datos públicas como: montos de contrato, método de selección, número de oferentes, RUC, fecha de actividad económica del proveedor adjudicado, entre otros.

Exclusiones: No se incluirán datos confidenciales de investigaciones en curso, identificadores personales no públicos, o información clasificada. No se realizarán investigaciones forenses o análisis de documentos privados o archivos PDFs que suelen acompañar y adjuntar en cada proceso en el sistema SOCE. Actualmente la Plataforma de Datos Abiertos no tiene datos o información de este tipo de documentación por cada proceso, por lo que se descarta cualquier tipo de análisis referente a esta materia.

Modelos comparados: Se entrenarán al menos 3 algoritmos de ML, y se consolidarán las métricas de rendimiento de cada uno de ellos, para entregar un resumen que permita evidenciar cual modelo es el mejor con este tipo de datos.

CAPITULO 2:

2. REVISIÓN DE LITERATURA

2.1. Estado del Arte

Se revisaron algunos trabajos de investigación en donde los esfuerzos se han centrado en la temática relacionada con este proyecto de titulación, es decir, encontrar anomalías atípicas en contratación pública. Todos estos trabajos evidencian que es posible aplicar técnicas de análisis de datos, así como modelos de aprendizaje de máquina (ML) no supervisados con el objetivo de hallar un conjunto de características o patrones que lleven a sospechar, por lo menos, de que se tratan de contratos que tienen información que debe ser investigada o contrastada, ya que no siguen las pautas normales de la mayoría de los procesos contractuales.

Carrillo Bustos y Fernández Fernández (2025) elaboraron un estudio orientado a evaluar el potencial analítico de los datos abiertos publicados por el SERCOP entre los años 2015 al 2025, recordando que es la misma fuente se ha usado en el proyecto. Usaron la información proveniente de las etapas de: licitación, adjudicación, proveedores y contratos mediante el identificador OCID. En su trabajo aplicaron procesos de limpieza, transformación de variables, procesamiento de lenguaje natural (NLP) sobre las descripciones de los procesos y técnicas de reducción de

dimensionalidad PCA y análisis factorial; todo ello con el objetivo de ejecutar algoritmos de agrupamiento no supervisado como: K-Means, GMM, DBSCAN y clúster jerárquico. Sus resultados permiten identificar perfiles de contratación pública incluyendo contratos con: baja competencia, de larga duración, montos altos, pocos oferentes, así como agrupaciones consideradas como normales. Los autores de este trabajo concluyen que los métodos no supervisados constituyen una alternativa importante cuando no se cuenta con etiquetas previas que permitan entrenar modelos supervisados, ya que aportan con la identificación de patrones ocultos, así como la construcción de indicadores de riesgo, los mismos que pueden servir para apoyar procesos de fiscalización.

Torres Berrú (2023) elaboró una tesis doctoral orientada en la detección automática de fraude y corrupción en contratos públicos en Ecuador mediante técnicas de minería de datos y de texto. Su trabajo propone una metodología integral para analizar los datos estructurados de contratación pública, así como información textual que se generan durante los procesos contractuales, para ello usó técnicas de aprendizaje automático y procesamiento de lenguaje natural (NLP). La fuente de datos fue obtenida por medio de web scraping sobre el portal SERCOP, específicamente sobre el sistema SOCE. Sus resultados muestran que la combinación de minería de datos, aprendizaje automático y análisis textual permite identificar patrones asociados a posibles riesgos de corrupción, alcanzando indicadores significativos en la detección de comportamientos atípicos. La investigación concluye que entre el 30% y el 35% de los procesos analizados presentaban indicios de posibles irregularidades, con lo que se demuestra el potencial de las técnicas de inteligencia artificial para fortalecer los mecanismos de transparencia, fiscalización y control en las compras públicas. Asimismo, destaca la importancia de utilizar enfoques supervisados y no supervisados para descubrir estructuras ocultas en grandes volúmenes de información gubernamental.

Araujo et al. (2024) desarrollaron una investigación orientada a la detección de propuestas anómalas en procesos de contratación pública del estado de Paraíba, Brasil, utilizando técnicas de aprendizaje automático no supervisado. Su trabajo parte de la premisa de que determinadas características de las ofertas pueden revelar posibles irregularidades asociadas con sobreprecios, colusión, direccionamiento de contratos o conflictos de interés. Para ello, los autores analizaron datos históricos de licitaciones públicas entre 2014 y 2023, integrando información proveniente de organismos gubernamentales brasileños y construyendo variables relacionadas con distancia geográfica, participación individual de proveedores, coocurrencia entre empresas participantes, antigüedad empresarial y diferencias entre montos ofertados y presupuestos estimados. Usaron tres algoritmos de detección de anomalías: Isolation Forest, Local Outlier Factor (LOF) y One-Class Support Vector Machine (One-Class SVM), comparando su capacidad para identificar contratos atípicos. Los resultados mostraron que One-Class SVM obtuvo el mejor desempeño con un ROC-AUC de 0.821, seguido por Isolation Forest con 0.769, mientras que LOF presentó un rendimiento considerablemente inferior con 0.406. Es importante señalar que dicho estudio empleó la métrica ROC-AUC debido a la disponibilidad de una variable objetivo que permitió contrastar las predicciones del modelo con una clasificación conocida. En contraste, la presente investigación no dispone de etiquetas que identifiquen previamente procesos contractuales regulares o irregulares, por lo que la evaluación se realizó mediante métricas apropiadas para aprendizaje no supervisado. Los autores concluyen que los algoritmos de aprendizaje automático constituyen herramientas efectivas para apoyar auditorías y fortalecer los mecanismos de transparencia y control en la contratación pública.

El estudio de García Rodríguez et al. (2022) evalúa la capacidad de distintos algoritmos de aprendizaje automático para detectar colusión en procesos de contratación pública. La investigación utiliza seis bases de datos reales provenientes de Brasil, Italia, Japón, Suiza y Estados Unidos, que contienen información de más de 9.700 licitaciones y más de 64.000 ofertas

históricas, incluyendo casos previamente identificados como colusivos por organismos de control y tribunales de justicia. En su trabajo evalúa once algoritmos de machine learning, incluyendo Random Forest, Extra Trees, AdaBoost, Gradient Boosting, SVM, K-Nearest Neighbors, Redes Neuronales y Naive Bayes, evaluando su desempeño bajo distintos escenarios de disponibilidad de información. Los resultados demuestran que los métodos Ensemble obtienen el mejor desempeño para detectar comportamientos anómalos asociados a colusión, alcanzando niveles de precisión superiores al 80% y reduciendo significativamente los errores de clasificación cuando se complementan con variables estadísticas derivadas de las ofertas recibidas.

También se resalta las iniciativas ejecutadas por la sociedad civil en Ecuador, en donde se puede ver el aporte de organismos como la Academia y sectores de la ciudadanía para construir plataformas tecnológicas en donde se muestran banderas o indicadores que permiten clasificar o ponderar a los contratos en escalas de riesgo. La metodología detrás de estos proyectos mayoritariamente se realiza a nivel de base de datos, en donde se plasma el marco teórico o ficha del indicador el mismo que permite de alguna manera etiquetar a los contratos. Pero no existe a nivel oficial del ente rector el SERCOP (Servicio de Contratación Pública en el Ecuador) u otro organismo gubernamental una fuente de datos oficial en donde se haya etiquetado a los procesos contractuales como anómalos. Por lo tanto, los esfuerzos de identificar a aquellos contratos con anomalías todavía no han sido tomados en cuenta por las entidades de gobierno encargadas de la administración y gestión de la ejecución de los contratos a nivel estatal. Entre los proyectos o iniciativas ecuatorianas orientadas a la transparencia y análisis de contratación pública, se citan los siguientes:

- **Proyecto KAPAK:** Impulsado por la Universidad San Francisco de Quito (USFQ) con el apoyo de la Cooperación Alemana para el Desarrollo (GIZ). Esta plataforma fue concebida con el objetivo de fortalecer la transparencia y facilitar el análisis de los procesos de contratación pública mediante herramientas de visualización, análisis de datos y monitoreo ciudadano. La

plataforma consolida información obtenida mediante técnicas de web scraping desde el Sistema Oficial de Contratación Pública del Ecuador (SOCE), y la transforma en indicadores que permiten identificar patrones de comportamiento de entidades contratantes, proveedores y procesos de adquisición. Su metodología incorpora procesos de extracción, depuración y estructuración de información provenientes de los registros públicos de contratación, facilitando la exploración de tendencias y posibles alertas de riesgo (Universidad San Francisco de Quito & GIZ, 2023).

- **Proyecto Banderas Rojas:** esta iniciativa es el sistema de Banderas Rojas desarrollado por el proyecto Contratos Transparentes Ecuador. Este enfoque se basa en la identificación de indicadores de riesgo previamente definidos que permiten alertar sobre posibles irregularidades en los procesos de contratación pública. La metodología implementada utiliza reglas de negocio y criterios derivados de estándares internacionales de transparencia para identificar situaciones potencialmente riesgosas, tales como la concentración de adjudicaciones en determinados proveedores, modificaciones contractuales inusuales, reducida concurrencia de oferentes o patrones atípicos en los procesos de compra. Cada bandera constituye una señal de alerta que orienta el trabajo de auditoría y control, sin que ello implique necesariamente la existencia de actos de corrupción o incumplimientos normativos (Fundación Ciudadanía y Desarrollo, 2025).
- **Plataforma de Datos Abiertos de Contratación Pública del SERCOP:** Esta es utilizada como principal fuente de información para este proyecto. La iniciativa forma parte de la adopción del estándar internacional Open Contracting Data Standard (OCDS), promovido por la Open Contracting Partnership para mejorar la transparencia, accesibilidad y reutilización de la información contractual. La plataforma pone a disposición conjuntos de datos históricos relacionados con planificación, procedimientos de contratación, adjudicaciones, contratos y ejecución contractual, estructurados bajo estándares abiertos que facilitan su explotación mediante herramientas de análisis de datos, inteligencia de negocios y aprendizaje automático (SERCOP, 2026b).

En resumen, los antecedentes científicos revisados, junto con las iniciativas de transparencia y apertura de datos desarrolladas tanto a nivel nacional como internacional, constituyen un soporte importante para el proyecto. Estas experiencias no solo validan la eficacia del uso de técnicas de aprendizaje automático para el análisis de la contratación pública, sino que también muestran las oportunidades y desafíos existentes en la detección temprana de riesgos contractuales. De manera particular, la disponibilidad de datos abiertos, la adopción de estándares como OCDS y los avances alcanzados por la academia y la sociedad civil han proporcionado los insumos necesarios para desarrollar una propuesta basada en modelos de detección de anomalías, orientada a contribuir al fortalecimiento de la transparencia, la eficiencia y la rendición de cuentas en los procesos de contratación pública ecuatoriana.

2.2. Marco Teórico

La detección de anomalías constituye una rama del aprendizaje automático orientada a identificar observaciones que presentan comportamientos significativamente diferentes respecto al patrón predominante de un conjunto de datos o datasets. Estas observaciones, conocidas como anomalías, outliers o valores atípicos, pueden corresponder a errores de medición, eventos poco frecuentes, comportamientos inusuales o situaciones potencialmente riesgosas que merecen su respectivo análisis adicional. En escenarios como la contratación pública, donde generalmente no existen etiquetas que permitan identificar de manera explícita procesos normales o irregulares, la detección de anomalías se convierte en una herramienta especialmente útil para identificar contratos cuyas características difieren significativamente del comportamiento histórico observado.

Según Chandola et al. (2009), una anomalía puede definirse como una observación que se desvía de manera considerable respecto al comportamiento esperado de la mayoría de los datos, de tal forma que genera sospechas acerca de que fue producida por un mecanismo diferente al resto de observaciones.

Las anomalías son de carácter puntual cuando presentan valores significativamente diferentes respecto al resto de los datos. También las anomalías pueden ser normales en un contexto y anómalas en otro, en la Ecuación 1 se muestra una observación individual.

Ecuación 1

Pertenencia de una observación a la distribución normal

$$x_i \notin D_{normal}$$

Donde:

- x_i: observación individual
- D_{normal}: distribución predominante de los datos

Las anomalías colectivas corresponden a grupos de observaciones que, analizadas individualmente, parecen normales, pero que en conjunto muestran un comportamiento atípico. Entre los métodos para detectar outliers se cita a los siguientes:

- **Métodos estadísticos**: Asumen que los datos siguen una distribución probabilística conocida, como Z-score:

Ecuación 2

Z-Score

$$Z_{i=}\frac{x_{i-\mu}}{\sigma}$$

Donde:

- μ: media
- σ: desviación estándar

  Generalmente: $|z_i| > 3$ se considera un posible outlier.

- **Métodos basados en distancias**: Una observación es anómala si se encuentra lejos de sus vecinos más cercanos, como Distancia Euclidiana:

Ecuación 3

Distancia Euclidiana

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- **Métodos basados en densidad**: Evalúan la densidad local alrededor de una observación. Una observación ubicada en una región con baja densidad respecto a sus vecinos puede considerarse anómala. El algoritmo Local Outlier Factor (LOF) pertenece a esta categoría.
- **Métodos basados en aislamiento**: Suponen que las anomalías son más fáciles de aislar que las observaciones normales. El algoritmo Isolation Forest pertenece a esta categoría.

2.2.1. Isolation Forest. Isolation Forest fue propuesto por Liu et al. (2008) y constituye uno de los algoritmos más utilizados para detección de anomalías en conjuntos de datos de alta dimensionalidad. La idea fundamental consiste en que las observaciones anómalas requieren menos particiones aleatorias para ser aisladas. Es un potente algoritmo de detección de anomalías que utiliza árboles de aislamiento (isolation trees) para identificar anomalías en los datos. Este algoritmo aísla las anomalías mediante la partición aleatoria del espacio de datos y la medición de la profundidad de aislamiento de cada punto. Sus funciones se describen a continuación: (Barua et al., 2024, p. 352).

1. Seleccionar aleatoriamente características y dividir los puntos dentro de sus rangos.
2. Dividir los datos en dos ramas según la división elegida.
3. Repetir los pasos 1 y 2 recursivamente hasta alcanzar el aislamiento o una profundidad máxima.
4. Cada punto de datos tiene una longitud de ruta que representa el número de divisiones necesarias para aislarlo.
5. Calcula el promedio de las longitudes de ruta de un punto de datos en todos los árboles de

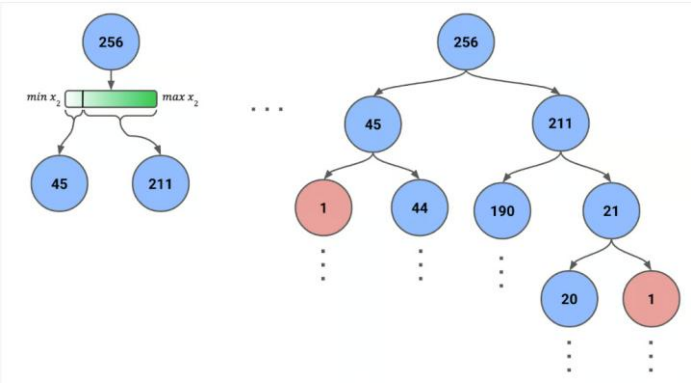
aislamiento del bosque.

6. Los puntos con longitudes de ruta más cortas (más fáciles de aislar) se consideran más anómalos.

A diferencia de los árboles de decisión tradicionales, cuyo objetivo consiste en maximizar la separación entre clases mediante criterios como la ganancia de información o el índice de Gini, Isolation Forest realiza las particiones completamente al azar. La capacidad del algoritmo para detectar anomalías surge precisamente del hecho de que las observaciones poco frecuentes tienden a quedar aisladas en menos particiones que aquellas pertenecientes a regiones densamente pobladas.

Figura 2

Representación gráfica del algoritmo Isolation Forest



Nota. Reproducido de "Guía del Bosque del Aislamiento: Explicación e implementación en Python", por C. O'Sullivan, 2024, DataCamp.

La Figura 2 muestra el principio de funcionamiento de Isolation Forest mediante un árbol de aislamiento. El nodo superior representa el conjunto completo de observaciones, mientras que en cada nivel del árbol se selecciona aleatoriamente una variable y un punto de corte para dividir el conjunto de datos en dos subconjuntos. Este proceso continúa de forma recursiva hasta que cada observación queda completamente aislada en un nodo hoja. Las observaciones representadas en color rojo corresponden a puntos que fueron aislados tras pocas particiones, lo que indica una

longitud de camino corta y, por tanto, una mayor probabilidad de ser consideradas anomalías. En contraste, las observaciones ubicadas en regiones densas del espacio de datos requieren un mayor número de divisiones para quedar aisladas, generando longitudes de camino más largas y siendo interpretadas como observaciones normales. Los números mostrados dentro de cada nodo representan la cantidad de observaciones que permanecen agrupadas después de cada partición. A medida que el árbol se profundiza, el número de registros disminuye hasta que una observación queda aislada en un nodo hoja. Cuando dicho aislamiento ocurre tras pocas divisiones, el algoritmo interpreta que la observación presenta características significativamente diferentes del resto de los datos. El proceso de aislamiento mostrado es para un único árbol; sin embargo, el algoritmo construye un bosque compuesto por múltiples árboles de aislamiento. El puntaje final asignado a cada observación se obtiene a partir de la longitud promedio del camino recorrida en todos los árboles, lo que reduce la variabilidad asociada a las particiones aleatorias y proporciona una estimación más estable del grado de anomalía.

Cada árbol selecciona aleatoriamente una variable y un punto de corte hasta aislar completamente las observaciones. La longitud esperada de camino representa la longitud del camino requerido para aislar una observación.

Ecuación 4

Longitud esperada del camino

$$h(x)$$

La longitud promedio del camino de aislamiento observada en la Figura 2 constituye la base para el cálculo del puntaje de anomalía. Cuanto menor sea la profundidad necesaria para aislar una observación dentro de los diferentes árboles del bosque, mayor será el valor del puntaje de anomalía asignado por el algoritmo. El score de anomalía se define como:

Ecuación 5

Score de una anomalía

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Donde:

- $E\big(h(x)\big)$: longitud promedio de aislamiento
- $c(n)$: factor de normalización

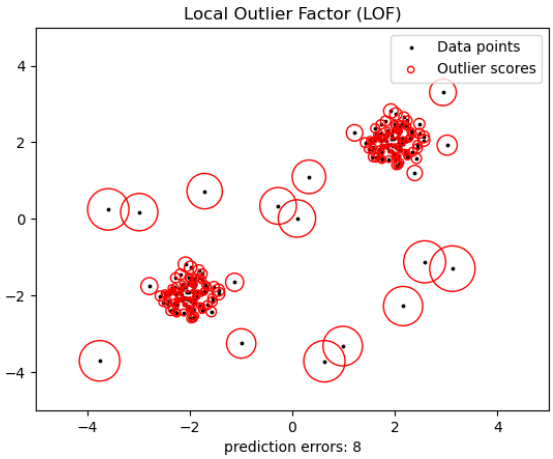
La interpretación de la anomalía cuando $s(x) \approx 1$ entonces hay una alta probabilidad de anomalía. Cuando $s(x) \approx 0.5$ el comportamiento es normal. El puntaje de anomalía toma valores entre 0 y 1. Valores cercanos a 1 indican que la observación fue aislada rápidamente en la mayoría de los árboles del bosque y, por tanto, presenta una alta probabilidad de corresponder a un comportamiento atípico. Valores cercanos a 0.5 representan observaciones cuyo comportamiento es consistente con el resto del conjunto de datos. En la práctica, el umbral de decisión depende del parámetro de contaminación (contamination) definido durante el entrenamiento del modelo.

Se escogió el algoritmo Isolation Forest porque identifica eficientemente observaciones globalmente aisladas mediante particiones aleatorias.

2.2.2. Local Outlier Factor (LOF). LOF fue propuesto por Breunig et al. (2000). Es otro potente algoritmo para la detección de anomalías, que utiliza el concepto de densidad local para identificar puntos de datos que se desvían de su entorno. Se utiliza para calcular la relación entre la densidad local de un punto y la densidad de sus vecinos.

Figura 3

Visualización de la detección de anomalías con Local Outlier Factor



Nota. Reproducido de "Outlier detection with Local Outlier Factor (LOF)", por Scikit-learn developers, 2026, Scikit-learn 1.9.0 documentation.

La Figura 3 ilustra el principio de funcionamiento del algoritmo Local Outlier Factor (LOF). Los puntos negros representan las observaciones del conjunto de datos, mientras que los círculos rojos indican el puntaje de anomalía asignado por el algoritmo. El tamaño del círculo es proporcional al valor del LOF: cuanto mayor es el círculo, mayor es la probabilidad de que la observación corresponda a un valor atípico. Puede observarse que los puntos ubicados dentro de agrupamientos densos presentan círculos pequeños, debido a que su densidad local es similar a la de sus vecinos. En cambio, las observaciones aisladas o localizadas en regiones de menor densidad muestran círculos de mayor tamaño, reflejando una diferencia significativa respecto a su entorno inmediato y siendo identificadas como posibles anomalías. A diferencia de los métodos basados únicamente en distancia, LOF evalúa la densidad relativa de cada observación respecto de sus vecinos más cercanos. Esto permite detectar anomalías locales que podrían pasar inadvertidas para otros algoritmos cuando los datos contienen regiones con diferentes niveles de concentración.

Valores altos de LOF indican posibles anomalías. LOF compara la densidad local de un punto de datos (su vecindario) con las densidades locales de sus vecinos. Se consideran anomalías

los puntos con una densidad local significativamente menor, lo que indica que se encuentran en regiones con baja densidad en comparación con sus vecinos. El mecanismo de funcionamiento y los pasos necesarios para su correcto funcionamiento se describen a continuación: (Barua et al., 2024, p. 352).

1. Definir el vecindario: Elija un parámetro “k” que represente el número de vecinos más cercanos (NN) a considerar para cada punto de datos.
2. Calcular la distancia de accesibilidad: Medir la distancia de accesibilidad desde cada punto a sus vecinos, teniendo en cuenta tanto la distancia entre ellos como la densidad de sus vecindarios.
3. Calcular la densidad local: Estimar la densidad local de cada punto invirtiendo la distancia de alcanzabilidad promedio de su kNN.
4. Calcular LOF: Dividir la densidad local de un punto por la densidad local promedio de sus kNN.
5. Identificación de anomalías: Los puntos con valores LOF significativamente más altos (superiores a un umbral predefinido) se consideran anomalías potenciales, ya que se encuentran en regiones menos densas en comparación con su entorno.

La distancia alcanzable (Reachability Distance) evita que pequeñas variaciones entre observaciones extremadamente cercanas produzcan estimaciones inestables de densidad. Para ello, considera la mayor distancia entre la separación real de dos observaciones y la distancia al k-ésimo vecino del punto de referencia, se define como:

Ecuación 6

Distancia Alcanzable

$$reach_{dist_k}(A, B) = \max(k\_distance(B), d(A, B)).$$

A partir de la distancia alcanzable se calcula la densidad local alcanzable (Local Reachability Density, LRD), que representa qué tan densamente rodeada se encuentra una

observación respecto de sus vecinos más próximos. Valores elevados indican regiones densas, mientras que valores bajos corresponden a observaciones más aisladas, la densidad local alcanzable se define como:

Ecuación 7

Densidad local alcanzable

$$LDR_k(A) = \left(\frac{\sum B \in N_k(A)^{reach\_dist_k(A,B)}}{|N_k\ (A)|} \right)^{-1}$$

Local Outlier Factor compara la densidad local de una observación con la densidad promedio de sus vecinos. Cuando ambas densidades son similares, el valor del LOF será cercano a uno; por el contrario, cuando la densidad de la observación es considerablemente menor que la de su vecindario, el valor del LOF aumenta indicando un posible comportamiento anómalo, se define como:

Ecuación 8

Valor calculado de LOF

$$LOF_k(A) = \frac{\sum B \in N_k(A) \frac{LDR_k\ (B)}{LDR_k\ (A)}}{|N_k\ (A)|}$$

La interpretación cuando $LOF \approx 1$ la observación presenta una densidad similar a la de sus vecinos y, por tanto, corresponde a un comportamiento esperado, cuando $LOF > 1$ la densidad local comienza a diferir de la de su entorno inmediato, indicando una posible anomalía y cuando $LOF \gg 1$ la observación presenta una densidad significativamente menor que la de sus vecinos, siendo considerada una fuerte candidata a anomalía, recordando que el símbolo >> en machine learning significa mucho mayor que uno.

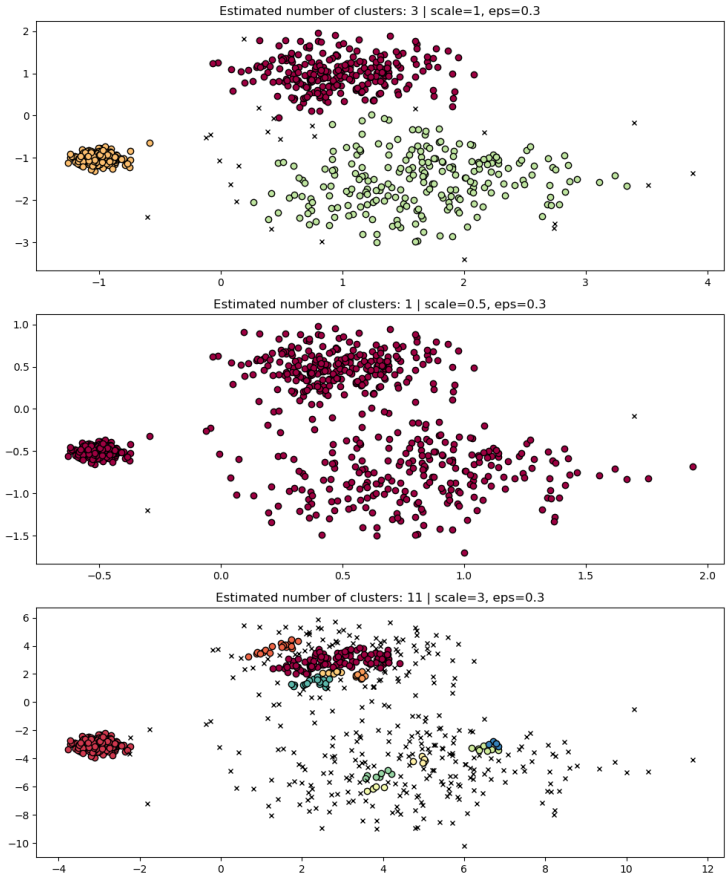
Se eligió Local Outlier Factor, LOF, porque permite detectar anomalías locales comparando la densidad de cada proceso contractual con la de su entorno inmediato.

2.2.3. HDBSCAN. Su siglas provienen del inglés *Hierarchical Density-Based Spatial Clustering of Applications with Noise*. Fue propuesto por Campello et al. (2013) y constituye una extensión de DBSCAN que permite detectar agrupaciones de densidad variable. Su principal ventaja consiste en que puede identificar grupos con diferentes niveles de densidad sin requerir un único valor fijo del parámetro ε (epsilon), limitación presente en DBSCAN. Para lograrlo, HDBSCAN construye una jerarquía de agrupamientos considerando múltiples niveles de densidad y posteriormente selecciona únicamente aquellos clústeres que demuestran mayor estabilidad a lo largo de dicha jerarquía. Las observaciones que no pertenecen a ningún grupo estable son clasificadas como ruido (noise), característica que convierte al algoritmo en una herramienta especialmente útil para la detección de anomalías. En el contexto de la presente investigación, esta característica resulta particularmente relevante debido a que los procesos de contratación pública no presentan una distribución homogénea. Existen modalidades de contratación, entidades y proveedores con comportamientos muy diferentes entre sí, generando agrupamientos de distinta densidad. En este escenario, HDBSCAN ofrece una ventaja frente a otros algoritmos de agrupamiento al adaptarse automáticamente a dichas variaciones y permitir identificar procesos que no pertenecen a ningún patrón estable, los cuales pueden representar comportamientos potencialmente atípicos.

