



MAESTRÍA EN CIENCIA DE DATOS Y MÁQUINAS DE APRENDIZAJE CON MENCIÓN EN INTELIGENCIA ARTIFICIAL

**Trabajo previo a la obtención de título de
Magister en Ciencia de Datos y Máquinas de
Aprendizaje con mención en Inteligencia
Artificial**

Cajas Tapia Kevin Alexander

Chavez Casares Lenin Fernando

Conde Zhingre Lorena Elizabeth

Lastre Baquerizo David Jamil

ANDRÉS ROBERTO NAVAS CASTELLANOS

PABLO RENATO ESCOBAR VALLEJOS

Modelos predictivos para detección de señales de riesgo de
corrupción en licitaciones municipales de Pichincha

Certificación de autoría

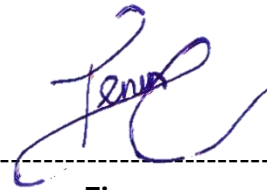
Nosotros, **Cajas Tapia Kevin Alexander, Chavez Casares Lenin Fernando, Conde Zhingre Lorena Elizabeth y Lastre Baquerizo David Jamil**, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y 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.



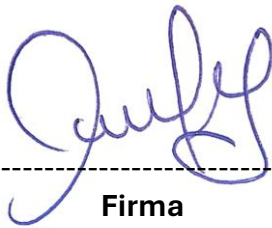
Firma

Cajas Tapia Kevin Alexander



Firma

Chavez Casares Lenin Fernando



Firma

Conde Zhingre Lorena Elizabeth



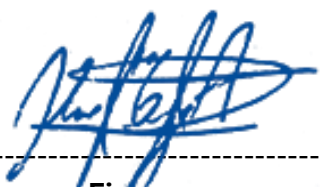
Firma

Lastre Baquerizo David Jamil

Autorización de Derechos de Propiedad Intelectual

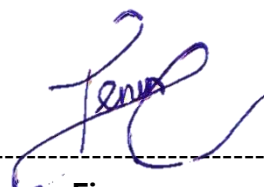
Nosotros, **Cajas Tapia Kevin Alexander, Chavez Casares Lenin Fernando, Conde Zhingre Lorena Elizabeth y Lastre Baquerizo David Jamil**, en calidad de autores del trabajo de investigación titulado ***Modelos predictivos para detección de señales de riesgo de corrupción en licitaciones municipales de Pichincha***, 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.

24. Julio. Quito, (julio 2026)



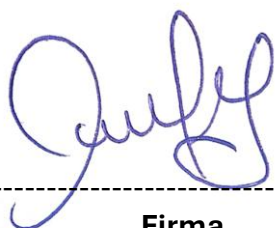
Firma

Cajas Tapia Kevin Alexander



Firma

Chavez Casares Lenin Fernando



Firma

Lorena Elizabeth Conde Zhingre

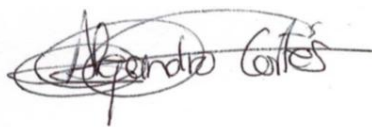


Firma

Lastre Baquerizo David Jamil

Aprobación de dirección y coordinación del programa

Nosotros, **Director Alejandro Cortés López de EIG y Coordinadora Karla Estefanía Mora Cajas de UIDE**, declaramos que: **Cajas Tapia Kevin Alexander, Chavez Casares Lenin Fernando, Conde Zhingre Lorena Elizabeth y Lastre Baquerizo David Jamil**, 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

A todos quienes han apoyado y motivado nuestra formación profesional, a nuestros docentes quienes nos motivaron a transformar los datos para fomentar la transferencia en las instituciones públicas.

AGRADECIMIENTOS

A nuestra familia por ser parte fundamental en este proceso, a la Universidad Internacional del Ecuador por su compromiso con la formación de profesionales de excelencia.

RESUMEN

La contratación pública en Ecuador posee un conjunto de procesos mediante los cuales las instituciones públicas adquieren bienes y/o servicios, consultorías entre otros, su propósito es garantizar la transparencia, igualdad de oportunidades, eficiencia, todo ello basado en la Ley organiza del Sistema Nacional de Contratación Pública. El manejo de los recursos del estado a través del proceso de contratación más comunes como subasta inversa, licitaciones, ínfima cuantía, entre otros genera posibilidades para irregularidades como colusión, fraccionamiento y sobre costos. Detectar estas anomalías a tiempo es primordial para la gobernanza pública, pero las auditorías tradicionales dependen de verificaciones manuales y denuncias, un proceso lento y reactivo por naturaleza. Frente a este impedimento, la presente investigación emplea técnica de ciencias de datos para consolidar los mecanismos de control y supervisión del Estado.

El objetivo general es identificar señales estadísticas de riesgo en los procesos de contratación pública otorgados, con una evaluación en los gobiernos municipales de la Provincia de Pichincha. Para ello se utilizaron datos abiertos de Servicio Nacional de Contratación Pública (SERCOP) correspondientes al periodo abarcado entre enero 2022 y diciembre de 2025, bajo un planteamiento cuantitativo y analítico.

La recopilación de datos se organizó en función de variables claves de rendimiento, como los montos de los contratos, el número de oferentes, la concentración de adjudicaciones, los tiempos de procesos y la recurrencia de contratistas específicos. Sobre la base de esta información, se crearon indicadores específicos, medir el riesgo definiendo una variable binaria objetiva: 0 para procesos sin señal de riesgo y 1 para procesos con señal de riesgo.

Se entrenaron varios algoritmos de aprendizaje para resolver esta tarea de clasificación entre ellos Random Forest, XGBoost, Regresión Logística y Máquinas de Vectores de Soporte (SVM). El desarrollo requirió una extensa limpieza de datos, variables y normalización, así como una estricta separación de los datos de entrenamiento y prueba, y realizar una validación cruzada para evitar el sobreajuste.

El rendimiento de los clasificadores se evaluó utilizando métricas sólidas listas para usar, entre las que sobresalen el Recall, el F1-score y la curva ROC-AUC. El modelo LightGBM, combinado con estrategias de balanceo de clases, obtuvo los mejores resultados predictivos para distinguir procesos normales de anómalos, superando a Random Forest, XGBoost, Regresión Logística y SV. La salida principal del sistema es una probabilidad matemática de riesgo destinado a cada licitación; es imprescindible indicar que el modelo no determina la existencia de corrupción, si no que trabaja como un sistema de alerta temprana basado en desviaciones estadísticas.

En síntesis, esta herramienta tecnológica maximiza la supervisión del Estado y facilita el control social al priorizar las auditorías hacia los contratos realmente sospechosos. Esta variación metodológica admite transformar la fiscalización reactiva en una planificación analítica de prevención continua.

Palabras Claves: Contratación pública, Ciencia de Datos, Detección de anomalías, Random Forest, Gobernanza digital, Sistemas de alerta temprana, Datos abiertos.

ABSTRACT

Public procurement in Ecuador involves vast financial resources, which creates opportunities for irregularities such as collusion, splitting of contracts, and cost overruns. Detecting these anomalies in a timely manner is essential for public governance, but traditional audits rely on manual verifications and complaints a process that is slow and reactive by nature. To address this challenge, this study employs data science techniques to strengthen the government's control and oversight mechanisms.

The overall objective is to identify statistical risk signals in awarded public procurement processes, with a focus on municipal governments in the Province of Pichincha. To this end, open data from the National Public Procurement Service (SERCOP) covering the period from January 2023 to December 2025 was used, employing a quantitative and analytical approach.

Data collection was organized around key performance variables, such as contract amounts, the number of bidders, the concentration of awards, process durations, and the recurrence of specific contractors. Based on this information, specific indicators were created to measure risk by defining an objective binary variable: 0 for processes without a risk signal and 1 for processes with a risk signal.

Several machine learning algorithms were trained to solve this classification task, including Random Forest, XGBoost, Logistic Regression, and Support Vector Machines (SVM). The development process required extensive data cleaning, variable handling, and normalization, as well as strict separation of the training and test datasets and cross-validation to avoid overfitting.

The performance of the classifiers was evaluated using robust, off-the-shelf metrics, notably Recall, F1-score, and the ROC-AUC curve. Random Forest and XGBoost achieved the best predictive results for distinguishing normal processes from anomalous ones. The system's primary output is a mathematical risk probability assigned to each bid; it is essential to note that the model does not determine the existence of corruption, but rather functions as an early-warning system based on statistical deviations.

In summary, this technological tool maximizes government oversight and facilitates social control by prioritizing audits of contracts that are genuinely suspicious. This methodological shift makes it possible to transform reactive oversight into analytical planning for continuous prevention.

Keywords: Public Procurement, Data Science, Anomaly Detection, Random Forest, Digital Governance, Early Warning Systems, Open Data.

TABLA DE CONTENIDOS (Índice)

CAPITULO 1:	13
1. INTRODUCCIÓN.....	13
1.1. Definición del Proyecto	13
1.2. Justificación e importancia del Trabajo de Investigación	15
1.3. Alcance.....	17
1.3.1. Alcance Operativo (Modalidad de Contratación).....	17
1.3.2. Alcance temporal.....	18
1.3.3. Alcance Institucional y Geográfico	18
1.3.4. Alcance Tecnológico	19
1.3.5. Límites del Sistema	19
1.4. Objetivos.....	20
1.4.1. Objetivo General	20
1.4.2. Objetivo Especifico	20
CAPITULO 2:	21
2. Revisión de Literatura	21
2.1. Estado del Arte.....	21
2.1.1. Estudios Recientes Sobre Corrupción en Contratación Pública.....	21
2.1.2. Trabajos relacionados con detección de indicadores de riesgo en contrataciones públicas.....	21
2.1.3. Analítica de datos aplicada a la contratación Pública.	23
2.1.4. Machine learning y la detección de señales estadísticas de riesgo.	24
2.1.5. Modelos de clasificación supervisada aplicados a fraude, corrupción o riesgo.....	25
2.2. Marco Teórico	26
2.2.1. Importancia de la Contratación Pública	26
2.2.2. Tipos de Contratación Pública	26
2.2.3. Procesos de Licitación Pública y Adjudicados.....	28
2.2.4. Riesgos en Procesos de Contratación Pública.....	29
2.2.5. Riesgos por fase de contratación.....	31
2.2.6. Tipologías de fraude y riesgo	31
2.2.7. Señales e indicadores de riesgo en Contratación Pública.	33
2.2.8. Machine Learning y sus Modelos	38
2.2.8.1. Modelos Supervisado	38
2.2.8.2. Aprendizaje No Supervisado	40

2.2.8.3. Aprendizaje Automático Predictivo y Preventivo	40
2.2.8.4. Algoritmos de ensamblaje y Clasificación Supervisada.....	41
2.2.9. Analítica de Datos en Contratación Pública.	42
2.2.10. Indicadores y Métricas de Evaluación.	44
CAPITULO 3:	46
3. DESARROLLO.....	46
3.1. Metodológico CRISP-DM	46
3.2. Comprensión del Negocio (Business Understanding)	47
3.3. Preparación de los Datos.....	50
3.3.1. Selección de los Datos.....	50
3.3.2. Limpieza de los Datos	52
Duplicidad.	52
Valores Faltantes.....	53
Corrección de errores e inconsistencia.	53
Valores Nulos.....	54
Valores atípicos.....	56
3.3.3. Construcción de datos	57
3.3.4. Integración de Datos.....	59
3.3.5. Formateo de Datos	59
3.4. Modelado	62
3.4.1. Selección de la técnica de modelado.....	62
Desviación estándar y puntaje Z.....	65
Coeficiente de variación.	66
Percentiles.....	66
Monto adjudicado atípico.	66
Variación entre presupuesto y adjudicación (sobrepeso relativo).....	66
Baja competencia.....	67
Duración del proceso.	67
Concentración por proveedor y concentración de mercado.	67
Recurrencia entidad-proveedor.	68
Criterios de clasificación del riesgo.	68
3.4.2. Construcción del Modelo	69
3.5. Evaluación	71
3.6. Metodología Aplicada.....	75

4. ANÁLISIS DE RESULTADOS.....	77
4.1. Pruebas de Concepto y preparación del entorno analítico	77
4.2. Ingeniería de Variables Basada en Literatura.....	78
4.3. Resultados de las pruebas.	80
4.4. Interpretabilidad e Importancia de Variables	87
5. CONCLUSIONES Y RECOMENDACIONES	91
5.1. Conclusiones.....	91
5.2. Recomendaciones	93
6. REFERENCIAS BIBLIOGRÁFICAS	95

LISTA DE TABLAS (Índice de tablas)

Tabla 1 Codificación de la variable objetivo	14
Tabla 2 Variables y descripción	49
Tabla 3 Criterios para selección de datos	50
Tabla 4 Variables por grupos	51
Tabla 5 Indicadores derivados que operacionalizan las banderas	58
Tabla 6 Integridad de datos.....	59
Tabla 7 Codificación de variable categórica.....	61
Tabla 8 Variables estadísticas y criterios de riesgo para la construcción de variables objetivo	64
Tabla 9 Medidas estadísticas empleadas para el cálculo de indicadores de riesgo	65
Tabla 10 Criterios de clasificación del riesgo por indicador	68
Tabla 11 Regla general de clasificación del riesgo.....	68
Tabla 12 Distribución del conjunto de datos por nivel de riesgo	72
Tabla 13 Distribución de los conjuntos de entrenamiento y prueba.....	72
Tabla 14 Resultados de validación cruzada por modelo (RepeatedStratifiedKFold, k=5, 3 repeticiones).....	73
Tabla 15 Resultados del estudio de ablación (LightGBM, conjunto de prueba).....	75
Tabla 16. Matriz de alineación teórica: Indicadores del modelo predictivo frente a banderas rojas	79
Tabla 17 Métricas de evaluación en el conjunto de prueba independiente (LightGBM, n=534)....	80

LISTA DE FIGURAS (Índice de figuras)

Figura 1 Tipos de contrataciones	27
Figura 2 Fases CRISP-DM	47
Figura 3 API SERCOP	48
Figura 4 Duplicados de datos	52
Figura 5 Detección de inconsistencias	54
Figura 6 Semáforo porcentaje de valores nulos.....	55
Figura 7 Variables con valores atípicos	57
Figura 8 Composición del conjunto de datos	61
Figura 9 Arquitectura del Modelo	70
Figura 10 Flujo metodológico para la selección del modelo definido	71
Figura 11 Matriz de confusión.....	74
Figura 12 Matriz de confusión del modelo LightGBM.....	82
Figura 13 Mapa de calor del reporte de clasificación por nivel de riesgo.....	83
Figura 14 Curva Precisión-Recall para clases Alto+Crítico (PR-AUC = 0,954)	83
Figura 15 Gráfico de radar comparativo del rendimiento de los modelos de ensamble en las principales métricas de evaluación	86
Figura 16 Comparación de métricas de rendimiento por modelo	87
Figura 17 Top 20 variables de mayor importancia predictiva del modelo LightGBM	89
Figura 18 Interfaz Ingesta y Curaduría de Datos.....	90
Figura 19 Análisis exploratorio de datos	90

CAPITULO 1:

1. INTRODUCCIÓN

1.1. Definición del Proyecto

En Ecuador la contratación pública es el principal mecanismo mediante el cual los gobiernos municipales ejecutan sus políticas públicas y administran los recursos del estado, dentro de las contrataciones se abordan las licitaciones que son procedimientos mediante el cual los gobiernos adquieren bienes, servicios u obras, invitando a proveedores a presentar sus ofertas para seleccionar las más convenientes según las necesidades y presupuesto (Torres Berrú, 2023). De acuerdo con la Contraloría General del Estado (2023), estos procesos tienen como objetivo garantizar el uso eficiente de los recursos públicos, basando en los principios de transparencia, igualdad, legitimidad y eficiencia. En la actualidad los procesos de licitaciones involucran un alto volumen de recursos involucrados lo que hace que se puedan presentar prácticas irregulares que afecten la eficiencia, transparencia y legalidad de la gestión pública. Adicionalmente, existe una carencia de herramientas automatizadas; los métodos de auditoría actuales son predominantemente manuales y reactivos, por lo que resultan insuficientes ante el volumen de transacciones. Este estudio busca cubrir ese vacío científico introduciendo un enfoque analítico predictivo que permita a las entidades públicas priorizar revisiones de forma escalable y objetiva.

En la provincia de Pichincha, los gobiernos municipales realizan constantemente procesos de licitación para la adquisición de bienes, servicios y ejecución de obras públicas, en la actualidad existen marcos normativos y regulatorios que buscan garantizar la transparencia y eficiencia como la ley organiza del sistema de Contratación pública (LOSNCP) y la regulación del servicio Nacional de Contratación Pública (SERCOP), sin embargo investigaciones y la percepción ciudadana evidencian riesgos asociados a la corrupción administrativa, frente a ello el Estado Ecuatoriano promueve la apertura de datos públicos como estrategia para fortalecer la

transparencia y el control social. Sin embargo, estos datos no han sido analizados ni tratados para identificar señales de riesgo en los procesos de licitación a nivel municipal.

Basado en ello, nuestro problema de investigación se define como la necesidad de desarrollar un modelo analítico que permitan clasificar los procesos de licitación municipal basados en indicadores estadísticos objetivos.

El proyecto busca categorizar procesos existentes según la clasificación: modelos con señal de riesgos, modelos sin señal de riesgo, implementado con ello un modelo de clasificación supervisada, donde cada proceso adjudicado constituye una unidad de análisis, fortaleciendo de esta manera el principio de transparencia, igualdad, legitimidad y eficiencia. Para el entrenamiento del modelo la variable objetiva se define como:

Tabla 1

Codificación de la variable objetivo

Codificación	Descripción
0	Procesos sin señal de riesgo (Bajo / Medio)
1	Procesos con señal de riesgo (Alto / Crítico)

Para la salida principal del modelo, se considera la probabilidad asociada a la clase de riesgo, logrando así una evaluación continua del nivel de riesgo de cada proceso de licitación. El procesamiento de los datos abiertos permitirá detectar posibles señales de riesgo de corrupciones administrativas en los procesos de licitación de los gobiernos municipales de la provincia de Pichincha.

Nuestra pregunta de investigación se centra en ¿De qué manera el análisis de datos abiertos del SERCOP con técnicas de aprendizaje automático permite identificar patrones de posibles señales estadísticas de riesgo administrativa dentro de los procesos de licitación de los gobiernos municipales de la provincia de Pichincha?

Nuestro proyecto se limita a los procesos de licitación adjudicados por los gobiernos municipales de la provincia de Pichincha desde el 1 de enero del 2022 al 31 de enero del 2024, considerando únicamente procesos adjudicados bajo las modalidades licitación pública esto permite la comparabilidad estadística.

Acorde al SERCOP los datos utilizados están en formato CSV o Excel, para el tratamiento de los datos se aplicará: Limpieza y depuración, estandarización de variables, clasificación de variables categóricas y construcción de variables derivadas.

1.2. Justificación e importancia del Trabajo de Investigación

La contratación pública constituye uno de los principales mecanismos de asignación de recursos del estado. En los municipios, los procesos de licitación determinan la ejecución de obras, servicios e inversiones que impacta directamente en la calidad de vida de la ciudadanía, todos estos recursos convierten a la contratación pública susceptible a señales estadísticas de riesgo.

Varios informes nacionales e internacionales muestran que las prácticas irregulares en procesos de contratación provocan sobrecostos, desviaciones presupuestarias (Decarolis & Giorgiantonio, 2022), esto también se evidencia dentro del contexto nacional, donde análisis de auditoría y estudios de fraude contractual en gobiernos autónomos descentralizados evidencian sobrepagos recurrentes en las adquisiciones locales (Torres, 2023). En la provincia de Pichincha, los municipios manejan importantes presupuestos para obras, adquisiciones y servicios estratégicos; Sin embargo, los mecanismos de control tradicionales suelen ser manuales y limitados, lo que dificulta la detección temprana de infracciones.

En este contexto, el trabajo actual está motivado por la necesidad de incluir herramientas de ciencia de datos e inteligencia artificial que permite analizar grandes cantidades de información disponible en portales de datos abiertos. Si bien estos datos son públicos, rara vez se utilizan para generar inteligencia operativa que facilite la detección de patrones de riesgo relacionados con posibles prácticas corruptas, tales como: adjudicaciones repetidas de contratos a los mismos

proveedores, montos atípicos, procesos con baja competencia, tiempos inusualmente cortos o largos retrasos en las etapas de licitación, colusión entre postores, entre otros.

Esta investigación desarrolla un enfoque metodológico basado en clasificación supervisada que permitan identificar posibles riesgos en los procesos de licitaciones municipal. Iniciando desde la construcción de indicadores estadísticos que permitan operacionalizar la variable de riesgo de manera cuantificable y no subjetiva.

La importancia de este estudio es su potencial impacto en la transparencia y la gobernanza pública, ofreciendo un método replicable que puede aplicarse no sólo en Pichincha, sino también en otros municipios del país. Es, asimismo un aporte académico que integra enfoques informáticos con problemas sociales y administrativos que demuestran la importancia del análisis algorítmico de la contienda contra la corrupción.

Por último, esta investigación promueve el uso activo de los datos abiertos como recurso estratégico para el control social y la mejora de la gobernanza estatal. El uso de técnicas avanzadas de análisis de datos apoya a crear un entorno institucional más transparente, eficiente y justo, también muestra que la tecnología puede convertirse en esencial en la prevención y detección de infracciones administrativas. Por tanto, la investigación es pertinente, oportuna y socialmente valiosa

Como recalca Torres Berrú (2023), en Ecuador, las compras públicas personifican un espacio crítico de asignación de recursos y, en consecuencia, un punto sensible frente a riesgos de colusión, direccionamiento, fraccionamiento, baja competencia y otros comportamientos irregulares.

El enfoque provincial (Pichincha) es significativo puesto que permite analizar heterogeneidad institucional y territorial: municipios con distintas capacidades administrativas, niveles de contratación, tipos de obra/servicio y dinámicas de proveedores. Esta demarcación faculta construir comparaciones, detectar outliers y proponer umbrales más realistas que los

promedios nacionales. Además, al usar datos abiertos, el proyecto se alinea con principios de gobierno abierto: reutilización de datos públicos para control social, periodismo de datos, investigación académica y mejora institucional.

En conclusión, el proyecto tiene relevancia social porque la corrupción en contratación pública se relaciona con pérdidas económicas, deterioro de servicios, inequidad y desconfianza institucional. Proporciona una herramienta basada en datos abiertos y metodologías de Ciencia de Datos que contribuye a edificar un ecosistema donde la prueba cuantitativa sea parte de la conversación pública. En suma, la naturaleza del proyecto es coherente con una tesis de maestría orientada a impacto público, combinando rigurosidad técnica, reproducibilidad y utilidad práctica en el contexto ecuatoriano y específicamente en los gobiernos municipales de Pichincha.

1.3. Alcance

Este estudio tiene como punto de partida la intersección entre la ciencia de datos y gobernanza digital, por lo que debemos establecer límites para garantizar la viabilidad tanto del modelo predictivo como a nivel estadístico. A continuación, se detallan los límites de acuerdo con los siguientes parámetros:

1.3.1. Alcance Operativo (Modalidad de Contratación)

En el presente estudio solamente se analizarán los procedimientos de contratación adjudicados mediante **licitación pública** (proceso que abarcan bienes, obras y servicios). Se excluyen otras modalidades presentes en la SERCOP como ínfima cuantía, menor cuantía y subasta inversa electrónica. Las razones por las que se excluyen dichas modalidades se deben a que las licitaciones públicas concentran gran parte del presupuesto del Estado Ecuatoriano con especificaciones técnicas de cumplimiento complejas, por lo que la probabilidad de que exista sobrecostos, manipulación en el proceso o prácticas colusorias aumenta. Además, el proceso de licitación pública se rige por un ciclo de vida extenso que abarca las fases precontractuales, contractuales y evaluación técnica. Este proceso multietapa genera varios estados de alta

dimensión. A partir de esto, extraemos métricas altamente granulares, como la latencia entre la publicación de la licitación la apertura de las ofertas, la dispersión de las ofertas y las desviaciones presupuestarias posteriores a la adjudicación, cruciales para diseñar las señales de alerta que entrenan nuestros algoritmos de clasificación supervisada.

1.3.2. Alcance temporal

Se restringe el periodo de análisis a los procesos de contratación pública publicados y adjudicados estrictamente entre el 1 de enero de 2022 y el 31 de diciembre de 2024. Este período de tres años se seleccionó deliberadamente. Este análisis captura un ciclo regulatorio altamente representativo bajo SERCOP, justificado por los siguientes parámetros.

Mitigación de anomalías atípicas (valores atípicos históricos), aislar este período específico nos permite analizar el comportamiento habitual de la administración pública dentro de una época pospandémica estabilizada. Excluimos deliberadamente los años fiscales 2020 y 2021. Durante este período, las declaraciones de emergencia provocadas por la COVID-19 obligaron a una relajación de los protocolos de contratación pública introduciendo mecanismos de tramitación acelerada y altamente irregulares que distorsionarían gravemente la base matemática de un modelo de aprendizaje automático calibrado para detectar la corrupción en condiciones administrativas normales.

Tres años consecutivos de registros transaccionales proporcionan datos suficientemente representativos para entrenar y validar nuestros modelos de aprendizaje conjunto (específicamente Random Forest y XGBoost). Esta ventana temporal garantiza que los patrones de riesgo extraídos capturen las realidades contemporáneas de las operaciones municipales. De esta manera, evitamos datos atípicos introducidos por pandemia.

1.3.3. Alcance Institucional y Geográfico

Se limita la investigación a los procesos de contratación pública ejecutados exclusivamente por los Gobiernos Autónomos Descentralizados (GAD) de la provincia de Pichincha. Esto abarca

nueve entidades públicas fundamentales: el GAD Provincial de Pichincha, la Municipalidad del Distrito Metropolitano de Quito (DMQ) y siete GAD cantonales (Cayambe, Mejía, Pedro Moncayo, Pedro Vicente Maldonado, Puerto Quito, Rumiñahui y San Miguel de los Bancos). El límite propuesto por estos cantones tiene como objetivo la comparativa entre estos cantones manteniendo la heterogeneidad.

1.3.4. Alcance Tecnológico

Como fuente de información se utilizan los datos abiertos del Servicio Nacional de Contratación Pública (SERCOP), que se rigen con el protocolo internacional OCDS 1.1 (Open Contracting Data Standard). Adicional, el foco se fijó en las fases de preprocesamiento, depuración (limpieza) e implementación de clasificadores supervisados de árboles de decisión (Random Forest y XGBoost), estas fases corresponden al ciclo de vida de ML (Machine Learning) con datos estructurados.

1.3.5. Límites del Sistema

Se delimita este sistema computacional propuesto estrictamente de forma analítica, probabilística y preventiva. Debido a que se genera una señal de riesgo que no es nada más que una probabilidad matemática de riesgo para cada licitación, basada únicamente en anomalías estadísticas. En ningún caso estos resultados computacionales constituyen un veredicto legal, una auditoría financiera vinculante ni una prueba judicial de culpabilidad. Funciona como un mecanismo de alerta temprana. Está diseñado para ayudar a los organismos de supervisión a identificar riesgos.

1.4. Objetivos

1.4.1. Objetivo General

Identificar señales de riesgos de corrupción en los procesos de licitación adjudicados de los gobiernos municipales de Pichincha mediante el análisis de datos abiertos con modelos de clasificación supervisada.

1.4.2. Objetivo Especifico

- Depurar los datos abiertos provenientes del SERCOP correspondientes a los procesos de licitación adjudicados por los gobiernos autónomos descentralizados de Pichincha durante el periodo 2022-2024.
- Construir indicadores estadísticos de riesgo contractual empleando métricas fundamentadas en percentiles y desviaciones estándar.
- Interpretar el desempeño de modelos de clasificación supervisada orientados a estimar la probabilidad de riesgos en los procesos de licitación.
- Validar el modelo de clasificación mediante las métricas Recall, F1-score y ROC-AUC para verificar su capacidad de generalización.

CAPITULO 2:

2. Revisión de Literatura

2.1. Estado del Arte

2.1.1. Estudios Recientes Sobre Corrupción en Contratación Pública

La contratación pública representa una de las actividades gubernamentales más susceptibles a la corrupción y el fraude, debido al alto volumen de recursos económicos involucrados y la complejidad de sus procesos. Existen diferentes estudios de mapeo sistemático como el de Schneider dos Santos et al. (2025), que demuestra que la literatura científica experimenta un crecimiento exponencial en la búsqueda de metodologías cuantitativas capaces de suplantar las auditorías tradicionales, por auditorías automatizadas y tiempos de respuesta mínimos.

Con estos antecedentes, Lyra et al. (2022) propone un análisis donde se identifican posibles modalidades más recurrentes en riesgos de corrupción estudiadas a nivel global como por ejemplo el favoritismo político y el desvío de fondos, Según Gnaldi y Del Sarto (2024), en procesos en tiempos de emergencia los marcos regulatorios tienden a flexibilizarse y los mecanismos de control tienden a debilitarse, esto incrementa significativamente la vulnerabilidad de los sistemas de contratación.

2.1.2. *Trabajos relacionados con detección de indicadores de riesgo en contrataciones públicas.*

Algunas literaturas, para la detección de irregularidades, usan habitualmente el diseño y validación de "indicadores de riesgo" (red flags). Esto consiste en calcular índices compuestos a partir de señales transaccionales. Como ejemplo de esto es esta el Índice de Riesgo de Corrupción (CRI, por sus siglas en inglés), utilizado ampliamente en la Unión Europea. Según Longobucco y Ferwerda (2025), este índice sintetiza cinco señales críticas:

- La presencia de licitaciones con un único postor.

- Periodos extremadamente cortos para la presentación de ofertas.
- Modificaciones o líneas de tiempo irregulares en las decisiones.
- Uso injustificado de procedimientos no competitivos (como la adjudicación directa).
- Elevada concentración de contratos solo entre compradores y vendedores específicos.

Según, Decarolis y Giorgiantonio (2020) en el artículo sobre señales de alerta de corrupción en la contratación públicas analizaron este espectro de alertas analizando licitaciones italianas y generando variables claves directamente de los procesos operativas de las fuerzas del orden y del poder judicial. Sin embargo, todavía se requiere un sustento metodológico profundo para evitar falsos positivos.

Gnaldi y Del Sarto (2023) en sus investigaciones proponen la evaluación y validación de estas banderas rojas bajo el marco de la Teoría de Respuesta al Ítem (IRT), validando que el riesgo de corrupción posee una estructura compuesta por varias dimensiones que evidencian de manera precisa las prácticas que limitan transparencia en los procesos de contrataciones.

Actualmente las técnicas de ciencia de datos fortalecen la auditoría tradicional hacia modelos predictivos y preventivos en tiempos más cortos. Los datos son los activos más importantes de las organizaciones, las instituciones en su principio de transparencia divulgan datos abiertos los cuales son accesibles a todo público, estos datos pueden ser analizados y estudiados para detectar señales de riesgos de posibles fraudes de forma eficiente y a corto plazo.

- Según, Aldana et al. (2022) en México utilizaron un "hyper-forest" que es un ensamble de Random Forest para predecir posibles contratos corruptos. Este modelo se basa en árboles que gestionan con éxito conjuntos de datos desbalanceados.

- La experiencia de Colombia presenta una evidencia relevante según Gallego et al. (2021) y Umbarila Suárez (2024) en su investigación aplicaron Gradient Boosting Machine (GBM) para detectar variables estadísticas de riesgo, sobrecostos y retrasos. Rodríguez Arévalo (2020) alcanzó una precisión superior al 90% con CatBoost al predecir adiciones contractuales en Bogotá.
- Otra estrategia desarrollada por Villa Pedroza (2021) implementó técnicas de sobremuestreo (SMOTE) y selección de características (CFS) con algoritmos Naïve Bayes, SVM y Random Forest. Este enfoque clasifica el riesgo de incremento en el valor de licitaciones de obra pública.
- En la investigación de Torres-Berru y López Batista (2021) en su investigación se enfocaron con SOM y K-Means en Ecuador para detectar anomalías en la calificación de procesos. Posteriormente usaron SVM como modelo semi-supervisado y PCA para identificar posibles irregularidades en los procesos. El sistema alcanzó un 95% de precisión. En México, Álvarez Suárez y Nieto Morales (2025) usaron la técnica de agrupamiento jerárquico con la distancia de Jaccard, para clasificar los contratos según los niveles de riesgo sin requerir que los datos estén etiquetados.
- En la investigación de Colonnelli et al. (2022) analizaron el uso de SuperLearner para auditar procesos en el país de Brasil. Así mismo en Italia De Blasio et al. (2020) implementaron árboles de clasificación para detectar posibles delitos de corrupción. Finalmente, en otra investigación de Decarolis y Giorgiantonio (2022) estudiaron proyectos de infraestructura mediante algoritmos de regularización Lasso y Ridge.

2.1.3. Analítica de datos aplicada a la contratación Pública.

Los procesos de contratación pública está dentro del proceso digitalización, así como la adopción de normativas de datos abiertos que faciliten el uso de esos datos para el procesamiento masivo de información gubernamental. Vaqueiro et al. (2023) menciona que, a pesar de la disponibilidad legal de los datos en portales abiertos, una gran parte de la información estratégica se encuentra en formatos de texto no estructurado o portales fragmentados, lo que minimiza el control social y los procesos de auditoría, por ello el uso de técnicas de ciencia de datos es primordial para el tratamiento y análisis de datos.

Frente a ello y para dar una solución problemática de integración de datos, se han desarrollado Sistemas de Soporte a las decisiones (DSS) basados consolidación y la integridad de los datos. Velasco et al. (2021) en su investigación estructura un marco que integra procesos robustos de extracción, transformación y carga (ETL) estos combinados con análisis estadísticos para identificar desvíos de recursos públicos, tomando como base esquemas reales de corrupción institucionalizada. En la actualidad, la analítica predictiva moderna no se limita únicamente a las bases de datos tabulares tradicionales; por ejemplo las investigaciones actuales (Adelanwa et al., 2023; Wachs et al., 2019) propone tratar el proceso de contratación con la ciencia de redes bipartitas donde existe la interacción entre entidades compradoras y adjudicatarias, logrando así visualizar la centralización del mercado y descubrir posibles patrones de colusión que la estadística descriptiva tradicional pasa por alto.

2.1.4. Machine learning y la detección de señales estadísticas de riesgo.

El avance de la analítica de datos estática hacia el aprendizaje automático o Machine Learning (ML) ha logrado ampliar la capacidad de las entidades de control para anticipar y predictivas en tiempo real, posibles riesgos de corrupción, bajo estos escenarios los datos etiquetados de posible corrupción con uso del aprendizaje no supervisado y la detección de anomalías juegan un rol estelar. El estudio de Torres-Berru y López Batista (2021) detallan el uso de minería de datos para identificar distorsiones y manipulación en los parámetros de calificación

técnica de los pliegos de contratación pública, logrando reforzar que los algoritmos para detección de anomalías pueden descubrir direccionamientos de contratos. De igual manera, Majilana (2025) validó la efectividad del algoritmo Isolation Forest acoplado a través de valores SHAP para validar transacciones públicas sin requerir un conocimiento previo del fraude, lo que facilita priorizar las alertas para luego ser analizadas por los auditores.

Complementariamente el aprendizaje automático basado en grafos (Graph Machine Learning) tiene alta relevancia por su capacidad para modelar para manejo de estructuras de datos representadas mediante grafos. Schneider dos Santos et al. (2025) en su estudio propone el uso de redes neuronales de grafos para robustecer la identificación de corrupción, seleccionando características topológicas directamente de las redes de relaciones de empresas implicadas en grandes escándalos de corrupción (como la operación Car Wash o Lava Jato), lo que incrementa a gran escala el rendimiento de los modelos en comparación con el uso exclusivo de datos tabulares.

2.1.5. Modelos de clasificación supervisada aplicados a fraude, corrupción o riesgo.

Cuando se dispone de registros históricos de empresas sancionadas, contratos investigados o auditorías con dictámenes negativos, la clasificación supervisada se posiciona como la herramienta más precisa para predecir el riesgo antes de la adjudicación. Longobucco y Ferwerda (2025) evaluaron de manera comparativa seis algoritmos supervisados (Distributed Random Forest, Gradient Boosting Machine, Deep Learning, Regularized Logistic Regression, Naive Bayes y Support Vector Machines) entrenados para clasificar contratos de alto riesgo, demostrando la alta precisión de los modelos de ensamble cuando se alimentan con datos limpios y registros sobre los beneficiarios reales (beneficial ownership) de las empresas.

Otra línea de investigación relacionada con los modelos híbridos, Ndolo y Kasyoka (2025) en su investigación integraron la Regresión Logística seleccionado por su claridad en la interpretación de coeficientes a nivel de políticas públicas, con el algoritmo Random Forest, capaz de identificar interacciones no lineales entre las variables, los resultados encontrados detallaron

que las ofertas de un único contratista y los historiales de denuncias contra los proveedores constituyen el indicador con mayor precisión para predecir la probabilidad de fraude. En las propuestas para detección de anomalías en concursos públicos, Araújo et al. (2024) usa el modelo One-Class SVM por su mejor desempeño en conjunto de datos desproporcionados entre sus clases. En el caso de Medina-Hernández et al. (2025) demuestra que el trabajar con clasificadores supervisados entrenados a partir de patrones de comportamiento históricos con datos de proveedores formalmente sancionados por el estado permite anticipar posibles casos de corrupción en procesos de contratación pública.

2.2. Marco Teórico

2.2.1. Importancia de la Contratación Pública

La contratación pública es importante para permitir que el estado adquiera servicios y bienes. Esto representa un porcentaje bastante alto en el gasto gubernamental. En países que conforman la Organización para la Cooperación y el Desarrollo Económicos (OCDE), este monto equivale al 13% PIB. La correcta gestión de este presupuesto, especialmente en obras crea un impacto en el desarrollo económico e impulso del comercio. La contratación pública al ser tan fundamental para la sociedad y manejar muchas transacciones y montos alto de dinero, puede ser mas vulnerable y propenso a irregularidades. Las compras públicas presentan señales estadísticas de riesgo e ineficiencia presupuestaria. Esto ocurre por el capital gestionado, la asimetría de información y la complejidad técnica de los contratos (Rodríguez Arévalo, 2020; Aldana et al., 2022). La corrupción genera competencia desigual, desvío de recursos y favoritismos propios del contratante. Esta práctica limita el acceso a servicios, genera competencia desleal y reduce el valor social del gasto público (Aldana et al., 2022; Álvarez Suárez & Nieto Morales, 2025).

2.2.2. Tipos de Contratación Pública

En el Ecuador, el marco normativo regulado por la Ley Orgánica del Sistema Nacional de Contratación Pública (LOSNCP) y supervisado por el Servicio Nacional de Contratación Pública (SERCOP) establece y clasifica los procedimientos de contratación pública en varias categorías fundamentales, dependiendo de la naturaleza del objeto del contrato (bienes, servicios, obras o consultoría) y del monto presupuestario estimado.

Figura 1

Tipos de contrataciones



Nota. Tomado de SERCOP (2023).

Las contrataciones de Régimen común se aplican cuando los bienes o servicios no son normalizados (tienen características complejas o específicas) o para la ejecución de obras. Su elección depende directamente del presupuesto referencial del proyecto, el cual se calcula anualmente según los coeficientes del Presupuesto Inicial del Estado (PIE).

- **Ínfima Cuantía:** Para compras pequeñas y de ejecución inmediata (cuyo valor no supere el coeficiente de 0.000002 del PIE). No requiere un proceso concursal complejo.
- **Menor Cuantía:** Se prioriza la contratación con micro y pequeñas empresas, artesanos o actores de la economía popular y solidaria local. Aplica para bienes/servicios no normalizados u obras de montos bajos.

- **Cotización:** Para proyectos de mediana escala. Se invita a proveedores que cumplan con el perfil técnico para que presenten sus propuestas.
- **Licitación:** Es el proceso más grande y riguroso. Se utiliza para la construcción de grandes obras o la adquisición de bienes y servicios complejos que superan los montos de los procesos anteriores (SERCOP, 2023).

Entre las modalidades del régimen común, esta investigación se centra en Licitaciones, ya que corresponde a procesos de contratación de mayor complejidad y cuantía económica, así mismo se cuenta con una mayor disponibilidad de información.

2.2.3. Procesos de Licitación Pública y Adjudicados

La licitación pública es un procedimiento de régimen común de carácter abierto, diseñado para garantizar la máxima concurrencia, transparencia y competencia entre los oferentes.

- **Ámbito de aplicación:** Se utiliza principalmente para la contratación de obras, o para la adquisición de
 - bienes y servicios no normalizados cuyo presupuesto referencial excede los montos máximos fijados anualmente por el Servicio Nacional de Contratación Pública (Asamblea Nacional, 2008).
- **Criterios de selección:** A diferencia de procesos dinámicos como la subasta inversa donde la prioridad absoluta es la puja hacia el precio más bajo, la licitación pondera de forma multidimensional la calidad, experiencia del oferente y propuesta económica con base en los pliegos (Armas, 2019).

Segun la Ley organica del Sistema Nacional de Contratación Pública (LOSNCP) a traves del Servicio Nacional de Contratación Pública (SERCOP), se divide en 3 fases fundamentales administradas en el Sistema Oficial de contratación del Estado (SOCE):

- **Fase Preparatoria (Planificación):**

- Plan anual de Contratación (PAC)
- Estudios previos, es decir especificaciones técnicas o términos de referencia, estudios de mercado y presupuesto referencial.
- Certificación presupuestaria
- Pliegos y resoluciones.
- Fase precontractual (Selección)
 - Convocatoria en el portal del SERCOP/SOCE
 - Preguntas y respuestas
 - Presentación de ofertas
 - Calificación de ofertas
 - Adjudicación
- Fase contractual (Ejecución)
 - Firma del contrato
 - Ejecución y administración
 - Recepción y liquidación

2.2.4. Riesgos en Procesos de Contratación Pública

La literatura sobre la integridad pública identifica comúnmente 3 categorías principales de riesgo en los procesos de contratación pública: corrupción, colusión e ineficiencia administrativa (OCDE,2021).

La clasificación simplifica el análisis conceptual, pero la revisión de expedientes auditados por la Contraloría General del Estado afirma que, en la práctica estas anomalías rara vez se presentan de forma aislada. Para este proyecto, identificarlos con precisión es indispensable, ya que de esto depende la construcción de variables capaces de representar patrones reales y relevantes para el modelo de clasificación.

Corrupción, Ocurre cuando existe un interés del funcionario público hacia un proveedor, mediante acciones como el cohecho, el peculado o el tráfico de influencias. La ley Orgánica de la Contraloría General del Estado establece que los hallazgos con posible relevancia penal deben ser dirigidos a la Fiscalía, mientras que la Contraloría actúa sobre responsabilidades administrativas y civiles culposas (Asamblea Nacional, 2002). Se define este riesgo desde una perspectiva analítica siendo el quiebre de una imparcialidad en una relación vertical funcionario-proveedor; en las bases de datos de contratación este quiebre no aparece de forma evidente, sino que debe detectarse a partir de señales indirectas: pliegos con especificaciones técnicas o términos de referencia excesivamente dirigidas, calificaciones que se apartan de los criterios publicados o adjudicaciones metódicamente cercanas al presupuesto referencial sin sustento técnico.

Colusión, se ubica en la relación horizontal entre proveedores (Avila Ordóñez, 2017) la define como el acuerdo entre dos o más oferentes orientado a distorsionar artificialmente el resultado de un proceso de contratación, Estas acciones están prohibidas por el artículo 11 de la Ley Orgánica de Regulación y Control del Poder de Mercado. En Ecuador, los modelos más comunes son las ofertas complementarias (participación simulada para aparentar competencia), la rotación de adjudicaciones entre un grupo reducido de empresas y la supresión de ofertas, donde un proveedor no compite para favorecer a otro (Ávila Ordóñez, 2017). La concentración de mercado no es una prueba de colusión por si sola, pero si genera condiciones óptimas para que ocurra. Cuando pocas empresas acumulan un número elevado de adjudicaciones dentro de una misma institución pública disminuye la presión competitiva y aumenta el riesgo de comportamientos coordinados.

En cantones de Pichincha, actividades de limpieza vial o servicios de alimentación institucional muestran en ocasiones una distribución donde proveedores recurrentes concentran la mayoría de las adjudicaciones, esto es un patrón que requiere atención del análisis de riesgo.

Ineficiencia administrativa, es un riesgo menos visible por que aparentemente no pretende intencionalmente causar daño o beneficiar indebidamente a un proveedor. Se origina de fallas de gestión como planificación insuficiente, pliegos mal elaborados, comisiones técnicas sin capacitación adecuada o sistemas de información que no se integra entre si y su resultado puede compararse con una práctica irregular, es decir: una especificación deficiente reduce oferentes, una evaluación mal hecha favorece accidentalmente a un proveedor, una mala ejecución de la planificación obliga a contrataciones emergentes. El manual de Buenas Prácticas en la Contratación Publica documenta casos donde ambos componentes: intencionalidad y negligencia se enlazan, lo que obliga a interpretar cualquier anomalía detectada en los datos con cautela, sin asumirla como evidencia directa de corrupción.

2.2.5. Riesgos por fase de contratación

Para el modelado de datos resulta útil relacionar las 3 categorías de riesgo con las fases del proceso de contratación. Durante la fase preparatoria y precontractual pueden aparecer direccionamiento de especificaciones técnicas, fraccionamiento de contratos, clausulas candados o restrictivas y descalificaciones injustificadas.