La Figura 4 está dividida en tres ilustraciones, donde en la primera ilustración los datos presentan una escala homogénea, permitiendo que HDBSCAN identifique correctamente tres agrupamientos principales y clasifique como ruido únicamente aquellas observaciones aisladas. En la segunda ilustración, al reducir la escala de los datos, las distancias entre observaciones disminuyen y la densidad aparente aumenta, ocasionando que el algoritmo considere la mayoría de los puntos como un único agrupamiento. En la tercera ilustración, el incremento de la escala provoca una mayor dispersión entre las observaciones, permitiendo identificar múltiples agrupamientos de menor tamaño y un mayor número de puntos clasificados como ruido.

Figura 4

Invarianza de escala en el algoritmo de agrupamiento HDBSCAN



Nota. Reproducido de "Demo of HDBSCAN clustering algorithm", por Scikit-learn developers, 2026, Scikit-learn 1.9.0 documentation.

La distancia mutua de alcanzabilidad (Mutual Reachability Distance) redefine la distancia entre dos observaciones considerando no solamente la distancia euclidiana entre ellas, sino también la densidad local de cada punto. De esta manera, regiones con diferente densidad pueden compararse dentro de un mismo espacio métrico, permitiendo construir una representación jerárquica más estable de los datos:

Ecuación 9

Distancia mutua de alcanzabilidad

$$mreach_k(a,b) = \max\{core_k(a), core_k(b), d(a,b)\},$$

Donde:

- d(a,b): distancia entre los puntos.
- $core_k(a)$: distancia al k-ésimo vecino más cercano de a.
- $core_k(b)$: distancia al k-ésimo vecino más cercano de b.

La densidad local se estima de forma inversamente proporcional a la distancia del núcleo (core distance). En consecuencia, cuanto menor sea la distancia al vecino más cercano requerido para satisfacer el criterio de densidad, mayor será la densidad estimada alrededor de la observación, se define como:

Ecuación 10

Criterio de densidad

$$Density(x) \propto \frac{1}{core\_distance(x)}$$

En la detección del ruido los puntos que no pertenecen a ningún clúster estable son etiquetados como label = -1 y pueden interpretarse como anomalías.

Se eligió HDBSCAN porque detecta agrupamientos con diferentes niveles de densidad e identifica ruido de forma natural.

2.2.4. Ensemble Learning para Detección de Anomalías. Los métodos Ensemble combinan múltiples modelos para producir una decisión más robusta. La hipótesis es que cada algoritmo detecta diferentes tipos de anomalías. Pueden definirse por la siguiente fórmula, que fue la utilizada en este proyecto. Para otros escenarios puede variar dependiendo del número de algoritmos que se usen.

Ecuación 11

Score para el método de Ensemble utilizado con IF, LOF y HDBSCAN

$$Score_{Ensemble} = \frac{Score_{IF} + Score_{LOF} + Score_{HDBSCAN}}{3}$$

El objetivo de los métodos de conjunto o ensemble es combinar diferentes clasificadores en un metaclasificador que tenga un mejor rendimiento de generalización que cada clasificador individual. Estos métodos utilizan el principio de votación por mayoría. La votación por mayoría simplemente significa que se seleccionó la etiqueta de clase que ha sido predicha por la mayoría de los clasificadores, es decir, que ha recibido más del 50 por ciento de los votos. Estrictamente hablando, el término "votación por mayoría" se refiere únicamente a configuraciones de clases binarias. Sin embargo, es fácil generalizar el principio de votación por mayoría a configuraciones multiclase, lo que se conoce como votación por pluralidad. (Raschka et al., 2022, p. 205).

2.2.5. Métricas de Evaluación. Entre las métricas utilizadas se tienen a:

Silhouette Score. Mide qué tan bien separadas se encuentran las observaciones respecto a los clústeres identificados, viene dado por:

Ecuación 12

Silhouette Score

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Donde:

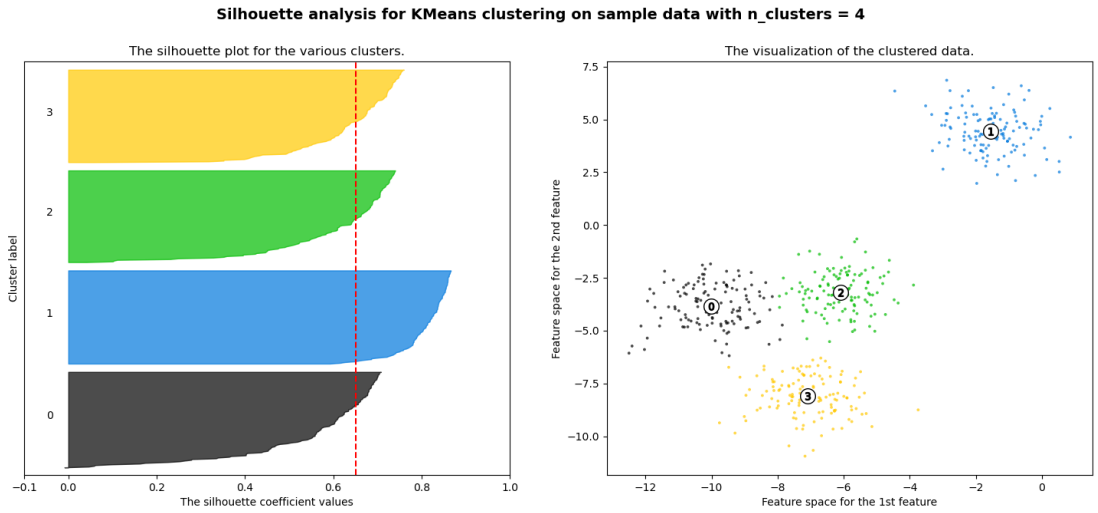
a(i): es la distancia promedio dentro del mismo clúster.

b(i): es la distancia promedio al clúster más cercano.

El rango está definido como: $-1 \leq s(i) \leq 1$, los valores cercanos a 1 indican buena separación.

Figura 5

Ejemplo de Análisis Silhouette para KMeans



Nota. Reproducido de "Selecting the number of clusters with silhouette analysis on KMeans clustering", por Scikit-learn developers, 2026, Scikit-learn 1.9.0 documentation.

La Figura 5 muestra un ejemplo del análisis de Silhouette para un conjunto de datos agrupado mediante K-Means. En el panel izquierdo se representa el coeficiente de Silhouette para cada observación, agrupadas según el clúster al que pertenecen. El ancho de cada banda refleja el número de observaciones del clúster, mientras que la longitud horizontal indica el valor del coeficiente de Silhouette. La línea vertical punteada representa el valor promedio del coeficiente para todos los datos. Valores cercanos a 1 indican que las observaciones están bien asignadas a su grupo y claramente separadas de los clústeres vecinos; valores próximos a 0 sugieren solapamiento entre agrupamientos, mientras que valores negativos indican que una observación podría haber sido asignada al clúster incorrecto. El panel derecho presenta la distribución espacial de los cuatro clústeres identificados. Se observa una separación clara entre los grupos, lo que explica los elevados valores del coeficiente de Silhouette obtenidos para la mayoría de las observaciones. Este ejemplo ilustra cómo la métrica permite evaluar simultáneamente la cohesión interna de cada grupo y la separación respecto de los demás agrupamientos.

Davies-Bouldin Index (DBI). Mide la compactación interna y separación entre clústeres, viene dado por:

Ecuación 13

Davies-Bouldin Index, DBI

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq 1}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

Donde:

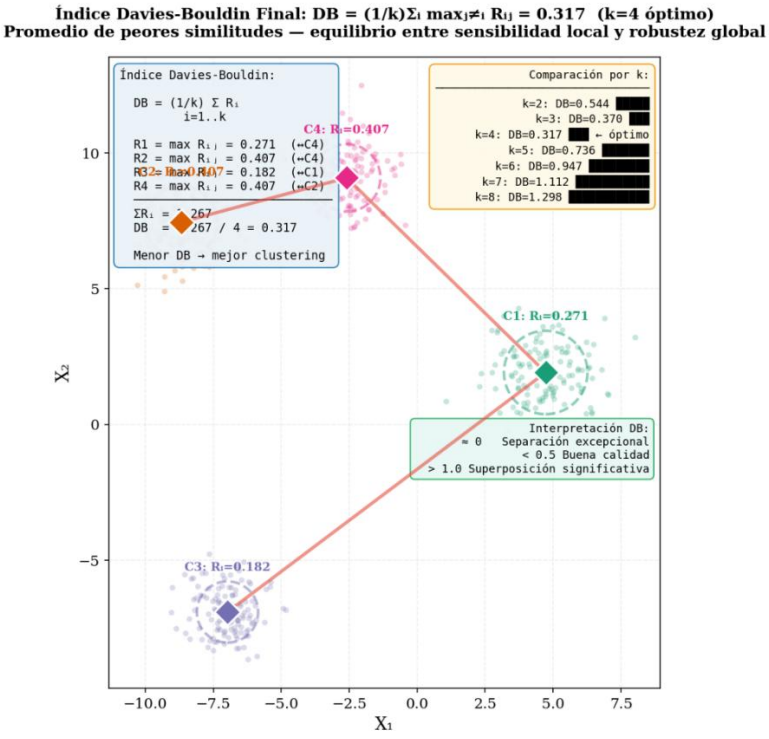
S_i: dispersión interna del clúster.

M_{ij}: distancia entre centroides

Cuando $DBI \rightarrow 0$ indica mejor agrupamiento.

Figura 6

Ejemplo de DBI variando los valores de k



Nota. Reproducido de "Davies bouldin score (Scikit-learn)", por Picard, s.f., CDS Institute – Data Science Python Blog.

La Figura 6 ilustra el cálculo del índice de Davies–Bouldin para un conjunto de datos compuesto por cuatro clústeres. Cada agrupamiento posee una dispersión interna representada mediante el radio de influencia alrededor de su centroide, mientras que las líneas entre centroides representan la distancia existente entre los diferentes grupos. El índice compara simultáneamente la compactación interna de cada clúster con su separación respecto de los demás. En el ejemplo mostrado se obtiene un DBI = 0.317, valor considerado bajo, indicando que los clústeres presentan buena compactación interna y adecuada separación entre sí. En términos generales, cuanto menor es el valor del índice de Davies–Bouldin, mejor es la calidad del agrupamiento obtenido.

Separación promedio entre scores normales y anómalos. Se define como:

Ecuación 14

Separación promedio entre scores normales y anómalos

$$\Delta = \bar{S}_{anomalias} - \bar{S}_{normales}$$

En donde:

- Mayor Δ
- Mejor capacidad discriminativa

Esta métrica suele usarse como indicador práctico de separación entre poblaciones. A diferencia del Silhouette Score y del índice de Davies–Bouldin, que evalúan la estructura de los agrupamientos, esta métrica analiza directamente la capacidad del modelo para diferenciar observaciones normales de aquellas consideradas potencialmente anómalas mediante sus puntajes de riesgo. Una mayor diferencia promedio entre ambos grupos indica una mejor capacidad discriminativa del algoritmo.

Coeficiente de Cohen's Kappa (κ). Mide el grado de concordancia entre dos clasificadores más allá del azar, viene dado por:

Ecuación 15

Coeficiente de Cohen's Kappa

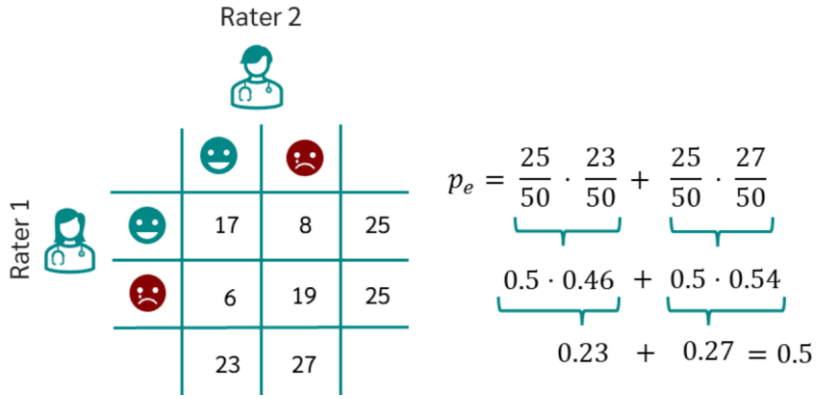
$$k = \frac{p_o - p_e}{1 - p_e}$$

Donde:

- p_o: acuerdo observado
- p_e: acuerdo esperado por azar

Figura 7

Calificaciones de dos evaluadores, uso del coeficiente Cohen's Kappa



Nota. Reproducido de "Kappa de Cohen – Explicación sencilla", por Numiqo Team, 2026, Numiqo: Online statistics calculator.

La Figura 7 presenta un ejemplo del cálculo del coeficiente de Cohen's Kappa a partir de la clasificación realizada por dos evaluadores independientes. La matriz de contingencia resume el número de observaciones sobre las cuales ambos evaluadores coincidieron y aquellas en las que existieron discrepancias. A partir de esta información se calcula el acuerdo observado (p_o) y el acuerdo esperado por azar (p_e) , obteniéndose finalmente el coeficiente de Cohen (κ). El principal aporte de esta métrica consiste en que descuenta el acuerdo atribuible al azar, proporcionando una medida más robusta de concordancia que el porcentaje simple de coincidencias. En esta investigación el coeficiente κ fue empleado para evaluar el grado de coincidencia entre los

algoritmos de detección de anomalías al identificar procesos contractuales potencialmente riesgosos.

Landis y Koch (1977) proponen una escala para interpretar los valores del coeficiente de Cohen's Kappa, clasificando la concordancia desde pobre hasta casi perfecta, la que se puede ver en la Tabla 3.

Tabla 3

Escala propuesta para interpretar los valores del coeficiente de Cohen's Kappa

k	Interpretación
< 0	Sin acuerdo
0 – 0.20	Leve
0.21 – 0.40	Aceptable
0.41 – 0.60	Moderado
0.61 – 0.80	Sustancial
0.81 – 1.00	Casi perfecto

Nota. Adaptado de *Interpretación Kappa de Cohen* (Numiqo Team, 2026)

Las métricas descritas en esta sección constituyen el marco metodológico utilizado para evaluar el desempeño de los algoritmos de detección de anomalías desarrollados en esta investigación. Su aplicación permitió comparar los modelos desde diferentes perspectivas, considerando la calidad del agrupamiento, la capacidad de discriminación entre observaciones normales y anómalas, así como el grado de concordancia entre los algoritmos. En el capítulo siguiente estas métricas se emplean para validar los resultados obtenidos y seleccionar la estrategia más adecuada para la construcción del modelo de scoring de riesgo.

CAPITULO 3:

3. DESARROLLO

3.1. Desarrollo del Trabajo

La selección de la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) responde a que esta constituye uno de los marcos de trabajo más ampliamente utilizados para el desarrollo de proyectos de Ciencia de Datos y Minería de Datos. Su enfoque iterativo permite abordar de forma organizada todas las etapas del ciclo analítico, desde la comprensión del problema hasta el despliegue de los resultados. Esta característica resulta especialmente adecuada para la presente investigación, ya que el desarrollo del sistema no se limitó al entrenamiento de modelos de machine learning, sino que incluyó la integración de datos provenientes de múltiples fuentes, la construcción de un repositorio de información, el diseño de variables para el análisis, la evaluación de algoritmos de detección de anomalías y la implementación de una aplicación para la visualización de los resultados.

Para desarrollar el proyecto se utilizó la metodología CRISP-DM. En la Figura 8 se pueden visualizar las diferentes etapas que la conforman. (Ramjan y Sunkpho, 2023, p. 6).

Figura 8

Proceso de Análisis con Técnicas de Data Mining

Nota. Adaptado de "Introduction to data mining", por S. Ramjan y J. Sunkpho, 2023, en Principles and theories of data mining with RapidMiner (p. 6), IGI Global.

Esta metodología es de amplio uso en proyectos de Ciencia de Datos y Minería de Datos, y sus etapas brindan un marco estructurado.

Tabla 4

Actividades desarrolladas por Fase

Fase	Actividades desarrolladas
Business Understanding	Definición del problema de investigación y de los objetivos para detectar anomalías en procesos contractuales del sector salud.
Data Understanding	Descarga y exploración de archivos JSON (OCDS) del SERCOP y archivos CSV del SRI; análisis de calidad de los datos.
Data Preparation	Procesos ETL, integración de fuentes, limpieza, transformación, construcción de variables (feature engineering), almacenamiento en DuckDB, Parquet y PostgreSQL.
Modeling	Entrenamiento de los modelos Isolation Forest, HDBSCAN, LOF y construcción del modelo ensemble para generar el puntaje de riesgo.
Evaluation	Comparación de modelos mediante Silhouette Score, índice de Davies–Bouldin, separación de puntajes de anomalía y coeficiente de Cohen (κ); análisis e interpretación de resultados.
Deployment	Implementación de la base de datos PostgreSQL, creación de vistas materializadas y desarrollo de una aplicación web en Streamlit para consulta, análisis y visualización del puntaje de riesgo.

Estas fases, mostradas en la Tabla 4, fueron adaptadas al contexto del análisis de contratación pública ecuatoriana y usadas como guía en todo el proceso de construcción del sistema analítico y de detección de anomalías.

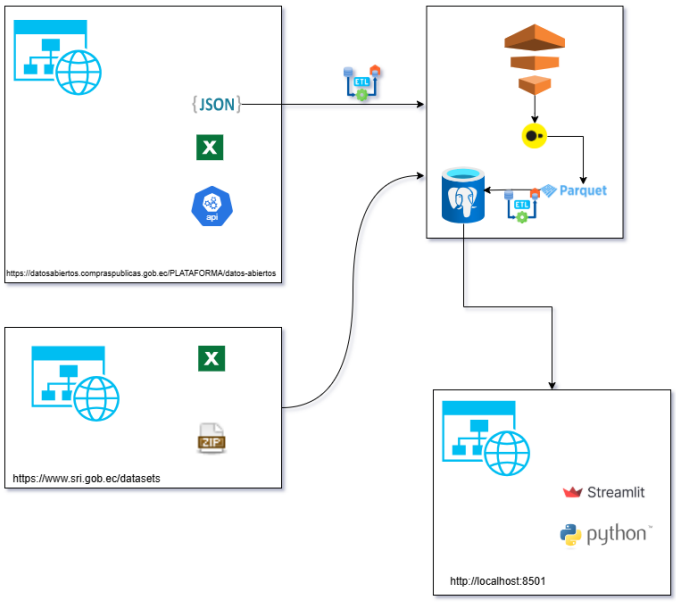
El proyecto se enfocó en aprovechar los datos que se encuentran disponibles en la plataforma de Contrataciones Abiertas Ecuador – OCDS (SERCOP, 2026c), la cual provee información de los procesos contractuales desde el año 2015 en adelante. Con base en las notas metodológicas que se encuentran en el portal, se identifica como fuentes primarias para la publicación de los datos abiertos de contratación pública a los siguientes portales:

1. Sistema Oficial de Contratación Pública
2. Módulo de Compras Corporativas del Sector Salud
3. Módulo de Compras Corporativas de Alimentos Escolar

Los datos abiertos son extraídos, procesados y transformados de las fuentes primarias, de manera automatizada, con la finalidad de brindar autenticidad en la información. La frecuencia de extracción debe ser diaria. Dentro de este marco se ha definido la siguiente arquitectura para aprovechar estos datos y construir una aplicación web que tiene la siguiente arquitectura:

Figura 9

Arquitectura de la solución propuesta



La arquitectura propuesta en la Figura 9 fue diseñada con el propósito de soportar el procesamiento eficiente de grandes volúmenes de datos provenientes de diversas fuentes abiertas, garantizando escalabilidad, rendimiento y facilidad de mantenimiento. Para ello se seleccionaron herramientas especializadas que cumplen funciones específicas dentro del flujo de procesamiento: JSON como formato de intercambio de datos, DuckDB para el procesamiento analítico en memoria de archivos de gran tamaño, Parquet como formato columnar optimizado para almacenamiento y consultas analíticas, PostgreSQL como repositorio central de información

integrada y Streamlit como plataforma para el despliegue interactivo de resultados y modelos de aprendizaje automático. La selección de estas herramientas responde tanto a sus capacidades técnicas como a su integración dentro de un flujo moderno de analítica de datos.

Se usan archivos en formato JSON, "JavaScript Object Notation o Notación de Objetos JavaScript", formato que permite representar un objeto JavaScript como una cadena de texto la misma que puede ser almacenada o compartida, (Freeman y Robson, 2024, p. 596), disponibles para su descarga en el portal de datos abiertos. El formato JSON fue seleccionado porque constituye el estándar utilizado por la Plataforma de Datos Abiertos del SERCOP y permite representar estructuras jerárquicas de información de manera ligera y fácilmente procesable mediante herramientas de Ciencia de Datos. El proceso de descarga envuelve no solo bajarse los archivos sino descomprimirlos y usar DuckDB (Apéndice B) debido a su capacidad para consultar directamente archivos JSON y Parquet sin requerir previamente su carga en una base de datos relacional. Su arquitectura orientada al procesamiento analítico (OLAP) permite ejecutar consultas SQL sobre archivos de gran tamaño utilizando eficientemente la memoria disponible, reduciendo significativamente los tiempos de procesamiento durante las tareas de extracción y transformación de datos. El uso de archivos en formato Parquet permitió reducir considerablemente el espacio de almacenamiento requerido y acelerar las consultas posteriores, debido a que almacena la información de forma columnar, característica especialmente ventajosa para procesos analíticos donde únicamente se consultan determinadas variables del conjunto de datos, es de muy amplio uso en procesamiento de Big Data, siendo utilizados por Apache Spark. Sus eficientes técnicas de compresión y codificación ofrecen la ventaja de reducir los requisitos de almacenamiento y mejorar el rendimiento de las consultas. (Needham et al., 2024, p. 108). Dependiendo del año que se escoja en procesar, un archivo JSON puede tener un tamaño de más de 1 GB. El trabajo investigativo se centró en cubrir los datos comprendidos entre los años 2020 al 2025, tomando en cuenta que durante este rango de tiempo el mundo atravesó la pandemia de Covid 19, y existe

evidencias de presunta corrupción en contratos públicos relacionados el sector de la salud, como los que la Contraloría General de Estado investigó y emitió glosas a entidades públicas. (Primicias, 2022).

Cabe notar que el aplicativo construido se puede alimentar de información compatible de cualquier año, y entrenarse con una mayor cantidad de años o modificarse para cubrir una mayor cantidad de sectores en caso de ser necesario. La aplicación web usa PostgreSQL como repositorio central debido a su robustez, soporte para grandes volúmenes de información, compatibilidad con consultas SQL complejas y capacidad para implementar vistas materializadas, las cuales mejoran el desempeño de las consultas realizadas desde la aplicación desarrollada, y para la interfaz de usuario se usa Streamlit por permitir construir aplicaciones analíticas interactivas utilizando Python, facilitando la integración directa con los modelos de machine learning y con la base de datos PostgreSQL sin requerir el desarrollo de aplicaciones web tradicionales.

Tabla 5

Tamaño de los archivos JSON

Año	Tamaño archivo JSON (GB)
2020	1.04
2021	1.24
2022	1.55
2023	1.50
2024	1.54
2025	1.47

El proceso de ETL (Extracción, Transformación y Carga) toma como insumo a los archivos JSON, cuyos tamaños se muestran en la Tabla 5, a los que se realizan transformaciones para convertir la información no estructurada y almacenarla en archivos parquet.

El patrón ETL tiene su origen en las primeras aplicaciones de almacenamiento de datos y todavía se utiliza ampliamente en la actualidad. Con este patrón, los datos se extraen de los

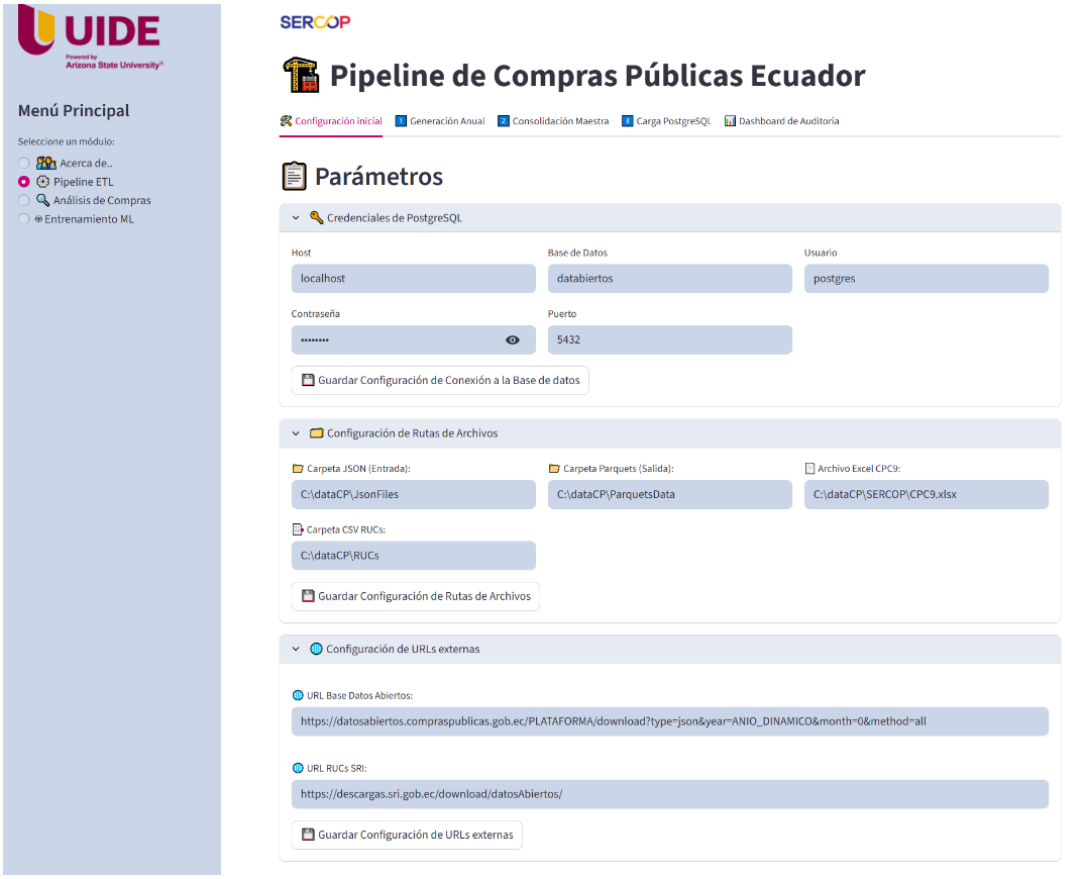
sistemas de origen, como bases de datos operativas o archivos; se formatean, limpian y transforman según los requisitos del negocio; y finalmente se cargan en las tablas de destino como Snowflake o cualquier otro repositorio (Ferle, 2025, p. 210). De los archivos JSON descargados se extraen las secciones más importantes y almacenan en archivos parquet, archivos que se comportan como una tabla de una base de datos, por lo que por medio del motor DuckDB se pueden efectuar consultas sobre estas nuevas fuentes de datos temporales; de allí que una parte importante del proyecto se basa en trabajar con SQL y estos archivos parquet para consolidar la información, y posteriormente cargarla a una base de datos PostgreSQL (ver Apéndice B); donde se construye un esquema relacional con tablas que tienen su estructura similar a los archivos parquet. Un proceso similar es el que se efectúa con la fuente primaria de datos del SRI en donde se obtienen datasets en formato CSV y ZIP de archivos que contienen información de RUCs de empresas y personas naturales. Toda esta información se centraliza en la base de datos relacional, permitiendo obtener acceso a datos que ayudan a construir los reportes y facilitan el monitoreo que se muestra en la aplicación que construida utilizando Streamlit (ver Apéndice B), además en PostgreSQL se crean vistas materializadas, las cuales proporcionan la información necesaria para crear datasets que son utilizados para entrenar los modelos de machine learning que fueron propuestos para el proyecto actual, a los cuales se hace referencia en el capítulo 1 dentro de los objetivos específicos.

3.2. Descripción del Aplicativo

El pipeline comienza con la configuración de variables de ambiente.

Figura 10

Configuración de variables de ambiente



En la Figura 10 se muestran las secciones en donde se ingresan valores de configuración, como las credenciales para poder conectarse a la base de datos PostgreSQL, las rutas de los archivos, y las direcciones URL de los portales de datos abiertos.Tabla 12

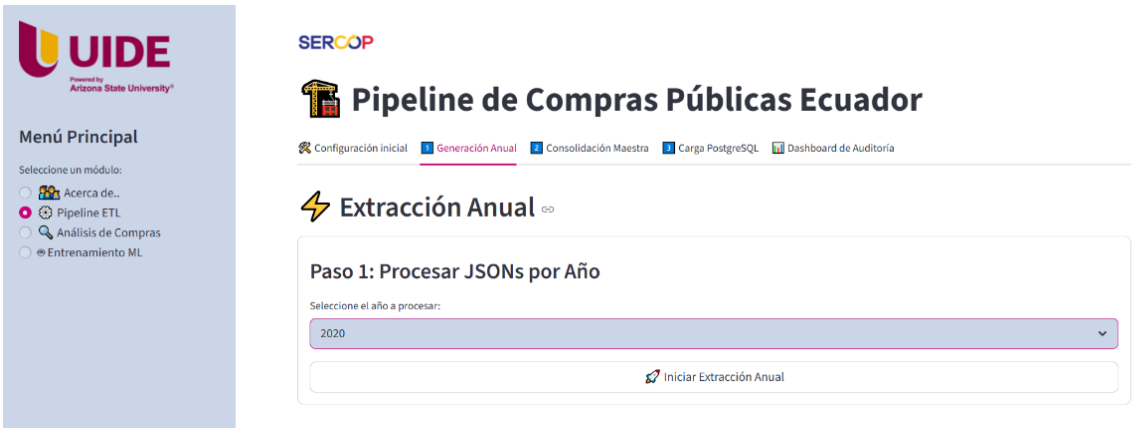
Explicación de resultados obtenidos utilizando Isolation Forest

Para el funcionamiento del aplicativo es necesario tener acceso a una base de datos PostgreSQL, así como haber creado una base de datos inicial, para el proyecto se utilizó la versión 17, siendo recomendable usar la versión 17 o superior. Luego se encuentra la configuración de rutas de los archivos, en donde se parametriza la ruta o carpeta en donde se descargarán los archivos en formato JSON desde la Plataforma de Datos Abiertos. Se debe, de igual manera, especificar la ruta en donde se crearán los archivos en formato parquet, los mismos que contienen información relacionada a las secciones más relevantes encontradas dentro de los archivos JSON.

También se define la carpeta en donde se descargarán los archivos en formato CSV que contendrán los RUCs por provincia y que se encuentran disponibles en datos abiertos del Servicio de Rentas Internas (SRI). Adicional a esto, es necesaria la descarga de un archivo de forma manual que contiene la información de CPC a nivel 9 (SERCOP, 2026e). Tanto para los archivos JSON como los CSV del SRI se cuenta con las direcciones URL, las mismas que se parametrizan en esta parte del proceso, y que sirven para que el proceso de extracción se lo haga de manera automatizada. En el caso de que tanto el SRI como datos abiertos cambien sus direcciones URL, será necesario modificar estas referencias para que el proceso se ejecute con éxito. Esta configuración se la hace básicamente una sola vez, considerando que el usuario puede crear los ambientes que considere pertinente (bases de datos de prueba o producción, así como rutas temporales).

Figura 11

Descarga de archivos JSON por año



El siguiente paso es extraer la información de los archivos en formato JSON, para ello se escoge el año particular que se quiere procesar, y se pulsa el botón **Iniciar Extracción Anual**, mostrado en la Figura 11. El resultado serán archivos en formato parquet, mostrados en la Tabla 6, los cuales son grabados en la ruta de salida que se definió en el paso anterior.

Tabla 6

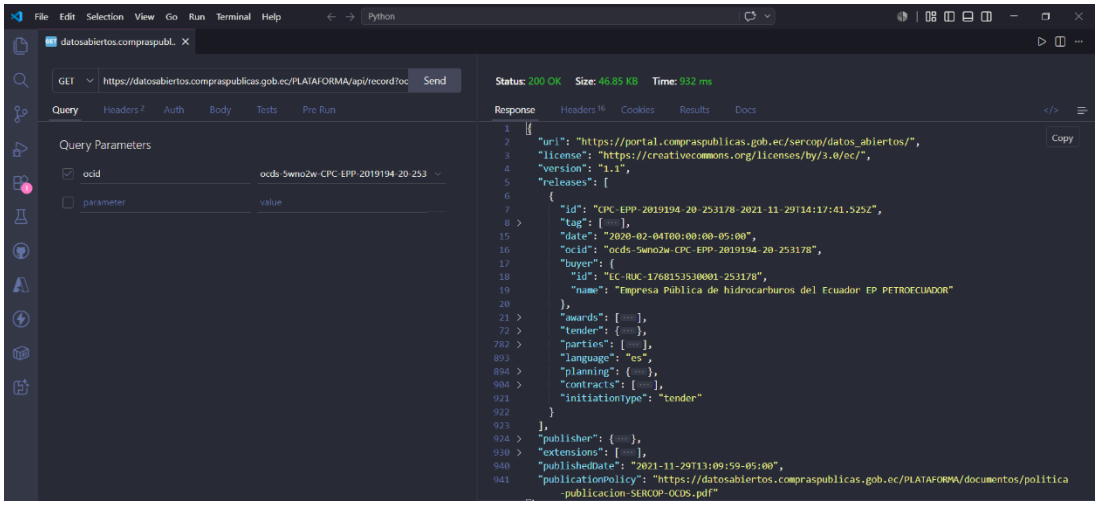
Contenido de los archivos parquet

Archivo parquet por año	Descripción
releases.parquet	Contiene información de las licitaciones o contratos públicos
tenderers.parquet	Contiene información de los licitadores
tender_items.parquet	Contiene información de los ítems o productos ofrecidos por los licitadores
awards.parquet	Contiene información de las adjudicaciones
awards_items.parquet	Contiene información de los ítems o productos adjudicados
parties.parquet	Contiene información más detallada de las partes que interviene en un contrato (desde el contratante y los proveedores que participaron)
contracts.parquet	Contiene información del contrato público

La creación de los archivos parquet se hizo tomando como referencia la estructura de los archivos JSON de origen, los que pueden ser analizados usando un sin número de herramientas como VIM, Visual Studio Code, o cualquier otro programa que facilite la lectura de texto.

Figura 12

Revisión de la estructura del archivo JSON con Visual Studio Code



También es posible analizar respuestas en formato JSON obtenidas directamente del servidor utilizando clientes de tipo API como Postman, Insomnia, Bruno, Apidog, etc., e incluso agregando complementos a Visual Studio Code, como lo muestra la Figura 12. Actualmente cuando se habla de APIs se hace referencia a APIs web, en donde un cliente se encuentra en un lado de la interfaz y envía solicitudes, mientras que un servidor (o servidores) se encuentra en el otro lado de la interfaz y responde a la solicitud. (Westerveld, 2024, p. 2).

En la Figura 13 se puede apreciar el resultado luego de haber buscado un contrato en particular mediante su código OCID, y cómo Visual Studio Code muestra el archivo JSON devuelto como respuesta, permitiendo analizar su estructura. La Plataforma de Datos Abiertos también dispone de información en formato CSV, la cual puede ser abierta directamente en Excel, teniendo el archivo que es del interés del presente trabajo las secciones: releases, plannig, tender, etc. En la plataforma también se dispone de un API que permite obtener información de manera individual (por contrato), la misma que podría ser usada en otros escenarios. Para este trabajo se utilizará la forma más completa y actualizada que es trabajar directamente con los archivos JSON.

Una vez que se han procesado los años seleccionados, se tendrá un archivo parquet por cada uno de ellos.

Figura 13

Creación de archivos maestros

El siguiente paso en el pipeline diseñado consiste en consolidar toda la información descargada desde los archivos JSON de datos abiertos, así como los archivos CSV que contienen información de RUCs. Al final de este proceso se tendrán archivos parquet maestros que contienen los datos consolidados y que serán usados para llenar las tablas y vistas materializadas de la base de datos en PostgreSQL.

Luego de esto se debe cargar la información a la base de datos en PostgreSQL, y en caso de reprocesos y ejecutar nuevamente este paso, se eliminará previamente los datos en las tablas relacionadas, así como se borrarán las vistas materializadas. En la Figura 14 se muestra este proceso.

Figura 14

Carga final a la base de datos

Finalmente se muestra un Dashboard o tablero en donde se presentan las métricas de este pipeline. Se anota que, durante todo este proceso y sus pasos, se graban pistas de auditoría en la base de datos PostgreSQL, las mismas que son usadas para alimentar KPIs y gráficos que se muestran en este tablero de control.

Figura 15

Métricas del Proceso ETL

En la Figura 15 se puede apreciar métricas como: el número total de registros procesados, recordando que se manejan millones de registros, el tiempo que se demoró en los procesamientos de transformación y subida de datos. Esto permite tener una visión y ajustar cualquier inconveniente para mejorar los procesos.

Ahora se describe la sección del Análisis de Compras, el mismo que contiene un Dashboard inicial o tablero de control, en donde se pueden apreciar por año algunos KPIs como: Monto total adjudicado, total de contratos firmados, número de entidades compradoras y número de proveedores adjudicados. Se cuenta con filtros que permiten explorar los datos ya sea por año, por el enfoque de compra: el mismo que puede ser todas o averiguar por compras que contienen insumos médicos; también se puede seleccionar el tipo de contratación. Con todos estos insumos se verán gráficos de la Evolución histórica de la Contratación Pública basada en los años que se

procesaron, la estructura del gasto según el tipo de contratación, el top 20 de entidades gubernamentales que gastaron más por el año en particular escogido, así como el top 20 de proveedores adjudicados en ese mismo año. También se muestra el top 20 de los proveedores con mayores contratos, así como una tabla resumen de los proveedores adjudicados. Toda esta información se la obtiene producto de los datos subidos en la base de datos en PostgreSQL, se puede comparar con las métricas y gráficos que el mismo portal de datos abiertos suministra, es decir, de alguna manera se presenta la misma información, pero con otro tipo de gráficos y con más detalle.

En la

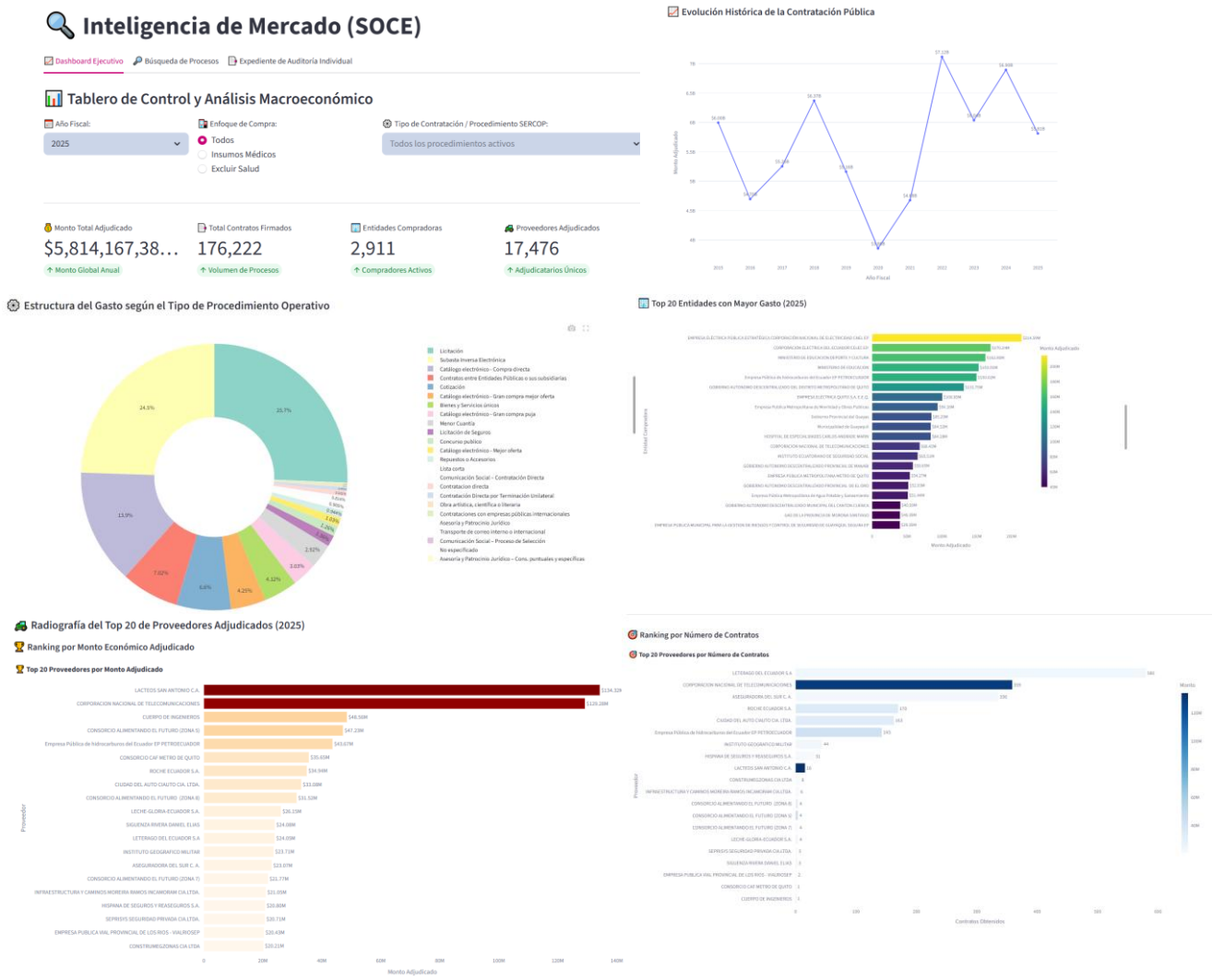
Figura 16 se muestra una serie de reportes obtenidos de la base de datos con la información que ya se ha subido, siendo los gráficos y tablas mostradas:

- Selección de año fiscal, mostrando monto adjudicado, contratos firmados, entidades compradoras, y proveedores adjudicados
- Evolución histórica de la contratación pública
- Estructura del gasto según el tipo de procedimiento operativo
- Top 20 entidades con mayor gasto del año seleccionado
- Top 20 proveedores adjudicados del año seleccionado
- Ranking por número de contratos
- Tabla resumen de los top 20 proveedores por monto

Se cuenta con la funcionalidad de búsqueda de procesos, en donde el usuario puede aplicar una serie de filtros como: código OCID, año de procesamiento, tipo de contratación, etc. El resultado se muestra en forma de tabla con un límite de 500 registros como máximo, la misma que se puede exportar en formato CSV.

Figura 16

Exploración de datos en la base



Ver tabla resumen del Top 20 de proveedores

Proveedor	RUC	Contratos Obtenidos	Monto Adjudicado	Monto Texto
LACTEOS SAN ANTONIO C.A.	0390011024001	16	134320091.7285	$134.32M
CORPORACION NACIONAL DE TELECOMUNICACIONES	1768152560001	359	129275998.8008	$129.28M
CUERPO DE INGENIEROS	1768007040001	1	48555272.33	$48.56M
CONSORCIO ALIMENTANDO EL FUTURO (ZONA 5)	1793219395001	4	47234362.0143	$47.23M
Empresa Pública de hidrocarburos del Ecuador EP PETROECUADOR	1768153530001	143	43668396.0463	$43.67M
CONSORCIO CAF METRO DE QUITO	1792508215001	1	35650000	$35.65M
ROCHE ECUADOR S.A.	1790475689001	170	34941034.9	$34.94M
CIUDAD DEL AUTO CIAUTO CIA. LTDA.	1891748376001	163	33083604.2755	$33.08M
CONSORCIO ALIMENTANDO EL FUTURO (ZONA 8)	1793219396001	4	31519326.4348	$31.52M
LECHE-GLORIA-ECUADOR S.A.	1790663973001	4	26147492.1402	$26.15M

En la Figura 17 se muestra la interacción a través de filtros de búsqueda para obtener información de contratos.

Figura 17

Búsqueda de Contratos

Finalmente, se cuenta con una ficha o expediente por cada contrato, en donde se simuló la funcionalidad de banderas rojas, similar a lo que se puede encontrar en portales en donde se muestran indicadores para detección de alertas en la contratación pública, (Fundación Ciudadanía y Desarrollo, s.f.) tomando en cuenta dos variables: la antigüedad del proveedor en base a la información obtenida del SRI en el proceso de ETL de RUCs, ésta se calcula en número de días entre la fecha inicial de actividades económicas y la fecha de firma del contrato. La otra variable

es el número de días transcurridos entre la fecha de planeación del proceso y la fecha de la firma del contrato. En la Figura 18 se muestra información particular de un contrato,

Figura 18

Expediente de Auditoría Individual

Antes de pasar con la siguiente sección, es importante señalar que, una vez consolidada la información en la base de datos, el siguiente paso consistió en construir un conjunto de datos (dataset) destinado al entrenamiento de los modelos de machine learning. Para comprender este proceso, es necesario considerar la estructura de información disponible en la base de datos, la cual se presenta en la Tabla 7.

Tabla 7

Número de procesos por año

Año	Número de procesos
2015	280643
2016	279636
2017	334275
2018	357687
2019	275055
2020	160676
2021	167058
2022	212324
2023	217913
2024	219185
2025	176222

En total un universo de 2680674 registros o contratos entre los años 2015 hasta el 2025. El proyecto se basó en estudiar el periodo de tiempo comprendido entre los años 2020 y 2025, por lo tanto, hay un universo de 1153378 contratos. De este universo se tiene que filtrar solo a aquellos procesos que involucran a insumos médicos, y para ello se hizo uso del CPC 9 y/o su categoría.

Para agilitar la consulta de contratos se creó la vista materializada **mv_universo_contratos**. (Ver Apéndice C), la cual provee de todo el universo de contratos. En todas las vistas materializadas que se construyeron se usó CTEs (Common Table Expression o Expresión Común de Tabla), las mismas que son un resultado temporal obtenido de una sentencia SQL o consulta o query. Esta sentencia puede contener instrucciones SELECT, INSERT, UPDATE o DELETE. La vida útil de una CTE es igual a la vida útil de la consulta, y para hacerla persistente en la base de datos se la encapsula dentro de la creación de la vista materializada; de esta manera los datos siempre estarán listos sin tener que procesar la consulta para obtener la información. (Ferrari y Pirozzi, 2023, p. 145).

Aquí se puede visualizar una sección en donde se califica o categoriza al proceso contractual si involucra algún código CPC relacionado a insumos médicos. (SERCOP, 2026e). En el portal de Compras Públicas se cuenta con acceso al clasificador central de productos o conocido como CPC (SERCOP, 2026e), donde se encuentran las categorías mostradas en la Tabla 8:

Tabla 8

Ejemplo de categorías de productos

Categoría CPC	Descripción
2411	ALCOHOL ETILICO SIN DESNATURALIZAR CON UNA CONCENTRACION ALCOHOLICA EN VOLUMEN DEL 80% O MAS
2412	ALCOHOL ETILICO Y OTROS ALCOHOLES DESNATURALIZADOS DE CUALQUIER CONCENTRACION ALCOHOLICA
352	PRODUCTOS FARMACEUTICOS
481	APARATOS MEDICOS Y QUIRURGICOS Y APARATOS ORTOPEDICOS

En el caso de que existiera otro CPC (SERCOP, 2026d) relacionado con el ámbito médico se procedería a ajustar en código. Cabe notar que en el sistema SOCE en cada proceso no existe una etiqueta o categoría que indica explícitamente si el contrato público es de carácter médico.

Para el caso de las entidades públicas, no se cuenta con un identificador explícito que permita clasificarlas por áreas de interés. Entonces se realizó una clasificación en base al nombre de la entidad, previamente se crea una tabla o entidad a la que se le insertan algunos patrones basados en expresiones regulares (regex), como ejemplo se tiene la tabla **diccionario_categorias_salud**. (Ver Apéndice C)

Entonces se crea un universo de entidades objetivo, a las que las se clasificó en base a su nombre, y en el caso de no pertenecer al ámbito “SALUD”, se las etiqueta como “NO SALUD”.

Con esto no se eliminan datos, solo se clasifican de una mejor manera para el propósito de analizar los contratos médicos. (Ver Apéndice C)

El siguiente paso consistió en crear un dataset al que se le aplicó algunas limpiezas necesarias como son: filtrar solo aquellos procesos en donde se tenga al monto adjudicado (award_amount) mayor a cero, así como que el ID o RUC del proveedor no sea nulo; con eso se eliminan procesos "basura" (Ver Apéndice C).

Producto de esta limpieza se tiene un total de 981792 registros entre los años 2020 al 2025, y un total de 2413139 de procesos entre el intervalo de años comprendidos entre el 2015 al 2025. Finalmente se crea el dataset limpio mv_dataset_base_limpio (Ver Apéndice C).

Aquí lo que se procura es filtrar aquellos procesos contractuales en donde el proveedor adjudicado tenga una fecha de inicio de actividades registrada en el SRI, que el monto adjudicado sea mayor a cero, y que los tipos de contratación, tipo de institución, método de adquisición, fecha de planeación, fecha de la firma de contrato y la categoría de la compra no sean nulas. Se puede efectuar en este punto, o haberlo hecho en el Análisis Exploratorio de Datos (EDA). El resultado en número de registros es: un total de 328516 registros comprendidos entre el 2015 y 2025, y una cantidad de 166052 registros para el período entre el 2020 al 2025.

Finalmente se construye el conjunto de datos **mv_dataset_base_ml**, que contiene únicamente los contratos con insumos médicos y algunas columnas que se utilizaron para entrenar los distintos métodos de machine learning.

El resultado que se tiene es el siguiente: un universo de 64180 registros entre los años 2015 al 2025, y un total de **34745** registros para los años comprendidos entre el 2020 al 2025. Recordando que solo se toma en cuenta a los procesos que fueron categorizados como de insumos médicos. Este último dataset es el punto de partida para el entrenamiento de los modelos de machine learning.

Para obtener el dataset que se usó para el entrenamiento basta con exportar los datos de la vista materializa llamada mv_dataset_base_ml hacia un archivo csv filtrando los años entre 2020 y 2025.

```
copy (SELECT * FROM mv_dataset_base_ml where anio_proceso in
('2020','2021','2022','2023','2024','2025')) TO
'D:\dataCP\dbP\dataset_ml_2020_2025.csv' WITH CSV HEADER;
```

La siguiente sección denominada “Entrenamiento ML” está conformada por un resumen ejecutivo en donde se muestra los resultados de haber ejecutado el entrenamiento de los modelos no supervisados como: Isolation Forest, HDBSCAN y Local Outlier Factor (LOF) mediante un Notebook en Python dentro de la plataforma de Google Colab. Puntualmente se muestra que del universo de 34745 contratos se encontraron que 1738 tiene un riesgo alto, es decir, el sistema de scoring que se elaboró pudo catalogar al 5% aproximadamente de contratos con ese tipo de riesgo. También se muestra que el mejor modelo entrenado no supervisado fue Isolation Forest, y el patrón dominante encontrado es la concentración de contratos entre proveedores y las entidades públicas que les adjudican los contratos. En la Figura 19 se muestran los resultados del entrenamiento de los modelos de machine learning.

Figura 19

Resumen Ejecutivo del Entrenamiento ML

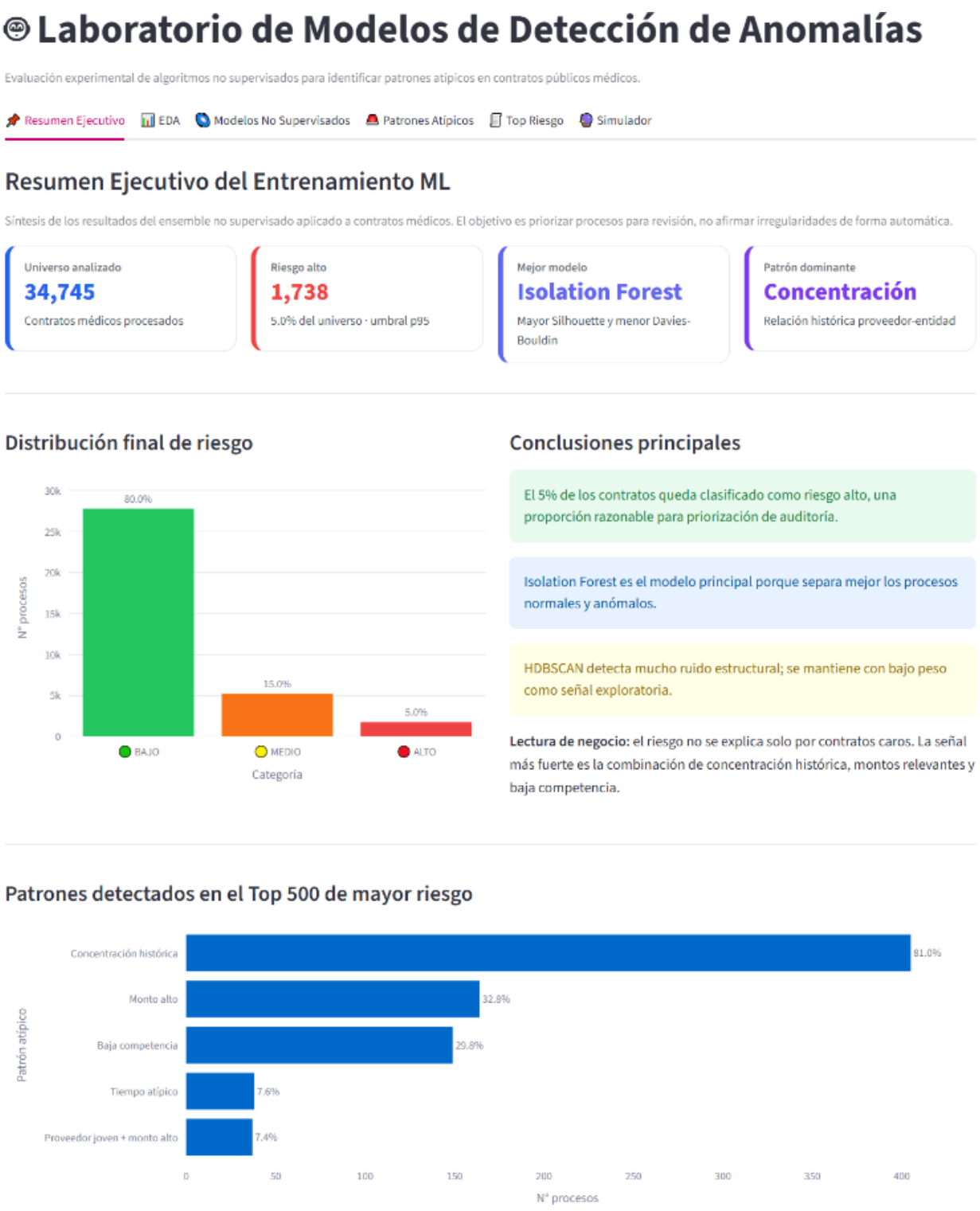


Figura 20

Análisis Exploratorio del Dataset de Contratos

En la Figura 20 se muestran los resultados obtenidos luego de haber cargado el dataset de 34745 contratos, al que se le aplica un EDA (Análisis Exploratorio de Datos). En la Figura 21 se muestra el análisis exploratorio que se aplica a los datos.