En la fase contractual o de ejecución pueden presentarse modificaciones injustificadas del objeto contractual, ampliaciones de plazo, incrementos de presupuesto, incumplimientos técnicos y pagos por trabajos no ejecutados. Durante la fase de cierre o post contractual puede detectarse incumplimiento en la liquidación, recepción de obras incompletas, liberación indebida de pólizas de cumplimiento o pagos finales sin la presentación de la documentación correspondiente. Estas señales son las que un modelo de clasificación supervisada puede aprender a reconocer para identificar procesos que necesitan auditoria más profunda.

2.2.6. Tipologías de fraude y riesgo

Favoritismo y direccionamiento: Se origina cuando los requisitos, parámetros de calificación CUMPLE/NO CUMPLE benefician a un proveedor en específico. Esto puede ocurrir

al solicitar personal técnico o condiciones que solo una empresa puede cumplir, dejando afuera a muchos proveedores, dando a lugar adjudicación de estas con precios elevados.

Restricción de la competencia y concentración: Se refiere al uso recurrente de modalidades poco competitivas es decir invitaciones a un número reducido de oferentes o adjudicaciones a los mismos proveedores. Esto no es prueba que afirme colusión, pero puede reducir la libre participación y favorecer acuerdos entre proveedores (Álvarez Suárez & Nieto Morales, 2025).

Sobrepregios y prorrogas injustificadas: Se refiere a los incrementos del presupuesto, contratos complementarios, ampliaciones de plazo y modificaciones dentro de la ejecución del contrato sin la debida justificación técnica. Estas modificaciones pueden ser usadas para extraer recursos de manera irregular o para corregir errores provenientes de una mala planificación (Scheller & Silva, 2017; Villa Pedroza, 2021; Rodríguez Arévalo 2020)

Empresas nuevas sin experiencia verificable: Comprende a la adjudicación de contratos a empresas o compañías recién creadas, con escaso historial de contratos, limitada capacidad técnica. Estas particularidades pueden asociarse con empresas creadas para simular servicios, canalizar recursos o ser fachada para los propietarios reales. Si bien es cierto la antigüedad de una empresa no se considera como prueba de fraude, es una señal más para un auditoria.

Relación con las variables del modelo: Con estas categorías y tipologías podemos diseñar variables que serán utilizadas por el modelo de clasificación. La concentración de contratos por proveedor puede asociar se con un posible riesgo de coordinación o dependencia institucional; las adjudicaciones con valor cercano al presupuesto referencial nos indica poca participación de proveedores y los plazos cortos pueden reflejar restricciones para la participación o una mala planificación de fechas de control del proceso.

También se puede tomar como variables el número de oferentes, recurrencia de adjudicaciones al mismo proveedor, diferencias del valor de presupuesto inicial y valor de adjudicación, las modificaciones dentro del contrato y los incrementos de plazo o valor.

Cabe recalcar que esto se toman como señales de alerta, no de pruebas de corrupción para una afirmación es necesario realizar una auditoría.

2.2.7. Señales e indicadores de riesgo en Contratación Pública.

La transformación desde los conceptos jurídicos hasta la creación de indicadores medibles constituye uno de los principales retos metodológicos en los proyectos de análisis de datos aplicado al control gubernamental. La Ley Orgánica de la Contraloría General del Estado se refiere a la evaluación de riesgos, pero no determina de forma exacta cómo deben evaluarse. De igual manera, la Ley Orgánica del Sistema Nacional de Contratación Pública establece principios generales, mientras que los manuales del SERCOP describen prácticas y posibles irregularidades principalmente mediante explicaciones narrativas.

Para aplicar técnicas de Ciencia de Datos, estas disposiciones y descripciones deben transformarse en variables numéricas que puedan ser procesadas por un modelo. Esta conversión requiere traducir conceptos como falta de competencia, direccionamiento, concentración o irregularidad temporal en características observables dentro de los datos de contratación pública.

Open Contracting Partnership plantea que una alternativa eficiente consiste en construir un conjunto estandarizado de señales de alerta o *red flags* a partir de la información publicada bajo el estándar Open Contracting Data Standard, conocido como OCDS. En 2024, esta organización presentó una librería de código abierto (Cardinal) diseñada para automatizar el cálculo de indicadores de riesgo sobre conjuntos de datos estructurados bajo dicho estándar (Open Contracting Partnership, 2024).

Los indicadores calculados pueden agruparse en cinco dimensiones principales. La primera corresponde a la competitividad, que considera aspectos como el número de oferentes calificados

y la frecuencia de procesos con un solo participante. La segunda se relaciona con la temporalidad e incluye los días transcurridos entre la publicación, la recepción de ofertas, la adjudicación y la firma del contrato.

La tercera dimensión analiza la concentración de las adjudicaciones, especialmente la participación acumulada de un proveedor dentro de una entidad contratante o categoría determinada. La cuarta examina variables presupuestarias, como la relación entre el valor adjudicado y el presupuesto referencial, además de la frecuencia y magnitud de las modificaciones contractuales. Finalmente, la dimensión relacionada con el comportamiento del proveedor considera su recurrencia como único oferente y su participación simultánea en diferentes procedimientos.

Aplicada a las licitaciones municipales de Pichincha correspondientes al periodo 2022-2024, esta metodología permite construir un vector de indicadores para cada proceso. En términos estructurales, cada procedimiento puede representarse como una fila dentro de la base de datos, mientras que las columnas contienen variables binarias asociadas con las diferentes señales de alerta.

A estos indicadores se incorporan variables de contexto, como el cantón contratante, la categoría del objeto contractual, el monto referencial, el Registro Único de Contribuyentes del adjudicatario y el número de oferentes habilitados. Esta información permite comparar cada proceso con otros procedimientos que presentan características similares.

La presencia aislada de una señal de alerta rara vez permite concluir que existe una irregularidad. El modelo de clasificación debe aprender combinaciones de indicadores que, de acuerdo con la evidencia histórica disponible, se encuentren asociadas con procedimientos confirmados como riesgosos o irregulares (Fortuny et al., 2023; Torres Berrú, 2023).

Un plazo demasiado reducido puede limitar la participación de otros proveedores debido al corto tiempo para la preparación de la oferta o un proveedor tuvo acceso anticipado a la publicación del proceso

El número reducido de oferentes constituye otra señal relevante. Los procedimientos con menos de tres participantes efectivamente calificados presentan una presión competitiva menor y pueden generar condiciones favorables para que la adjudicación se aproxime al presupuesto referencial. Esta señal aumenta el nivel de riesgo cuando coincide con plazos cortos, requisitos restrictivos o adjudicaciones recurrentes (OECD, 2021).

La concentración de contratos se presenta cuando un mismo proveedor acumula una proporción considerable del monto adjudicado por una institución durante un periodo determinado. Para este análisis, puede establecerse una señal cuando un mismo RUC concentra más del 25 % del valor contratado por una entidad dentro de una categoría y ejercicio fiscal.

Una participación elevada puede responder a la especialización o capacidad técnica de una empresa. No obstante, también puede sugerir dependencia institucional, barreras de entrada, direccionamiento de requisitos o relaciones recurrentes entre comprador y proveedor. Por esta razón, debe analizarse en conjunto con el número de competidores, la frecuencia de participación y la tasa de éxito del adjudicatario (Universidad Nacional de San Martín, 2023).

Calidad y validación de los datos. La confiabilidad de las señales de alerta depende directamente de la calidad de los datos utilizados para calcularlas. El Sistema Oficial de Contratación Pública del Ecuador publica información abierta bajo el formato OCDS desde 2021; sin embargo, se encontraron bastantes campos vacíos, nombres de instituciones mal registrados, etc.

Estas limitaciones pueden alterar los resultados del análisis. Por ejemplo, un campo vacío en el número de oferentes no debe interpretarse automáticamente como un proceso con un único participante. De igual manera, dos formas distintas de escribir el nombre de una misma institución

pueden provocar que el sistema las considere entidades diferentes, afectando el cálculo de concentración.

Por esta razón, antes de construir los indicadores es necesario ejecutar un proceso de limpieza, normalización y validación. Esto incluye la homologación de nombres, el tratamiento de valores ausentes, la corrección de fechas, la eliminación de registros duplicados y la verificación de la coherencia entre las diferentes etapas del procedimiento.

Existe un proyecto de modernización del SERCOP promovido por el Banco Interamericano de Desarrollo, que plantea disminuir estas grietas mediante el intercambio sistemático de información con la Contraloría General del Estado y el desarrollo de herramientas inteligentes para la detección de riesgos.

Alcance de las señales de alerta. Es importante recalcar que una red flag o alerta no es una acusación directa de un caso de corrupción, un valor atípico no es el equivalente automático a una irregularidad por ejemplo un municipio rural que ha enfrentado una emergencia puede registrar plazos reducidos en la adjudicación de un proceso por motivos reales y transparentes, un proveedor con una amplia especialización o experiencia también puede concentrar un gran número de contratos adjudicados.

Por lo tanto, estos indicadores se deben utilizar como instrumentos para profundizar una auditoria en estos casos mas no como afirmación de un caso de corrupción. De esta forma no gastamos recursos limitados en procesos que no presentan señales de corrupción pudiendo concentrarse en procesos que si tengan señales relevantes (Banco Mundial, 2025).

Variables predictoras. Estas variables se basan en la evidencia técnica y la disponibilidad de la información sea de API, CSV o cualquier tipo de BDD. Algunos estudios afirman que las características de la relación entidad-proveedor pueden tener mayor capacidad predictiva que las especificaciones técnicas del contrato (Aldana et al., 2022)

Con lo antes expuesto el modelo incorpora los siguientes indicadores:

Ausencia de competencia. Incluye los procesos con un solo oferente, un número reducido de participantes habilitados, adjudicaciones directas recurrentes y una limitada participación en los sondeos de mercado. También se toman en cuenta los procesos adjudicados reiteradamente al mismo proveedor (Salazar et al., 2024; Álvarez Suárez & Nieto Morales, 2025).

Anomalías temporales. Se toman en cuenta los plazos de publicación extremadamente cortos, los periodos atípicos entre la recepción de ofertas y la adjudicación (Salazar et al., 2024).

Relación presupuestaria. Analiza la cercanía entre el valor adjudicado y el presupuesto referencial, la diferencia porcentual entre ambos valores y la recurrencia de adjudicaciones próximas al límite presupuestado.

Perfil del proveedor. Considera la antigüedad de la empresa, la diversidad atípica de sus actividades económicas, la obtención de contratos de alto valor poco después de su creación, la frecuencia con la que participa como único oferente y su tasa adjudicación.

Relación entre comprador y proveedor. Examina la recurrencia de adjudicaciones entre una misma entidad y un mismo proveedor, la concentración de contratos en determinadas unidades compradoras y la continuidad de estas relaciones a lo largo del tiempo.

Modificaciones contractuales. Incluye incrementos del valor contratado, ampliaciones de plazo, contratos complementarios y cambios posteriores en el objeto de contratación.

Anomalías financieras y de ejecución. Comprende las inconsistencias detectadas durante la ejecución del contrato, como pagos superiores a los valores facturados o contratados, desembolsos realizados fuera de los plazos esperados y ausencia de pólizas o garantías de cumplimiento en contratos que, por su monto o naturaleza, deberían contar con estos mecanismos de respaldo (Álvarez Suárez & Nieto Morales, 2025). Para representar este riesgo, el modelo puede analizar el valor inicial y final del contrato, los montos pagados, las diferencias entre facturación y desembolso, las modificaciones presupuestarias y los tiempos transcurridos entre el

registro de las obligaciones y la realización de los pagos (Rodríguez Arévalo, 2020; Villa Pedroza, 2021).

2.2.8. Machine Learning y sus Modelos

Machine learning o aprendizaje automático, es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos estadísticos, diseñados para emular la inteligencia humana aprendiendo del entorno (Mitchell, 2014). Estos algoritmos permiten: aprender de los datos, identificar patrones y tomar decisiones o hacer predicciones para mejorar su rendimiento en una tarea específica. Según Tom Mitchell definió el Machine Learning como: "Un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y una medida de rendimiento P , si su rendimiento en las tareas en T , tal como se mide por P , mejora con la experiencia E " (Mitchell, 1997).

Dentro de machine learning se encuentran las siguientes clasificaciones: Aprendizaje supervisado, no supervisado, Aprendizaje semi-supervisado, aprendizaje por refuerzo. Utilizamos el enfoque supervisado cuando trabajamos con muestras que ya traen una "solución" o etiqueta incorporada. De este modo, el algoritmo aprende comparando sus predicciones con los resultados reales ya conocidos. El aprendizaje no supervisado utiliza información no etiquetada. (Bobadilla, 2020).

2.2.8.1. Modelos Supervisado

El aprendizaje supervisado es un enfoque del aprendizaje automático en el que un algoritmo infiere una función matemática a partir de un conjunto de datos de entrenamiento previamente etiquetado.

De acuerdo con Mitchell (1997), el problema del aprendizaje supervisado se puede definir matemáticamente de la siguiente manera: dado un conjunto de entrenamiento de N pares de ejemplos, $D = \{(X_{(1)} Y_{(1)}), (X_{(2)} Y_{(2)}), \dots, (X_{(N)} Y_{(N)})\}$ donde X es el vector de características (variables independientes o de entrada) y $Y_{(1)}$ es la etiqueta o respuesta deseada

(variable dependiente o de salida) y, el objetivo del algoritmo es aproximar una función $f: X \rightarrow Y$ tal que $f(x)$ prediga con la mayor precisión posible el valor de y para observaciones futuras y no vistas.

El ciclo de vida de un modelo supervisado se fundamenta en la generalización. No se busca que el modelo "memorice" los datos, sino que aprenda los patrones subyacentes.

Este proceso se divide en varias fases (Hastie, Tibshirani, & Friedman, 2009):

Conjunto de Entrenamiento (Training Set): El modelo recibe los datos de entrada junto con sus etiquetas correctas. A través de un proceso iterativo de optimización, el algoritmo ajusta sus parámetros internos.

- **Función de Pérdida (Loss Function):** Durante el entrenamiento, el algoritmo calcula el error entre su predicción y el valor real y . La función de pérdida penaliza los errores, y el objetivo del algoritmo (usando técnicas como el Descenso del Gradiente) es minimizar esta métrica matemática.
- **Validación y Prueba (Testing/Validation Set):** Una vez entrenado, el modelo se evalúa en un conjunto de datos "ciego" (que no vio durante el entrenamiento) para medir su capacidad de generalización y evitar el sobreajuste (overfitting).

El aprendizaje supervisado comprende métodos que aprenden a partir de datos etiquetados y suele dividirse en dos categorías principales: clasificación y regresión.

- **Clasificación:** Se utiliza cuando la variable de salida es discreta o categórica. El objetivo es asignar una observación a una o varias clases predefinidas. Ejemplo: Predecir si un correo electrónico es "Spam" o "No Spam". En este enfoque, el objetivo principal es construir un modelo que aprenda la distribución de las etiquetas de clase en función de las características predictoras.

- **Regresión:** Se utiliza cuando la variable de salida es continua o numérica. El modelo busca predecir un valor real. Ejemplo: Predecir el precio de una vivienda (en dólares) en función de sus metros cuadrados y ubicación.

La clasificación busca identificar a qué categoría pertenece una observación, mientras que la regresión estima valores numéricos continuos a partir de variables de entrada (Bishop, 2006; Alpaydin, 2021).

2.2.8.2. Aprendizaje No Supervisado

El aprendizaje no supervisado es un enfoque de machine learning que trabaja con datos sin etiquetas previas, es decir, sin que el sistema conozca de antemano qué casos son fraudulentos o no fraudulentos. En contratación pública es especialmente útil porque muchos municipios no tienen bases históricas con procesos ya clasificados judicial o administrativamente como fraude, corrupción o colusión (Lyra et al., 2022).

2.2.8.3. Aprendizaje Automático Predictivo y Preventivo

La auditoría tradicional en contratación pública generalmente funciona de manera reactiva, es decir, revisa los expedientes cuando el proceso ya terminó o cuando existe una denuncia. Aunque este tipo de revisión es importante desde el ámbito jurídico y de control, pueden resultar limitadas frente a la gran cantidad de información que generan los sistemas de contratación pública. Es decir, el aprendizaje automático permite complementar la auditoría tradicional mediante un enfoque más predictivo y preventivo, ya que ayuda a identificar patrones de riesgo y priorizar los casos que requieren mayor atención antes de realizar una revisión manual completa.

Los modelos de aprendizaje automático permiten analizar una gran cantidad de procesos adjudicados y detectar combinaciones de variables que pueden estar relacionadas con señales de alerta o red flags. Entre estas señales se encuentran la baja competencia, existencia de un único oferente, montos adjudicados fuera de lo normal, adjudicaciones muy cercanas al presupuesto

referencial, pazos demasiados cortos, concentración repetida de contratos entre una misma entidad y un mismo proveedor, así como documentación incompleta dentro del proceso.

2.2.8.4. Algoritmos de ensamblaje y Clasificación Supervisada

La clasificación supervisada permite entrenar un modelo a partir de datos previamente etiquetados, para que luego pueda clasificar nuevos registros. En esta investigación cada proceso de contratación pública se representa mediante variables numéricas, categóricas y variables construidas a partir de la información disponible. La variable objetivo corresponde al nivel de riesgo asignado a cada proceso. Debido a que las señales de riesgo no siempre tienen una relación simple o lineal con las variables del proceso, se seleccionan modelos de ensamble basados en árboles de decisión.

Random Forest es un modelo que combina varios árboles de decisión entrenados sobre diferentes muestras del conjunto de datos. Cada árbol analiza los datos desde una perspectiva distinta y luego el modelo final integra sus resultados mediante votación. Esta forma de trabajo ayuda a reducir el sobreajuste y hace que el modelo sea más estable que un árbol de decisión individual. Además, al introducir aleatoriedad tanto en las observaciones como en las variables utilizadas, los árboles no dependen todos de la misma estructura de datos

En este proyecto se utiliza Random Forest con `class_weight='balanced_subsample'`, ya que esta configuración permite ajustar los pesos de las clases en cada muestra de entrenamiento. Esto es importante porque los procesos con riesgo alto o crítico suelen ser menos frecuentes que los procesos de bajo riesgo. Al dar mayor peso a estas clases minoritarias, el modelo mejora su capacidad para reconocer casos relevantes dentro del análisis.

XGBoost y LightGBM también son modelos de ensamble basados en árboles, pero pertenecen a la familia de gradiente boosting. A diferencia de Random Forest, donde los árboles se construyen de forma independiente, en estos modelos los árboles se entrenan de manera secuencial. Es decir cada nuevo árbol intenta corregir los errores cometidos por los árboles

anteriores. Esta característica permite mejorar progresivamente la capacidad predictiva del modelo, especialmente cuando existen relaciones complejas entre las variables.

XGBoost se utiliza porque es un algoritmo robusto para problemas de clasificación con datos tabulares. Además, incorpora mecanismos de regularización y optimización que ayudan a controlar el sobreajuste y mejorar el rendimiento del modelo. Para este proyecto se configura con pesos balanceados por observación, con el fin de dar mayor importancia a las clases menos frecuentes durante el entrenamiento.

LightGBM, también trabaja bajo la lógica de gradient boosting, pero está diseñado para entrenar modelos de forma más rápida y eficiente. Esto lo convierte en una alternativa adecuada cuando se manejan conjuntos de datos con varias variables o mayor volumen de registros. En este proyecto se mantiene porque combina eficiencia computacional, buen desempeño predictivo y soporte para el tratamiento de clases desbalanceadas.

2.2.9. Análítica de Datos en Contratación Pública.

La analítica de datos no se reduce a estadística descriptiva ni a la generación de tableros visuales. Conformar el procedimiento sistemático por medio del cual datos crudos se transforman en evidencia accionable para la toma de decisiones (Fortuny et al., 2023). En el contexto de la fiscalización gubernamental, esta explicación tiene consecuencias prácticas: el indicador no es el producto final, sino la materia prima de una decisión auditar o no auditar un expediente, abrir o no abrir un examen especial, priorizar o no un proceso para comprobación.

La taxonomía clásica diferencia cuatro niveles. La analítica descriptiva responde a “qué pasó” y se concreta en estadísticos resumen y visualizaciones; la diagnóstica responde a “por qué pasó” e inserta análisis de correlaciones y desagregaciones por categoría; la predictiva contesta a “qué va a pasar” mediante modelos de aprendizaje automático que evalúa probabilidades; la prescriptiva responde a “qué deberíamos hacer” y combina modelos con reglas de decisión y técnicas de optimización (OECD, 2021). El presente proyecto efectúa en los niveles descriptivos y

predictivo, dejando la dimensión prescriptiva a la deliberación posterior de los órganos de control, donde la decisión final implica consideraciones jurídicas y políticas que rebasan el alcance del modelo.

El Sistema de Control, Fiscalización y Auditoría del Estado, definido por la LOCGE, prepara el control gubernamental ecuatoriano alrededor de un eje fundamentalmente reactivo. Las auditorías financieras, de gestión, de aspectos ambientales, de obras públicas se ejecutan sobre procesos ya consumados; los exámenes especiales investigan hechos ocurridos; las responsabilidades administrativas, civiles culposas y penales se determinan ex post (Asamblea Nacional, 2002, arts. 18-28). La facultad de la Contraloría caduca a los siete años de realizado el acto, y las obligaciones civiles culposas establecen a los diez años desde la ejecutoria de la resolución (Asamblea Nacional, 2002, arts. 71-75). Este diseño normativo es sólido en términos de debido proceso, pero deja un margen amplio para que el daño se materialice antes de que el control lo localice.

El cambio hacia un seguimiento preventivo apoyado en inteligencia artificial no requiere modificar la LOCGE. Requiere, en cambio, articular las facultades de control existentes con herramientas capaces de operar sobre el flujo de datos en tiempo casi real. El Proyecto de Modernización SERCOP financiado por el BID, la iniciativa Smart Citizen Control elaborada por la Universidad San Francisco de Quito (Fortuny et al., 2023) y la investigación doctoral de Torres Berrú (2023) sobre minería de datos en SERCOP convergen en un mismo diagnóstico: el Ecuador dispone de los datos, dispone de las facultades de control y dispone de los modelos. Lo que falta es el puente operativo entre las tres capas, y es allí donde el presente trabajo ambiciona aportar.