Tabla 9

Tipos de variables del dataset utilizado para el entrenamiento

Categoría de variable	**Variables principales**	**Tipo de dato**	**Tratamiento aplicado**	**Finalidad en el modelo**
Identificadores	ocid, tender_id, award_id, supplier_ruc, buyer_ruc	Texto	Excluidas del entrenamiento	Identificación y trazabilidad de los registros
Variables monetarias	award_amount, monto_proveedor_entidad_anio, monto_proveedor_entidad_hist	Numérico continuo	Transformación log1p y estandarización	Reducir asimetría y efecto de valores extremos
Variables de frecuencia	contratos_proveedor_entidad_hist, licitacion_numero_proveedores	Numérico discreto	Transformación log1p y estandarización	Representar intensidad de contratación y competencia
Variables temporales	dias_maduracion_empresa, dias_entre_planeacion_firma_contrato	Numérico continuo	Transformación log1p y estandarización	Detectar patrones temporales atípicos
Variables categóricas	Tipo de procedimiento, provincia, tipo de contratación, etc.	Categórico	Codificación (*One-Hot Encoding*)	Incorporar características estructurales de los procesos
Variable objetivo	No aplica	—	No utilizada	La investigación emplea aprendizaje no supervisado, por lo que no existen etiquetas de fraude o irregularidad.

Figura 21

EDA - Tipos de datos

La Tabla 9 resume la clasificación de las variables empleadas durante el análisis exploratorio de datos (EDA), así como el tratamiento aplicado previo al entrenamiento de los modelos de aprendizaje automático. Las variables identificadoras fueron excluidas del proceso de entrenamiento por no aportar información predictiva, mientras que las variables numéricas relacionadas con montos, frecuencias y tiempos fueron transformadas mediante la función logarítmica log1p y posteriormente estandarizadas para disminuir la influencia de distribuciones altamente asimétricas y valores extremos. Las variables categóricas fueron codificadas mediante técnicas de representación adecuadas para permitir su incorporación a los algoritmos de detección de anomalías. Finalmente, debido a que el estudio se desarrolló bajo un enfoque de aprendizaje no supervisado, no se dispuso de una variable objetivo que identificara procesos contractuales fraudulentos o irregulares, lo que justificó el uso de técnicas orientadas a la identificación de patrones atípicos.

Para obtener las métricas de comparación de los modelos no supervisados se usan los resultados obtenidos en el entrenamiento de cada uno de ellos, con el siguiente código en Python:

```
METRICS_SAMPLE = min(10000, N)
np.random.seed(42)
m_idx = np.random.choice(N, size=METRICS_SAMPLE, replace=False)
X_m   = X_scaled[m_idx]

metrics = {}
for name, label_col, score_col in [
    ('Isolation Forest', 'iso_label',    'iso_score'),
    ('HDBSCAN',          'dbscan_label', 'dbscan_score'),
    ('LOF',              'lof_label',    'lof_score'),
]:
    labels      = df_model[label_col].iloc[m_idx].values
    unique_lbls = np.unique(labels)
    if len(unique_lbls) > 1:
        try:
            sil = silhouette_score(X_m, labels, sample_size=5000,
random_state=42)
            cal = calinski_harabasz_score(X_m, labels)
            dav = davies_bouldin_score(X_m, labels)
        except Exception:
            sil, cal, dav = np.nan, np.nan, np.nan
    else:
        sil, cal, dav = np.nan, np.nan, np.nan
    scores = df_model[score_col].iloc[m_idx]
    m_norm = scores[labels==1].mean()  if (labels==1).any()  else np.nan
    m_anom = scores[labels==-1].mean() if (labels==-1).any() else np.nan
    sep    = m_anom - m_norm if not np.isnan(m_anom) else np.nan

    metrics[name] = {
        'N° anomalías (total)':      (df_model[label_col]==-1).sum(),
        '% anomalías':               f"{(df_model[label_col]==-
1).sum()/N:.1%}",
        'Silhouette (↑)':            round(sil,4) if not np.isnan(sil) else
'N/A',
        'Calinski-Harabasz (↑)':     round(cal,2) if not np.isnan(cal) else
'N/A',
```

```
        'Davies-Bouldin (↓)':        round(dav,4) if not np.isnan(dav) else
'N/A',
        'Score medio normal':        round(m_norm,4) if not np.isnan(m_norm)
else 'N/A',
        'Score medio anómalo':       round(m_anom,4) if not np.isnan(m_anom)
else 'N/A',
        'Separación Δ (↑)':          round(sep,4)  if not np.isnan(sep)  else
'N/A',
        'Peso en ensemble':          {'Isolation
Forest':W_ISO,'HDBSCAN':W_HDBSCAN,'LOF':W_LOF}[name],
    }
metrics_df = pd.DataFrame(metrics).T
print(f'Métricas calculadas sobre muestra de {METRICS_SAMPLE:,} registros')
metrics_df
```

Esto permite contrastar y comparar los resultados con el método Ensemble que junta los resultados de cada uno de los algoritmos. En la Figura 22 se muestra la comparación de los modelos de ML.

Figura 22

Comparación de Modelos No Supervisados

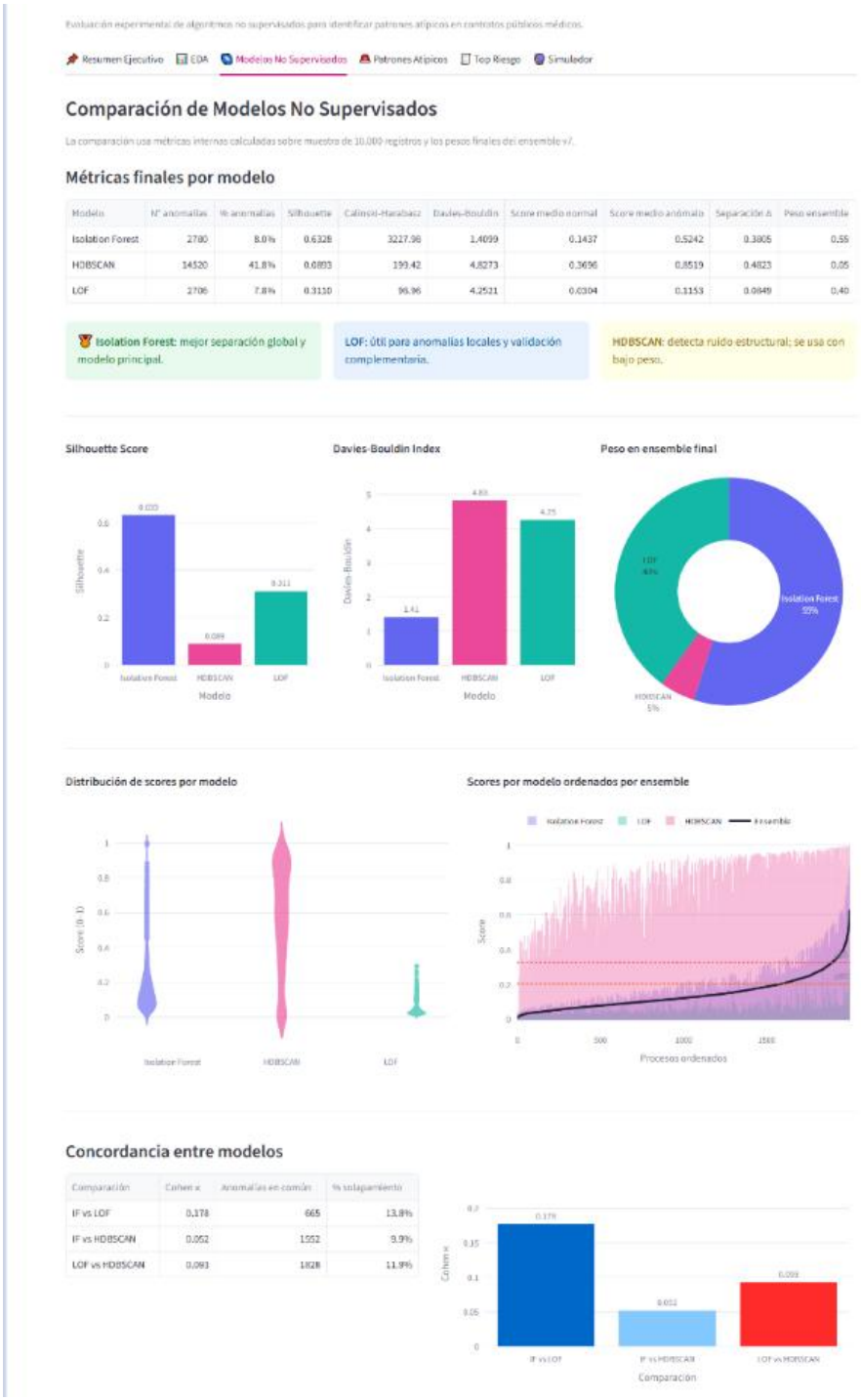
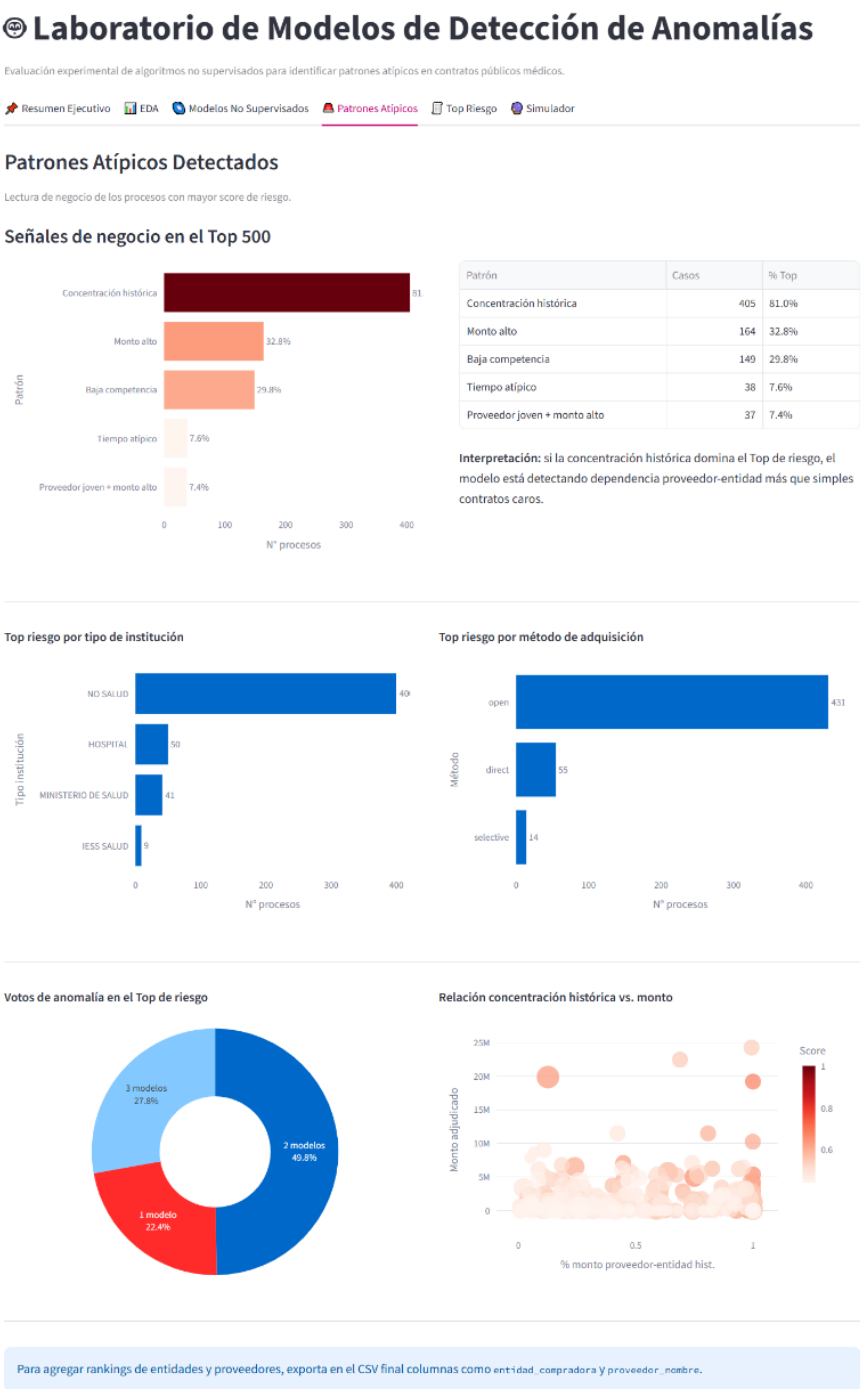


Figura 23

Patrones Atípicos Detectados



En la Figura 23 se presenta un resumen del conjunto de patrones atípicos detectados en los contratos públicos analizados.

Figura 24

Top Procesos de Riesgo

En la Figura 24 se presenta la interacción con la aplicación web para mostrar los 500 contratos medidos en su riesgo en forma descendente. Aquí se pueden aplicar algunas funcionalidades de filtros para encontrar o segmentar información. Finalmente se incluyen dos ilustraciones que permiten exponer a un simulador de riesgo la misma que se muestra en la Figura 25, en donde se puede escoger algunos parámetros como: monto adjudicado, antigüedad del proveedor adjudicado, número de oferentes que participaron en el proceso contractual, y algunas otras variables; el resultado se muestra en la Figura 26 como una alerta sobre la simulación en base a los datos ingresados, y saber si el proceso simulado tiene un riesgo alto, medio o bajo.

Figura 25

Simulador de Predicción

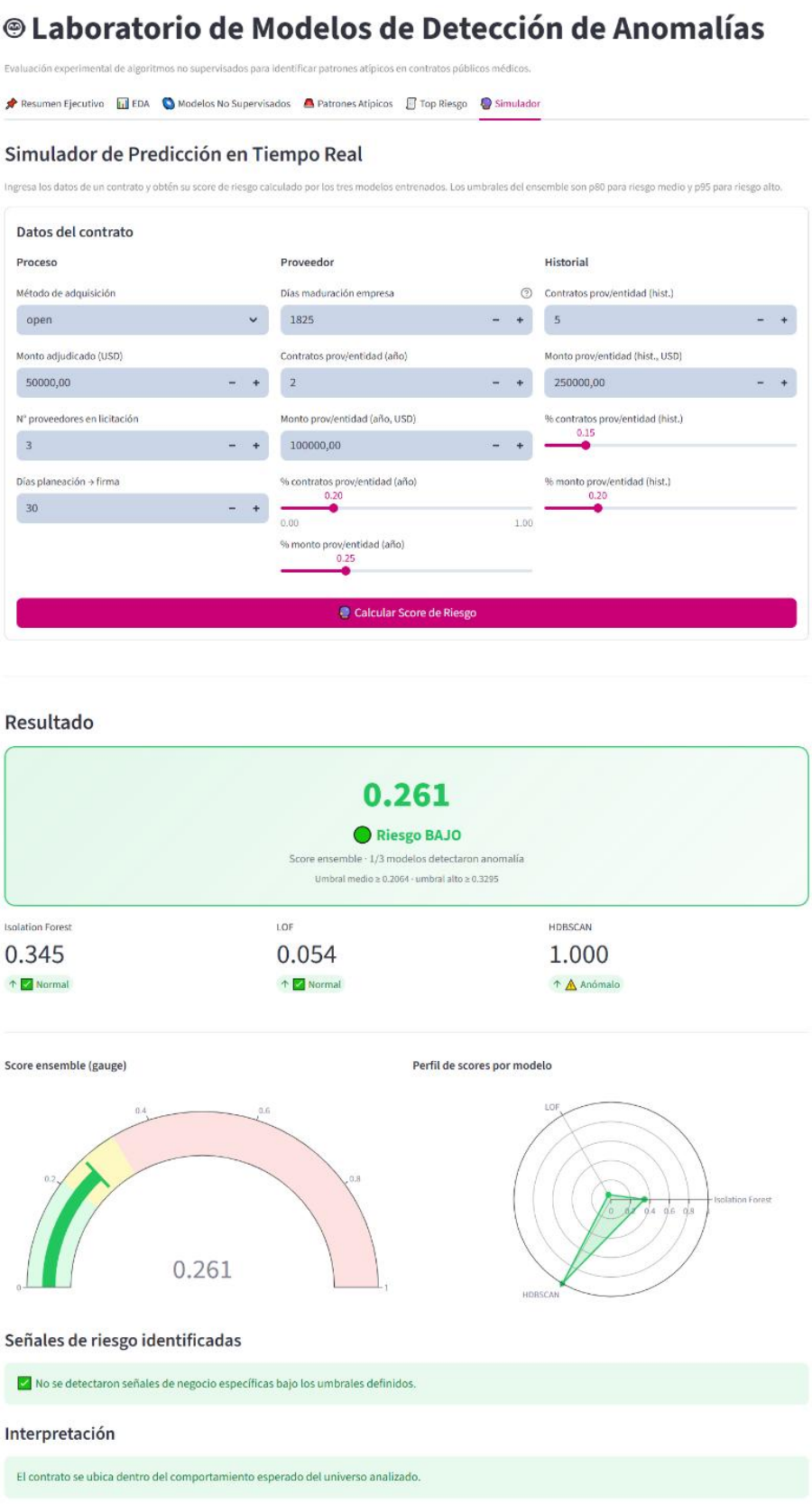
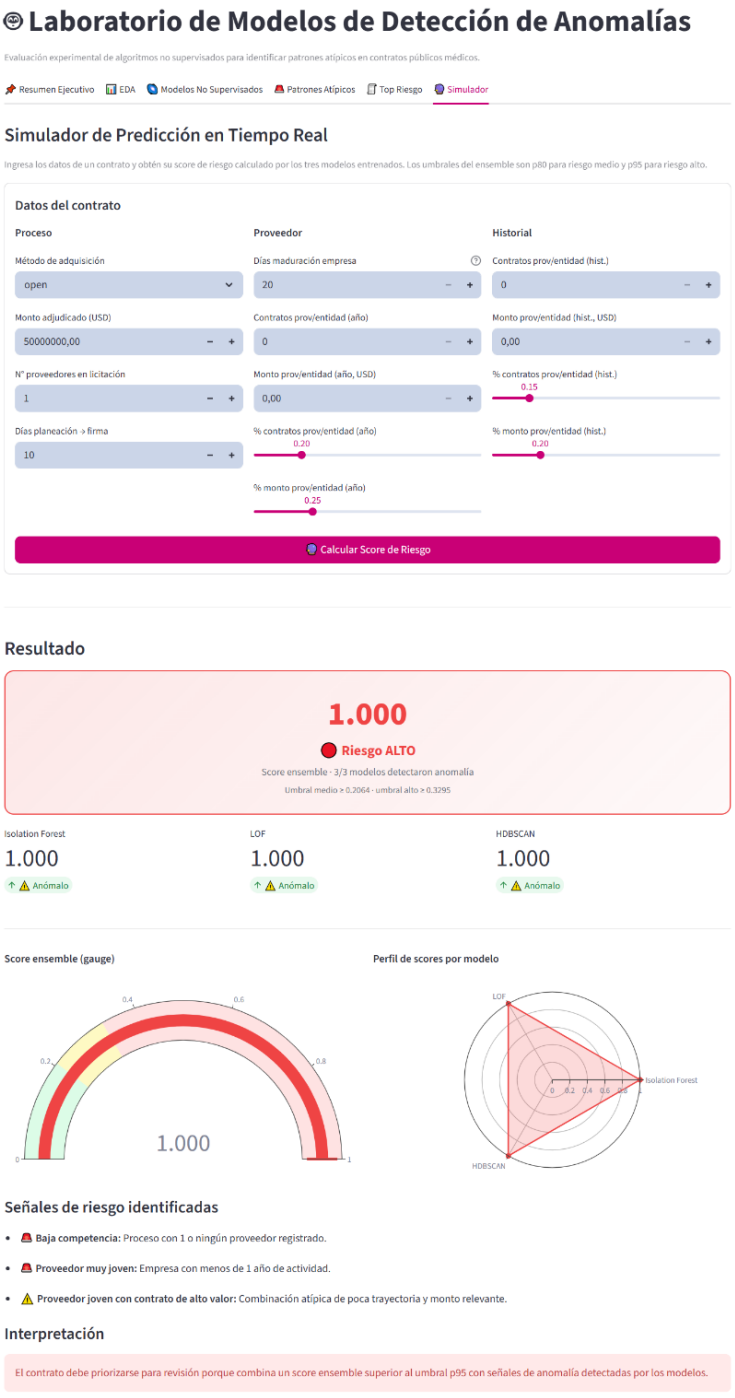


Figura 26

Simulador de Predicción – Riesgo Alto



CAPITULO 4:

4. ANÁLISIS DE RESULTADOS

Los resultados presentados en este capítulo permiten verificar el cumplimiento de los objetivos específicos de la investigación.

Tabla 10

Variables Numéricas del dataset

Variable	Descripción	Unidad o escala	Relevancia para la detección de anomalías
award_amount	Valor económico adjudicado al contrato.	USD	Permite identificar procesos con montos significativamente superiores o inferiores al comportamiento esperado.
licitacion_numero_proveedores	Número de proveedores que participaron en el proceso contractual.	Número de oferentes	Una baja competencia puede constituir un indicador de riesgo en determinados procesos de contratación.
monto_proveedor_entidad_anio	Monto total adjudicado a un mismo proveedor por una entidad durante el año del proceso.	USD	Permite identificar concentraciones inusuales de contratación en un período determinado.
monto_proveedor_entidad_hist	Monto histórico acumulado contratado entre una entidad y un proveedor.	USD	Evalúa relaciones comerciales recurrentes que podrían representar patrones atípicos de contratación.
contratos_proveedor_entidad_hist	Número histórico de contratos adjudicados entre una entidad y un proveedor.	Número de contratos	Complementa la variable anterior al medir la frecuencia de contratación entre ambas partes.
dias_maduracion_empresa	Tiempo transcurrido desde la constitución legal de la empresa hasta la adjudicación del contrato.	Días	Permite detectar empresas recientemente creadas que obtienen contratos en plazos inusualmente cortos.
dias_entre_planeacion_firma_contrato	Número de días entre la planificación del proceso y la firma del contrato.	Días	Valores extremadamente bajos o elevados pueden evidenciar comportamientos atípicos en la ejecución del proceso contractual.

En primer lugar, se evidencia la consolidación y preparación del repositorio de datos mediante el análisis exploratorio de la información obtenida del dataset preparado. Posteriormente,

se presentan los resultados derivados del entrenamiento de los algoritmos de detección de anomalías y de la construcción del modelo de scoring de riesgo. A continuación, se analizan los procesos contractuales identificados como atípicos a partir de variables económicas, temporales y relacionales. Finalmente, se compara el desempeño de los modelos mediante métricas de validación para aprendizaje no supervisado y se discuten los principales hallazgos obtenidos, los cuales sirven de base para las conclusiones y recomendaciones presentadas en el capítulo siguiente.

La Tabla 10 presenta las variables numéricas seleccionadas para el entrenamiento de los modelos de detección de anomalías. Estas variables fueron elegidas por representar características económicas, temporales y relacionales de los procesos de contratación pública que, de acuerdo con la literatura especializada y el conocimiento del dominio, pueden reflejar patrones atípicos asociados a posibles riesgos de irregularidad. Antes del entrenamiento, todas las variables fueron sometidas a un proceso de transformación mediante la función log1p y posteriormente estandarizadas, con el propósito de reducir la asimetría de sus distribuciones y minimizar la influencia de valores extremos sobre los algoritmos de aprendizaje no supervisado.

En el conjunto de datos existen 34745 procesos contractuales relacionados con productos médicos entre los años 2020 al 2025. Se observa una marcada heterogeneidad en las variables monetarias, justificada por desviaciones estándar significativamente superiores a las medias en varios casos. Es común en procesos de contratación pública evidenciar adquisiciones con montos bajos y considerablemente altos. De igual manera varias variables presentan distribuciones altamente asimétricas, con una concentración importante de observaciones en valores bajos y una pequeña cantidad de registros con valores extremadamente altos.

4.1. Tipos de Variables Utilizadas

4.1.1. Variables Relacionadas con la Competencia. La variable licitacion_numero_proveedores tiene una mediana de 3 proveedores y un promedio de 3.83

participantes por proceso o contrato. Esto indica que la mayoría de los procedimientos reciben una cantidad limitada de ofertas, aunque existen casos excepcionales que alcanzan hasta 57 proveedores que participan. Desde un punto de vista de análisis de riesgo, se podría pensar o creer que una baja concurrencia de proveedores puede constituir una señal relevante relacionada a la competitividad dentro de los procesos de contratación.

4.1.2. Variables Monetarias. La variable award_amount presenta los siguientes resultados (medidos en USD), observándose una diferencia significativa entre la media y la mediana, lo que indica una distribución fuertemente sesgada hacia la derecha. Este comportamiento evidencia la existencia de un número reducido de contratos de elevado valor económico que incrementan considerablemente el promedio de la distribución. Un patrón similar se observa en las variables monto_base_modelo, monto_proveedor_entidad_anio, monto_total_entidad_anio, monto_proveedor_entidad_hist y monto_total_entidad_hist.

- Promedio: 93101.76
- Mediana: 26729.10
- Percentil 75: 70868.75
- Valor máximo: 24249958.99

No obstante, la presencia de valores extremos no implica necesariamente la existencia de irregularidades, ya que algunos contratos pueden corresponder legítimamente a procesos de gran magnitud, como adquisiciones consolidadas de medicamentos, equipamiento hospitalario especializado, construcción o adecuación de infraestructura sanitaria, especialmente durante el período analizado (2020–2025), en el cual se registraron contrataciones extraordinarias asociadas a la emergencia sanitaria provocada por la COVID-19. Por esta razón, el monto adjudicado no fue considerado de manera aislada como un indicador de riesgo, sino que fue analizado conjuntamente con variables de competencia, temporalidad y relaciones entre entidades y proveedores dentro del modelo de detección de anomalías.

4.1.3. Antigüedad Empresarial. La variable días_maduracion_empresa (medida en días) muestra que la mayoría de los proveedores poseen varios años de experiencia, sin embargo, también se observan aquellos que tienen 32 días de antigüedad; de igual manera, podría considerarse un factor de riesgo cuando se podrían combinar procesos con montos altos y empresas con pocos días de antigüedad.

- Promedio: 5206.07
- Mediana: 4906
- Percentil 75: 7264

4.1.4. Concentración Proveedor – Entidad. Las variables: porcentaje_monto_proveedor_entidad_anio y porcentaje_monto_proveedor_entidad_hist miden la concentración del gasto público de las entidades en determinados proveedores, aquí se puede observar:

- La mediana anual es aproximadamente 3,23%
- La mediana histórica es aproximadamente 0,98%

No obstante, ambas variables alcanzan valores máximos cercanos al 100% lo que indica que existen casos donde casi todo el gasto de una entidad se concentra en un único proveedor. Este tipo de comportamiento es ventajoso para los modelos de detección de anomalías ya que una marcada concentración puede significar la existencia de patrones atípicos en la dinámica de la contratación pública.

4.1.5. Historial de Contrataciones. Las variables: total_contratos_entidad_hist y contratos_proveedor_entidad_hist permiten evaluar la relación histórica entre proveedores y entidades contratantes. Los resultados que se tienen con ellas son:

- La gran mayoría de relaciones entre proveedores y entidades tienen pocos contratos acumulados
- Existen casos extremos con cientos o miles de contratos históricos

De aquello se puede indicar que la variabilidad es consistente con el fenómeno de concentración observado en las variables monetarias y puede facilitar la identificación de proveedores con niveles inusualmente altos de participación dentro de una entidad en particular.

4.2. Dispersión y potencial presencia de valores atípicos

Las mayores desviaciones estándar se pueden observar en las variables relacionadas con:

- Montos adjudicados
- Montos históricos entre proveedores y entidades
- Montos acumulados por entidad

Aquí se puede también observar que existe una elevada dispersión, lo que muestra la existencia de observaciones alejadas del comportamiento promedio. Es por ello por lo que se justifica la selección de los modelos de detección de anomalías como:

- Isolation Forest
- HDBSCAN
- Local Outlier Factor (LOF)

Al usar estos algoritmos se puede determinar si los valores extremos corresponden a posibles patrones atípicos o por el contrario se trata de comportamientos propios del mercado.

4.3. Transformación Logarítmica de Variables Numéricas

El análisis exploratorio evidenció que varias variables cuantitativas presentan distribuciones fuertemente asimétricas hacia la derecha, debido a la presencia de un reducido número de contratos de elevado valor económico. Esta característica puede afectar el desempeño de algunos algoritmos de detección de anomalías al incrementar la influencia de los valores extremos. Para reducir dicha asimetría se aplicó la transformación logarítmica mediante la función:

Ecuación 16

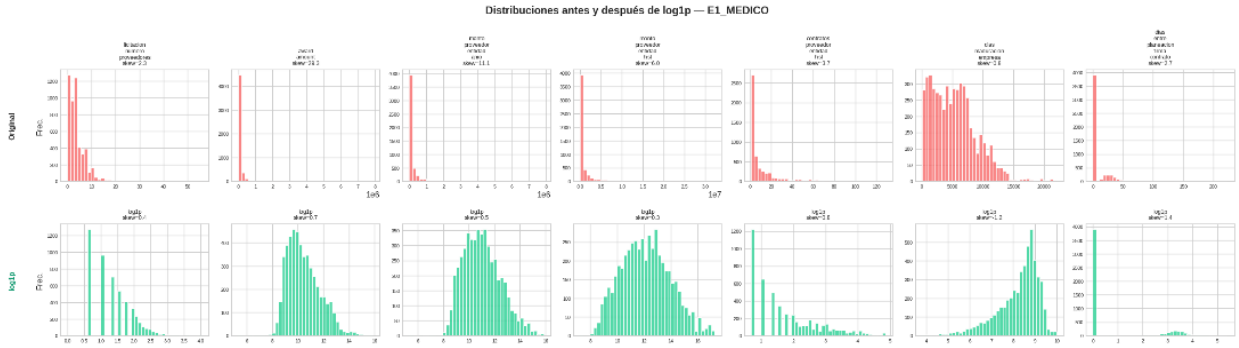
Transformación logarítmica utilizada para reducir la asimetría

$$x' = \log(1 + x)$$

Se seleccionó la función log1p porque permite transformar variables que contienen valores iguales a cero, evitando los problemas matemáticos asociados al cálculo del logaritmo de cero, además de comprimir el rango de valores elevados sin alterar el orden relativo de las observaciones. Si bien existen otras transformaciones como Box-Cox o Yeo-Johnson, la primera requiere variables estrictamente positivas y la segunda incorpora una parametrización adicional que no resultó necesaria, dado que las variables monetarias del estudio presentan únicamente valores no negativos y una marcada asimetría positiva. En consecuencia, log1p constituye una alternativa simple, robusta y ampliamente utilizada en Ciencia de Datos para estabilizar la varianza y reducir el efecto de valores extremos antes del entrenamiento de modelos de aprendizaje automático. La transformación no se aplicó indiscriminadamente, sino únicamente a las variables donde el análisis exploratorio evidenció una asimetría significativa. En la Figura 27 se muestra el resultado de la aplicación de la transformación logarítmica.

Figura 27

Distribuciones antes y después de log1p



La transformación de datos es usada en muchos campos de la estadística y el análisis de datos para mejorar la calidad de los datos o lograr que el análisis sea más fácil. Esto se logra aplicando funciones matemáticas lineales y no lineales a los datos. Dentro de las técnicas no

lineales se encuentran el uso de la transformación logarítmica, la misma que convierte o transforma los datos que siguen una distribución asimétrica hacia la derecha (right-skewed distribution) en una distribución más simétrica al tomar el logaritmo de los valores de los datos (Subramanian, 2025, pp. 164–165).

El resultado de aplicar esta transformación se observa en la reducción significativa de la asimetría en prácticamente todas las variables. Los principales efectos son:

- Se puede comprender mejor a los valores extremos: Los contratos con montos muy altos dejan de dominar la escala de la variable; con esto se evita que unos pocos registros de contratos extremos condicionen el comportamiento de los algoritmos.
- Se aproxima a distribuciones simétricas: Las variables transformadas muestran formas más cercanas a distribuciones unimodales y aproximadamente normales.
- Se conserva el orden relativo: La transformación no altera la jerarquía de los registros. Un contrato de mayor valor seguirá siendo mayor que otro de menor cuantía. De igual manera se mantienen las relaciones relativas necesarias para los modelos de ML.
- Implicación para la detección de anomalías: La normalización de la escala es particularmente importante para algoritmos basados en distancia o densidad como: LOF y HDBSCAN; debido a que estos modelos son sensibles a variables con rangos muy diferentes. De igual manera la transformación ayuda a que Isolation Forest identifique patrones atípicos estructurales y no solamente observaciones con montos muy altos.

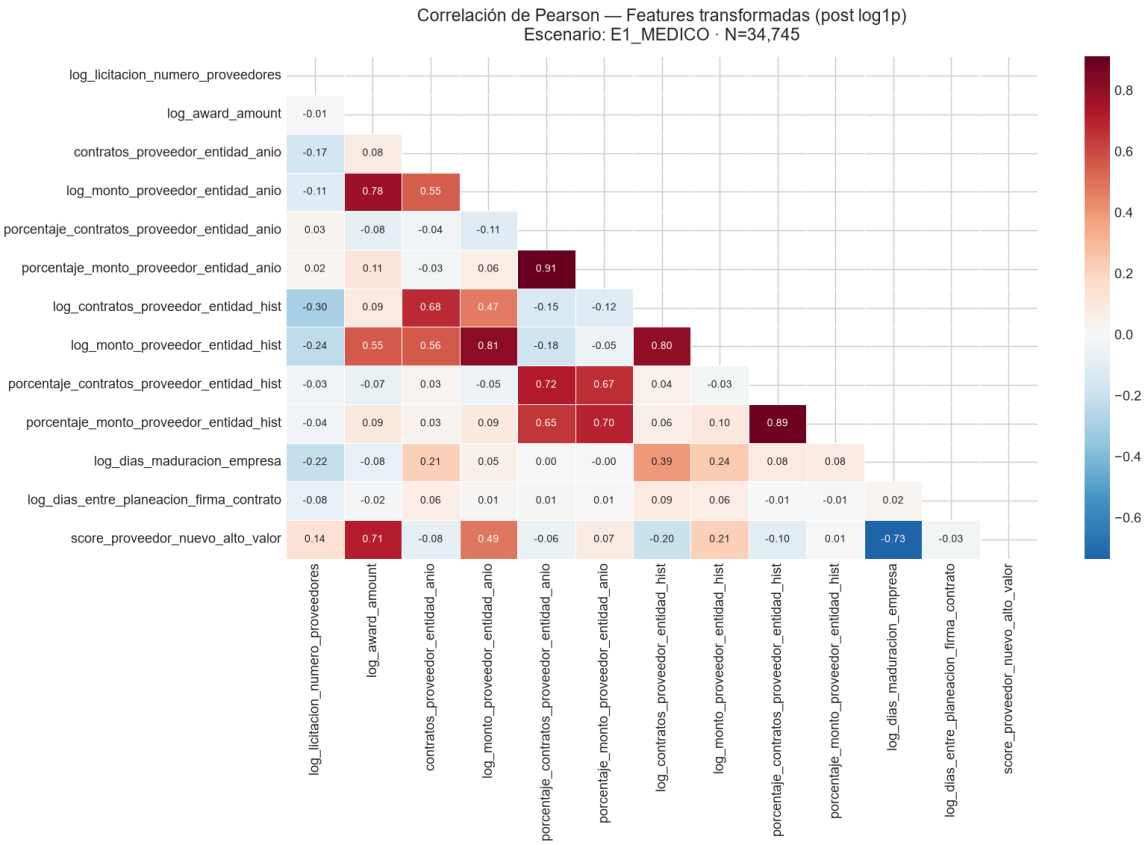
4.4. Matriz de correlación de Pearson

En la Figura 28 se presenta la matriz de correlación de Pearson correspondiente a las variables transformadas mediante la función log1p(), con el propósito de identificar relaciones lineales entre ellas. La transformación logarítmica aplicada previamente permitió reducir la fuerte asimetría observada en las variables monetarias y disminuir la influencia de valores extremos, condiciones que favorecen una interpretación más estable del coeficiente de Pearson. Si bien este

coeficiente asume relaciones aproximadamente lineales y es sensible a observaciones extremas, su utilización en esta investigación tiene un carácter exploratorio, orientado a identificar dependencias, redundancias y posibles patrones estructurales entre las variables utilizadas en el entrenamiento de los modelos. Aquí se aprecia el mapa de calor de correlaciones, esto ayuda a identificar dependencias, redundancias y posibles patrones estructurales dentro del dataset. Se recuerda que la matriz es cuadrada y simétrica, en donde cada fila y cada columna representa una variable diferente, el valor en la intersección de cada fila y columna es la correlación entre las dos variables. En la práctica es común mostrar sola la parte inferior o superior de la matriz, los elementos de la diagonal principal también suelen omitirse ya que todos tiene el valor de 1 lo que significa las correlaciones de las variables consigo mismas. (Tabachnick y Fidell, 2014, p. 45)

Figura 28

Correlación entre variables para el Escenario 1



Nota. Para generar la matriz de correlaciones se tomaron en cuenta tres escenarios.

El escenario 1 mostrado en la Figura 28 considera los contratos con insumos médicos, el escenario 2 contratos sin insumos médicos, y el escenario 3 todos los contratos del sector salud.

La gran mayoría de correlaciones, mostradas en la Tabla 11, se encuentran dentro de rangos moderados, lo que facilitará la detección de anomalías, ya que las variables o features aportan información complementaria. Básicamente se resumen en tres grupos de variables:

- Variables de concentración proveedor y entidad
- Variables de historial de contratación
- Variables económicas relacionadas a montos adjudicados

Tabla 11

Correlaciones encontradas entre las variables – Escenario E1_MEDICO

Grupo de variables	Coeficiente	Explicación
Relación entre concentración monetaria y contractual	r = 0.91	Uno de los hallazgos más relevantes que se encuentran es la fuerte correlación entre: porcentaje_contratos_proveedor_entidad_anio y porcentaje_monto_proveedor_entidad_anio. Esta correlación muestra que los proveedores que reciben una mayor proporción de contratos dentro de una entidad también suelen concentrar una mayor proporción del presupuesto adjudicado. En contratación pública este resultado es consistente con escenarios en donde ciertos proveedores mantienen relaciones periódicas y de impacto económico relevante con determinadas entidades.
Constancia histórica de las relaciones entre proveedores y entidades	r = 0.89	Las variables: porcentaje_contratos_proveedor_entidad_hist y porcentaje_monto_proveedor_entidad_hist tiene una correlación elevada. Esto indica que la concentración observada en un período específico suele mantenerse a lo largo del tiempo, es decir, las

Grupo de variables	Coeficiente	Explicación
		relaciones de contratación entre proveedores y entidades tienden a presentar cierta estabilidad histórica. Dentro del análisis de patrones atípicos esto es muy importante porque permite distinguir entre relaciones históricamente consolidadas y concentraciones inusuales o recientes.
Historial contractual y montos adjudicados	r = 0.80	Se observa una correlación significativa entre las variables: log_contratos_proveedor_entidad_hist y log_monto_proveedor_entidad_hist. Este resultado indica que los proveedores con mayor número de contratos históricos suelen acumular también los mayores montos adjudicados; esto es consistente con las leyes del mercado económico en donde se puede evidenciar que a medida que aumenta la frecuencia de contratación también aumenta el volumen económico asociado.
Relación entre experiencia previa y actividad anual	r = 0.78	Entra las variables: contratos_proveedor_entidad_anio y log_monto_proveedor_entidad_anio también se puede ver una correlación alta. Esto indica que la actividad contractual y el valor adjudicado evoluciona conjuntamente. De todas maneras, la correlación no alcanza valores que se acerquen a 1, lo que revela que existen proveedores con pocos contratos, pero con montos elevados; así también se encontrarán proveedores con numerosos contratos con pequeños montos adjudicados.
Antigüedad empresarial o económica	r = 0.39	La variable: log_dias_maduracion_empresa presenta correlaciones relativamente bajas con la mayoría de las otras variables. De esas relaciones se destaca la que mantiene con la variable: log_contratos_proveedor_entidad_hist. Esto indica

Grupo de variables	Coeficiente	Explicación
		que las empresas más antiguas tienden a acumular más experiencia en contratación pública.
Proveedor nuevo con alto valor	r = 0.71 r = - 0.73	La variable: score_proveedor_nuevo_alto_valor muestra una correlación positiva relativamente alta con log_award_amount y una correlación negativa con log_dias_maduracion_empresa. Estos resultados indican que la construcción de esta variable fue acertada conceptualmente, ya que se busca identificar proveedores con poca trayectoria y que recibieron adjudicaciones relevantes. Las correlaciones enseñan que la métrica captura ambas dimensiones, por un lado, montos altos adjudicados y baja antigüedad empresarial.
Ausencia de multicolinealidad		Si bien existen correlaciones elevadas entre algunas variables de concentración, no se observan coeficientes cercanos a 1 (positivo o negativo) que indiquen duplicación total de información. Esto resulta útil en el aprendizaje no supervisado porque se conserva la diversidad de información del conjunto de variables, además que se reduce el riesgo de que una sola dimensión domine el cálculo de distancias o densidades; de esta manera los algoritmos pueden capturar anomalías multidimensionales más complejas.

4.5. Análisis de Resultados

4.5.1. Resultados del Modelo Isolation Forest. Con el objetivo de identificar procesos de contratación pública con patrones atípicos, se aplica el algoritmo Isolation Forest sobre las variables transformadas y normalizadas del conjunto de datos, ésta poderosa técnica de machine learning funciona dividiendo el conjunto de datos en subconjuntos más pequeños según ciertos

valores de variables para facilitar la predicción de la clasificación o regresión en cada subconjunto. Con este algoritmo se aprovecha la facilidad con la que se puede separar un valor atípico de la mayoría de los casos, basándose en las diferencias entre sus valores o combinaciones de valores. El algoritmo registra el tiempo que tarda en separar un caso de los demás y colocarlo en su propio subconjunto. Cuanto menor sea el esfuerzo necesario para separarlo, mayor será la probabilidad de que se trate de un valor atípico. Como medida de dicho esfuerzo, Isolation Forest genera una medida de distancia: cuanto menor sea la distancia, mayor será la probabilidad de que el caso sea un valor atípico. (Mueller y Massaron, 2024, cap. 16, p. 303)

En la Figura 29 se puede visualizar simultáneamente la distribución del riesgo, la separación entre contratos normales y los anómalos, así como las características que explican dicho riesgo. En la Tabla 12 se explican los resultados obtenidos.

Figura 29

Isolation Forest - Análisis de Riesgo en Contratos Públicos

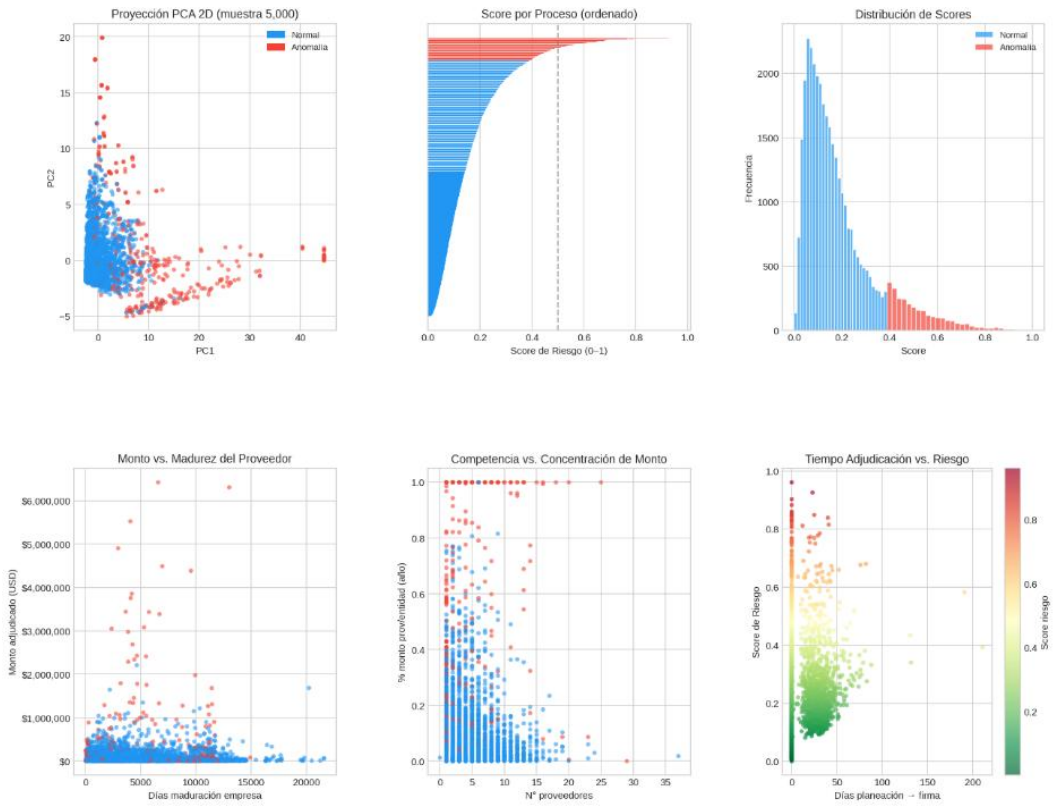


Tabla 12

Explicación de resultados obtenidos utilizando Isolation Forest

Gráfico	Resultados obtenidos
Distribución General del Riesgo Detectado (gráfico superior central)	El gráfico ordena los procesos según su score de riesgo. Se observa una distribución continua donde: La mayoría de los contratos presenta valores de riesgo bajos Se evidencia una cola de observaciones con scores progresivamente mayores Se ve también que un subconjunto reducido de procesos supera el umbral de anomalía definido por el modelo
Distribución de Scores (Gráfico superior derecho)	La distribución muestra una fuerte concentración de observaciones en rangos de riesgo bajos. La mayor densidad se encuentra entre 0.05 y 0.25 mientras que los contratos públicos clasificados como anómalos aparecen por encima de 0.40. El patrón observado sugiere que las anomalías identificadas no constituyen simples fluctuaciones aleatorias, sino observaciones diferenciadas del comportamiento general. La separación visual entre ambas distribuciones indica que el modelo logró discriminar adecuadamente entre contratos normales y los potencialmente atípicos.
Separación en el Espacio PCA (Gráfico superior izquierdo)	La proyección PCA (Principal Component Analysis o Análisis de Componentes Principales) permite visualizar los contratos en dos dimensiones conservando la mayor cantidad posible de variabilidad de los datos originales. Aquí se puede observar que: Los contratos normales (azul) se concentran en una región compacta. Las anomalías (rojo) aparecen dispersas alrededor de los bordes del espacio. Los casos de mayor riesgo se localizan en regiones de baja densidad

Gráfico	Resultados obtenidos
	Este comportamiento coincide exactamente con la teoría de Isolation Forest. El gráfico indica que el modelo no está generando anomalías arbitrariamente, sino identificando regiones poco pobladas del espacio multidimensional.
Monto Adjudicado vs. Madurez Empresarial (Gráfico inferior izquierdo)	Aquí se puede observar que: La mayoría de los contratos normales corresponden a proveedores con varios años de existencia y montos moderados. Muchas anomalías aparecen asociadas a contratos de alto valor económico. Existen proveedores relativamente jóvenes que reciben adjudicaciones millonarias. Estos casos coinciden con uno de los patrones de riesgo definidos durante la construcción de variables: Proveedor nuevo + contrato de alto valor Esto no implica necesariamente irregularidad, pero sí un escenario suficientemente inusual para justificar una revisión o auditoria adicional.
Competencia vs. Concentración de Monto (Gráfico inferior central)	Aquí se puede ver las relaciones entre: Número de proveedores participantes. Porcentaje del monto concentrado en un proveedor. También se observa que numerosas anomalías se ubican en escenarios caracterizados por: baja competencia y alta concentración del gasto.
Tiempo de Adjudicación vs. Riesgo (Gráfico inferior derecho)	En el gráfico se muestra una relación positiva entre: Días transcurridos entre planificación y firma del contrato. Score de riesgo. Se observa que los valores de riesgo más elevados suelen concentrarse en procesos con tiempos extremadamente cortos o largos respecto al comportamiento típico. Esto sugiere que el modelo está capturando desviaciones respecto a los ciclos administrativos normales de contratación.

En términos generales, Isolation Forest procesó 34745 procesos contractuales sin requerir etiquetado previo, con lo que se demuestra que el algoritmo tiene la capacidad para identificar observaciones atípicas en el dataset de estudio.

Para el entrenamiento del algoritmo Isolation Forest se configuraron los hiperparámetros mostrados en la

Tabla 13. Se utilizaron 200 árboles de aislamiento (n_estimators = 200) con el fin de obtener una estimación estable del puntaje de anomalía sin incrementar excesivamente el costo computacional. El parámetro max_samples se estableció en 5000 observaciones, permitiendo que cada árbol se construya sobre una muestra aleatoria suficientemente representativa del conjunto de datos y favoreciendo la diversidad del bosque. Asimismo, el parámetro contamination se fijó en 0.08 (8%), valor obtenido tras un análisis de sensibilidad que permitió evitar una sobreestimación del número de anomalías y obtener una proporción de procesos de alto riesgo más consistente con el comportamiento observado en los datos. Finalmente, se empleó random_state = 42 para garantizar la reproducibilidad del experimento y n_jobs = -1 para aprovechar todos los núcleos disponibles durante el entrenamiento.

Tabla 13

Hiperparámetros utilizados en Isolation Forest

Hiperparámetro	Valor utilizado	Justificación
n_estimators	200	Proporciona estabilidad en el cálculo del puntaje de anomalía manteniendo un tiempo de entrenamiento adecuado.
max_samples	5000	Permite construir cada árbol utilizando una muestra representativa, reduciendo el costo computacional en conjuntos de datos grande.
contamination	0.08 (8%)	Define el porcentaje esperado de anomalías; se seleccionó tras un análisis de sensibilidad para evitar una sobreestimación de procesos atípicos.
random_state	42	Garantiza la reproducibilidad de los resultados obtenidos.
n_jobs	-1	Utiliza todos los núcleos disponibles del procesador para acelerar el entrenamiento.

El valor del parámetro contamination fue seleccionado después de realizar un análisis de sensibilidad considerando los valores 5%, 8%, 10% y 15%. Como resultado, el valor de 8% produjo una separación más consistente entre observaciones normales y anómalas, evitando al mismo tiempo clasificar un número excesivo de procesos como riesgosos. Isolation Forest requiere la definición de varios hiperparámetros que influyen en el proceso de detección de

anomalías. En esta investigación dichos parámetros fueron seleccionados considerando el tamaño del conjunto de datos y un análisis de sensibilidad sobre el porcentaje esperado de anomalías, buscando un equilibrio entre estabilidad, capacidad de discriminación y eficiencia computacional.

4.5.2. Resultados del Modelo HDBSCAN. Con el objetivo de identificar agrupaciones naturales dentro de los procesos de contratación pública y detectar observaciones que no pertenecen a ninguna estructura de comportamiento estable, se aplicó el algoritmo HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise o Agrupamiento Espacial Jerárquico Basado en Densidad de Aplicaciones con Ruido). La detección de valores atípicos basados en clústeres utiliza algoritmos de agrupamiento estándar como:k-means, DBSCAN o HDBSCAN entre otros. En este caso HDBSCAN es similar a DBSCAN, pero admite agrupamiento jerárquico, y es más resistente al ruido. (Kennedy, 2025, p. 71).

HDBSCAN es un algoritmo de agrupamiento diseñado para descubrir clústers en conjunto de datos basándose en la distribución de densidad de los puntos de datos. En comparación con otros métodos de agrupamiento, no se requiere especificar el número de clústers con anterioridad, lo que lo hace más adaptable a diferentes datasets. Utiliza regiones de alta densidad para identificar clústers y considera los puntos aislados o de baja densidad como ruido. Este modelo es especialmente útil para conjuntos de datos con estructuras complejas o densidades variables, ya que crea un árbol jerárquico de clústeres que permite a los usuarios examinar los datos con diferentes niveles de granularidad. (GeeksforGeeks, 2025).

En la Figura 30 se pueden apreciar los resultados de la aplicación de este modelo, y en la Tabla 14 la explicación de los resultados obtenidos.