La diferencia entre control reactivo y monitoreo preventivo no es retórica. Un examen especial promovido por la Contraloría dieciocho meses después de adjudicado un contrato puede determinar responsabilidades, pero la obra ya está construida o, peor, no construida y los recursos ya están comprometidos. Una alerta producida por un modelo de clasificación al momento mismo

de la convocatoria, en cambio, habilitas intervenciones tempranas: solicitar aclaraciones a la entidad contratante, suspender la apertura, ampliar plazos de presentación de ofertas. La analítica, entendida así, no compite con la auditoría tradicional: la complementa, desplazando el centro de gravedad desde la sanción retrospectiva hacia la prevención prospectiva (Open Contracting Partnership, 2022).

Existe, sin embargo, una restricción epistémica que conviene explicitar antes de cerrar este apartado. Un modelo predictivo aprende del pasado y reproduce, por tanto, los sesgos del pasado. Si los procesos históricamente etiquetados como riesgosos provienen de una muestra sesgada, por ejemplo, auditorías concentradas en cantones de mayor visibilidad mediática o en categorías de objeto contractual políticamente sensibles, el modelo heredará esa concentración y subdetectará en los segmentos infravigilados. A esto se suma el problema del desbalance de datos: en cualquier muestra municipal pichinchana, los procesos con señal de riesgo confirmada constituyen una minoría frente al universo de licitaciones limpias, lo que sesga las funciones de pérdida estándar hacia la clase mayoritaria. La analítica responsable exige, en consecuencia, técnicas de remuestreo SMOTE, undersampling estratificado, ajuste de pesos por clase, y una auditoría continua del propio modelo mediante análisis de feature importance, evaluación de equidad por subgrupo cantonal y monitoreo de drift temporal (Datalat, s.f.; Torres Berrú, 2023).

2.2.10. Indicadores y Métricas de Evaluación.

En problemas de riesgo y auditoría, la métrica Accuracy no siempre refleja el verdadero desempeño del modelo. Si la mayoría de los procesos pertenecen a la clase de bajo riesgo, un modelo podría obtener una exactitud alta clasificando casi todos los casos como Bajo. Sin embargo, ese resultado no sería útil para esta investigación, porque el objetivo principal es identificar procesos con señales de riesgo Alto o Crítico. Por esta razón, la evaluación debe enfocarse en métricas que permitan medir la utilidad real del modelo para priorizar revisiones.

Precision permite conocer qué proporción de las alertas generadas por el modelo corresponde realmente a procesos de riesgo. En auditoría pública, una mayor precisión significa que los casos priorizados contienen más señales relevantes, reduciendo la cantidad de falsos positivos y evitando una carga innecesaria para los equipos de control.

Recall mide la capacidad del modelo para capturar los procesos que realmente presentan riesgo. Esta métrica es clave porque, en un contexto de control público, dejar fuera un proceso crítico puede ser más grave que revisar un caso que finalmente no presente irregularidades. Por eso, un modelo con bajo Recall podría no cumplir adecuadamente con el propósito de priorización.

F1-Score resume el equilibrio entre Precision y Recall. Esta métrica es importante porque permite balancear dos necesidades: detectar una cantidad suficiente de procesos riesgosos y, al mismo tiempo, evitar un exceso de alertas poco relevantes. En esta tesis se prioriza especialmente el F1-Score de la clase agregada Alto+Crítico, ya que representa el grupo de mayor interés para auditoría.

PR-AUC, o área bajo la curva Precision-Recall, permite evaluar el comportamiento del modelo en diferentes umbrales de decisión. Esta métrica es especialmente útil cuando los casos de riesgo son minoritarios, porque se enfoca en la calidad de las predicciones positivas. En este proyecto, PR-AUC permite comparar qué tan bien Random Forest, XGBoost y LightGBM ordenan los procesos con mayor probabilidad de presentar señales de riesgo.

En conjunto, Precision, Recall, F1-Score y PR-AUC permiten evaluar los modelos desde una perspectiva práctica y accionable. La finalidad no es afirmar jurídicamente que existe corrupción, sino construir un sistema de priorización transparente, reproducible e interpretable que apoye la toma de decisiones en procesos de control público. (Fawcett, 2006; Hossin & Sulaiman, 2015).

CAPITULO 3:

3. DESARROLLO

3.1. Metodológico CRISP-DM

El presente proyecto tomo como marco metodológico CRISP-DM es el acrónimo en inglés de Cross-Industry Standard Process for Data Mining (Proceso Estándar de la Industria para la Minería de Datos). De acuerdo con Chapman et al. (2000), consiste en un modelo de proceso de minería de datos no propietario, neutral respecto a las herramientas analíticas y de libre acceso, diseñado para servir como guía estructurada en la ejecución de proyectos dentro de las organizaciones.

El ciclo de vida definido por CRISP-DM está enfocado en la exploración y análisis de la información estructurándose en seis fases principales. El proceso no es estrictamente secuencial o lineal, sino cíclico e iterativo, lo que significa que los equipos de trabajo avanzan y retroceden constantemente entre las etapas en función de las necesidades del negocio y de los hallazgos en los datos (Dobre, 2025).

Figura 2*Fases CRISP-DM*

Nota. Adaptado de Chapman et al. (2000).

3.2. Comprensión del Negocio (Business Understanding)

La presente investigación se desarrolla en el contexto de la contratación pública ecuatoriana, un ámbito de alta relevancia económica y social debido al volumen de recursos públicos que administra y a su impacto directo sobre la provisión de bienes, servicios y obras para la ciudadanía.

El proyecto se centra en identificar posibles señales de señales estadísticas de riesgo en los procesos de licitación adjudicados por los Gobiernos Autónomos Descentralizados Municipales de la provincia de Pichincha mediante el análisis de datos abiertos provenientes del Servicio Nacional de Contratación Pública (SERCOP). Se centra específicamente a los años del 2022- 2025 en procesos de Licitaciones.

En Ecuador, el Servicio Nacional de Contratación Pública (SERCOP) publica información contractual mediante el estándar internacional Open Contracting Data Standard (OCDS), proporcionando acceso a datos estructurados sobre las diferentes etapas del ciclo de contratación.

Los datos corresponden al Portal de Datos Abiertos del Servicio Nacional de Contratación Pública (SERCOP) del Ecuador, accesible a través de una interfaz programática de aplicaciones (API) la cual permite acceder de manera automática a los registros de contratación pública datosabiertos.compraspublicas.gob.ec.

Figura 3

API SERCOP

```
# --- API SERCOP ---
"API_BASE": "https://datosabiertos.compraspublicas.gob.ec/PLATAFORMA/api",
"ENDPOINT_SEARCH": "search_ocds",
"ENDPOINT_RECORD": "record",
```

La Figura 3 detalla la API donde se realiza consulta con endpoint de búsqueda (search_ocds) y devuelve el listado de identificadores OCID y endpoint record para la descarga del archivo consolidado.

El SERCOP constituye la entidad rectora de la contratación pública ecuatoriana y, en su rol de productor primario de información transaccional, expone los procesos de contratación adheridos al Open Contracting Data Standard (OCDS), un estándar internacional de transparencia desarrollado por la Open Contracting Partnership.

La descripción de los datos reporta las propiedades superficiales del corpus: volumen, estructura, formato y significado de los atributos (Chapman et al., 2000). La unidad de análisis es el proceso de contratación, identificado unívocamente por el campo ocid. En este contexto se analiza los procesos de contratación ejecutados por los Gobiernos Autónomos Descentralizados (GAD) de la provincia de Pichincha durante el período 2022–2024, con el propósito de detectar patrones relacionados con baja competencia, concentración de proveedores, anomalías temporales y comportamientos atípicos en los montos adjudicados.

El conjunto de datos para análisis quedó conformado por 2111 procesos únicos y 33 variables, atribuidas a cada entidad de los GAD.

Tabla 2*Variables y descripción*

Variable	Tipo	Descripción
Ocid	Texto	Identificador único del proceso (OCDS)
release_id	Texto	Identificador de la entrega (release)
release_date	Fecha	Fecha de publicación de la entrega
release_tags	Texto	Etapas del ciclo contenidas en el paquete
buyer_id	Texto	Identificador de la entidad compradora
buyer_name	Texto	Nombre del GAD comprador
buyer_region	Texto	Región del comprador (Pichincha)
buyer_locality	Texto	Localidad del comprador
supplier_id	Texto	Identificador del proveedor adjudicado
supplier_name	Texto	Razón social del proveedor
tender_id	Texto	Identificador del procedimiento de licitación
tender_title	Texto	Objeto o título del proceso
tender_status	Texto	Estado del proceso (active / complete)
tender_procurementMethod	Texto	Método de contratación (categoría OCDS)
tender_procurementMethodDetails	Texto	Detalle del procedimiento (SERCOP)
tender_period_start	Fecha	Inicio del periodo de licitación
tender_period_end	Fecha	Cierre del periodo de licitación
award_id	Texto	Identificador de la adjudicación
award_status	Texto	Estado de la adjudicación
award_date	Fecha	Fecha de adjudicación
dias_tender_a_award	Numérico	Días entre cierre y adjudicación (derivada)
tender_value_amount	Numérico	Presupuesto referencial del proceso
award_value_amount	Numérico	Monto adjudicado
tender_value_currency	Texto	Divisa del presupuesto (USD)
award_value_currency	Texto	Divisa del monto adjudicado (USD)
tender_numberOfTenderers	Numérico	Número de oferentes participantes
n_suppliers	Entero	N.º de proveedores adjudicados (derivada)
n_awards	Entero	N.º de adjudicaciones del proceso (derivada)
n_releases_en_paquete	Entero	N.º de releases del paquete (derivada)
buyer_country	Texto	País del comprador (EC)
is_imputed_tender_numberOfTenderers	Binaria	Marca de imputación del n.º de oferentes
is_imputed_tender_value_amount	Binaria	Marca de imputación del presupuesto
outlier_award_value_amount_iqr	Binaria	Marca de valor atípico por IQR

3.3. Preparación de los Datos

Luego de la exportación de los datos de la página de datos abiertos en formato json que son los archivos base para el procesamiento de la información, a continuación, se describe cada subtarea de la preparación de datos.

3.3.1. Selección de los Datos

La selección de los datos consiste en determinar, de manera justificada respecto de los objetivos de la investigación, qué registros (filas) y qué atributos (columnas) integran el conjunto de datos analítico, considerando criterios de pertinencia conceptual, calidad y restricciones técnicas (Chapman et al., 2000). El Data set descargado corresponde a todos los procesos de contratación pública a nivel nacional, a esto se aplicaron criterios de inclusión y exclusión para delimitar el universo de procesos del GAD de la provincia de Pichincha durante la ventana fiscal del 2022-2024.

Tabla 3

Criterios para selección de datos

Criterio	Definición operativa
Inclusión	Procesos cuyo comprador (buyer) pertenece a un GAD de la provincia de Pichincha, verificado por buyer_region y, complementariamente, por expresión regular sobre buyer_name; procesos con etapa de adjudicación (award) registrada; periodo de observación 2022–2024.
Exclusión	Procesos sin identificador ocid o sin identificación del comprador (buyer_id, buyer_name); procesos de entidades no pertenecientes a la jurisdicción de Pichincha; registros duplicados por OCID derivados de la estrategia de descubrimiento multi-consulta.

El filtrado geográfico se diseñó como un esquema de dos capas: una capa primaria que consulta el campo buyer_region en busca de la cadena «Pichincha» y una capa secundaria de

respaldo, basada en coincidencia por expresión regular sobre buyer_name. Como parte del proceso para depuración de la data, se descartaron procesos que no constan con un OCID (Open Contracting ID) que es el identificador único de proceso, se eliminaron procesos que no contaba con buyer_id y buyer_name sin estos campos no es posible verificar si el proceso pertenece a un municipio de Pichincha, al aplicar este criterio, finalmente se obtuvo un conjunto de 33 variables detalladas en la Tabla 4 y se encuentran organizadas en grupos acorde a su función dentro de cada análisis.

Tabla 4

Variables por grupos

Grupo	Variables principales
Identificación	ocid, tender_id, award_id
Entidad compradora	buyer_id, buyer_name, buyer_region, buyer_locality
Proveedor	supplier_id, supplier_name
Modalidad	tender_procurementMethod, tender_procurementMethodDetails
Fechas	release_date, tender_period_start, tender_period_end, award_date
Valores económicos	tender_value_amount, award_value_amount
Competencia	tender_numberOfTenderers
Calidad de datos	is_imputed_tender_numberOfTenderers, is_imputed_tender_value_amount
Outliers previos	outlier_award_value_amount_iqr

Las variables de la Tabla 3 fueron seleccionadas en base a criterios identificados por la literatura de variables seleccionadas como factores asociados al riesgo de corrupción dentro de la contratación pública. Particularmente se seleccionaron variables asociadas a valor adjudicado del contrato, cantidad de oferentes y la modalidad de contratación (Fazekas, Tóth & King, 2016; Open

Contracting Partnership, 2021). Así mismo se incorporaron variables de identificación y de contexto institucional necesarias para construir métricas utilizadas en el presente análisis.

3.3.2. Limpieza de los Datos

Comprende las operaciones orientadas a elevar la calidad del conjunto hasta el nivel exigido por las técnicas analíticas previstas, abordando duplicados, inconsistencias de tipo, valores faltantes y valores atípicos (Chapman et al., 2000; García et al., 2015).

Duplicidad. El campo ocid es la llave primaria del estándar OCDS, y debe ser único por proceso.

Figura 4

Duplicados de datos

```

dups_ocid = df_licitaciones_pichincha["ocid"].duplicated(keep=False).sum()
dups_completos = df_licitaciones_pichincha.duplicated(keep=False).sum()

print(f"Filas con OCID duplicado      : {dups_ocid:,}")
print(f"Filas idénticas en todas las columnas: {dups_completos:,}")

if dups_ocid > 0:
    print("\nMuestra de duplicados por OCID:")
    print(df_licitaciones_pichincha[df_licitaciones_pichincha["ocid"].duplicated(keep=False)]
          .sort_values("ocid").head(6)[["ocid", "buyer_name", "release_date", "tender_status"]])

    # Deduplicación: nos quedamos con el release más reciente por OCID
    antes = len(df_licitaciones_pichincha)
    df_licitaciones_pichincha = (
        df_licitaciones_pichincha
        .sort_values("release_date")
        .drop_duplicates(subset=["ocid"], keep="last")
        .reset_index(drop=True)
    )
    print(f"\n✓ Deduplicado: {antes:,} → {len(df_licitaciones_pichincha):,} filas")
else:
    print("\n✓ Sin duplicados de OCID. Dataset limpio en ese frente.")
30] ✓ 0.0s

.. Filas con OCID duplicado      : 0
   Filas idénticas en todas las columnas: 0

   ✓ Sin duplicados de OCID. Dataset limpio en ese frente.

```

La Figura 4 muestra que el conjunto de datos no tiene valores duplicados, garantizando de esta forma la unicidad y consistencia de la información, basado en identificador único de contratación (OCID).

Valores Faltantes. Se realiza el análisis de valores faltantes mediante la función `isnull()` de Pandas, que calcula el número total de valores nulos por variable y el porcentaje de ausencia respecto al total del conjunto de datos.

```
# Total de faltantes

total_faltantes = df_licitaciones_pichincha.isnull().sum().sum()

# Identificación de valores faltantes por variable

faltantes = df_licitaciones_pichincha.isnull().sum()

faltantes = faltantes[faltantes > 0]

print(faltantes)
```

Datos con 0 valores faltantes.

Corrección de errores e inconsistencia. Se realiza un análisis orientado a identificar diferencias de escritura, uso inconsistente de mayúsculas y minúsculas, así como la presencia de espacios en blanco tanto al inicio como al final de los registros.

Los datos obtenidos muestran la existencia de un número bajo de inconsistencias asociadas a espacios adicionales en los nombres, como se muestra en la Figura 5, donde la presencia de un espacio al final del texto genera una representación distinta de un mismo valor.

Figura 5*Detección de inconsistencias*

Valores con espacios extra (3):
['ASOCIACION DE SERVICIOS ALIMENTACION GOURMET JUNTOS POR EL BIENESTAR SOCIAL AGOURMET ']

Posibles diferencias de escritura:
Output is truncated. View as a [scrollable element](#) or open in a [text editor](#). Adjust cell output [settings](#)...

	Original	Normalizado
0	DATAPRO S.A	DATAPRO S.A
1	ABSEG SEGURIDAD Y AMBIENTE CIA LTDA	ABSEG SEGURIDAD Y AMBIENTE CIA LTDA
2	ESPINOSA PROAÑO EDISON GUSTAVO	ESPINOSA PROAÑO EDISON GUSTAVO
3	PROTELCOTELSA S.A.	PROTELCOTELSA S.A.
4	ANDINOPARADISE TRAVEL CIA. LTDA.	ANDINOPARADISE TRAVEL CIA. LTDA.
5	CAMARA DE INDUSTRIAS Y COMERCIO ECUATORIANO ALEMANA	CAMARA DE INDUSTRIAS Y COMERCIO ECUATORIANO ALEMANA
6	TEDEIN S.A.	TEDEIN S.A.
7	CV SISTEMS & COPIERS CS&C CIA. LTDA.	CV SISTEMS & COPIERS CS&C CIA. LTDA.
8	TUTASI BASANTES DARWIN ROLANDO	TUTASI BASANTES DARWIN ROLANDO
9	GODOY PROAÑO FREDY NEPTALI	GODOY PROAÑO FREDY NEPTALI
10	COMPAÑIA GENERAL DE COMERCIO COGECOMSA S. A.	COMPAÑIA GENERAL DE COMERCIO COGECOMSA S. A.
11	CRUZ ROJA ECUATORIANA JUNTA PROVINCIAL DE PICHINCHA	CRUZ ROJA ECUATORIANA JUNTA PROVINCIAL DE PICHINCHA
12	GRANDA JARAMILLO DARWIN RODRIGO	GRANDA JARAMILLO DARWIN RODRIGO
13	TRANSPORTE ESCOLAR TOURESTUDIANTIL S. A.	TRANSPORTE ESCOLAR TOURESTUDIANTIL S. A.
14	GUERRERO BRAVO PABLO FERNANDO	GUERRERO BRAVO PABLO FERNANDO
15	JIMENEZ RIVADENEIRA MAXIMILIANO DAVID	JIMENEZ RIVADENEIRA MAXIMILIANO DAVID
16	BELTEM SECURITY CIA. LTDA.	BELTEM SECURITY CIA. LTDA.
17	EMPRESA DE INVESTIGACION DE MERCADOS Y ANALISIS PIVOT CIA.LTDA.	EMPRESA DE INVESTIGACION DE MERCADOS Y ANALISIS PIVOT CIA.LTDA.
18	Flores Castillo Marco Eduardo	FLORES CASTILLO MARCO EDUARDO
19	SINELET S.A.	SINELET S.A.

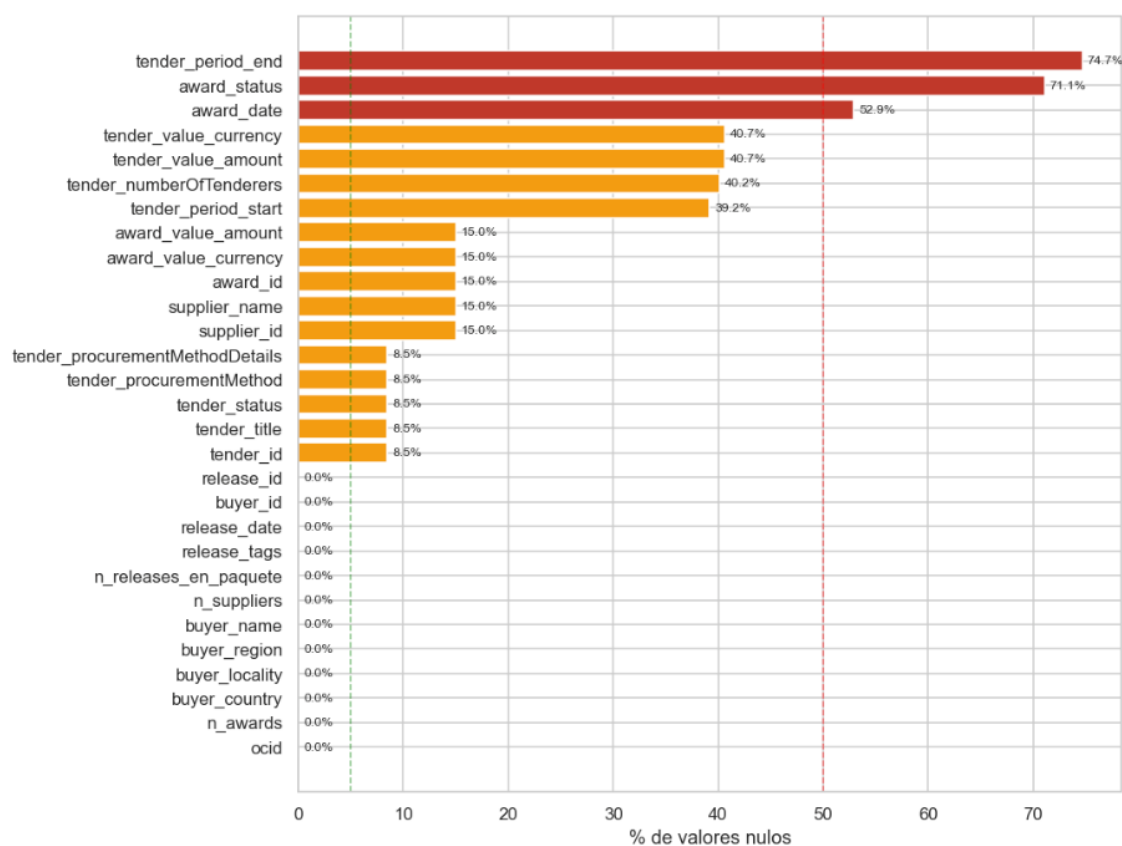
Así mismo la Figura 5 detalla posibles diferencias de escritura en determinadas variables, sin embargo, solo corresponden a variaciones de formato, esto demuestra que nuestro data ser presenta consistencia de datos.

Valores Nulos. Se calculó el porcentaje de valores nulos presente en cada variable del conjunto de datos, como se muestra en la Figura 6, donde se visualiza un esquema tipo semáforo para identificar la calidad de los datos, los umbrales están en base a criterios empleados en procesos de preparación de datos y análisis exploratorio. Se considero que para variables con menos del 5% de valores faltantes presentan una ausencia de información mínima, por lo que su impacto en el análisis es reducido, variables con más del 5% - 50% son clasificadas como criticidad moderna, en este caso si puede afectar moderadamente el análisis de los datos, por lo que resulta necesario evaluar estrategias de tratamiento, tales como la imputación de valores o la conservación de la variable en función de su relevancia para el estudio. Las variables con más del 50% de valores faltantes son consideradas de alta criticidad, bajo este criterio la calidad y

representatividad de las variables puede verse comprometidas. Este criterio permitió identificar de forma objetiva las variables que requerían mayor atención durante la fase de limpieza y preparación de los datos.

Figura 6

Semáforo porcentaje de valores nulos



La Figura 6 muestra cuales son las variables: tender_period_end (74,7 %), award_status (71,1 %) y award_date (52,9 %) que fueron clasificadas como de alta criticidad, al superar el umbral del 50 % de ausencia de información.

Se identificó un conjunto de variables con niveles moderados de valores faltantes, entre las que destacan tender_value_currency, tender_value_amount, tender_numberOfTenderers y tender_period_start, cuyos porcentajes de ausencia se situaron alrededor del 40 %.

Sin embargo, las variables relacionadas con la identificación de los procesos de contratación y de las entidades contratantes presentaron una completitud del 100 %, por lo que no requirieron intervenciones asociadas al tratamiento de datos faltantes.

Valores atípicos. Se aplicó en método IQR de Rango Intercuartílico (Interquartile Range) que es una técnica utilizada en el análisis exploratorio de datos debido a su robustez frente a distribuciones asimétricas y a la presencia de valores extremos.

El procedimiento consistió en calcular, para cada variable numérica, los siguientes estadísticos:

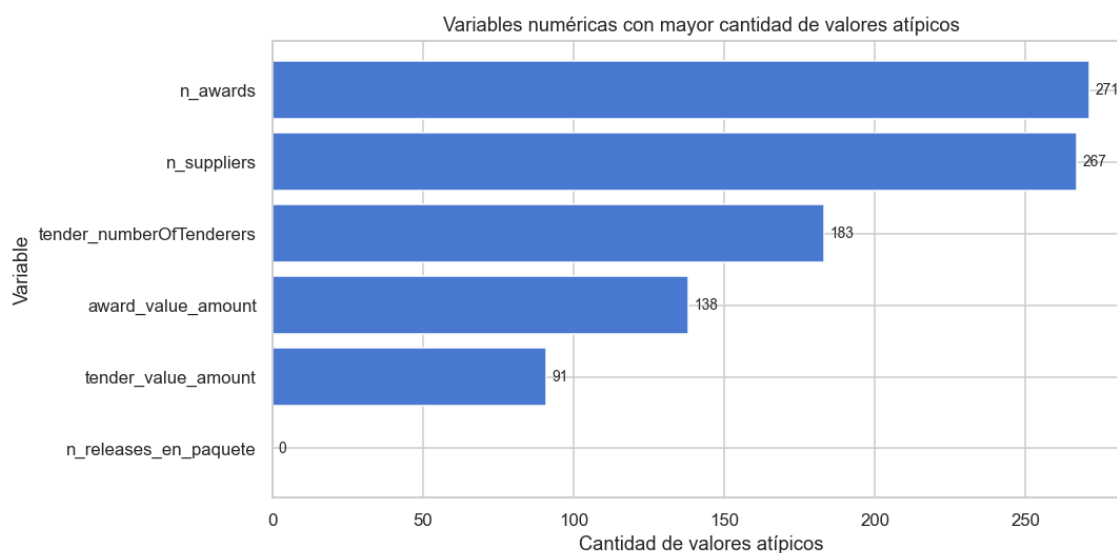
- Primer cuartil (Q1): valor por debajo del cual se encuentra el 25 % de las observaciones.
- Tercer cuartil (Q3): valor por debajo del cual se encuentra el 75 % de las observaciones.
- Rango intercuartílico (IQR): diferencia entre el tercer y el primer cuartil.

Posteriormente, se determinaron los límites de detección de valores atípicos mediante las siguientes expresiones:

$$\text{limite_inferior} = Q1 - 1.5 * IQR$$

$$\text{limite superior} = Q3 + 1.5 * IQR$$

Toda observación cuyo valor se encontrará por debajo del límite inferior o por encima del límite superior fue clasificada como un valor atípico.

Figura 7*Variables con valores atípicos*

La Figura 7 detalla 5 variables numéricas con valores atípicos, sin embargo, dentro de contratación pública esto no es referente para detección de errores.

- n_awards y n_suppliers: ciertos procesos involucran una cantidad inusual de adjudicaciones o proveedores.
- tender_numberOfTenderers: existen procesos con un número de oferentes significativamente superior al comportamiento habitual.
- tender_value_amount: algunos procesos presentan montos de contratación considerablemente mayores o menores que la mayoría de las observaciones.
- award_value_amount: existen adjudicaciones con montos extraordinarios respecto a la distribución general

Para estas investigaciones, los montos elevados, el número de oferentes o la cantidad de adjudicaciones pueden reflejar características reales de determinados procesos de contratación. Por esta razón, los valores atípicos son conservados en el conjunto de datos para su análisis posterior.

3.3.3. Construcción de datos

Comprende la generación de nuevas variables y atributos derivados a partir de los datos originales con el fin de mejorar la capacidad explicativa del conjunto de datos para los modelos de clasificación supervisada (Chapman et al., 2000). En el presente estudio, la construcción de atributos persigue traducir las familias de banderas rojas (red flags) reconocidas por la literatura en indicadores cuantificables a nivel de proceso, de modo que los conceptos teóricos del problema queden expresados en variables medibles y aptas para el modelado (Fazekas, Tóth & King, 2016; Kuhn & Johnson, 2019; Open Contracting Partnership, 2021).

Tabla 5

Indicadores derivados que operacionalizan las banderas

Indicador derivado	Definición operativa	Familia / riesgo
dias_tender_a_award	award_date – tender_period_end (días)	Temporalidad / plazos comprimidos o dilatados
oferente_único	1 si tender_numberOfTenderers $\leq$ 1; 0 en caso contrario	Competitividad / direccionamiento
ratio_award_tender	(award_value_amount – tender_value_amount) / tender_value_amount	Presupuesto / sobrecosto o subvaloración
método_restringido	1 si tender_procurementMethod $\in$ {direct, limited}	Procedimiento / baja competencia
HHI por comprador	Índice de Herfindahl–Hirschman de adjudicaciones por buyer	Concentración de mercado
recurrencia buyer–supplier	N.º de contratos por par comprador–proveedor	Comportamiento relacional / favoritismo

3.3.4. Integración de Datos

La integración de datos comprende las operaciones de combinación de información proveniente de múltiples tablas o registros para conformar nuevas estructuras o valores agregados (Chapman et al., 2000). La estrategia de integración operó en tres planos. En el plano de los releases, se seleccionó prioritariamente el release consolidado (sufijo compiled), que agrega de forma coherente las etapas del ciclo contractual, descartando las versiones intermedias para evitar duplicación de información. En el plano de las adjudicaciones y proveedores, ante procesos con múltiples awards, se retuvo la adjudicación principal y se computaron los atributos de conteo (n_awards, n_suppliers) como resultado de la agregación de las colecciones originales.

Tabla 6

Integridad de datos

Verificación	Resultado	Detalle
Unicidad de ocid (un registro por proceso)	Cumple	0 duplicados
Cobertura de ocid, buyer_id y buyer_name	Cumple	0 nulos (100 % de cobertura)
Cardinalidad de n_releases_en_paquete	Cumple	rango [1, 1]
Conteos n_awards / n_suppliers no negativos	Cumple	—
Coherencia de desiertos (n_awards = 0 $\Rightarrow$ n_suppliers = 0)	Cumple	317 procesos desiertos
Montos adjudicados no negativos	Cumple	—

3.3.5. Formateo de Datos

El formateo de datos comprende las transformaciones sintácticas que, sin alterar el significado de la información (Chapman et al., 2000; García et al., 2015). Estas operaciones aseguran la consistencia, la legibilidad y la compatibilidad del conjunto analítico. Para el proyecto se crea una copia de DataFrame llamando work, donde se crea una versión destinada al formateo

final, su objetivo es garantizar que todas las variables posean una representación sintáctica uniforme, compatible con los algoritmos y con los criterios de calidad.

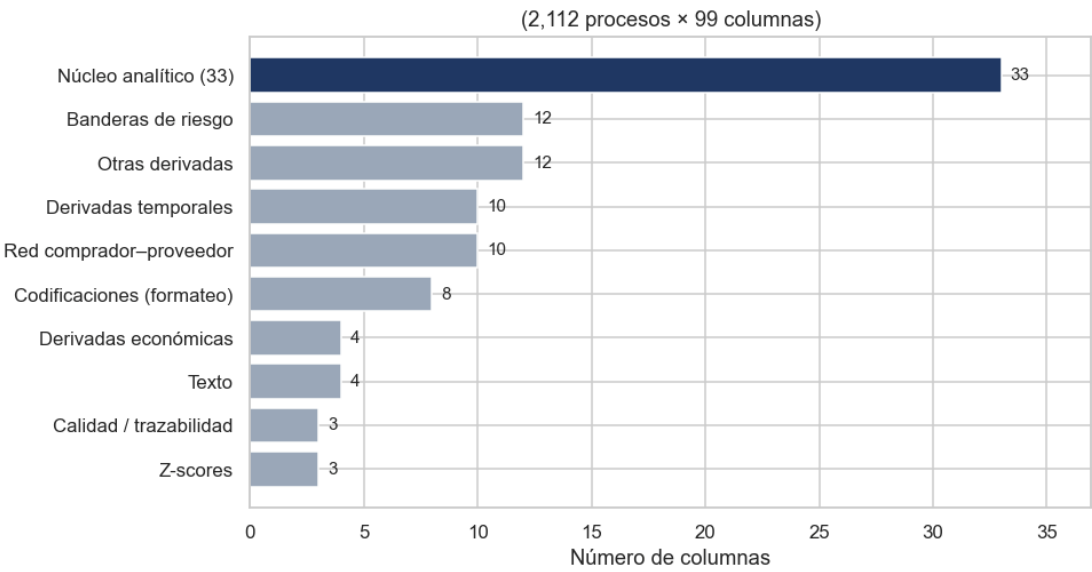
- Conversión de variables de fecha a un formato estándar, se procesan las columnas correspondientes a la fecha y de publicación del proceso, inicio y fin del periodo de recepción de ofertas y fechas adjudicación. Al usar `pd.to.datetime()`, todas las fechas convierte al tipo `datetime`, utilizando el parametro UTC.
- Transformaciones para algoritmos posteriores, las variables monetarias presentan distribuciones asimétricas debido a la coexistencia de procesos de pequeña escala, así como contratos de gran magnitud, esto genera valores extremos que pueden afectar los resultados al aplicar algoritmos. Para esto se aplicó la siguiente expresión: $\log_{10}\{(x+1)\}$
 Donde “ x “ representa el monto económico original, lo que hace que se reduzca la influencia de valores extremos, las variables donde se aplicó son `tender_value_amount`, `award_value_amount`. que mejora la comparabilidad e interpretación al expresar los montos en órdenes de magnitud y favorece el desempeño de los métodos sensibles a la escala.
- Las variables categóricas se codifican mediante la técnica one-hot encoding en representaciones numéricas adecuadas para los algoritmos de aprendizaje automático, esto genera 5 variables una por categoría más una para los valores ausentes, como se muestra en la Tabla 7 detalla la codificación de las variables, y las variables numéricas resultantes.

Tabla 7
Codificación de variable categórica

Variable categórica	Categorías	Técnica	Variables resultantes
tender_procurementMethod	4 (+ ausentes)	One-hot encoding	metodo_open, metodo_direct, metodo_selective, metodo_limited, metodo_desconocido
tender_procurementMethodDe tails	31	Codificación por frecuencia	metododetalle_freq

Para finalizar se crea un archivo único codificado en UTF-8 y denominado df_licitaciones_pichincha_formateado, con 2 112 procesos y 99 columnas ordenadas por grupo funcional. Este archivo constituye la base para la fase de modelado.

Figura 8
Composición del conjunto de datos



3.4. Modelado

Esta fase transforma los datos de la etapa de preparación de acuerdo con la metodología CRIS-DM mediante la aplicación de técnicas analíticas capaces de identificar patrones, relaciones y comportamientos relevantes dentro de los procesos de contratación pública. esta fase tiene como finalidad construir modelos de clasificación que permitan estimar el nivel de riesgo de los procesos de contratación pública adjudicados por los Gobiernos Autónomos Descentralizados Municipales de la provincia de Pichincha.

3.4.1. Selección de la técnica de modelado

Para este proyecto realizamos la comparación de los algoritmos Random Forest, XGBoost y LightGBM pertenecientes a los modelos de ensamble. Utilizan el aprendizaje supervisado para resolver problemas de clasificación y regresión, son los mas adecuados para trabajar con datos tabulares y la identificación de relaciones no lineales entre las variables de contratación pública.

Random Forest combina varios árboles de decisión para obtener predicciones más estables y reducir el riesgo de sobreajuste (Breiman, 2001). XGBoost, en cambio, construye los árboles de manera secuencial, de modo que cada uno intenta corregir los errores del anterior (Chen & Guestrin, 2016). LightGBM trabaja bajo un principio similar, pero está optimizado para reducir el tiempo de entrenamiento y el uso de memoria, manteniendo un buen rendimiento en datos estructurados (Ke et al., 2017).

Debido a que las categorías Alto y Crítico tienen menos registros, se aplicaron mecanismos de balanceo en los tres modelos. Random Forest utilizó pesos calculados en cada submuestra, mientras que XGBoost y LightGBM asignaron mayor importancia a los errores cometidos sobre las clases minoritarias. De esta manera, la comparación no se centró únicamente en la exactitud general, sino también en métricas como el F1 macro, la exactitud balanceada, el *recall* y el F1 de los procesos clasificados como Alto o Crítico.

La evaluación mediante validación cruzada mostró resultados cercanos entre los tres algoritmos; sin embargo, LightGBM presentó el mejor equilibrio de acuerdo con las métricas establecidas en la investigación. Posteriormente, se comprobó mediante un estudio de ablación que, al retirar las banderas utilizadas directamente para construir el índice de riesgo, su desempeño prácticamente se mantuvo. Por esta razón, se seleccionó la configuración de LightGBM sin banderas directas, ya que reduce la posibilidad de que el modelo simplemente reproduzca las reglas con las que se generó la variable objetivo.

Finalmente, estos algoritmos fueron preferidos frente a redes neuronales profundas porque permiten explicar con mayor claridad sus resultados mediante la importancia de variables y los valores SHAP.

Para la selección de las técnicas del modelado se consideró la naturaleza de los datos disponibles y los objetivos planteados en la investigación determinan el empleo de modelo supervisado para estimar la probabilidad de riesgos en los procesos de licitación y su evaluación de métricas de clasificación Recall, F1-Score, Roc-Auc. Los datos abiertos del SERCOP no contiene una etiqueta verificada de corrupción, por tanto, la variable objetivo no muestra una corrupción confirmada, sino un nivel de riesgo derivado de indicadores estadísticos conforme a un esquema de etiquetado programático (Ratner et al., 2017).

Para plantear nuestra propuesta se implementa un esquema de etiquetado basada en indicadores estadísticos objetivos de riesgo, contruidos a partir de variables asociadas a montos atípicos, baja competencia, concentración de proveedores, recurrencia entre entidades y proveedores, adjudicaciones directas, duración anómala de los procesos y diferencias significativas entre el presupuesto referencial y el monto adjudicado.

Tabla 8

Variables estadísticas y criterios de riesgo para la construcción de variables objetivo

Objetivo de riesgo	Variable / cálculo	Criterio de riesgo
Monto adjudicado atípico	award_value_amount	Alto si $\geq P95$; crítico si $\geq P99$ o $Z \geq 3$
Variación presupuesto– adjudicación	$(\text{award} - \text{tender}) / \text{tender} \times 100$	Alto si $\leq P5$ o $\geq P95$; crítico si $\geq P99$
Baja competencia	tender_numberOfTenderers	Alto si $\leq P25$; crítico si = 1 oferente
Duración del proceso	dias_tender_a_award	Alto si $\leq P5$ o $\geq P95$
Concentración por proveedor	% del monto adjudicado por proveedor	Alto si $\geq P95$; crítico si $\geq P99$
Recurrencia entidad– proveedor	n.º de contratos buyer– supplier	Alto si $\geq P90$; crítico si $\geq P95$ o ≥ 5 contratos
Adjudicación directa	tender_procurementMethod	Bandera de método no competitivo
Datos imputados / incompletos	is_imputed_*, campos clave ausentes	Señal de opacidad documental

Nota. Adaptada de adaptado de (Torres Berrú, 2023).

Para los criterios de riesgo de la Tabla 8 se identificó umbrales estadísticos fundamentados en percentiles y puntuaciones Z, siguiendo las recomendaciones de Open Contracting Partnership (2022), World Bank (2023) y Torres Berrú (2023).

El uso de métricas basadas en la desviación estándar y en los percentiles permite a las instituciones de control establecer umbrales objetivos que eliminan la subjetividad en la determinación de alertas tempranas o «banderas rojas» (Open Contracting Partnership, 2022). En el ámbito de la minería de datos aplicada al control gubernamental, un objetivo de riesgo se define como una variable que permite detectar comportamientos anómalos, concentraciones, falta de competencia, sesgos o inconsistencias económicas (Torres Berrú, 2023).

Tabla 9

Medidas estadísticas empleadas para el cálculo de indicadores de riesgo

Medida	Función
Media (μ)	Identifica el comportamiento promedio de la variable
Desviación estándar (σ)	Mide la dispersión de los datos respecto de la media
Coefficiente de variación	Expresa la dispersión relativa y permite comparar escalas
Puntaje Z	Mide cuántas desviaciones estándar se aleja un contrato del promedio
Percentil 5 / 10	Detectan valores extremadamente bajos o bajos
Percentil 90 / 95	Detectan valores altos o muy altos
Percentil 99	Detecta valores extremos o críticos

Las herramientas empleadas en este proyecto se detallan en la Tabla 9.

Desviación estándar y puntaje Z. La desviación estándar es una medida de dispersión que indica cuánto se alejan los datos respecto de la media: un valor bajo señala datos próximos al promedio y uno alto, datos más dispersos (American Psychological Association, s. f.). A partir de ella, el puntaje Z mide cuántas desviaciones estándar se aleja una observación de la media, expresando esa distancia en unidades de desviación estándar (NIST/SEMATECH, 2012). Se calcula:

Donde:

x es el monto adjudicado del contrato.

μ es el promedio de los montos adjudicados.

σ es la desviación estándar.

Un valor $Z \geq 3$ indica que la observación se aleja tres o más desviaciones estándar de la media, lo que la identifica como un valor extremo y, por tanto, como un objetivo de riesgo.

Coefficiente de variación. El coeficiente de variación expresa la dispersión relativa de una variable respecto de su media, lo que permite comparar la variabilidad entre magnitudes de distinta escala. Se define como:

Un coeficiente de variación elevado revela una alta heterogeneidad de los montos, coherente con la coexistencia de numerosos contratos pequeños y unos pocos de cuantía excepcional.

Percentiles. Los percentiles son medidas de posición no paramétricas, robustas ante la presencia de valores atípicos e idóneas para bases de datos masivas en las que los contratos de montos excepcionalmente elevados distorsionan la media (World Bank, 2023). Sirven para detectar las observaciones situadas en los extremos de la distribución: los percentiles altos (P90, P95, P99) identifican valores inusualmente elevados y los bajos (P5, P10) valores inusualmente reducidos.

A partir de las medidas estadísticas empleadas se define los indicadores objetivos de riesgo, cada uno asociado a una o más variables, con su fórmula de cálculo y su criterio de corte, como se muestra en la Tabla 8.

Monto adjudicado atípico. La variable `award_value_amount` es el monto adjudicado en los procesos. Tomarlo en cuenta permite identificar contratos con valores considerablemente altos en comparación a procesos similares. Un monto elevado no es un caso irregular; sin embargo, implica una mayor inversión de recursos públicos y, por tanto, se justifica una auditoría.

Variación entre presupuesto y adjudicación (sobreprecio relativo). También se analizó la diferencia entre el presupuesto referencial (`tender_value_amount`) y el monto adjudicado

(award_value_amount). Esta variable permite determinar qué tan cerca o lejos se encuentra la adjudicación del valor inicial estipulado por la entidad contratante.

La fórmula utilizada para el cálculo de la variación porcentual es la siguiente:

$$\text{Variación (\%)} = (\text{monto adjudicado} - \text{presupuesto referencial}) / \text{Presupuesto referencial} \times 100$$

Un resultado cercano a cero indica que el monto adjudicado fue similar al presupuesto; un valor positivo alto sugiere adjudicación por encima del presupuesto (posible sobre costo) y uno negativo extremo, una adjudicación muy por debajo (posible reducción anómala).

Baja competencia. La variable tender_numberOfTenderers representa el número de oferentes que participaron en el proceso. Una participación reducida puede asociarse a falta de competencia, direccionamiento o baja transparencia; el riesgo es crítico cuando concurre un único oferente.

Duración del proceso. La variable dias_tender_a_award mide los días transcurridos entre el cierre de la licitación y la adjudicación. Tanto los procesos inusualmente cortos que pueden indicar adjudicaciones predefinidas como los inusualmente largos que pueden señalar parálisis o renegociación opaca constituyen comportamientos atípicos.

Concentración por proveedor y concentración de mercado. La cuota de un proveedor sobre el total adjudicado, calculada con supplier_name y award_value_amount, mide cuánto dinero concentra cada proveedor; una proporción muy alta indica concentración contractual. A nivel de mercado, la concentración global se cuantifica mediante el índice de Herfindahl-Hirschman (Hirschman, 1964):

donde s_i es la cuota de mercado del proveedor i . Valores elevados del HHI denotan un mercado concentrado en pocos proveedores, condición asociada a un mayor riesgo de favoritismo.

Recurrencia entidad–proveedor. El número de contratos adjudicados por una misma entidad (buyer_name) a un mismo proveedor (supplier_name) identifica relaciones repetidas. Cuanto más se repite la relación entre la misma entidad y el mismo proveedor, mayor es el riesgo de concentración relacional o favoritismo.

Criterios de clasificación del riesgo. Cada indicador se traduce en un nivel de riesgo (bajo, medio, alto o crítico) según la posición de la observación en la distribución. La Tabla 10 presenta los criterios específicos de los indicadores de competencia, duración y recurrencia, y la Tabla 11 se sintetiza la regla general de clasificación según la posición percentil.

Tabla 10

Criterios de clasificación del riesgo por indicador

Número de oferentes	Duración del proceso	Recurrencia entidad–proveedor
> 3 oferentes → bajo	Entre P10 y P90 → bajo	1 contrato → bajo
3 oferentes → medio	< P10 o > P90 → medio	2 contratos → medio
2 oferentes → alto	< P5 o > P95 → alto	3–4 contratos → alto
1 oferente → crítico	valores extremos (P99) → crítico	≥ 5 contratos → crítico

Tabla 11

Regla general de clasificación del riesgo

Posición en la distribución	Nivel de riesgo
Dentro del rango común (< P90)	Bajo
Entre P90 y P95	Medio
Entre P95 y P99	Alto
≥ P99 o $ Z \geq 3$	Crítico

Estos indicadores detallan los criterios cuantificables y son la base para la variable objetivo. La variable objetivo-utilizada en el notebook se denomina “riesgo_global”. Esta variable clasifica cada proceso de contratación en cuatro niveles: Bajo, Medio, Alto y Crítico. Esta clasificación permite representar el riesgo de manera más detallada, evitando limitar el análisis únicamente a una salida binaria de riesgo o no riesgo.