Figura 30

HDBSCAN Segmentado - Clústeres y Anomalías en Contratos Públicos

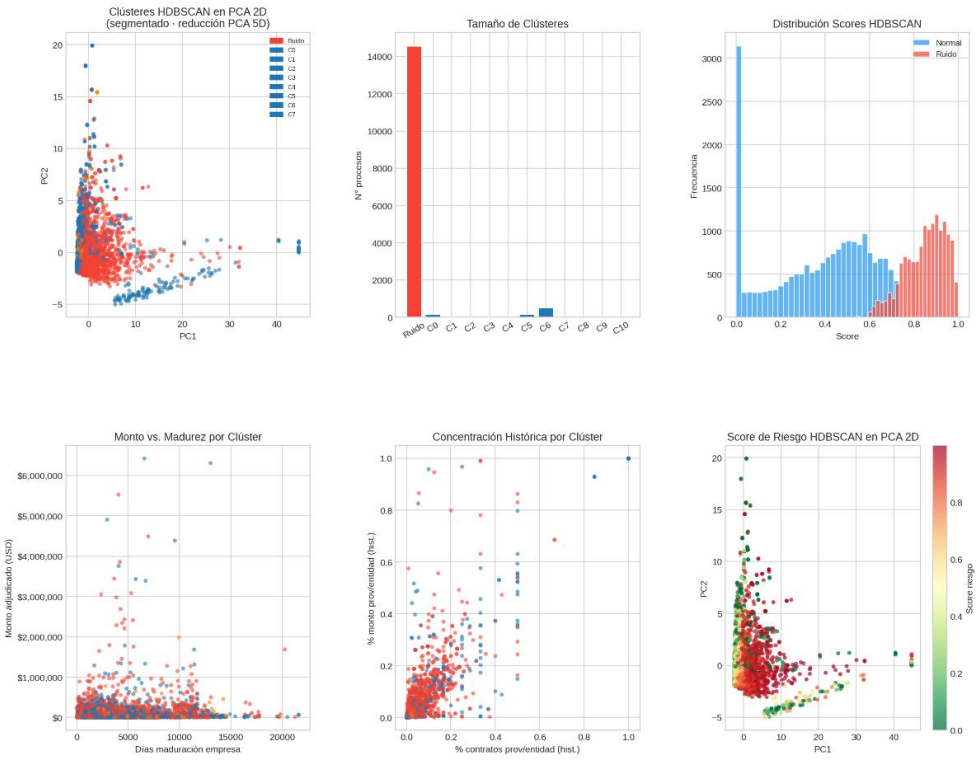


Tabla 14

Explicación de resultados obtenidos utilizando HDBSCAN

Gráfico	Resultados obtenidos
Formación de clústeres y ruido (Gráfico superior izquierdo)	La proyección PCA muestra la segmentación obtenida por el algoritmo, se observa que: El modelo identificó múltiples grupos diferenciados (C0–C10). La mayoría de las observaciones se concentran en una gran región central. Existe un número importante de registros clasificados como "ruido". Este resultado es relevante porque muestra que el mercado de contratación pública no es homogéneo, sino que está compuesto por varios patrones de distinta índole. En los

Gráfico	Resultados obtenidos
	clústers encontrados se puede interpretar como diferentes escenarios de contratación como: Compras pequeñas y recurrentes. Contrataciones de proveedores consolidados. Procesos con alta concentración. Contratos con grandes montos adjudicados. Por otro lado, las observaciones catalogadas como ruido representan procesos que no encajan claramente en ninguno de esos comportamientos habituales.
Distribución de tamaños de clúster **(Gráfico superior central)**	Aquí se observa una marcada desproporción entre el tamaño de los clústers detectados, el modelo identifica: Un gran conjunto clasificado como ruido. Algunos clústeres medianos. Varios clústeres pequeños y especializados.
Distribución de Scores HDBSCAN **(Gráfico superior derecho)**	La distribución de scores muestra una clara separación entre: Observaciones normales. Observaciones clasificadas como ruido. Se aprecia que los contratos normales presentan valores preferentemente intermedios y el ruido se concentra en scores elevados. Esto sugiere que las observaciones identificadas como atípicas se encuentran alejadas de las regiones densas en donde se ubican la mayoría de los procesos contractuales. En comparación con Isolation Forest, aquí el riesgo está asociado a la falta de pertenencia a un grupo estructurado de contratos similares.
Monto adjudicado y madurez empresarial **(Gráfico inferior izquierdo)**	Aquí la relación entre monto adjudicado y antigüedad empresarial muestra una elevada heterogeneidad dentro de los clústeres. Se observa que: Los contratos de alto valor económico aparecen distribuidos en distintos grupos. Algunos registros con montos millonarios quedan aislados respecto al comportamiento dominante.

Gráfico	Resultados obtenidos
	Existen proveedores relativamente jóvenes asociados a adjudicaciones significativas. Esto confirma que el monto adjudicado por sí solo no explica completamente la estructura de los datos, por lo que el modelo está considerando simultáneamente múltiples dimensiones para determinar la pertenencia a cada clúster.
Concentración histórica proveedor–entidad (Gráfico inferior central)	Aquí se observa que: Participación histórica del proveedor. Concentración del gasto adjudicado. Algunos clústeres muestran niveles elevados de concentración histórica, mientras que otros representan relaciones más diversificadas. Desde el punto de vista de riesgo contractual, los puntos alejados de la tendencia principal podrían pertenecer a: Proveedores dominantes dentro de determinadas entidades. Relaciones contractuales altamente concentradas. Escenarios de dependencia institucional respecto a un número reducido de adjudicatarios.
Distribución espacial del riesgo (Gráfico inferior derecho)	La proyección PCA coloreada por score de riesgo permite visualizar dónde se localizan las observaciones más atípicas. Aquí se puede observar que: Los mayores niveles de riesgo aparecen en regiones periféricas. Las zonas centrales presentan scores bajos. El riesgo aumenta conforme disminuye la densidad local.

Los resultados sugieren que los procesos considerados de mayor riesgo no necesariamente corresponden a contratos con montos extraordinarios, sino a procesos que presentan combinaciones de características poco frecuentes respecto a los patrones habituales del mercado.

HDBSCAN aporta una visión estructural del riesgo, permitiendo identificar contratos que operan fuera de los grupos naturales de comportamiento observados en el sistema de contratación pública.

El algoritmo requiere la definición de hiperparámetros relacionados con la densidad mínima necesaria para formar clústeres, en la Tabla 15 se muestran los hiperparámetros usados en este algoritmo. En esta investigación se empleó una estrategia de agrupamiento segmentado según el método de adquisición, permitiendo adaptar el tamaño mínimo de los clústeres a las características de cada modalidad de contratación pública. Adicionalmente, antes del entrenamiento se aplicó un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad del conjunto de datos a cinco componentes principales, mejorando la eficiencia computacional sin perder la estructura general de la información. A diferencia de una implementación convencional de HDBSCAN, donde se emplea un único valor para min_cluster_size, en esta investigación dicho parámetro se ajustó de acuerdo con el método de adquisición de cada proceso contractual. Esta estrategia permitió adaptar el algoritmo a las diferencias de tamaño existentes entre las distintas modalidades de contratación pública, evitando que segmentos con pocos registros fueran clasificados de manera sistemática como ruido.

Tabla 15

Hiperparámetros utilizados en HDBSCAN

Hiperparámetro	Valor utilizado	Justificación
min_cluster_size	Variable por segmento (15, 8, 3 y 3)	Se adaptó al tamaño de cada modalidad de contratación para obtener agrupamientos representativos.
min_samples	5	Controla el nivel de densidad requerido para formar un clúster y mejora la robustez frente al ruido.
metric	Manhattan	Adecuada para variables escaladas y menos sensible a valores extremos que la distancia euclidiana.
PCA previo	5 componentes	Reduce la dimensionalidad y mejora el rendimiento computacional.
core_dist_n_jobs	-1	Utiliza todos los núcleos disponibles del procesador.

Respecto a la elección de la distancia Manhattan, se recuerda que HDBSCAN admite diferentes métricas de distancia (euclidiana, Manhattan, Minkowski, entre otras). Se seleccionó Manhattan porque, tras aplicar RobustScaler y log1p, las variables conservaban distribuciones heterogéneas y aún podían existir diferencias importantes entre dimensiones. La distancia Manhattan suele ser más robusta que la euclidiana frente a estas condiciones y reduce la influencia de observaciones extremas sobre el cálculo de las distancias.

4.5.3. Resultados del Modelo LOF. Con el objetivo de identificar procesos contractuales cuyo comportamiento difiere significativamente del de sus vecinos más próximos, se

aplicó el algoritmo Local Outlier Factor (LOF) o Factor de Valores Atípicos Locales. LOF es probablemente el método basado en densidad más conocido. Evalúa cada punto estimando la densidad en su ubicación, utilizando las distancias a los puntos más cercanos. Cuantos menores sean las distancias a los vecinos, mayor será la densidad en la ubicación del punto; cuanto mayor sea la distancia, menor será la densidad. Para puntuar cada punto, LOF compara la densidad de cada punto con la densidad de los demás puntos en su vecindario. Es decir, utiliza un estándar local, no global, para determinar qué nivel de densidad es normal (Kennedy, 2025, cap. 5, p. 132).

La Figura 31 muestra los principales resultados obtenidos, y en la Tabla 16 su explicación.

Figura 31

Local Outlier Factor - Anomalías por Densidad Local

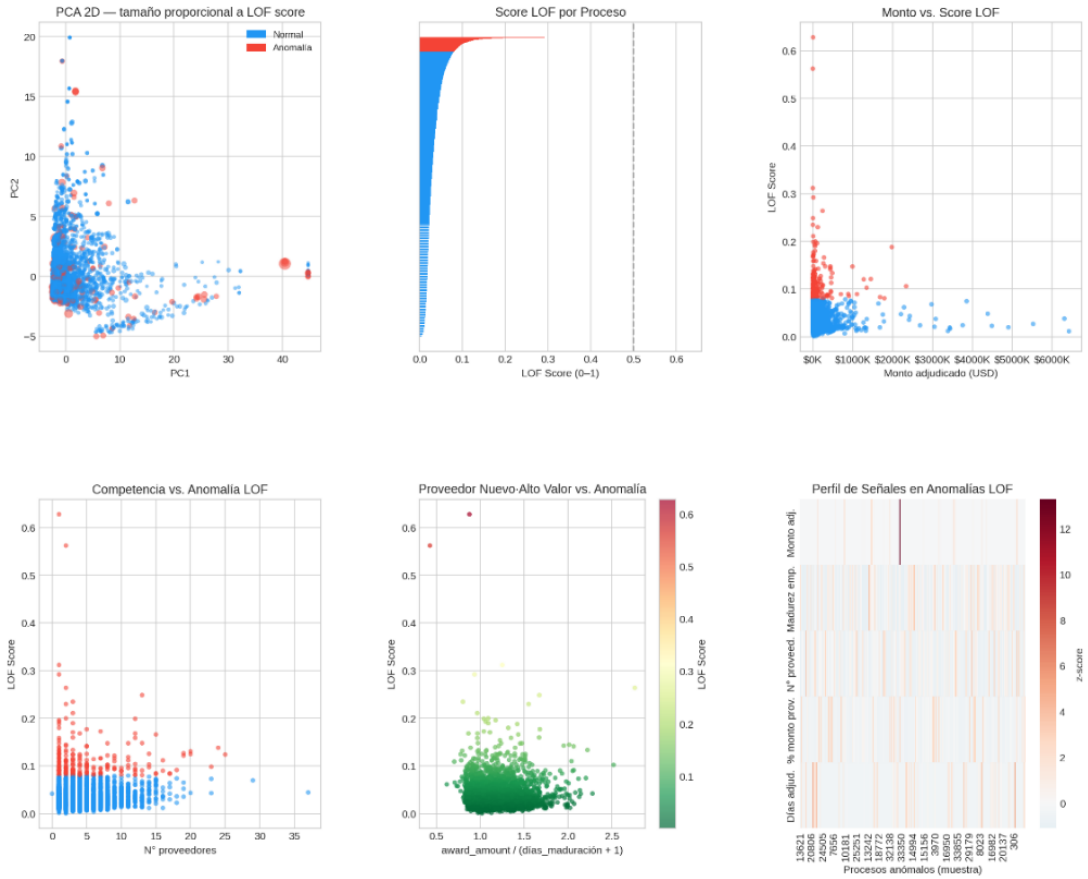


Tabla 16

Explicación de resultados obtenidos utilizando Local Outlier Factor

Gráfico	Resultados
Distribución espacial de las anomalías (Gráfico superior izquierdo)	La proyección PCA muestra la ubicación de los contratos clasificados como normales y anómalos, aquí se observa que: La mayoría de los procesos se concentra en regiones densas del espacio. Las anomalías aparecen principalmente en zonas periféricas. Existen algunos puntos claramente aislados respecto al resto de observaciones. En comparación con Isolation Forest, donde las anomalías pueden encontrarse distribuidas en múltiples regiones, LOF resalta las observaciones que presentan un comportamiento inusual dentro de su contexto local.
Distribución de Scores LOF (Gráfico superior central)	La distribución de puntuaciones muestra una fuerte concentración de observaciones con scores bajos y una pequeña cola de valores elevados. También se observa que: La gran mayoría de contratos presentan comportamientos consistentes con sus vecinos. Un subconjunto reducido recibe puntuaciones significativamente mayores. Las anomalías identificadas representan una fracción pequeña del universo estudiado.
Relación entre monto adjudicado y anomalía (Gráfico superior derecho)	El gráfico muestra la relación entre el monto adjudicado y el score LOF. Un hallazgo interesante es que los mayores scores de anomalía no necesariamente corresponden a los contratos de mayor valor económico, por el contrario que: Muchos contratos con montos moderados presentan scores elevados. Algunos contratos millonarios permanecen dentro de rangos normales.

Gráfico	Resultados
Competencia y anomalía local (Gráfico inferior izquierdo)	La relación entre número de proveedores participantes y score LOF evidencia un patrón relevante. Las anomalías se concentran principalmente en escenarios con: Baja competencia. Número reducido de oferentes. Comportamientos alejados de la tendencia predominante. Esto muestra que el modelo está capturando configuraciones contractuales poco frecuentes dentro de los procesos de baja concurrencia.
Proveedor nuevo y contrato de alto valor (Gráfico inferior central)	Este gráfico analiza una de las variables diseñadas específicamente para esta investigación: Proveedor Nuevo – Alto Valor Los mayores scores LOF aparecen concentrados en valores elevados de este indicador. Este hallazgo valida empíricamente la utilidad de la variable construida durante la etapa de ingeniería de características. Los resultados sugieren que: Proveedores con poca trayectoria y adjudicaciones significativas constituyen un patrón poco frecuente dentro de los datos. El modelo reconoce estas observaciones como desviaciones respecto al comportamiento habitual del mercado. Esto no implica necesariamente irregularidades, pero si una alerta o señal que justifica un análisis o auditoria adicional a los procesos.
Perfil de señales presentes en las anomalías (Gráfico inferior derecho)	El mapa de calor resume la intensidad relativa de las variables para una muestra de procesos identificados como anómalos. Aquí se observa que: No existe una única variable responsable de todas las anomalías. Diferentes contratos presentan combinaciones distintas de señales de riesgo. Algunas anomalías están asociadas a montos elevados. Otras responden a concentración histórica, madurez empresarial reducida o baja competencia.

El algoritmo LOF requiere definir parámetros relacionados con el cálculo de la densidad local de las observaciones. En esta investigación se empleó una implementación orientada a la detección de novedades (novelty detection), permitiendo entrenar el modelo sobre una muestra representativa del conjunto de datos y posteriormente aplicarlo al universo completo de procesos contractuales.

Tabla 17

Hiperparámetros utilizados en LOF

Hiperparámetro	Valor utilizado	Justificación
n_neighbors	min(20, max(5, N//5000))	Ajusta automáticamente el número de vecinos de acuerdo con el tamaño del conjunto de datos, garantizando suficiente información local sin incrementar innecesariamente el costo computacional.
contamination	0.08	Define el porcentaje esperado de anomalías empleado para establecer el umbral de clasificación.
novelty	True	Permite entrenar el modelo una sola vez y posteriormente evaluar nuevos registros, característica necesaria para su integración con la aplicación desarrollada.
metric	Euclidean	Calcula las distancias entre observaciones para estimar la densidad local.
n_jobs	-1	Aprovecha todos los núcleos disponibles del procesador durante el entrenamiento.

En la Tabla 17 se muestran los hiperparámetros usados con este algoritmo. Esta estrategia permitió reducir el costo computacional sin afectar el proceso de evaluación de anomalías sobre el conjunto de datos consolidado.

Debido al tamaño del conjunto de datos, el entrenamiento del modelo se realizó sobre una muestra aleatoria representativa, mientras que la predicción de anomalías se efectuó posteriormente sobre la totalidad de los registros utilizando procesamiento por lotes (batch processing) de 10 000 observaciones. Esta estrategia permitió disminuir el consumo de memoria y el tiempo de procesamiento, manteniendo la capacidad del algoritmo para asignar un puntaje de anomalía a cada proceso contractual. El parámetro n_neighbors no fue establecido mediante un valor fijo, sino calculado dinámicamente en función del tamaño del conjunto de datos, restringiendo su valor entre 5 y 20 vecinos. Esta estrategia permitió mantener un equilibrio entre la representatividad del entorno local de cada observación y la eficiencia computacional del algoritmo. La distancia euclidiana (metric='euclidean') se utilizó después de aplicar RobustScaler a las variables. Esto es importante porque la distancia euclidiana es sensible a las diferencias de escala entre variables. Al normalizar previamente los datos con RobustScaler, esa limitación se reduce significativamente y la distancia resulta adecuada para estimar la densidad local.

Los resultados obtenidos indican que los contratos considerados atípicos por LOF presentan configuraciones poco frecuentes respecto a procesos con características similares. Entre las principales señales observadas se destacan a:

- Baja competencia.
- Alta concentración proveedor y entidad.
- Proveedores relativamente nuevos.
- Combinaciones inusuales de experiencia, monto y concentración contractual.

En resumen, se puede diferenciar a los tres modelos usados en la Tabla 18:

Tabla 18

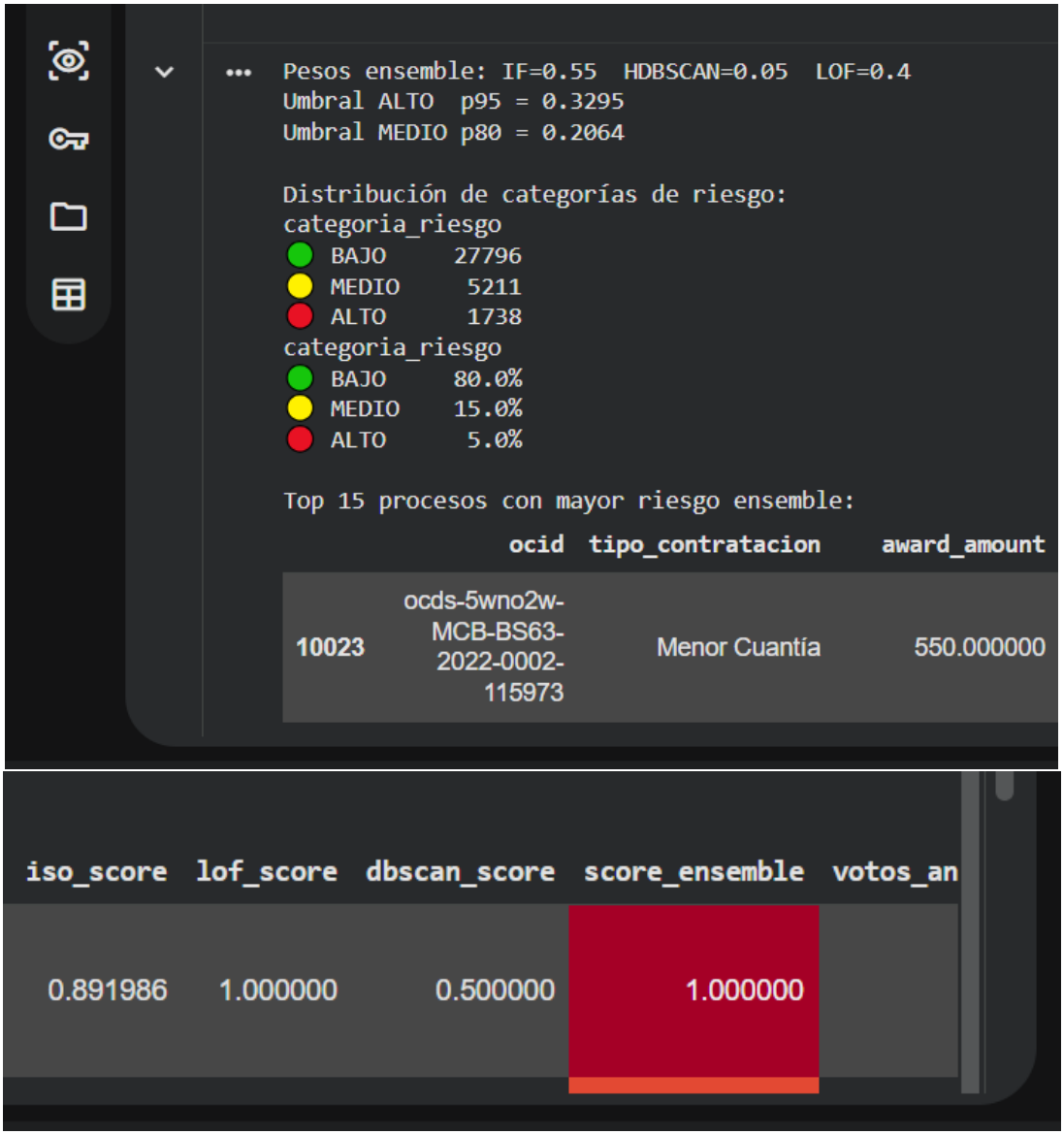
Comparación entre modelos

Modelo	Criterio	Alcance
Isolation Forest	Facilidad de aislamiento	Identifica anomalías globales
HDBSCAN	Pertenencia a regiones densas	Detecta observaciones fuera de comunidades naturales
LOF	Diferencia de densidad respecto a vecinos cercanos	Detecta anomalías locales

4.5.4. Resultado del Modelo Ensemble. El objetivo de construir este modelo ensemble o modelo de aprendizaje por conjuntos es combinar los resultados de Isolation Forest, HDBSCAN y LOF en un único score de riesgo. Se busca identificar a los procesos contractuales con mayor riesgo relativo dentro del dataset analizado. Un emsemble es un método popular que se utiliza para mejorar la precisión de diversos algoritmos usados en la minería de datos o data mining. Se combinan las salidas de múltiples algoritmos base para crear una salida unificada. La idea básica de este enfoque es que algunos algoritmos tendrán un buen rendimiento en un subconjunto específico de puntos, mientras que otros tendrán un mejor rendimiento en otros subconjuntos. No obstante, la combinación de conjuntos suele ofrecer un rendimiento más robusto en general gracias a su capacidad para combinar las salidas de múltiples algoritmos (Aggarwal, 2017, cap. 6, p. 185). En la Figura 32 se visualizan los resultados.

Figura 32

Modelo ensemble



La combinación se realizó mediante una ponderación de los scores normalizados obtenidos en cada modelo:

- Isolation Forest: 55%
- LOF: 40%
- HDBSCAN: 5%

La fórmula general utilizada fue:

Ecuación 17

Score para el método de Ensemble utilizado para el proyecto

$$Score_{Ensemble} = 0.55(IF) + 0.40(LOF) + 0.05(HDBSCAN)$$

La mayor ponderación asignada a Isolation Forest y LOF responde a que ambos algoritmos mostraron una mejor capacidad para aislar patrones atípicos dentro del conjunto de datos, mientras que HDBSCAN se utilizó principalmente como mecanismo complementario de validación estructural. La distribución de categorías de riesgo tuvo los valores mostrados en la Tabla 19.

Tabla 19

Distribución de categorías

Categoría	Procesos	Porcentaje
Bajo	27796	80%
Medio	5211	15%
Alto	1738	5%

Los resultados demuestran una distribución altamente concentrada en la categoría de bajo riesgo. Para la definición de los umbrales de riesgo el modelo utilizó percentiles de la distribución de scores para establecer las categorías de riesgo como se pueden visualizar en la Tabla 20.

Tabla 20

Percentiles y observaciones por categoría

Categoría	Percentil	Observación
Riesgo Alto	$P_{95} = 0.3295$	Todo proceso con score superior a este valor fue clasificado como riesgo alto.
Riesgo Medio	$P_{80} = 0.2064$	Los procesos ubicados entre los percentiles 80 y 95 fueron clasificados como riesgo medio
Riesgo Bajo		Corresponde al 80% restante de contratos.

La utilización de percentiles presenta varias ventajas como:

- Se adapta automáticamente a la distribución de los datos.
- Evita establecer umbrales arbitrarios.
- Facilita la comparación entre diferentes períodos o sectores.
- Mantiene una proporción controlada de alertas.

El proceso presenta simultáneamente:

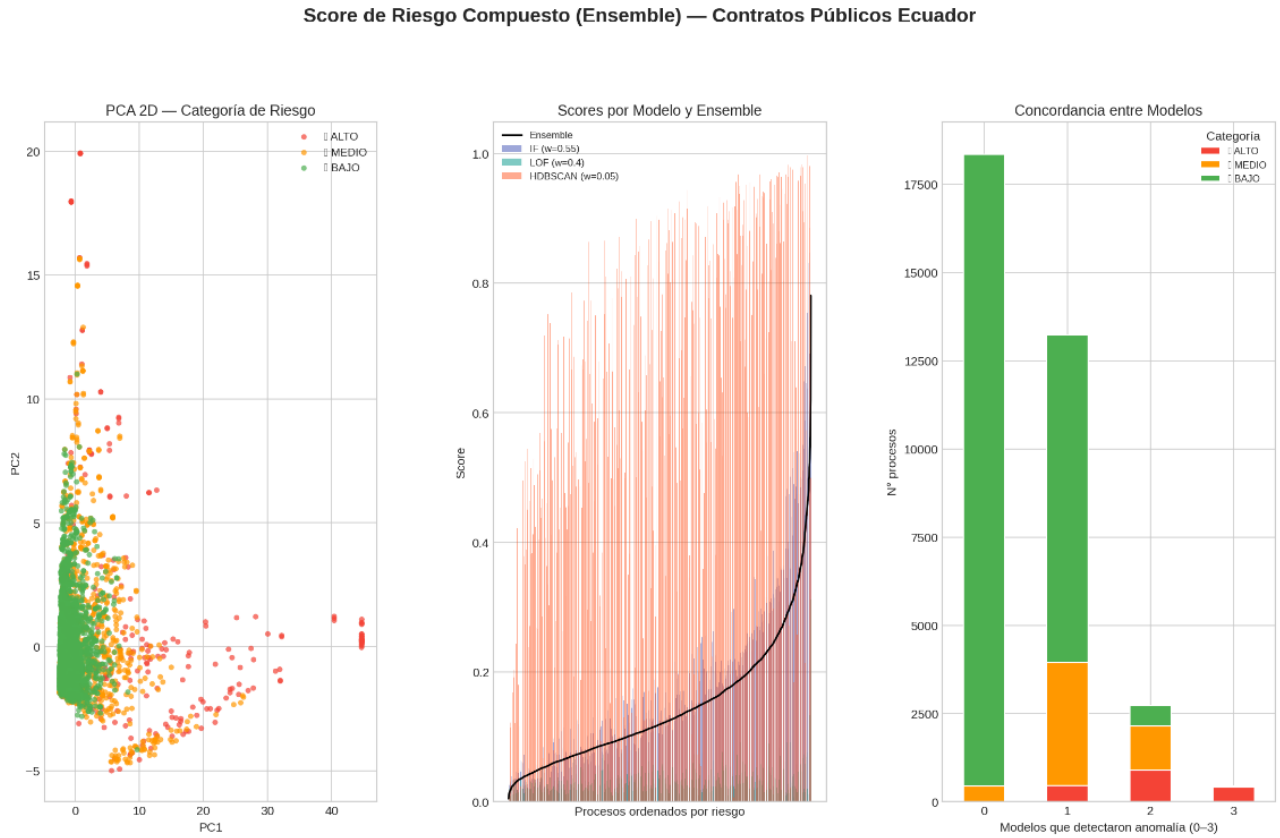
- Score muy elevado en Isolation Forest.
- Score máximo en LOF.
- Evidencia parcial de anomalías en HDBSCAN.

La coincidencia entre múltiples algoritmos incrementa la confianza en la detección. Los resultados indican que aproximadamente el 5% de los procesos analizados presentan características suficientemente inusuales para justificar una revisión más detallada. Cabe indicar que un score alto no implica necesariamente una señal de fraude, corrupción o irregularidad, sino más bien que el proceso presenta un comportamiento estadísticamente diferente respecto al patrón predominante observado en el dataset.

4.5.5. Análisis del Score de Riesgo Compuesto del Modelo Ensemble. Con el objetivo de obtener una evaluación más robusta del riesgo contractual, se consolidaron los resultados de Isolation Forest, Local Outlier Factor (LOF) y HDBSCAN en un score compuesto de riesgo. En la Figura 33 se aprecia la distribución espacial de las categorías de riesgo, la contribución relativa de cada modelo y el nivel de concordancia existente entre los detectores de anomalías.

Figura 33

Score de Riesgo Compuesto



Los gráficos de la Figura 33 son explicados en la

Tabla 21.

Tabla 21

Resultados de Score de riesgos compuestos

Gráfico	Resultado
Distribución Espacial de las Categorías de Riesgo (Gráfico PCA 2D – Izquierda)	La proyección PCA permite visualizar la ubicación relativa de los contratos clasificados como: Riesgo Bajo (verde) Riesgo Medio (naranja) Riesgo Alto (rojo) También se observa que: Los contratos de bajo riesgo forman el núcleo principal de la distribución. Los contratos de riesgo medio aparecen en regiones de transición. Los contratos de alto riesgo se concentran principalmente en zonas periféricas y menos densas.
Evolución de los Scores por Modelo (Gráfico Central)	Este gráfico ordena todos los procesos según su score final de riesgo y muestra simultáneamente la contribución de: Isolation Forest LOF HDBSCAN Score Ensemble final La línea negra representa el score compuesto, aquí se observan tres fenómenos relevantes: Crecimiento progresivo del riesgo Diferentes sensibilidades entre modelos Estabilidad del score ensemble
Concordancia entre Modelos (Gráfico Derecho)	Este gráfico muestra cuántos algoritmos coincidieron en señalar un contrato como potencial anomalía. El eje horizontal representa: 0 modelos detectan anomalía. 1 modelo detecta anomalía.

Gráfico	Resultado
	2 modelos detectan anomalía. 3 modelos detectan anomalía. Los resultados muestran varios hallazgos importantes como: Mayoría de contratos sin señales de riesgo. Riesgo intermedio asociado a detecciones parciales. Alto riesgo asociado al consenso.

Los resultados sugieren que el riesgo contractual no puede explicarse mediante una única característica o un único algoritmo. Los contratos clasificados como de alto riesgo suelen presentar simultáneamente:

- Comportamientos atípicos globales.
- Densidades locales inusuales.
- Baja pertenencia a estructuras de contratación recurrentes.

Esto refuerza la idea de que las anomalías relevantes en contratación pública son fenómenos multidimensionales.

4.5.6. Comparación Cuantitativa de los Modelos de Detección de Anomalías.

Dado que la presente investigación aborda un problema de aprendizaje no supervisado, no se dispone de una variable objetivo que permita comparar las predicciones de los modelos con una clasificación previamente conocida. Por esta razón, no es metodológicamente apropiado emplear métricas de evaluación propias del aprendizaje supervisado, como Accuracy, Precision, Recall, F1-Score o ROC-AUC. En su lugar, se utilizaron métricas que permiten evaluar la calidad del agrupamiento, la capacidad de discriminación entre observaciones normales y anómalas, así como el grado de concordancia entre los diferentes modelos de detección de anomalías.

Tabla 22

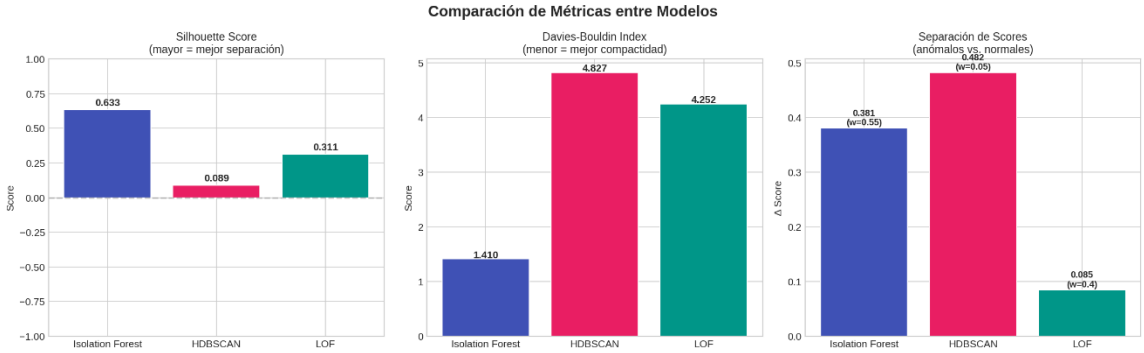
Comparación de Métricas entre Modelos

Gráfico	Resultado
Silhouette Score (Gráfico izquierdo)	El coeficiente de Silhouette evalúa qué tan bien separadas están las observaciones pertenecientes a diferentes grupos. Su valor oscila entre: $-1 \leq S \leq 1$ Donde: Valores cercanos a 1 indican excelente separación. Valores próximos a 0 indican superposición. Valores negativos sugieren clasificación deficiente. Los resultados obtenidos para Silhouette son: Isolation Forest con un valor de 0.633 LOF con un valor de 0.311 HDBSCAN con un valor de 0.089
Davies-Bouldin Index (Gráfico central)	El índice Davies-Bouldin mide simultáneamente: Compactación interna de grupos. Separación entre grupos. A diferencia de Silhouette en donde los valores menores son mejores, aquí se tiene que: $DBI \rightarrow 0$ Esto implica que hay agrupaciones más compactas y mejor diferenciadas. Los resultados obtenidos para Davies-Bouldin son: Isolation Forest con un valor de 1.410 LOF con un valor de 4.252 HDBSCAN con un valor de 4.827
Separación entre Scores (Gráfico derecho)	Esta métrica evalúa la diferencia promedio entre: Score de anomalías. Score de observaciones normales. Los resultados obtenidos para Separación son: HDBSCAN con un valor de 0.482 Isolation Forest con un valor de 0.381 LOF con un valor de 0.085

En la Figura 34 se visualizan los resultados obtenidos para Isolation Forest, HDBSCAN y Local Outlier Factor (LOF), los gráficos se encuentran explicados en la Tabla 22.

Figura 34

Comparación de Métricas entre Modelos



Con el objetivo de evaluar el desempeño relativo de los algoritmos utilizados en esta investigación, se compararon tres métricas complementarias:

- Silhouette Score
- Davies-Bouldin Index
- Separación promedio entre scores de anomalías y observaciones normales

Los resultados muestran que cada una de las métricas empleadas evalúa un aspecto diferente del comportamiento de los modelos. Mientras el Silhouette Score y el índice de Davies–Bouldin analizan la calidad estructural de los agrupamientos obtenidos, la separación de puntajes evalúa la capacidad del modelo para diferenciar procesos potencialmente anómalos respecto del comportamiento predominante. Por su parte, el coeficiente de Cohen (κ) permite analizar la consistencia entre los algoritmos al identificar observaciones de riesgo. En conjunto, estas métricas proporcionan una evaluación integral del desempeño de los modelos, adecuada para problemas de aprendizaje no supervisado donde no se dispone de una variable objetivo. Cabe resaltar que ninguna de estas métricas por si solas determinan cual es el mejor algoritmo, por el contrario, son complementarias.

En conjunto, los resultados justifican la construcción de un modelo ensemble, capaz de integrar fortalezas complementarias y proporcionar una evaluación más robusta del riesgo contractual que cualquiera de los algoritmos considerados individualmente.

4.5.7. Concordancia entre Modelos de Detección de Anomalías. Con el objetivo de evaluar el grado de coincidencia entre los algoritmos empleados, se construyeron matrices de solapamiento entre Isolation Forest (IF), Local Outlier Factor (LOF) y HDBSCAN. También se calculó el coeficiente de Cohen's Kappa (κ), la misma que es una métrica muy utilizada para medir el nivel de acuerdo entre clasificadores más allá de la coincidencia esperada por azar. La Figura 35 muestra los resultados obtenidos y la Tabla 23 la explicación de los gráficos.

Figura 35

Comparación de Matrices de Riesgo entre Modelos

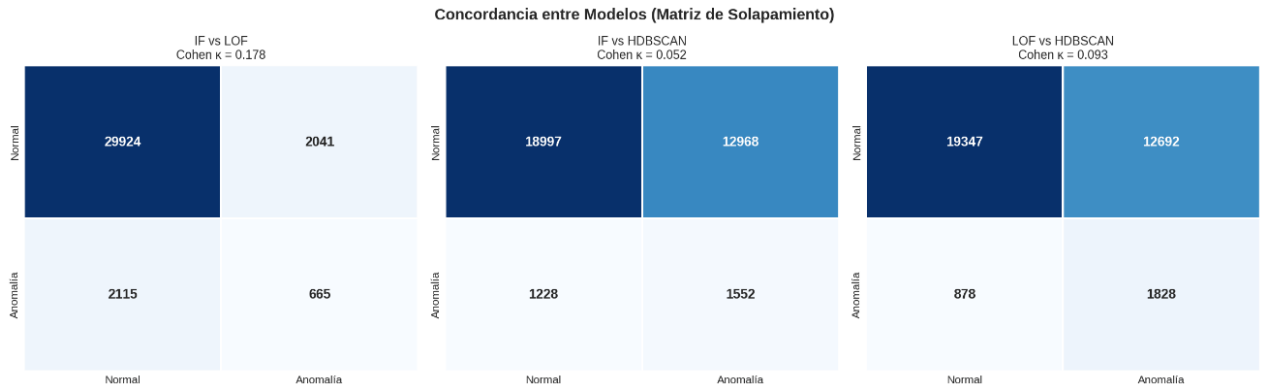


Tabla 23

Comparación de Matrices de Riesgo entre Modelos

Gráfico	Resultados
Acuerdo entre Isolation Forest (IF) y LOF (κ = 0.178) **La primera matriz compara las clasificaciones generadas por Isolation Forest y LOF.**	Aquí se observa que: • Ambos consideran normal a 29924 procesos • IF normal, LOF anomalía a 2041 procesos • IF anomalía, LOF normal a 2115 procesos • Ambos consideran anomalía a 665 procesos

Gráfico	Resultados
	La coincidencia en anomalías es relativamente reducida. Sin embargo, ambos algoritmos detectan observaciones atípicas, lo hacen utilizando criterios diferentes: • Isolation Forest se basa en aislamiento. • LOF se basa en densidad local. Por esta razón, muchas observaciones detectadas por uno de los modelos no son necesariamente identificadas por el otro. El valor $\kappa = 0.178$ corresponde a un acuerdo leve (*slight agreement*).
Acuerdo entre Isolation Forest y HDBSCAN ($\kappa = 0.052$) **La segunda matriz presenta el menor nivel de concordancia observado.**	Aquí se observa que: • Ambos consideran normales a 18997 procesos • IF normal, HDBSCAN anomalía a 12968 procesos • IF anomalía, HDBSCAN normal a 1228 procesos • Ambos anomalía a 1552 procesos La baja concordancia indica que HDBSCAN está identificando un conjunto de observaciones considerablemente distinto. Esto se debe a que este modelo clasifica como anomalías aquellos procesos contractuales que no pertenecen a regiones densas del espacio de datos. En contraste, IF identifica contratos fáciles de aislar independientemente de la densidad local.
Acuerdo entre LOF y HDBSCAN ($\kappa = 0.093$) **La tercera matriz muestra un nivel de acuerdo ligeramente superior al observado entre IF y HDBSCAN, aunque sigue siendo bajo.**	Aquí se observa que: • Ambos consideran normales a 19347 procesos • LOF normal, HDBSCAN anomalía a 12692 procesos • LOF anomalía, HDBSCAN normal a 878 procesos • Ambos anomalía a 1828 procesos Si bien ambos modelos se basan en conceptos relacionados con densidad, utilizan mecanismos diferentes: • LOF analiza la densidad relativa respecto a los vecinos inmediatos. • HDBSCAN analiza pertenencia a estructuras densas jerárquicas. Esto muestra la limitada coincidencia observada.

Los coeficientes obtenidos se pueden ver en la Tabla 24.

Tabla 24

Coeficientes obtenidos

Comparación	Valor de k
IF vs LOF	0.178
IF vs HDBSCAN	0.052
LOF vs HDBSCAN	0.093

La clasificación propuesta por Landis y Koch (1977) (Numiqo Team, 2026) para los valores obtenidos de k se muestra en la Tabla 25.

Tabla 25

Significado de los valores k

K	Significado
< 0.00	Sin acuerdo
0.00–0.20	Acuerdo leve
0.21–0.40	Acuerdo aceptable
0.41–0.60	Acuerdo moderado
0.61–0.80	Acuerdo sustancial
> 0.80	Acuerdo casi perfecto

Los tres modelos se ubican dentro de la categoría de: Acuerdo leve (slight agreement). Lo que no necesariamente implica un problema al tener una baja concordancia, al contrario, en detección de anomalías no supervisada suele interpretarse como una señal de que los algoritmos están capturando perspectivas complementarias del riesgo. Si los tres modelos detectaran exactamente los mismos contratos podría significar que:

- El ensemble aportaría poco valor adicional.
- Existiría redundancia metodológica.

CAPITULO 5:

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

Las conclusiones presentadas a continuación se estructuran en función de los objetivos específicos planteados en el Capítulo 1, con el propósito de evidenciar el grado de cumplimiento de la investigación y los principales hallazgos obtenidos durante el desarrollo del modelo de detección de anomalías aplicado a procesos de contratación pública del sector salud en Ecuador.

1. **Integración y preparación del repositorio de datos**: Se logró integrar y consolidar información proveniente de los portales de datos abiertos del SERCOP y del SRI, construyendo un repositorio unificado que permitió relacionar información contractual, económica y tributaria correspondiente al período 2020–2025. La incorporación de procesos ETL, DuckDB, archivos Parquet y PostgreSQL permitió disponer de una infraestructura adecuada para el procesamiento de grandes volúmenes de información y para la construcción del dataset utilizado durante el entrenamiento de los modelos de aprendizaje automático.
2. **Construcción del modelo de scoring de riesgo**: Se diseñó un modelo de scoring de riesgo basado en técnicas de aprendizaje no supervisado, con lo que se demuestra que es posible identificar procesos contractuales con comportamientos estadísticamente atípicos sin disponer de etiquetas previas de fraude, corrupción o irregularidad. Este resultado evidencia la viabilidad del uso de técnicas de detección de anomalías cuando no existe una variable objetivo que permita aplicar modelos supervisados.
3. **Identificación de patrones atípicos**: Los resultados demostraron que las variables derivadas de relaciones históricas entre proveedores y entidades contratantes aportan mayor capacidad explicativa que los montos adjudicados considerados de forma aislada. Variables como la concentración histórica de contratación, el porcentaje de contratación, la

antigüedad empresarial y la participación de proveedores nuevos en contratos de alto valor permitieron identificar configuraciones de riesgo que difícilmente serían detectadas mediante análisis tradicionales. En consecuencia, se concluye que el riesgo contractual constituye un fenómeno multidimensional cuya evaluación requiere considerar simultáneamente variables económicas, temporales, de competencia y relacionales.

4. **Evaluación comparativa de los modelos**: La comparación entre los algoritmos evidenció que cada uno aporta información complementaria sobre el comportamiento de los procesos contractuales. Isolation Forest obtuvo el mejor desempeño global en términos de separación de observaciones, Local Outlier Factor (LOF) permitió identificar anomalías locales que no fueron detectadas por los demás modelos y HDBSCAN aportó información estructural basada en la densidad de los datos. Ninguno de los algoritmos resultó suficiente por sí solo para capturar la totalidad de los patrones atípicos presentes en el conjunto de datos, justificando el empleo de un enfoque ensemble.
5. **Construcción del modelo Ensemble**: La integración de los tres algoritmos permitió construir un puntaje de riesgo más robusto que cualquiera de los modelos individuales. El modelo ensemble clasificó aproximadamente el 80 % de los procesos como riesgo bajo, el 15 % como riesgo medio y el 5 % como riesgo alto. Los procesos con mayor puntaje presentaron combinaciones de características relacionadas con baja competencia, alta concentración proveedor–entidad, proveedores con poca trayectoria, relaciones históricas dominantes y comportamientos temporales inusuales, confirmando que la combinación de múltiples variables mejora la capacidad para identificar comportamientos estadísticamente atípicos.
6. **Alcance y limitaciones de la investigación:** Los resultados obtenidos corresponden exclusivamente al análisis de procesos contractuales del sector salud en Ecuador durante el período 2020–2025. En consecuencia, la extrapolación del modelo hacia otros sectores

económicos, períodos diferentes o contextos institucionales requiere procesos adicionales de validación, ajuste y entrenamiento con información representativa de dichos escenarios. Asimismo, la investigación estuvo condicionada por la calidad y completitud de la información disponible en los portales de datos abiertos. Durante el proceso de integración se identificaron registros con información incompleta, particularmente en variables relacionadas con la firma de contratos, lo que constituye una limitación inherente a la fuente de datos utilizada. Adicionalmente, la ausencia de bases históricas con casos previamente confirmados como irregulares impidió realizar una validación mediante técnicas de aprendizaje supervisado y métricas como ROC-AUC, por lo que la evaluación se realizó utilizando métricas apropiadas para aprendizaje no supervisado.

7. **Beneficios potenciales de la propuesta**: El sistema desarrollado constituye una herramienta de apoyo para la supervisión y análisis de procesos de contratación pública, al proporcionar un mecanismo de priorización basado en comportamientos estadísticamente atípicos. No obstante, un puntaje elevado de riesgo no constituye evidencia de fraude, corrupción o ilegalidad, sino un indicador que permite orientar procesos posteriores de auditoría, fiscalización o análisis especializado. En consecuencia, el modelo debe entenderse como un mecanismo de alerta temprana que complementa, pero no reemplaza, los procedimientos de control establecidos por los organismos o entidades de control.

5.2. Recomendaciones

Incorporar información adicional proveniente de otras fuentes como: Contraloría General del Estado, Superintendencia de Compañías, SRI, Registro Mercantil. Esto permitiría fortalecer el estudio sobre la naturaleza de los proveedores. Se podría incorporar más variables o features que aporten no solo dimensionalidad a los análisis, sino que puedan enriquecer las relaciones entre las mismas.

Con el fin de contrastar los resultados del modelo sería ideal incorporar al estudio a expertos como auditores, analistas de compras públicas, organismos de control.

Para futuras investigaciones sería conveniente disponer de bases históricas que incluyan procesos previamente confirmados como irregulares, observados o sancionados por los organismos competentes. La disponibilidad de esta información permitiría evaluar el desempeño del modelo utilizando técnicas de aprendizaje supervisado y métricas como Precision, Recall, F1-Score y ROC-AUC, además de comparar los resultados con los obtenidos mediante aprendizaje no supervisado.

Se recomienda evaluar el modelo utilizando información correspondiente a otros sectores de la contratación pública y ampliar el horizonte temporal de análisis. Esto permitiría determinar el grado de generalización del modelo y analizar si los patrones de riesgo identificados para el sector salud durante el período 2020–2025 se mantienen en otros contextos de contratación pública.

Se recomienda evolucionar la aplicación desarrollada hacia una plataforma analítica de monitoreo continuo que procese periódicamente las nuevas contrataciones públicas y genere alertas tempranas sobre procesos con comportamiento potencialmente atípico. Para ello, sería conveniente incorporar prácticas de MLOps, incluyendo automatización de los procesos de extracción y preparación de datos, reentrenamiento periódico de los modelos, monitoreo del desempeño, control de versiones y despliegue automatizado. Estas capacidades favorecerían el mantenimiento y la actualización continua del sistema ante cambios en los patrones de contratación pública.

Se recomienda que el modelo de scoring de riesgo sea utilizado como una herramienta de apoyo dentro de los procesos institucionales de auditoría y supervisión de la contratación pública, complementando los mecanismos tradicionales de control. Su implementación debería orientarse a

priorizar procesos para revisión especializada o auditorias, sin sustituir el análisis técnico, jurídico o financiero realizado por los organismos de control.

Referencias Bibliográficas

Aggarwal, C. C. (2017). *Outlier analysis* (2nd ed.). Springer.

Araujo, H. R. F., Leite, P. F., Honorio, J. J. C. M., Souza, I. M. L., Albuquerque, D. W., & Santos, D. F. S. (2024). Detection of anomalous proposals in governmental bidding processes: A machine learning-based approach. *Proceedings of a scientific conference on machine learning and public procurement.*

Barua, T., Hiran, K. K., Jain, R. K., & Doshi, R. (2024). *Machine learning with Python*. Walter de Gruyter GmbH.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM SIGMOD Record, 29*(2), 93–104.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *PAKDD 2013* (pp. 160–172).

Carrillo Bustos, F. D., & Fernández Fernández, Y. (2025). Modelado predictivo de riesgos de ineficiencia y transparencia en la contratación pública ecuatoriana, basado en el análisis de patrones de datos clasificados del SERCOP. *Arandu UTIC, 12*(4).

Carmen, M. (2020, November 10). La corrupción en los hospitales públicos. *Diario El Mercurio*. https://elmercurio.com.ec/columnistas/2020/11/10/la-corrupcion-en-los-hospitales-publicos/

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58.

Contraloría General del Estado. (2020, December 31). En 2020, Contraloría reveló irregularidades en contrataciones de emergencia. https://www.contraloria.gob.ec/WFDescarga.aspx?id=2710&tipo=doc

El Comercio. (2020, November 17). Observan 202 contratos suscritos en la pandemia por anomalías en Ecuador. https://www.elcomercio.com/actualidad/seguridad/contratos-pandemia-posibles-anomalias-ecuador/

Ferle, M. (2025). *Snowflake data engineering*. Manning Publications.

Ferrari, L., & Pirozzi, E. (2023). *Learn PostgreSQL* (2nd ed.). Packt Publishing.

Freeman, E., & Robson, E. (2024). *Head first JavaScript programming* (2nd ed.). O'Reilly Media.

Fundación Ciudadanía y Desarrollo. (2025). *Banderas rojas*. Observatorio de Contratación Pública de Ecuador. https://www.contratostransparentes.ec/banderas-rojas

García Rodríguez, M. J., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E. D., & Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction, 133*, 104047.

GeeksforGeeks. (2025, July 23). Hierarchical density-based spatial clustering of applications with noise (HDBSCAN). https://www.geeksforgeeks.org/machine-learning/hdbscan/

Kennedy, B. (2025). *Outlier detection in Python*. Manning Publications.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Ley Orgánica del Sistema Nacional de Contratación Pública. (2025). *Ley Orgánica del Sistema Nacional de Contratación Pública*. Registro Oficial Suplemento No. 140, 7 de octubre de 2025.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the Eighth IEEE International Conference on Data Mining* (pp. 413–422).

Luis Garnica. (2026). Streamlit: La herramienta fácil para crear aplicaciones web de datos y ML. *Emprenderte*. https://emprenderte.co/ia-emprendedores/streamlit-la-herramienta-facil-para-crear-aplicaciones-web-de-datos-y-ml/

Mueller, J. P., & Massaron, L. (2024). *Python for data science for dummies* (3rd ed.). John Wiley & Sons.

Needham, M., Hunger, M., & Simon, M. (2024). *DuckDB in action*. Manning Publications.

Numiqo Team. (2026). Kappa de Cohen – Explicación sencilla. *Numiqo: Online statistics calculator*. https://numiqo.es/tutorial/cohens-kappa

O'Sullivan, C. (2024, September 25). Guía del Bosque del Aislamiento: Explicación e implementación en Python. *DataCamp*. https://www.datacamp.com/es/tutorial/isolation-forest

Picard. (s. f.). Davies Bouldin score (Scikit-learn). *CDS Institute – Data Science Python Blog*. https://datasciencepythonblog.net/davies-bouldin-score-scikit-learn/

Pykes, K. (2024). Introducción a DuckDB: ¿Qué es y por qué debes utilizarlo? *DataCamp*. https://www.datacamp.com/es/blog/an-introduction-to-duckdb-what-is-it-and-why-should-you-use-it

Primicias. (2022, July 8). Contraloría confirma irregularidades en la compra de pruebas Covid-19 en Quito. https://www.primicias.ec/noticias/sociedad/contraloria-glosa-pruebas-covid-municipio-quito/

Primicias. (2024, June 26). Gobierno pide examen especial a la compra de insumos médicos tras denuncia de Comisión Anticorrupción. https://www.primicias.ec/noticias/politica/comision-anticorrupcion-denuncia-irregularidades-compra-insumos-medicos/

Ramjan, S., & Sunkpho, J. (2023). Introduction to data mining. In *Principles and theories of data mining with RapidMiner* (p. 6). IGI Global.

Raschka, S., Liu, Y., & Mirjalili, V. (2022). *Machine learning with PyTorch and Scikit-learn*. Packt Publishing.

Riggs, S., & Ciolli, G. (2022). *PostgreSQL 14 administration cookbook*. Packt Publishing.