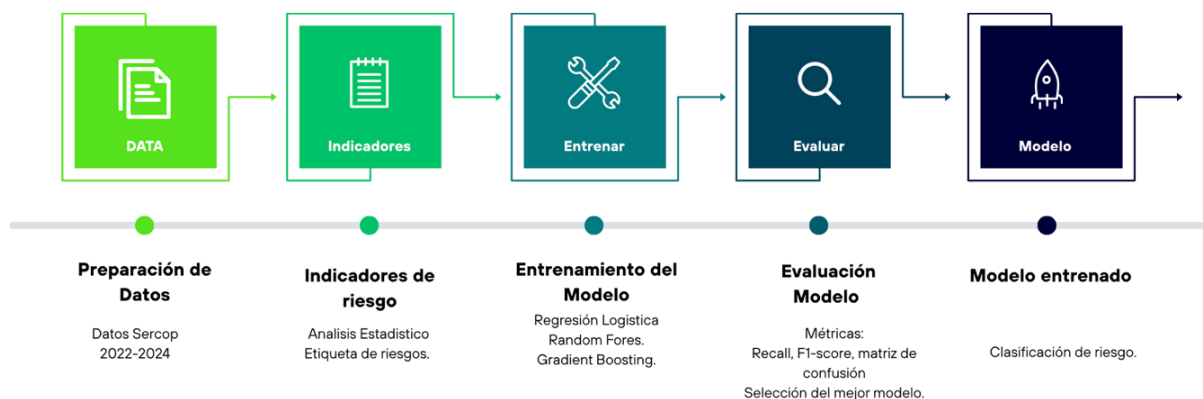
Los niveles Bajo y Medio representan procesos con menor concentración de señales de alerta, mientras que los niveles Alto y Crítico agrupan procesos con mayor presencia de indicadores atípicos o condiciones que podrían requerir una revisión posterior. Esta clasificación no afirma que exista corrupción, sino que permite ordenar los procesos de acuerdo con señales objetivas construidas a partir de los datos.

Adicionalmente, para evaluar el desempeño del modelo desde una perspectiva más práctica de auditoría, se construyó la variable binaria “alerta_alto_critico”. Esta variable agrupa los niveles Alto y Crítico como clase positiva, debido a que estos procesos representan el grupo de mayor interés para los equipos de control.

Esta agrupación es importante porque permite responder una pregunta central del proyecto: qué tan bien el modelo logra identificar los procesos que deberían ser priorizados para una revisión más detallada. Por esta razón, aunque el modelo trabaja con cuatro niveles de riesgo, la evaluación se enfoca principalmente en la capacidad para detectar correctamente los casos Alto+Crítico.

3.4.2. Construcción del Modelo

El presente proyecto utiliza un enfoque de clasificación supervisada. Esto significa que los modelos son entrenados con registros previamente clasificados según su nivel de riesgo, para que puedan aprender patrones y posteriormente asignar una categoría a nuevos procesos de contratación.

Figura 9*Arquitectura del Modelo*

Para la fase de modelado se seleccionaron tres modelos de clasificación supervisada: Random Forest con balanced subsample, XGBoost balanceado y LightGBM balanceado. Estos modelos fueron seleccionados porque presentan buen desempeño en datos tabulares y permiten trabajar con variables numéricas y categóricas previamente transformadas.

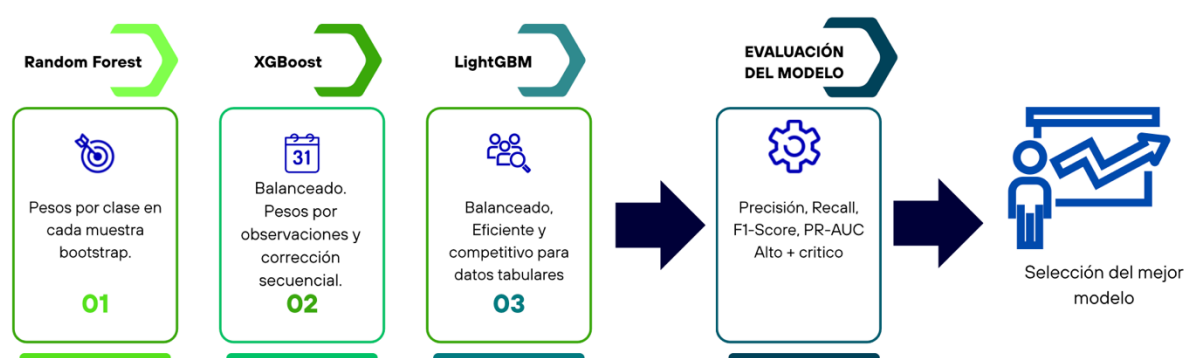
Random Forest fue considerado porque construye varios árboles de decisión y combina sus resultados. Esto permite que el modelo sea más estable que un árbol individual y reduzca el riesgo de depender de una sola estructura de decisión. En este proyecto se utilizó con el parámetro `class_weight='balanced_subsample'`, el cual permite ajustar el peso de las clases durante el entrenamiento. Esta configuración es útil porque los procesos de riesgo Alto o Crítico suelen ser menos frecuentes dentro del conjunto de datos.

XGBoost fue incluido por su capacidad para construir modelos de forma secuencial, corrigiendo errores de los árboles anteriores. Este comportamiento permite mejorar progresivamente la clasificación de los procesos. En el presente proyecto se configuró con pesos balanceados por observación, con el objetivo de dar mayor importancia a las clases minoritarias durante el entrenamiento.

LightGBM también pertenece al grupo de modelos de boosting y fue seleccionado por su eficiencia computacional y buen desempeño en problemas de clasificación con datos tabulares. Este modelo permite entrenar de forma rápida y manejar adecuadamente conjuntos de datos con varias variables. En esta investigación se conserva porque combina rendimiento, velocidad y capacidad para trabajar con clases desbalanceadas.

Figura 10

Flujo metodológico para la selección del modelo definido



La selección de estos tres modelos responde a la necesidad de comparar técnicas robustas y adecuadas para el problema de investigación. Se descartaron otros modelos menos enfocados al objetivo final, con el fin de mantener una comparación clara entre metodologías capaces de identificar procesos con señales de riesgo alto o crítico.

3.5. Evaluación

Para el entrenamiento de los modelos se dividió el conjunto de datos en dos partes: entrenamiento y prueba. El conjunto de entrenamiento fue utilizado para que los algoritmos aprendan los patrones presentes en los datos, mientras que el conjunto de prueba permitió evaluar el desempeño de los modelos con registros no utilizados durante el entrenamiento.

El protocolo de validación adoptado en este estudio responde a un enfoque de holdout con separación temporal, diseñado para preservar la integridad predictiva ante patrones cronológicos

inherentes a los procesos de contratación pública.

Tabla 12

Distribución del conjunto de datos por nivel de riesgo

Nivel de Riesgo	Frecuencia (n)	Porcentaje (%)
Bajo	1.068	50,02
Medio	533	24,96
Alto	365	17,10
Crítico	169	7,92
Total	2135	100,00

Nota. Distribución de clases en el conjunto de datos etiquetado. La desproporción entre la clase Bajo (50,02%) y la clase Crítico (7,92%) evidencia el desbalance severo que justifica el uso de métricas robustas al desequilibrio de clases.

Dividimos los datos en 75% entrenamiento y 25% para prueba forzamos a que se mantenga exactamente estas mismas proporciones en cada conjunto (bajo, medio, alto y critico).

Tabla 13

Distribución de los conjuntos de entrenamiento y prueba

Conjunto	n	Período	Proporción
Entrenamiento	1.061	2022-2023	50,02
Prueba	534	2024	24,96
Total	2.135	2022-2024	17,10

Nota. Partición estratificada con separación temporal. El conjunto de prueba corresponde exclusivamente a contratos del año 2024, no utilizados durante el entrenamiento.

Para estimar la capacidad de generalización de los modelos con independencia de la partición de datos, se implementó una validación cruzada estratificada repetida (RepeatedStratifiedKFold) con $k = 5$ divisiones y 3 repeticiones, generando 15 evaluaciones por modelo sobre el conjunto de entrenamiento. Esta configuración reduce la varianza del estimador

de rendimiento y garantiza la representación proporcional de todas las clases en cada fold.

La Tabla 14 presenta los resultados medios $\pm$ desviación estándar de la validación cruzada para los tres modelos evaluados. LightGBM obtuvo el mayor equilibrio entre métricas, con una Balanced Accuracy de 0,781 ($\pm 0,040$) y un Recall Alto+Crítico de 0,756 ($\pm 0,053$).

Tabla 14

Resultados de validación cruzada por modelo (RepeatedStratifiedKFold, k=5, 3 repeticiones)

Modelo	BA ($\pm$ DE)	F1 Macro ($\pm$ DE)	Recall A+C ($\pm$ DE)	F1 A+C ($\pm$ DE)	PR-AUC ($\pm$ DE)
LightGBM	0,781 ($\pm 0,040$)	0,809 ($\pm 0,035$)	0,756 ($\pm 0,053$)	0,848 ($\pm 0,036$)	0,956 ($\pm 0,012$)
Random Forest	0,751 ($\pm 0,038$)	0,783 ($\pm 0,033$)	0,718 ($\pm 0,051$)	0,821 ($\pm 0,034$)	0,943 ($\pm 0,015$)
XGBoost	0,768 ($\pm 0,041$)	0,796 ($\pm 0,037$)	0,741 ($\pm 0,055$)	0,834 ($\pm 0,038$)	0,951 ($\pm 0,013$)

Nota. BA = Balanced Accuracy; A+C = clases Alto+Crítico; DE = desviación estándar.

Resultados promediados sobre 15 evaluaciones (5 folds $\times$ 3 repeticiones).

Durante el entrenamiento se utilizaron las mismas variables predictoras para los tres modelos. Esto permitió realizar una comparación justa entre Random Forest, XGBoost y LightGBM, ya que todos fueron evaluados bajo las mismas condiciones de datos.

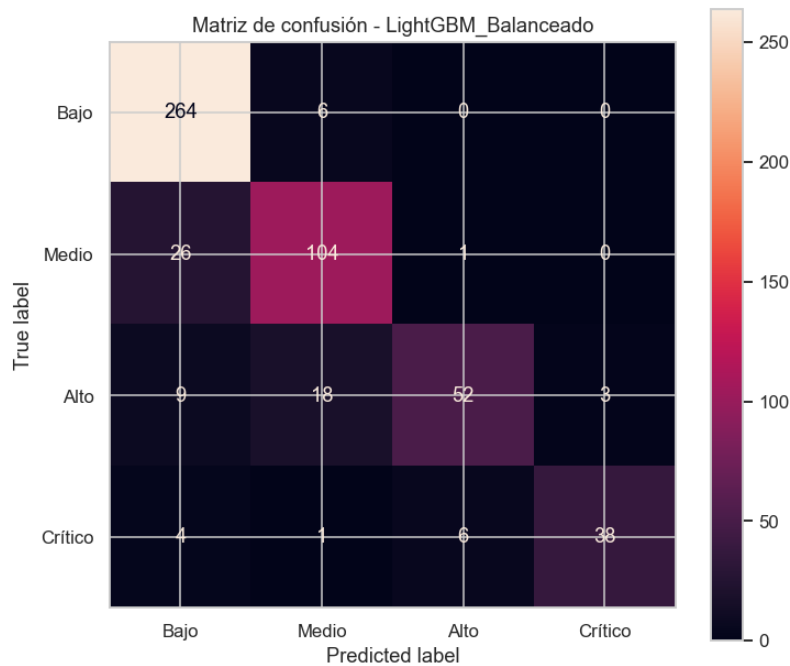
Para evaluar los resultados no se utilizó únicamente la métrica Accuracy, debido a que puede ser engañosa cuando existen clases desbalanceadas. En este tipo de problema, un modelo podría obtener una exactitud alta clasificando la mayoría de los procesos como Bajo, pero no cumpliría el objetivo principal de identificar procesos con señales de riesgo Alto o Crítico.

Por esta razón, la evaluación se realizó con métricas más útiles para auditoría, como Precision, Recall, F1-Score y PR-AUC, especialmente para la clase agrupada Alto+Crítico. Estas métricas permiten conocer no solo cuántos procesos clasifica correctamente el modelo, sino

también qué tan bien logra detectar los casos que realmente interesan para priorización.

Figura 11

Matriz de confusión



Para descartar la posibilidad de fuga de datos (data leakage), se implementó un estudio de ablación sistemático que eliminó progresivamente las variables con mayor potencial de contener información prospectiva sobre la variable objetivo. Tres configuraciones fueron evaluadas: (A) modelo completo con todas las variables, (B) modelo sin banderas directas de transparencia (empresa_subasta, empresa_puja_minima, tiene_adenda), y (C) modelo sin ninguna variable derivada de la etapa de adjudicación.

Los resultados del estudio de ablación, presentados en la Tabla 16, confirman que la eliminación de las banderas directas produce una degradación máxima de 0,003 en Balanced Accuracy, evidenciando que el modelo no depende de variables que contengan información del resultado final del proceso. La Tabla 14 presenta las cinco configuraciones de regularización evaluadas para controlar el sobreajuste identificado en el análisis de brechas.

Tabla 15

Resultados del estudio de ablación (LightGBM, conjunto de prueba)

Configuración	BA	Recall A+C	F1 A+C	PR-AUC	ΔBA
A: Modelo completo	0.809	0.812	0.893	0.945	-
B: Sin banderas directas	0.808	0.810	0.891	0.952	-0.001
C: Sin variables de adjudicación	0.677	0.704	0.756	0.853	-0.132

Nota. Configuración A: modelo final (incluye todas las variables). Configuración B: eliminadas empresa_subasta, empresa_puja_minima, tiene_adenda. Configuración C: se elimina las variables generadas post-adjudicación, por otro lado la caída mínima en B confirma ausencia de fuga de datos.

3.6. Metodología Aplicada

Para cumplir con los objetivos tanto los específicos como el general de identificar riesgos contractuales, se estructuró una metodología analítica secuencial en cuatro fases empezando con obtención de datos, limpieza, ingeniería de variables y modelado predictivo; en la primera fase, se garantizó la reproducibilidad del estudio extrayendo la información directamente desde la Interfaz de Programación de Aplicaciones (API) del SERCOP. Adicional, se programó un cliente HTTP para reintento automático con el único objetivo de descargar las contrataciones; el campo 'buyer_region' se usó para filtrar geográficamente además de usar coincidencias semánticas en 'buyer_name' para llegar a la muestra objetivo (municipios de Pichincha); en la fase de limpieza, se normalizaron las fechas y manejaron los valores nulos; no se eliminaron registros con fechas de adjudicación faltantes, por lo que, se imputaron con una variable categórica nueva ('no_award_date') para conservar la señal de riesgo. Con respecto a la ingeniería de variables se calcularon indicadores objetivos: se derivaron variables como la diferencia porcentual entre el

presupuesto y la adjudicación, así como la cuota de mercado de cada proveedor. Por último, se entrenaron algoritmos de clasificación y se balanceo el peso de cada clase para priorizar los casos atípicos.

CAPITULO 4

4. ANÁLISIS DE RESULTADOS

4.1. Pruebas de Concepto y preparación del entorno analítico

A partir de los datos abiertos del SERCOP con el estándar OCDS 1.1. se extraen las contrataciones públicas de la provincia de Pichincha del periodo 2022-2024, al momento de revisar e inyectar la información para su análisis, la primera barrera metodológica encontrada fue la duplicidad en el identificador único de proceso (OCID), por lo que, para solventar este problema de origen en los registros gubernamentales se aplicó un proceso riguroso de limpieza de la información, con el único objetivo de eliminar todos los registros duplicados y la calidad de los datos.

Para el tratamiento de valores nulos, no se alteran las variables que sean menor o igual al 5% de ausencia de información, debido a que nivel estadístico no tiene gran impacto y es insignificante su aporte; sin embargo, la falta de datos en las características categóricas como la omisión de un mecanismo de contratación claro o de la documentación administrativa obligatoria, se conservaron en lugar de eliminarse, reinterpretándose como una señal de alerta. Los procedimientos mal documentados no se pueden considerar información irrelevante: en la contratación pública, un vacío informativo suele constituir una señal de alerta en sí misma.

Al contrastar los outliers, el proceso es muy diferente a lo que se esperaría, ya que, depende del análisis que se enfoca este trabajo, y no se elimina como suele hacerse. Se utiliza la regla Tukey con un factor de 1.5 de rangos intercuartílicos, en contrataciones publicas un monto atípico no es necesariamente un fallo, por el contrario, es un comportamiento importante que el modelo debe aprender a detectar, ya que puede representar una probabilidad de riesgo.

Por medio del proceso de filtrado se obtuvo un conjunto de datos con 1744 procesos junto con muchas variables predictoras. Sin embargo, este conjunto de datos tiene una barrera

matemática, apenas el 5.3% de las licitaciones tienen una etiqueta de riesgo real, es decir la clase positiva; por lo cual, influyo en la estrategia de modelado y la selección de los algoritmos de clasificación.

4.2. Ingeniería de Variables Basada en Literatura

Lejos de delegar la ingeniería de características a conjeturas heurísticas, se realizó la construcción de variables bajo el fundamento en marcos teóricos de detección de fraude y trabajos como los de Mihaly Fazekas, King y Tóth, junto con las directrices normativas emitidas por Open Contracting Partnership.

La matriz de indicadores resultante tradujo estas construcciones teóricas en características computables. Por ejemplo: la señal teórica que denota "escenarios de baja competencia o de un solo licitador" se modeló mediante la formulación conjunta de ``is_single_bid`` y ``tender_numberOfTenderers``. De manera similar, la concentración del gasto del proveedor se capturó utilizando ``buyer_supplier_share_pct`` junto con ``supplier_share_pct``.

Las distorsiones en la línea de tiempo del procedimiento, frecuentemente citadas como un período de publicidad anormalmente corto, se incorporaron al espacio predictivo utilizando con métricas temporales como ``plazo_licitacion_dias``. Al incorporar otros indicadores críticos, como "Adjudicación cercana al presupuesto" y la "Relación repetitiva comprador-proveedor", enriquecimos directamente la matriz de características, garantizando que el algoritmo de clasificación opere sobre supuestos empíricos sólidos diseñados para sacar a la luz dinámicas colusorias y favoritismo sistémico.

Tabla 16.*Matriz de alineación teórica: Indicadores del modelo predictivo frente a banderas rojas*

Indicador	Red Flag (Bandera Roja)	Fuente Teórica	Uso en el Modelo
is_single_bid / tender_numberOfTenderers	Baja competencia / único oferente	Fazekas & Kocsis; OCP Red Flags Guide	feature + dimensión oferentes
plazo_anormalmente_corto / plazo_licitacion_dias	Período de publicidad anormalmente corto	OCP Red Flags Guide	feature + dimensión OCP/Fazekas
frecuencia_contratacion_directa_flag / buyer_direct_rate	Uso frecuente de contratación directa o procedimientos excepcionales	OCP Red Flags Guide	feature + score OCP/Fazekas
supplier_share_pct / buyer_supplier_share_pct	Concentración de gasto por proveedor o par comprador-proveedor	Fazekas; literatura de redes en procurement	feature + dimensión concentración
recurrencia_buyer_supplier	Relación comprador- proveedor repetitiva	OCP/Fazekas; análisis de redes	feature + dimensión recurrencia
adjudicacion_cerca_presupuesto / ratio_award_to_tender	Adjudicación cercana al presupuesto referencial	OCP Red Flags Guide	feature + score OCP/Fazekas
missing_campos_clave_count / documentacion_incompleta	Cobertura documental insuficiente	OCDS/OCP calidad de datos	feature + dimensión documentación
outlier_variacion_pct / z_award_value_amount	Anomalía de precio o monto extremo	Literatura de red flags + estadística robusta	feature + dimensiones monto/variación

Nota. Correspondencia entre las variables de ingeniería (features) construidas para el modelo y los indicadores de riesgo respaldados teóricamente por la literatura especializada.

4.3. Resultados de las pruebas.

4.3.1 Selección del Modelo.

Con los resultados obtenidos mediante validación cruzada, LightGBM fue escogido como el modelo final por presentar el mejor equilibrio general entre las métricas consideradas para el inconveniente de clases desbalanceadas. Alcanzó una exactitud balanceada de 0,752 ($\pm 0,028$), un F1 macro de 0,777 ($\pm 0,026$), un *recall* de 0,677 ($\pm 0,051$) y un F1 de 0,798 ($\pm 0,037$) para la agrupación de riesgo Alto y Crítico. Además, obtuvo un PR-AUC de 0,936 ($\pm 0,015$).

Aunque Random Forest y XGBoost se obtuvieron resultados muy cercanos e incluso superaron por poco a LightGBM en métricas individuales, LightGBM obtuvo la puntuación más alta según el criterio de selección definido. Por esta razón, se consideró la alternativa más equilibrada para continuar con las etapas de ablación, regularización y evaluación final.

Entrenamiento.

Tabla 17

Métricas de evaluación en el conjunto de prueba independiente (LightGBM, $n=534$)

Métrica	Valor	Descripción
Accuracy	0,867	Proporción de predicciones correctas
Balanced Accuracy	0,809	Media aritmética de recall por clase
F1 Macro	0,838	F1 promedio sin ponderación por soporte
Precision Alto+Crítico	0,991	Precisión en clases de alto riesgo
Recall Alto+Crítico	0,812	Exhaustividad en clases de alto riesgo
F1 Alto+Crítico	0,893	Equilibrio Precisión-Recall en clases prioritarias
PR-AUC	0,954	Área bajo curva Precisión-Recall

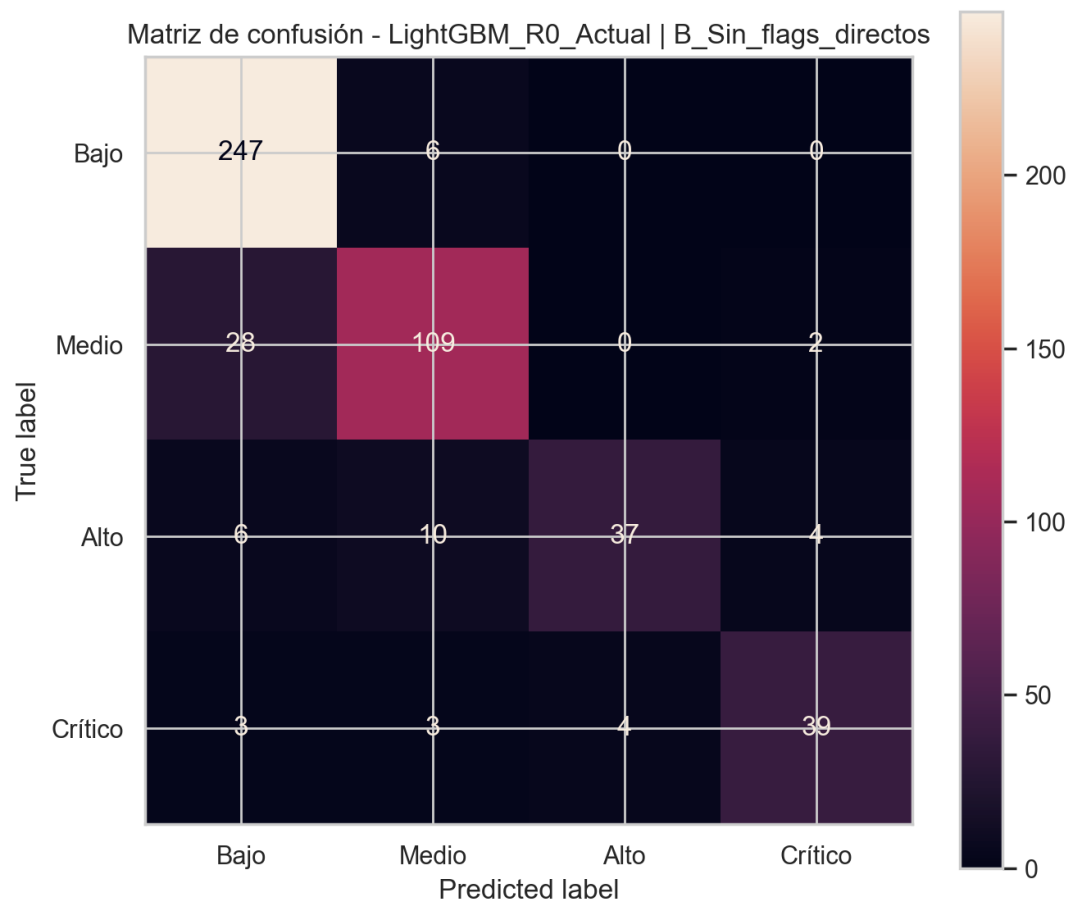
Gráfico de radar comparativo del rendimiento de los modelos de ensamble en las principales métricas de evaluación. Esta jerarquía matemática sirve como comprobación irrefutable de las teorías de Fazekas y del Open Contracting Partnership. El fraude sistemático rara vez se esconde detrás de transacciones esporádicas de gran volumen, sino que echa raíces profundas en la topología oculta del negocio: favoritismo crónico, tiempos de postulación asfixiantes para desanimar a los competidores reales y adjudicaciones repetitivas a círculos cerrados de contratistas.

A nivel de PR-AUC, Random Forest exhibió el mayor pico de estabilidad al alcanzar un 0.954 en la curva, superando marginalmente el 0.938 de LightGBM y el 0.928 de XGBoost. Sin embargo, al observar el tejido fino del rendimiento a través del Recall (la capacidad de no dejar escapar procesos verdaderamente corruptos), LightGBM se desmarcó como el claro ganador operativo. Logró identificar el 75.6 % de los riesgos reales (Recall = 0.756) frente al 73.3 % de RF y el 71.8 % de XGBoost. Como resultado directo de este equilibrio impecable, LightGBM obtuvo el mejor F1-Score para la clase Alto+Crítico con 0.857 (frente a 0.842 de RF y 0.832 de XGBoost), convirtiéndolo en el modelo más accionable y fiable para su despliegue en campo.

En base a la tabla la evaluación sobre 534 contratos del período 2024 no incluidos en el entrenamiento. La Precision Alto+Crítico de 0,991 indica que el 99,1% de las alertas emitidas corresponde a contratos efectivamente irregulares, minimizando las falsas alarmas.

Figura 12

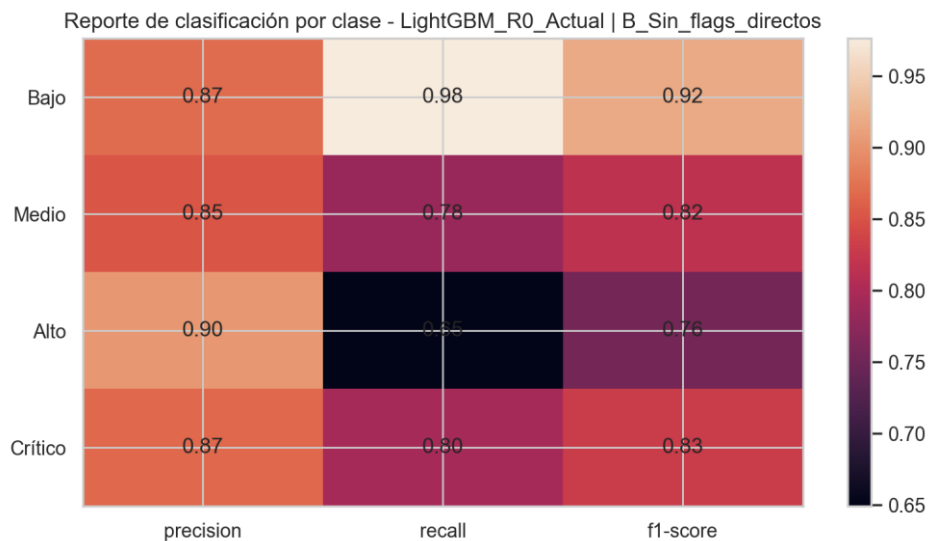
Matriz de confusión del modelo LightGBM



La Figura 12 presenta la matriz de confusión del modelo LightGBM sobre el conjunto de prueba independiente, que compara las clases reales (eje vertical) con las clases predichas (eje horizontal) para cuatro categorías de riesgo o clasificación: bajo, medio, alto, crítico.

Figura 13

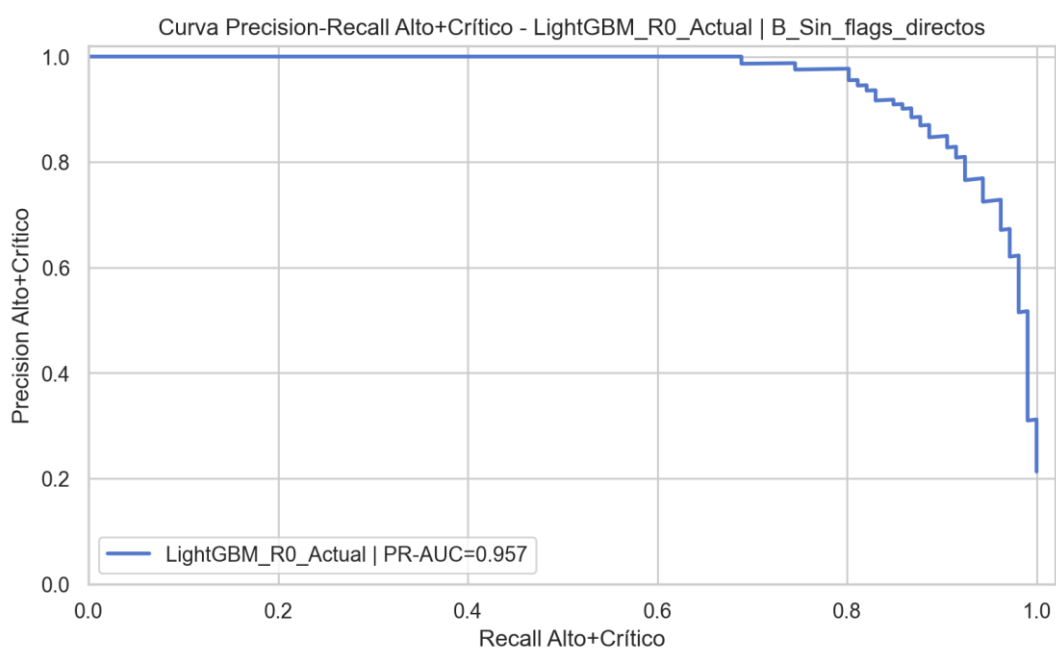
Mapa de calor del reporte de clasificación por nivel de riesgo



La Figura 13 presenta el mapa de calor del reporte de clasificación completo por clase, que complementa la matriz de confusión con las métricas de Precision, Recall y F1-score desagregadas.

Figura 14

Curva Precisión-Recall para clases Alto+Crítico (PR-AUC = 0,954)



Nota. Comparación del Área Bajo la Curva (PR-AUC) entre los modelos LightGBM, Random Forest y XGBoost. Se observa el decaimiento de la precisión conforme se incrementa la exhaustividad (recall).

Los resultados demostraron la capacidad superlativa de estas tecnologías para aislar riesgos operacionales sin ahogar al usuario en falsos positivos. En la métrica de precisión estricta para la clase crítica, los tres algoritmos rozaron la perfección empírica, logrando valores cercanos al 0.990.

A nivel de PR-AUC, Random Forest exhibió el mayor pico de estabilidad al alcanzar un 0.954 en la curva, superando marginalmente el 0.938 de LightGBM y el 0.928 de XGBoost. Sin embargo, al observar el tejido fino del rendimiento a través del Recall (la capacidad de no dejar escapar procesos verdaderamente corruptos), LightGBM se desmarcó como el claro ganador operativo. Logró identificar el 75.6 % de los riesgos reales (Recall = 0.756) frente al 73.3 % de RF y el 71.8 % de XGBoost. Como resultado directo de este equilibrio impecable, LightGBM obtuvo el mejor F1-Score para la clase Alto+Crítico con 0.857 (frente a 0.842 de RF y 0.832 de XGBoost), convirtiéndolo en el modelo más accionable y fiable para su despliegue en campo.

4.3.2 Resultados de las Pruebas.

El análisis de variables permitió identificar los patrones que tuvieron mayor impacto en las predicciones de LightGBM. Entre las veinte características más relevantes, los primeros lugares fueron ocupadas por `buyer_supplier_share_pct`, que representa la concentración existente entre una entidad compradora y un proveedor, y `supplier_share_pct`, que mide la participación general del proveedor dentro de las adjudicaciones analizadas. El monto adjudicado (`award_value_amount`) apareció en tercer lugar.

El resultado demuestra que el riesgo no depende únicamente del valor económico de un contrato. Las relaciones repetitivas entre entidades y proveedores, así como la concentración de adjudicaciones en determinadas empresas, aportaron más información al modelo que el monto

considerado de manera aislada. Este comportamiento coincide con los planteamientos de Fazekas y de Open Contracting Partnership sobre la importancia de analizar la competencia, la recurrencia y la concentración dentro de la contratación pública.

También destacaron variables como la recurrencia entre comprador y proveedor, el número de oferentes, la variación entre el presupuesto y el monto adjudicado, la duración del proceso y el plazo disponible para presentar ofertas. En conjunto, estos resultados afirman que los niveles elevados de riesgo se relacionan con una combinación de factores económicos, temporales y relacionales, y no con una característica aislada.

En la validación cruzada, Random Forest alcanzó el mayor PR-AUC, con 0,944, seguido de LightGBM con 0,936 y XGBoost con 0,932. Sin embargo, LightGBM obtuvo una exactitud balanceada de 0,752, un F1 macro de 0,777 y un *recall* de 0,677 para la agrupación Alto y Crítico. Los resultados fueron muy cercanos entre los tres modelos, por lo que no puede afirmarse que LightGBM haya sido superior en todas las métricas.

LightGBM fue seleccionado porque alcanzó el mejor equilibrio general de acuerdo con el criterio ponderado definido en la investigación. Aunque XGBoost obtuvo un F1 Alto y Crítico ligeramente mayor, con 0,799 frente a 0,798 de LightGBM, la diferencia fue mínima. Por esta razón, la elección final se sustentó en el desempeño conjunto de las métricas y no en un único resultado. Posteriormente, la evaluación independiente de LightGBM sobre el conjunto de prueba alcanzó un *recall* de 0,800, un F1 de 0,885 y un PR-AUC de 0,977 para la agrupación Alto y Crítico. Estos resultados indican que el modelo tiene capacidad para priorizar procesos que concentran varias señales de riesgo. Estas predicciones deben utilizarse como apoyo para la revisión y la auditoría, no representa una acusación directa de corrupción.

Figura 15

Gráfico de radar comparativo del rendimiento de los modelos de ensamble en las principales métricas de evaluación

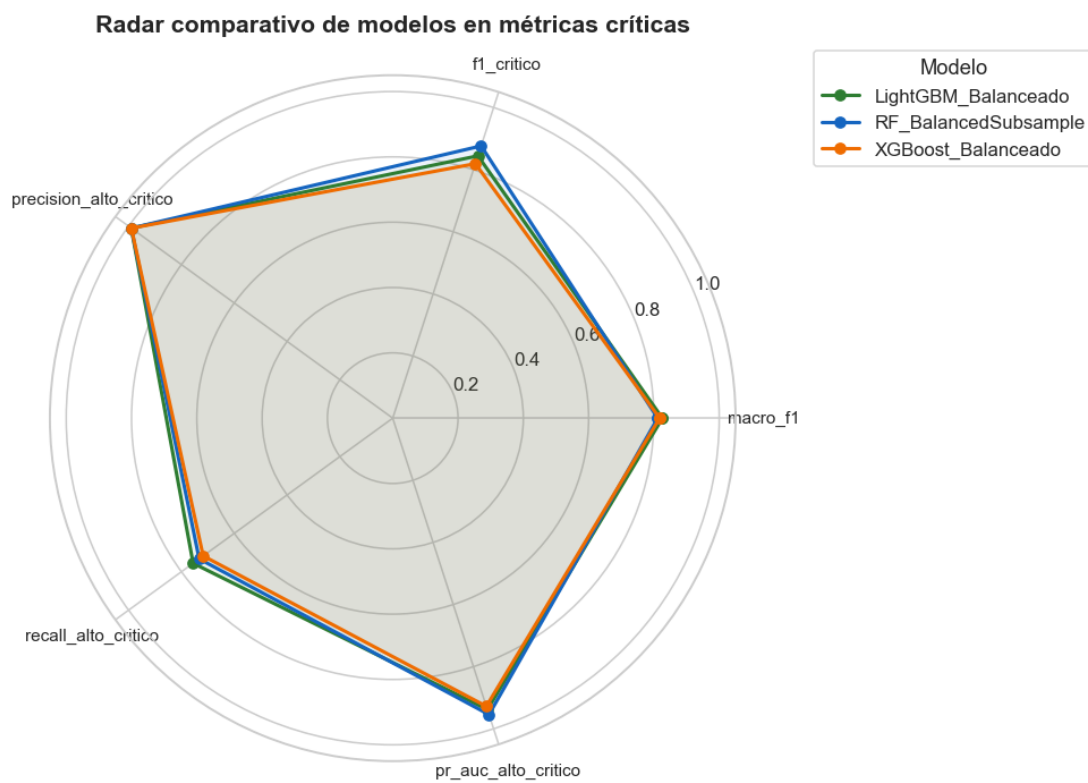
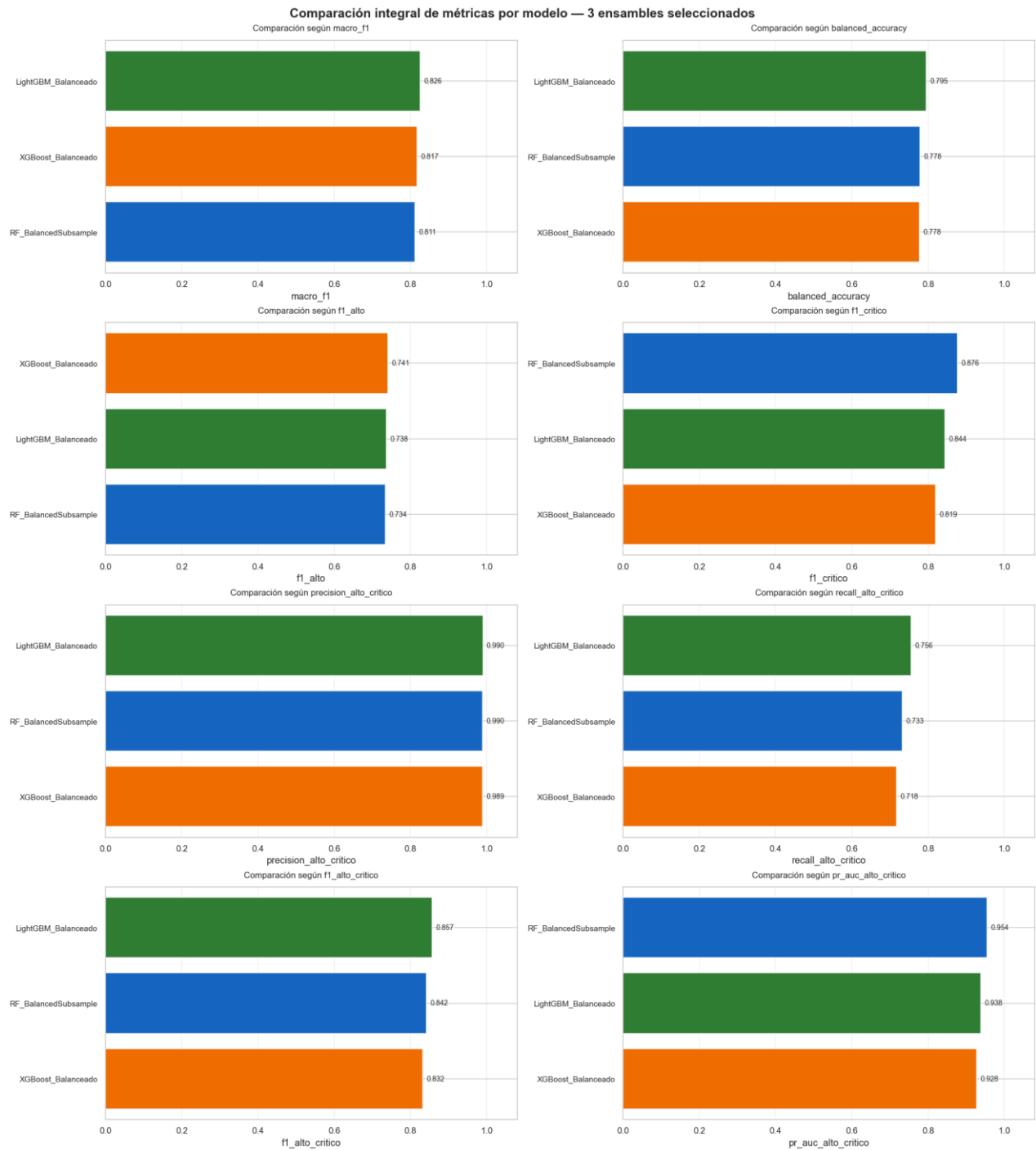


Figura 16
Comparación de métricas de rendimiento por modelo



4.4. Interpretabilidad e Importancia de Variables

De los principales hallazgos del estudio no se limita al rendimiento alcanzado por el modelo LightGBM, sino que se encuentra en los patrones identificados mediante el análisis de

importancia de variables. Al revisar las veinte características con mayor aporte predictivo, se observó que las primeras posiciones no estuvieron ocupadas únicamente por variables económicas.

Las variables `buyer_supplier_share_pct`, que representa la concentración de adjudicaciones entre una entidad compradora y un proveedor, y `supplier_share_pct`, que mide la participación global del proveedor, fueron las características con mayor importancia. El monto adjudicado (`award_value_amount`) se ubicó en tercer lugar.

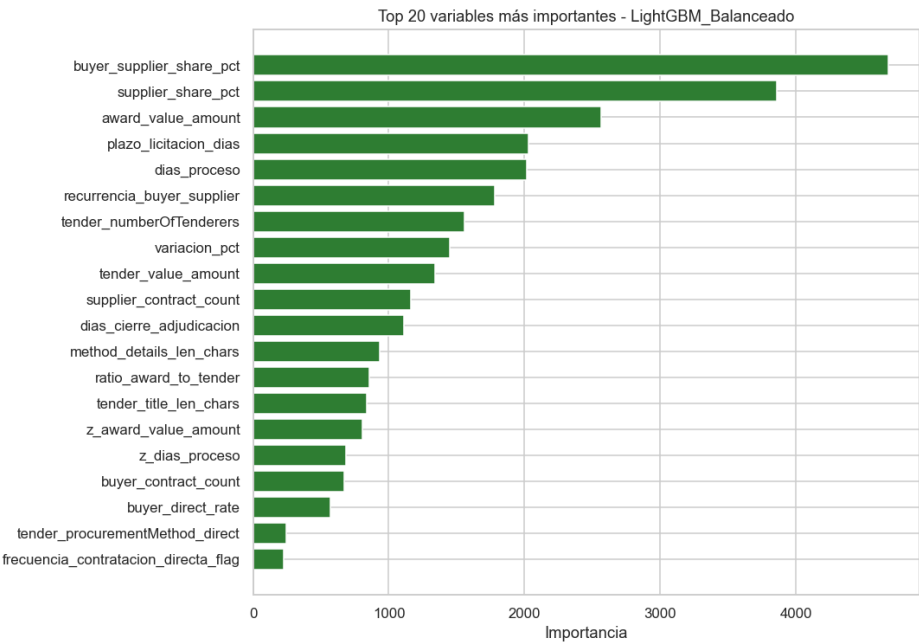
Este resultado sugiere que el nivel de riesgo no depende exclusivamente de que un contrato tenga un valor elevado. Las relaciones recurrentes entre compradores y proveedores, así como la concentración de contratos en determinadas empresas, aportaron más información al modelo que el monto analizado de manera aislada.

Entre las siguientes variables relevantes se encontraron el plazo disponible para presentar ofertas (`plazo_licitacion_dias`), la duración total del procedimiento (`dias_proceso`), la recurrencia contractual entre las partes (`recurrencia_buyer_supplier`) y el número de oferentes (`tender_numberOfTenderers`).

En conjunto, esta jerarquía respalda los planteamientos de Fazekas y de Open Contracting Partnership, quienes señalan que los riesgos en la contratación pública pueden manifestarse mediante patrones de concentración, baja competencia, relaciones repetitivas y plazos restrictivos. Por tanto, el análisis no debe enfocarse únicamente en contratos de gran valor, sino también en la estructura de las relaciones entre las entidades contratantes, los proveedores y las condiciones bajo las cuales se desarrollan los procedimientos.