Scikit-learn developers. (2026). Demo of HDBSCAN clustering algorithm — scikit-learn 1.9.0 documentation. *Scikit-learn documentation*. https://scikit-learn.org/stable/auto_examples/cluster/plot_hdbscan.html

Scikit-learn developers. (2026). Outlier detection with Local Outlier Factor (LOF). *Scikit-learn 1.9.0 documentation*. https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html

Scikit-learn developers. (2026). Selecting the number of clusters with silhouette analysis on KMeans clustering. *Scikit-learn 1.9.0 documentation*. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Segovia, J. (2018). Ventajas y desventajas de PostgreSQL. *TodoPostgreSQL*. https://www.todopostgresql.com/ventajas-y-desventajas-de-postgresql/

Servicio Nacional de Contratación Pública. (2026a). Acerca de la plataforma. https://datosabiertos.compraspublicas.gob.ec/PLATAFORMA/acerca

Servicio Nacional de Contratación Pública. (2026b). Cifras del Sistema Oficial de Contratación Pública del Ecuador. Portal Institucional SERCOP. https://portal.compraspublicas.gob.ec/sercop/cifras/

Servicio Nacional de Contratación Pública. (2026c). Contrataciones Abiertas Ecuador - OCDS. Datos abiertos de contratación pública. https://datosabiertos.compraspublicas.gob.ec/PLATAFORMA

Servicio Nacional de Contratación Pública. (2026d). Listado de CPC N9 [Visualización de datos de Power BI]. https://app.powerbi.com/view?r=eyJrIjoiYjRlZjg5YzItOWM4Ni00MGNkLWI1OGYtZD

A4MDAxNGYyNDQyIiwidCI6ImQ2NDk2NzM4LWY5MTItNGExZS04NDE1LTQwY2E2ZjRhOTRlZCJ9

Servicio Nacional de Contratación Pública. (2026e). Sistema Oficial de Contratación Pública. Búsqueda de códigos CPC en catálogo electrónico. https://www.compraspublicas.gob.ec/ProcesoContratacion/compras/RCC/RccFrmBuscarCpcEnCatalogo.cpe

Servicio Nacional de Contratación Pública. (2026f). Sistema Oficial de Contratación Pública. Búsqueda de productos con la clasificación central de productos (CPC) versión 1. https://www.compraspublicas.gob.ec/ProcesoContratacion/compras/CPC/index.cpe

Servicio de Rentas Internas. (s. f.). Base de datos catastro RUC por provincia: personas naturales y sociedades. https://www.sri.gob.ec/datasets

Subramanian, V. (2025). Applied machine learning for data science practitioners. John Wiley & Sons.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th ed.). Pearson.

Tezuysal, A., & Ahmed, I. (2024). Database design and modeling with PostgreSQL and MySQL. Packt Publishing.

Torres Berrú, Y. M. (2023). *Minería de datos y texto para la detección de fraude en contratos públicos: Caso de estudio Sistema Oficial de Contratación Pública de Ecuador* [Tesis doctoral, Universidad de Salamanca]. Gredos – Repositorio institucional de la Universidad de Salamanca. https://gredos.usal.es/handle/10366/157534

Torres-Berru, Y., & López Batista, V. F. (2021). Data mining to identify anomalies in public procurement rating parameters. *Electronics, 10*(22), 2873. https://doi.org/10.3390/electronics10222873

Transparency International. (2026, 10 febrero). *Corruption Perceptions Index 2025 – Ecuador*. Transparency International. https://www.transparency.org/en/cpi/2025/index/ecu

Transparency International. (2025, February 11). *The ABCs of the CPI: How the Corruption Perceptions Index is calculated*. Transparency International. https://www.transparency.org/en/news/how-cpi-scores-are-calculated

Treuille, A. (2023). *Streamlit for data science*. Packt Publishing.

Universidad San Francisco de Quito & Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ). (2023). *KAPAK: Transparencia en compras públicas*. https://kapak.usfq.edu.ec

Westerveld, D. (2024). *API testing and development with Postman* (2nd ed.). Packt Publishing.

World Bank. (2026, March 11). *Worldwide Governance Indicators [Data set]*. World Bank Data Catalog. https://datacatalog.worldbank.org/search/dataset/0038026/worldwide-governance-indicators

Apéndices

Apéndice A. Código Fuente

Para replicar el funcionamiento de la aplicación creada en Streamlit con Python, se debe acceder a la siguiente URL compartida, en donde se debe descargar la carpeta llamada Aplicacion, dentro de la misma se encuentran todos los programas Python y un archivo README.md que es una guía que muestra paso a paso la ruta para ejecutar la aplicación web. En la raíz de la carpeta compartida se encuentra el Notebook que se puede usar en Google Colab para entrenar los algoritmos que se usaron en el proyecto; de igual manera en el archivo de instrucciones se menciona de dónde se obtienen los datos, que es una vista materializada, y cómo deben ser leídos desde el notebook.

Apéndice B. Herramientas Utilizadas

DuckDB: Es un sistema de gestión de bases de datos gratuito, de código abierto, incrustado, en proceso, relacional, y de Procesamiento Analítico en Línea (OLAP). (Pykes, K., 2024). Su base de datos está diseñada para el análisis de datos, teniendo como una de sus

características ser increíblemente rápida. Adicional a esto, DuckDB funciona con un motor de consulta vectorial columnar, lo que le ayuda a hacer un uso eficiente de la caché de la CPU y a acelerar los tipos de respuesta para cargas de trabajo de consulta analítica. (Pykes, K., 2024). DuckDB es una base de datos analítica integrada moderna que se ejecuta de forma local y permite procesar y consultar gigabytes de datos desde diversas fuentes, como JSON o CSV, de manera eficiente (Needham et al., 2024, p. 2). DuckDB también permite la lectura y escritura de archivos parquet, integrándose con algunos lenguajes de programación, como lo es Python.

Streamlit: Es una herramienta de código abierto utilizada por científicos de datos y desarrolladores de aplicaciones crear aplicaciones web de forma sencilla, dado que se enfoca en la visualización de datos, permitiendo mostrar gráficamente los resultados de trabajos realizados con técnicas de Machine Learning. (Luis Garnica., 2026). Es una forma rápida de crear aplicaciones de datos, ya que se usan librerías del ecosistema de Python las mismas que permiten agilizar el desarrollo y entregar resultados analíticos, con experiencias interactivas complejas e iterar sobre modelos de ML. (Treuille, 2023, p. 1). Streamlit es compatible con bibliotecas de visualización populares en Python como Matplotlib y Plotly.

PostgreSQL: Es una base de datos de instalación gratuita e ilimitada, disponible para sistemas Unix, Linux, Windows y MacOS, adaptándose al CPU y la cantidad de memoria disponible de forma óptima. Tiene más de 20 años de desarrollo activo y en constante mejora. Brinda un entorno de alta disponibilidad. Además, cumple el estándar SQL ISO/IEC 9075:2011, facilitando la compatibilidad de sus scripts con los de otros motores de bases de datos. (Segovia, J., 2018). PostgreSQL es un servidor de bases de datos SQL avanzado, disponible en una amplia gama de plataformas. Una de sus principales ventajas es que es de código abierto, lo que significa que se dispone de una licencia muy permisiva para instalar, usar y distribuir PostgreSQL sin pagar ninguna tarifa ni regalías. (Riggs y Ciolli, 2022, p. 2).

Vistas materializadas: Las vistas materializadas en PostgreSQL son una potente herramienta para optimizar la recuperación de datos, especialmente útil en escenarios que requieren un acceso rápido a resultados de consultas complejas. A diferencia de las vistas estándar, que generan datos dinámicamente con cada consulta, las vistas materializadas almacenan físicamente los resultados igual que las tablas, lo que permite un acceso más rápido a los datos, aunque a costa de la actualización, que puede gestionarse mediante actualizaciones periódicas. (Tezuysal y Ahmed, 2024, p. 96).

Apéndice C. Código SQL

C.1. Vista Materializada mv_universo_contratos Creada con Código PL/psSQL en PostgreSQL

Las vistas materializadas fueron creadas tomando como referencia las actividades mencionadas en la Tabla 4, con el objetivo de tener filtrar la información con el objetivo de obtener la información necesaria para el análisis, y con el objetivo final de construir un dataset que contiene los contratos para entrenar los modelos de ML.

La vista **mv_universo_contratos** permite obtener todos los contratos que se tengan en la base de datos, y se obtiene la información relacionada de cada uno de ellos. En la Figura 12 se presenta la estructura JSON que tiene cada contrato; los atributos o llaves más relevantes se presentan como una tabla estructurada, y cada llave hija o anidada se representa como una columna. En esta vista se incluye la columna llamada proceso_con_insumos_medicos, que etiqueta a los contratos que tienen un producto o servicio en base al código CPC padre, considerando que en la tabla **awards_items** se almacena la información de los productos/servicios adjudicados en cada contrato. Tal como se detalla en las limitaciones, las fuentes de datos que utilizadas como el portal de datos abiertos o en el mismo portal del sistema SOCE del SERCOP, no existe una columna o etiqueta que permita identificar exactamente si un contrato es de un rubro o categoría como salud; esa es la razón por la que se creó dicha columna. Asimismo, en las

referencias se incluye la referencia al clasificador de CPC proporcionado por el SERCOP (SERCOP, 2026d) que permite navegar por los distintos niveles y leer sus descripciones.

```
-- ======================================================================
-- CREACIÓN DE VISTA MATERIALIZADA mv_universo_contratos
-- ======================================================================

DROP MATERIALIZED VIEW IF EXISTS mv_universo_contratos CASCADE;
CREATE MATERIALIZED VIEW mv_universo_contratos
AS WITH awards_summary AS (
        select a.anio_proceso, a.ocid,
            sum(a.amount) as award_amount
            from awards a
            group by a.anio_proceso,a.ocid
        ), awards_items_summary as (
            select b.anio_proceso,b.ocid,
                        max(limpiar_ruc(b.award_suppliers_id)) AS
award_suppliers_id,
                        max(b.award_suppliers_name) AS award_suppliers_name,
                        MAX(
                    CASE
                            WHEN b.item_clasificacion LIKE '2411%' THEN 1
                            WHEN b.item_clasificacion LIKE '2412%' THEN 1
                        WHEN b.item_clasificacion LIKE '352%' THEN 1
                        WHEN b.item_clasificacion LIKE '481%' THEN 1
                        ELSE 0
                    END
                ) AS proceso_con_insumos_medicos
                  from awards_items b
                  group by b.anio_proceso,b.ocid
        ),
        contract_summary AS (
         SELECT anio_proceso, ocid,
              SUM(monto_final) as contratacion_monto_total,
              MIN(contrato_fec_ini) as contratacion_periodo_inicio, -- Fecha
más antigua
              MAX(contrato_fec_fin) as contratacion_periodo_fin,    -- Fecha
más reciente
              SUM(contrato_duracion_dias) as
contratacion_duracion_total_dias,
```

```
              MIN(cast(fecha_firma as date)) as
contratacion_fecha_firma_primera - Se usa la primera firma
           FROM contracts
          GROUP BY contracts.anio_proceso,contracts.ocid
        ), sri_summary AS (
         SELECT ruc_sri."NUMERO_RUC",
            MIN(CAST("FECHA_INICIO_ACTIVIDADES" AS DATE)) as
fecha_inicio_real
           FROM ruc_sri
          GROUP BY ruc_sri."NUMERO_RUC"
        )
 SELECT DISTINCT r.ocid,
    regexp_replace(r.detalle_metodo_adquisicion, ' en el convenio.*', '') as
tipo_contratacion,
    r.fases_incluidas,
    limpiar_ruc(r.entidad_ruc_id) AS ruc_entidad_compradora,
    r.entidad_compradora,
    r.planning_monto_presupuesto AS planeacion_monto_presupuesto,
    r.numero_proveedores AS licitacion_numero_proveedores,
    r.tiene_preguntas,
    r.total_preguntas,
    r.categoria_compra_adquisiciones,
    r.metodo_adquisicion,
    r.criterios_adjudicacion,
    cast(r.fecha_release as date) as planeacion_fecha,
    cast(r.licitacion_fecha_inicial as date) as licitacion_fecha_inicial,
      cast(r.licitacion_fecha_final as date) as licitacion_fecha_final,
    r.licitacion_duracion_dias,
    r.licitacion_consultas_echa_inicial AS
licitacion_consultas_fecha_inicial,
    r.licitacion_consultas_fecha_final,
    r.licitacion_consultas_fecha_maxima,
    r.licitacion_consultas_duracion_dias,
    r.licitacion_pa_fin_periodo AS
licitacion_periodo_adjudicacion_fin_periodo,
    r.licitacion_pa_fecha_maxima AS
licitacion_periodo_adjudicacion_fecha_maxima,
    cs.contratacion_monto_total,
    cs.contratacion_periodo_inicio,
    cs.contratacion_periodo_fin,
    cs.contratacion_duracion_total_dias,
```

```
    cs.contratacion_fecha_firma_primera,
    ad.award_suppliers_id,
    ad.award_suppliers_name,
    ad.proceso_con_insumos_medicos,
    am.award_amount,
    rs.fecha_inicio_real AS sri_fecha_inicio_actividades,
    get_dias_lab(CAST(rs.fecha_inicio_real AS DATE), CAST(r.fecha_release AS
DATE)) as bandera_dias_ruc_firma_contrato,
      get_dias_lab(CAST(r.fecha_release AS DATE),
cs.contratacion_fecha_firma_primera) as
bandera_dias_planificacion_contratacion,
    r.anio_proceso
   FROM releases r
     LEFT JOIN contract_summary cs ON r.ocid = cs.ocid
     LEFT JOIN awards_summary am ON r.ocid = am.ocid
     left join awards_items_summary ad on r.ocid = ad.ocid
     LEFT JOIN sri_summary rs ON ad.award_suppliers_id = rs."NUMERO_RUC"
      --WHERE 'contract' = ANY (string_to_array(r.fases_incluidas, ', '))
WITH DATA;
```

C.2. Uso de expresiones para agrupar registros en un diccionario de datos

La tabla **diccionario_categorias_salud,** creada mediante el uso de expresiones regulares (regex), permite clasificar las entidades a partir de su nombre, en algunos grupos dentro del rubro salud como una ayuda visual, más que para análisis o considerar a esta variable como una que forme parte de los entrenamientos. Este procedimiento se realiza debido a que en las fuentes de datos utilizadas no existe ninguna variable que indique si la entidad pública pertenece a alguna categoría dentro de su ramo o actividad económica, haciendo necesario la identificación basada en el nombre de la institución.

```
INSERT INTO diccionario_categorias_salud (patron, 120ategoría_asignada)
VALUES
-- MINISTERIO DE SALUD
('AGENCIA DE ASEGURAMIENTO DE LA CALIDAD DE LOS SERVICIOS DE SALUD Y
MEDICINA PREPAGADA', 'MINISTERIO DE SALUD'),
('AREA.*SALUD', 'MINISTERIO DE SALUD'),
('DIRECCION.*SALUD', 'MINISTERIO DE SALUD'),
```

```
('DISTRIT(AL|O).*SALUD', 'MINISTERIO DE SALUD'),
('EMPRESA.*SALUD', 'MINISTERIO DE SALUD'),
('INSTITUTO.*SALUD PUBLICA', 'MINISTERIO DE SALUD'),
('MINISTERIO.*SALUD', 'MINISTERIO DE SALUD'),
('SERVICIO NACIONAL DE MEDICINA LEGAL Y CIENCIAS FORENSES', 'MINISTERIO DE
SALUD'),
('UNIDAD.*MEDICA', 'MINISTERIO DE SALUD'),
('ZONAL.*SALUD', 'MINISTERIO DE SALUD'),
-- CENTRO DE ESPECIALIDADES
('.*CENTRO DE ESPECIALIDADES.*', 'CENTRO DE ESPECIALIDADES'),
-- CENTROS DE SALUD
('CENTRO.*SALUD', 'CENTRO DE SALUD'),
-- CONSEJO DE SALUD
('.*CONSEJO NACIONAL DE SALUD.*', 'CONSEJO NACIONAL DE SALUD'),
-- MUNICIPIOS (SALUD)
('CONSEJO.*SALUD', 'MUNICIPIO SALUD'),
('RED MUNICIPAL.*SALUD', 'MUNICIPIO SALUD'),
('UNIDAD (METROPOLITANA|MUNICIPAL) DE SALUD', 'MUNICIPIO SALUD'),
-- UNIVERSIDADES
('FACULTAD.*MEDICAS', 'UNIVERSIDADES PUBLICAS SALUD'),
-- HOSPITALES / CLINICAS
('HOSPITAL|CLINICA', 'HOSPITAL'),
-- IESS
('IESS.*SALUD', 'IESS SALUD'),
('UNIDAD DE ATENCION MEDICA IESS', 'IESS SALUD'),
('UNIDAD DE ATENCI[OÓ]N M[EÉ]DICA IESS', 'IESS SALUD'),
-- PUESTOS
('PUESTO.*SALUD', 'PUESTO DE SALUD'),
-- UNIDADES
('UNIDAD.*SALUD', 'IESS SALUD');
```

C.3. Creación de vista materializada para las entidades objetivo

La vista materializada mv_entidades_objetivo se utiliza para focalizar el análisis en aquellos contratos de carácter médico, utilizando la columna en la que se ha etiquetado al contrato con insumos médicos; recordando que el proyecto delimita su alcance al análisis de los contratos médicos. En el caso de las entidades, no se aplica segmentación ni filtrado adicional, solo se

utiliza el clasificador en base al diccionario creado en el paso previo, aquellas entidades que no caen o empatan dentro de ese diccionario son calificadas como ajenas al sector salud. En síntesis, en esta etapa se incorpora un filtro en donde se evalúa uc.proceso_con_insumos_medicos = 1

```
DROP MATERIALIZED VIEW IF EXISTS mv_entidades_objetivo CASCADE;
CREATE MATERIALIZED VIEW mv_entidades_objetivo
AS
SELECT DISTINCT ON (uc.entidad_compradora)
    uc.entidad_compradora,
    uc.ruc_entidad_compradora,
    COALESCE(ds.categoria_asignada, 'NO SALUD') AS tipo_institucion_salud
FROM mv_universo_contratos uc
LEFT JOIN diccionario_categorias_salud ds
    ON uc.entidad_compradora::text ~* ds.patron
WHERE uc.proceso_con_insumos_medicos = 1
ORDER BY uc.entidad_compradora, ds.categoria_asignada ASC;
```

C.4. Creación de la vista materializada mv_dataset_base

Para esta vista materializada se realizaron algunos procesos de limpieza de datos, descartando aquellos contratos o registros que no tengan un ID de proveedor, y tomando en cuenta a aquellos contratos cuyo monto adjudicado sea mayor a cero. En la información que se procesó se encontró datos que no cumplían esas validaciones y no aportaban valor agregado para el análisis.

```
-- Se crea un dataset base, al que se le hace ciertas limpiezas
DROP MATERIALIZED VIEW IF EXISTS mv_dataset_base CASCADE;
CREATE MATERIALIZED VIEW mv_dataset_base
AS WITH datos_consolidados AS (
         SELECT uc.ocid,
            uc.tipo_contratacion,
            max(es.tipo_institucion_salud) AS tipo_institucion_salud,
            uc.tiene_preguntas,
            uc.licitacion_numero_proveedores,
            uc.categoria_compra_adquisiciones,
            uc.metodo_adquisicion,
            uc.contratacion_monto_total,
```

```
            uc.planeacion_fecha,
            uc.contratacion_fecha_firma_primera,
            uc.award_amount,
            uc.sri_fecha_inicio_actividades,
            uc.anio_proceso,
                  uc.ruc_entidad_compradora,
                  uc.award_suppliers_id,
                  proceso_con_insumos_medicos
           FROM mv_universo_contratos uc
             JOIN mv_entidades_objetivo es ON uc.ruc_entidad_compradora =
es.ruc_entidad_compradora
          WHERE uc.award_suppliers_id IS NOT NULL
          AND uc.award_amount > 0::double precision
          GROUP BY uc.ocid, uc.tipo_contratacion, uc.tiene_preguntas,
uc.licitacion_numero_proveedores,
          uc.categoria_compra_adquisiciones, uc.metodo_adquisicion,
uc.contratacion_monto_total, uc.planeacion_fecha,
          uc.contratacion_fecha_firma_primera, uc.award_amount,
uc.sri_fecha_inicio_actividades,uc.anio_proceso,uc.ruc_entidad_compradora,
           uc.award_suppliers_id,proceso_con_insumos_medicos
        )
 SELECT ocid,
    tipo_contratacion,
    tipo_institucion_salud,
    tiene_preguntas,
    licitacion_numero_proveedores,
    categoria_compra_adquisiciones,
    metodo_adquisicion,
    contratacion_monto_total,
    planeacion_fecha,
    contratacion_fecha_firma_primera,
    award_amount,
    sri_fecha_inicio_actividades,
    get_dias_lab(sri_fecha_inicio_actividades,
contratacion_fecha_firma_primera) AS dias_maduracion_empresa,
    get_dias_lab(planeacion_fecha, contratacion_fecha_firma_primera) AS
dias_entre_planeacion_firma_contrato,
    anio_proceso,
      ruc_entidad_compradora,
      award_suppliers_id,
      proceso_con_insumos_medicos
```

```
   FROM datos_consolidados ds
  ORDER BY planeacion_fecha
WITH DATA;
```

C.5. Creación de la vista materializada mv_dataset_base_limpio

Se colocan condiciones de tipo IS NOT NULL a varias columnas, con el objetivo de descartar a los contratos que no cumplan con las condiciones necesarias como: el proveedor adjudicado debe tener información de la fecha de actividades en el SRI ya que estas fechas son importantes para el análisis, permitiendo calcular la edad de la empresa proveedora, teniendo como premisa a priori validar o saber si existen contratos que se adjudican a empresas con poca experiencia o edad económicamente hablando. Las otras conficiones son intuitivas y fáciles de leer en base al nombre de las columnas.

```
-- Se crea el dataset limpio para entrenar los modelos de ML
DROP MATERIALIZED VIEW IF EXISTS mv_dataset_base_limpio CASCADE;
CREATE MATERIALIZED VIEW mv_dataset_base_limpio
AS SELECT ocid,
    tipo_contratacion,
    tipo_institucion_salud,
    tiene_preguntas,
    licitacion_numero_proveedores,
    categoria_compra_adquisiciones,
    metodo_adquisicion,
    planeacion_fecha,
    contratacion_fecha_firma_primera,
    award_amount,
    sri_fecha_inicio_actividades,
    dias_maduracion_empresa,
    dias_entre_planeacion_firma_contrato,
    anio_proceso,
      ruc_entidad_compradora,
      award_suppliers_id,
      proceso_con_insumos_medicos
   FROM mv_dataset_base mv
  WHERE sri_fecha_inicio_actividades IS NOT NULL
```

```
  AND award_amount IS NOT NULL
  AND tipo_contratacion IS NOT NULL
  AND tipo_institucion_salud IS NOT NULL
  AND metodo_adquisicion IS NOT NULL
  AND planeacion_fecha IS NOT NULL
  AND contratacion_fecha_firma_primera IS NOT NULL
  AND categoria_compra_adquisiciones IS NOT NULL
WITH DATA;
```

C.6. Creación de la vista materializada mv_dataset_base_ml

Esta vista es la más importante dentro del proyecto, es el compendio de los contratos a los que se ha aplicado el proceso de limpieza, y en donde se crean algunas columnas que permiten identificar, entre otros aspectos los montos adjudicados por entidad y por año, el monto al que le adjudicaron o que ganó un proveedor por año, número de contratos adjudicados por proveedor por año, etc.; es decir, se contabiliza o sumariza la información en base a las relaciones entre proveedores y entidades por año. De esta vista se realiza la exportación de datos considerando el año del procesamiento.

```
--Esta vista contiene todas las metricas y sirve para entrenar los modelos de
ML
DROP MATERIALIZED VIEW IF EXISTS mv_dataset_base_ml CASCADE;
CREATE MATERIALIZED VIEW mv_dataset_base_ml AS
WITH base AS (
    SELECT
        mv.*,
        COALESCE(mv.award_amount, 0)::numeric AS monto_base_modelo
    FROM mv_dataset_base_limpio mv
    WHERE mv.award_suppliers_id IS NOT NULL
      AND mv.ruc_entidad_compradora IS NOT null
      and mv.proceso_con_insumos_medicos = 1
),

entidad_anio AS (
    SELECT
        anio_proceso,
```

```
        ruc_entidad_compradora,
        COUNT(*) AS total_contratos_entidad_anio,
        SUM(monto_base_modelo)::numeric AS monto_total_entidad_anio
    FROM base
    GROUP BY anio_proceso, ruc_entidad_compradora
),

proveedor_entidad_anio AS (
    SELECT
        anio_proceso,
        ruc_entidad_compradora,
        award_suppliers_id,
        COUNT(*) AS contratos_proveedor_entidad_anio,
        SUM(monto_base_modelo)::numeric AS monto_proveedor_entidad_anio
    FROM base
    GROUP BY anio_proceso, ruc_entidad_compradora, award_suppliers_id
),

entidad_hist AS (
    SELECT
        ruc_entidad_compradora,
        COUNT(*) AS total_contratos_entidad_hist,
        SUM(monto_base_modelo)::numeric AS monto_total_entidad_hist
    FROM base
    GROUP BY ruc_entidad_compradora
),

proveedor_entidad_hist AS (
    SELECT
        ruc_entidad_compradora,
        award_suppliers_id,
        COUNT(*) AS contratos_proveedor_entidad_hist,
        SUM(monto_base_modelo)::numeric AS monto_proveedor_entidad_hist
    FROM base
    GROUP BY ruc_entidad_compradora, award_suppliers_id
)

SELECT
    b.ocid,
```

```
    b.tipo_contratacion,
    b.tipo_institucion_salud,
    b.tiene_preguntas,
    b.licitacion_numero_proveedores,
    b.categoria_compra_adquisiciones,
    b.metodo_adquisicion,
    b.planeacion_fecha,
    b.contratacion_fecha_firma_primera,
    b.award_amount,
    b.sri_fecha_inicio_actividades,
    b.dias_maduracion_empresa,
    b.dias_entre_planeacion_firma_contrato,
    b.anio_proceso,
    b.ruc_entidad_compradora,
    b.award_suppliers_id,
    b.monto_base_modelo,

    pea.contratos_proveedor_entidad_anio,
    pea.monto_proveedor_entidad_anio,
    ea.total_contratos_entidad_anio,
    ea.monto_total_entidad_anio,

    ROUND(
        pea.contratos_proveedor_entidad_anio::numeric
        / NULLIF(ea.total_contratos_entidad_anio, 0)::numeric,
        6
    ) AS porcentaje_contratos_proveedor_entidad_anio,

    ROUND(
        pea.monto_proveedor_entidad_anio::numeric
        / NULLIF(ea.monto_total_entidad_anio, 0)::numeric,
        6
    ) AS porcentaje_monto_proveedor_entidad_anio,

    peh.contratos_proveedor_entidad_hist,
    peh.monto_proveedor_entidad_hist,
    eh.total_contratos_entidad_hist,
    eh.monto_total_entidad_hist,
```

```
    ROUND(
        peh.contratos_proveedor_entidad_hist::numeric
        / NULLIF(eh.total_contratos_entidad_hist, 0)::numeric,
        6
    ) AS porcentaje_contratos_proveedor_entidad_hist,

    ROUND(
        peh.monto_proveedor_entidad_hist::numeric
        / NULLIF(eh.monto_total_entidad_hist, 0)::numeric,
        6
    ) AS porcentaje_monto_proveedor_entidad_hist

FROM base b
LEFT JOIN proveedor_entidad_anio pea
    ON b.anio_proceso = pea.anio_proceso
   AND b.ruc_entidad_compradora = pea.ruc_entidad_compradora
   AND b.award_suppliers_id = pea.award_suppliers_id

LEFT JOIN entidad_anio ea
    ON b.anio_proceso = ea.anio_proceso
   AND b.ruc_entidad_compradora = ea.ruc_entidad_compradora

LEFT JOIN proveedor_entidad_hist peh
    ON b.ruc_entidad_compradora = peh.ruc_entidad_compradora
   AND b.award_suppliers_id = peh.award_suppliers_id

LEFT JOIN entidad_hist eh
    ON b.ruc_entidad_compradora = eh.ruc_entidad_compradora

WITH DATA;
```