Figura 17

Top 20 variables de mayor importancia predictiva del modelo LightGBM



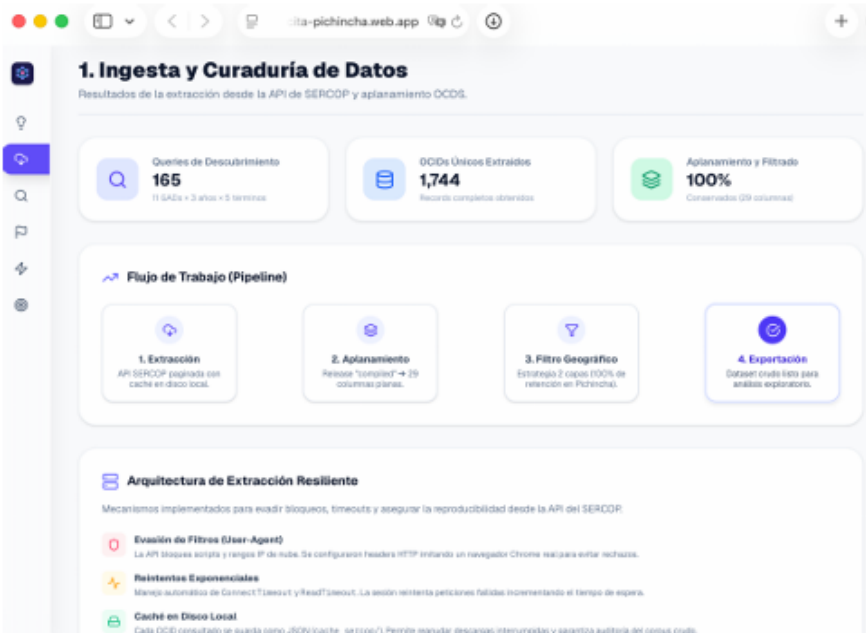
Nota. La variable que representa la concentración de adjudicaciones entre empresa y proveedor es claramente mayor que el valor del monto adjudicado.

Finalmente, los datos de mayor relevancia son presentados en el siguiente link:

<https://tesis-licita-pichincha.web.app>.

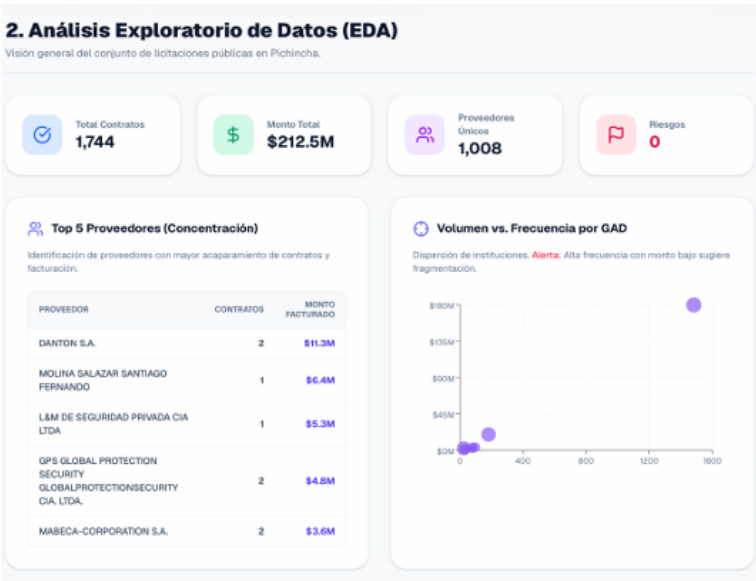
En esta página se detallan: La ingesta y curadoría de los datos, Análisis Exploratorio, Banderas, Modelado predictivo y finalizando con hallazgos.

Figura 18
Interfaz Ingesta y Curaduría de Datos



En la sección de análisis exploratorio Figura 19 se muestran los 5 proveedores que tienen mayor número de contratos asignados.

Figura 19
Análisis exploratorio de datos



5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

Como principal resultado, obtuvimos la respuesta a nuestra pregunta de investigación. Mediante técnicas de aprendizaje supervisado el análisis de los datos abiertos del nos permitió identificar y priorizar procesos de licitación municipal de los GAD de Pichincha, que presentan combinaciones atípicas de señales estadísticas de riesgo. Sin embargo, la clasificación generada no afirma la existencia de corrupción, simplemente nos da una sugerencia de que procesos deben ser auditados de manera prioritaria.

En la etapa de preparación de datos se demostró que la información publicada bajo el estándar OCDS (Open Contracting Data Estándar o en español Estándar de Datos para las Contrataciones Abiertas) puede utilizarse para construir modelos de clasificación, siempre que se realice el preprocesamiento de datos. El dataset analítico final utilizada para el modelado estuvo compuesta por 2.006 procesos, de los cuales 1.504 fueron empleados para el entrenamiento y 502 para la prueba independiente.

Mediante la validación cruzada estratificada repetida, LightGBM fue seleccionado como modelo final ya que presenta el mejor equilibrio general de acuerdo con el criterio definido en el estudio. Alcanzó una exactitud balanceada de 0,752, un F1 macro de 0,777, un recall de 0,677 y un F1 de 0,798 para la agrupación Alto+Crítico, además de un PR-AUC de 0,936. Las diferencias entre Random Forest y XGBoost fueron reducidas; estos 2 modelos superaron ligeramente a LightGBM en algunas métricas individuales. Sin embargo, la selección no se sustentó solo en superioridad absoluta, sino en el comportamiento conjunto de las métricas consideradas.

El estudio de ablación permitió corroborar que el desempeño se mantuvo al retirar las banderas utilizadas directamente para construir el índice de riesgo. Por esta razón, se seleccionó la configuración B_Sin_flags_directos, preparada por variables continuas, contextuales y relacionales. Posteriormente, la configuración de regularización R3_Moderado fue escogida

porque conservó un rendimiento similar y ayudó a controlar la complejidad del modelo. Estos resultados reducen la posibilidad de que LightGBM se limite a reproducir mecánicamente las reglas utilizadas para generar la variable objetivo.

Para la evaluación independiente, el modelo final alcanzó una exactitud balanceada de 0,807, un F1 macro de 0,831 y una precisión de 0,990 para las categorías Alto y Crítico. El recall de esta agrupación fue de 0,800, su F1 alcanzó 0,885 y el PR-AUC fue de 0,977. Los intervalos de confianza obtenidos mediante 1.000 réplicas bootstrap situaron el recall Alto + Crítico entre 0,736 y 0,864, el F1 entre 0,840 y 0,927 y el PR-AUC entre 0,963 y 0,989. Esto nos indica un comportamiento favorable, aunque también refleja que las métricas están sujetas a un nivel de incertidumbre que debe ser considerado al interpretar el modelo.

La validación temporal también respaldó la capacidad de generalización. Al entrenar con procesos de 2022 y 2023 y evaluar exclusivamente los registros de 2024, el modelo alcanzó una exactitud balanceada de 0,85, un F1 macro de 0,87, un recall Alto + Crítico de 0,77, un F1 de 0,86 y un PR-AUC de 0,97. Estos resultados son relevantes porque en 2024 el 60,42 % de los proveedores y el 49,27 % de las relaciones entre comprador y proveedor no habían sido observados durante el periodo de entrenamiento.

El análisis de importancia de variables demuestra que las características con mayor relevancia fueron la concentración entre una entidad y un proveedor (`buyer_supplier_share_pct`) y la participación global del proveedor (`supplier_share_pct`). El monto adjudicado se ubicó en tercer lugar, seguido de la recurrencia contractual, el número de oferentes, la variación presupuestaria y la duración del proceso.

En definitiva, el riesgo estadístico no se aplica solamente en el valor económico de una licitación, también se relaciona con la repetición de adjudicaciones entre proveedor-entidad, concentración del presupuesto de la entidad en determinados proveedores, poca participación de oferentes, variación en los presupuestos atípicos y las condiciones de cronograma de fechas del

proceso. El principal aporte del modelo es analizar estas señales para convertirlas en una herramienta de apoyo para la optimización de los recursos auditores hacia los procesos que necesitan atención prioritaria y detallada.

5.2. Recomendaciones

- Dada la carga de auditoría del gobierno, que actualmente depende en gran medida de revisiones aleatorias de documentos y consume importantes horas de trabajo, se recomienda que Contraloría General del Estado y otros órganos de auditoría adopten formalmente clasificadores de probabilidad como el desarrollado en este estudio. La implementación de un sistema impulsado por el modelo LightGBM capacitado permitiría a los equipos de auditoría priorizar la revisión de archivos por encima del umbral del riesgo relacional y optimizar el uso de los recursos limitados del país.
- Se recomienda la SERCOP fortalecer el control de calidad en el registro de sus interfaces de datos abiertos, especialmente en la asignación del identificador único OCID. Las duplicaciones encontradas en esta área no sólo afectan la presentación de informes básicos, sino que también limitan la capacidad de las organizaciones de la sociedad civil y de los analistas independientes para ejercer control social sobre el uso de los recursos públicos.
- Dado que el análisis de importancia de las variables encontró que la concentración empresa-proveedor y la presencia global de los proveedores eran los factores con mayor importancia predictiva, se recomienda que futuras direcciones de investigación exploren la representación de estos datos utilizando modelos gráficos. La introducción de Graph Neural Networks(GNN) permitiría un seguimiento más directo de los grupos de proveedores relacionados y las estructuras de

concentración, a las que el enfoque tabular actual sólo puede aproximarse indirectamente.

- Dado que la validación temporal mostró que el modelo mantiene desempeño favorable incluso cuando una fracción significativa de los proveedores y las relaciones comprador-proveedor no fueron observadas durante la capacitación, se recomienda establecer un esquema periódico de reentrenamiento del modelo esto preservaría su generalización en vista de la continua entrada de nuevos actores al sistema de contratación pública.
- Por último, el camino científico trazado en los resultados de este análisis prevé la inclusión de nuevos sectores tecnológicos en el futuro. Dado que la colusión y el favoritismo parecían ser fenómenos dominados por variables relacionales (quién hace tratos con quién), las direcciones de investigación futuras deberían migrar la información tabular a ecosistema modelo basados en grafos. La implementación de Redes Neuronales de Grafos (GNN) lograría rastrear clústeres ocultos de proveedores vinculados y monopolios estructurados, llevando la detección de anomalías a un estándar visual y estadístico revolucionario que la tabla tradicional jamás podrá alcanzar.

6. REFERENCIAS BIBLIOGRÁFICAS

- Adelanwa, A., Basnet, A., y Anene, U. N. (2023). Predictive analytics models for financial risk detection and fraud prevention in public systems. *International Journal of Advanced Multidisciplinary Research and Studies*, 3(6), 2663–2679.
<https://doi.org/10.62225/2583049X.2023.3.6.5969>
- Aldana, A., Falcón-Cortés, A., y Larralde, H. (2022). *A machine learning model to identify corruption in México's public procurement contracts*. arXiv.
<https://doi.org/10.48550/arXiv.2211.01478>
- Álvarez Suárez, D. E., y Nieto Morales, F. (2025). Detectando patrones de riesgo de corrupción: Análisis de contratos estatales y municipales en México mediante agrupamiento jerárquico. *Gestión y Análisis de Políticas Públicas*, (39), 21–43. <https://doi.org/10.24965/gapp.11518>
- American Psychological Association. (s. f.). Standard deviation. En *APA dictionary of psychology*.
<https://dictionary.apa.org/standard-deviation>
- Asamblea Nacional del Ecuador. (2002). *Ley Orgánica de la Contraloría General del Estado* (Registro Oficial Suplemento 595, 12 de junio de 2002; codificación actualizada). Contraloría General del Estado. <https://www.contraloria.gob.ec/>
- Asamblea Nacional del Ecuador. (2008). *Ley Orgánica del Sistema Nacional de Contratación Pública (LOSNCP)* (Registro Oficial Suplemento 395, 4 de agosto de 2008; codificación actualizada). Servicio Nacional de Contratación Pública.
https://portal.compraspublicas.gob.ec/sercop/wp-content/uploads/2021/04/losncp_actualizada1702.pdf
- Ávila Ordóñez, D. (2017). *La colusión en la contratación pública en el Ecuador* [Tesis de maestría, Universidad Andina Simón Bolívar – Sede Ecuador]. Repositorio Institucional

- UASB. <https://repositorio.uasb.edu.ec/bitstream/10644/5561/1/T2247-MDE-Avila-La%20colusi%C3%B3n.pdf>
- Banco Mundial. (2025). *Procurement regulations for IPF borrowers* (7.^a ed.). World Bank Group. <https://thedocs.worldbank.org/>
- Bobadilla, J. (2020). *Machine learning y deep learning: Usando Python, Scikit y Keras* (1.^a ed.). RA-MA Editorial.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., y Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.
- Chen, T., y Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- Datalat. (s. f.). *Investigación: Aporte desde las AI para el sistema de compras públicas del Ecuador con enfoque de género* [Proyecto]. <https://datalat.org/proyectos/>
- Dávila Castillo, M. R., Coka Flores, D. F., y Álvarez Tapia, M. E. (2021). Transparencia y control social en la contratación pública en Ecuador. *Universidad y Sociedad*, 13(S2), 385–391. <https://rus.ucf.edu.cu/index.php/rus/article/view/2324>
- Decarolis, F., y Giorgiantonio, C. (2020). *Corruption red flags in public procurement: New evidence from Italian calls for tenders* (Questioni di Economia e Finanza N.º 544). Bank of Italy.

- Dobre, C. (2025). *CRISP-DM: Los 6 pasos del proceso de data mining*. Smartup Digital.
- Fazekas, M., Tóth, I. J., y King, L. P. (2016). An objective corruption risk index using public procurement data. *European Journal on Criminal Policy and Research*, 22(3), 369–397.
- Fortuny, M., Guerrero, E., Riofrío, D., y Simon, F. (2023). Towards smart citizen control in public procurement: Ecuador's case study. En *2023 9th International Conference on eDemocracy and eGovernment (ICEDEG 2023)*. IEEE.
<https://doi.org/10.1109/ICEDEG58167.2023.10121991>
- Gallego, J., Rivero, G., y Martínez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting*, 37(1), 360–377. <https://doi.org/10.1016/j.ijforecast.2020.06.006>
- García, S., Luengo, J., y Herrera, F. (2015). *Data preprocessing in data mining*. Springer.
- García Torres, F. A., Pineda Arévalo, D. F., Hidalgo-Moreno, K., y Espinosa-Rodriguez, C. T. (2022). Los procesos de contratación pública en el Ecuador y su incidencia en los presupuestos del Sector Público durante los años 2020 y 2021. *Polo del Conocimiento*, 7(12), 692–709. <https://dialnet.unirioja.es/descarga/articulo/9227638.pdf>
- Gnaldi, M., y Del Sarto, S. (2023). *Validating corruption risk measures: A key step to monitoring SDG progress*. arXiv. <https://doi.org/10.48550/arXiv.2309.01462>
- Gnaldi, M., y Del Sarto, S. (2024). Measuring corruption risk in public procurement over emergency periods. *Social Indicators Research*, 172(3), 859–877.
<https://doi.org/10.1007/s11205-024-03331-w>
- Hirschman, A. O. (1964). The paternity of an index. *The American Economic Review*, 54(5), 761–762.

- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., y Liu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
- Longobucco, A., y Ferwerda, J. (2025). *Paper on the effectiveness of AI tools to detect corruption: A case study on Danish public procurements* (BridgeGap Deliverable 6.2). Utrecht University; European Commission Horizon Europe.
- Lyra, M. S., Damásio, B., Pinheiro, F. L., y Bacao, F. (2022). Fraud, corruption, and collusion in public procurement activities: A systematic literature review on data-driven methods. *Applied Network Science*, 7, Artículo 83. <https://doi.org/10.1007/s41109-022-00523-6>
- Majilana, R. (2025). Investigating unsupervised machine learning in public procurement fraud detection: A case study of POTRAZ using the Isolation Forest algorithm. *International Journal of Computer Science and Mobile Computing*, 14(7), 134–136. <https://doi.org/10.47760/ijcsmc.2025.v14i07.013>
- Medina-Hernández, M., Kertész, J., y Fazekas, M. (2025). *Learning from sanctioned government suppliers: A machine learning and network science approach to detecting fraud and corruption in Mexico*. arXiv. <https://doi.org/10.48550/arXiv.2512.19491>
- Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. *WIREs Computational Molecular Science*, 4, 468–481. <https://doi.org/10.1002/wcms.1183>
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Ndolo, M., y Kasyoka, P. (2025). A hybrid machine learning model for detecting and preventing corruption in Kenya's public procurement contracts. *Machine Learning Research*, 10(2), 45–58.

- NIST/SEMATECH. (2012). *e-Handbook of statistical methods*. National Institute of Standards and Technology. <https://www.itl.nist.gov/div898/handbook/>
- OECD. (2021). *OECD integrity review of the State of Mexico: Strengthening public sector integrity* (OECD Public Governance Reviews). OECD Publishing.
https://www.oecd.org/es/publications/2021/07/oecd-integrity-review-of-the-state-of-mexico_fc889b22.html
- Open Contracting Partnership. (2016, 11 de noviembre). *Red flags for integrity in public procurement*. <https://www.open-contracting.org/2016/11/11/red-flags-for-integrity-in-public-procurement/>
- Open Contracting Partnership. (2021). *Red flags for integrity: Giving the green-light to open data solutions* (OCP Technical Series).
- Open Contracting Partnership. (2022a, 10 de noviembre). *Democratizing public procurement in Ecuador*. <https://www.open-contracting.org/2022/11/10/democratizing-public-procurement-in-ecuador/>
- Open Contracting Partnership. (2022b). *Red flags and data analytics: Guidelines for identifying integrity risks in public procurement*. <https://www.open-contracting.org/resources/red-flags-data-analytics/>
- Open Contracting Partnership. (2024a, 12 de junio). *Cardinal: An open-source library to calculate public procurement red flags*. <https://www.open-contracting.org/2024/06/12/cardinal-an-open-source-library-to-calculate-public-procurement-red-flags/>
- Open Contracting Partnership. (2024b). *Red flags in public procurement: A guide to using data to detect and mitigate risks*.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pyle, D. (1999). *Data preparation for data mining*. Morgan Kaufmann.
- Rodríguez Arévalo, S. (2020). *Predicción de ineficiencias en la contratación pública de Bogotá* [Tesis de maestría, Universidad del Rosario].
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Salazar, A., Pérez, J. F., y Gallego, J. (2024). VigIA: Prioritizing public procurement oversight with machine learning models and risk indices. *Data & Policy*, 6, e75.
<https://doi.org/10.1017/dap.2024.83>
- Schneider dos Santos, E., Castro, M., y Carvalho, J. T. (2025). Improving public procurement collusion detection with graph-based machine learning methodologies. En *Anais da I Escola Regional de Aprendizado de Máquina e Inteligência Artificial da Região Sul (ERAMIA-RS 2025)*. Sociedade Brasileira de Computação.
<https://doi.org/10.5753/eramars.2025.16657>
- Schneider dos Santos, E., Machado dos Santos, M., Castro, M., y Carvalho, J. T. (2025). Detection of fraud in public procurement using data-driven methods: A systematic mapping study. *EPJ Data Science*, 14, Artículo 52. <https://doi.org/10.1140/epjds/s13688-025-00569-3>
- Servicio Nacional de Contratación Pública (SERCOP). (s. f.). *Normativa del Sistema Nacional de Contratación Pública* [Portal institucional].
<https://portal.compraspublicas.gob.ec/sercop/normativa/>

Servicio Nacional de Contratación Pública (SERCOP). (2017). *Manual de buenas prácticas en la contratación pública para el desarrollo del Ecuador*. Organización de los Estados Americanos.

https://www.oas.org/juridico/PDFs/mesicic5_ecu_panel5_SERCOP_3.2.1_man_bue_pr%C3%A1c_CP.pdf

Servicio Nacional de Contratación Pública (SERCOP). (2023a). *Normativa Secundaria del Sistema Nacional de Contratación Pública* (Resolución N.º R.E-SERCOP-2023-0134 de 1 de agosto de 2023, con reformas aplicadas hasta 2025).

<https://portal.compraspublicas.gob.ec/sercop/anexos/>

Servicio Nacional de Contratación Pública (SERCOP). (2023b). *Resolución N.º R.E-SERCOP-2023-0133: Estrategia Nacional de Compras Públicas con Enfoque de Género*.

Servicio Nacional de Contratación Pública (SERCOP). (2024). *Portal de Datos Abiertos*.

<https://datosabiertos.compraspublicas.gob.ec>

Superintendencia de Control del Poder de Mercado del Ecuador. (2019). *Análisis de concentraciones económicas: La experiencia de Ecuador*.

<https://www.sce.gob.ec/sitio/wp-content/uploads/downloads/2019/01/Libro-Analisis-de-concentraciones-econo%CC%81micas.pdf>

Torres Berrú, Y. (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.

<https://gredos.usal.es/handle/10366/157534>

- Torres-Berru, Y., y López Batista, V. F. (2021). Data mining to identify anomalies in public procurement rating parameters. *Electronics*, 10(22), Artículo 2873.
<https://doi.org/10.3390/electronics10222873>
- Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.
- Umbarila Suárez, M. F. (2024). Predicción del riesgo de corrupción en los procesos de contratación pública disponibles en la plataforma Datos Abiertos usando el modelo Gradient Boosting Machine. *Ingenio Libre*, 14(24). <https://doi.org/10.18041/2322-8415/ingelibre.2024.v14n24.12087>
- Universidad Nacional de San Martín. (2023). *Análisis de las compras a través del Servicio Nacional de Contratación Pública (SERCOP) en el periodo 2018-2019* [Tesis de maestría, Escuela de Política y Gobierno, UNSAM]. Repositorio Institucional UNSAM.
http://ri.unsam.edu.ar/bitstream/123456789/2281/1/TMAG_EPYG_2023_SSP.pdf
- Vaqueiro, R., Vargas, A., Escovedo, T., y Kalinowski, M. (2023). Machine learning applied to open government data for the detection of improprieties in the application of public resources. En *Proceedings of the XIX Brazilian Symposium on Information Systems (SBSI 2023)* (pp. 213–220). Association for Computing Machinery.
<https://doi.org/10.1145/3592813.3592908>
- Velasco, R. B., Carpanese, I., Interian, R., Paulo Neto, O. C. G., y Ribeiro, C. C. (2021). A decision support system for fraud detection in public procurement. *International Transactions in Operational Research*, 28(1), 27–47. <https://doi.org/10.1111/itor.12811>
- Villa Pedroza, F. I. (2021). *Métodos de aprendizaje automático para clasificación de riesgo de corrupción por adición en contratación de obras públicas mediante licitación: Caso de*

estudio en Colombia [Tesis de maestría, Universidad Internacional de Andalucía y Universidad de Huelva].

Wachs, J., Fazekas, M., y Kertész, J. (2019). *Corruption risk in contracting markets: A network science perspective*. arXiv. <https://doi.org/10.48550/arXiv.1909.08664>

Wirth, R., y Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. En *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29–39).

World Bank. (2023). *Public procurement analytics: Statistical indicators for market concentration, collusion, and corruption risks*. World Bank Group.