https://doi.org/10.69639/arandu.v12i4.1738

Modelado predictivo de riesgos de ineficiencia y transparencia en la contratación pública ecuatoriana, basado en el análisis de patrones de datos clasificados del SERCOP

Predictive modeling of inefficiency and transparency risks in Ecuadorian public procurement, based on the analysis of classified data patterns from SERCOP

Freddy Daniel Carrillo Bustos
freddy.daniel.carrillo.bustos@gmail.com
https://orcid.org/0009-0004-7868-5289
Universidad Iberoamericana del Ecuador
Ecuador – Ambato

Yasmany Fernández Fernández
yfernandez@doc.unibe.edu.ec
https://orcid.org/0000-0002-9530-4028
Universidad Iberoamericana del Ecuador
Ecuador - Quito

Artículo recibido: 18 septiembre 2025 -Aceptado para publicación: 28 octubre 2025
Conflictos de intereses: Ninguno que declarar.

RESUMEN

El presente estudio evalúa el potencial analítico de los datos abiertos publicados por el Servicio Nacional de Contratación Pública (SERCOP) de Ecuador en el periodo comprendido entre 2015 y 2025. El estudio planteado integró los dataset proporcionados por la plataforma oficial, aplicando procesos de análisis exploratorio de datos se depuraron los datos para realizar el proceso investigativo. En base a la información obtenida se aplican técnicas de Big Data, con las cuales se obtuvo patrones mediante técnicas de agrupación no supervisada; por lo que se obtienen diversas correlaciones entre variables como: presupuestos, tiempo y valoración descriptiva. Al realizar el análisis de las correlaciones obtenidas, se determina que la calidad de los datos abiertos de compras públicas en Ecuador es deficiente, por lo que se dificulta el cálculo de factores de riesgo asociados a cada proceso.

Palabras clave: contratación pública, datos abiertos, big data, calidad de datos, Ecuador

ABSTRACT

This study assesses the analytical potential of the open-data sets published by Ecuador's National Public Procurement Service (SERCOP) for the 2015–2025 period. The proposed study integrated the datasets provided by the official platform, applying exploratory data analysis processes and purifying the data to conduct the research. Based on the information obtained, Big Data

techniques were applied, obtaining patterns using unsupervised clustering techniques, yielding various correlations between variables such as budgets, time, and descriptive assessment.

Keywords: public procurement, open data, big data, data quality, Ecuador

INTRODUCCIÓN

La contratación pública, un pilar fundamental de la economía ecuatoriana, representa una porción significativa del presupuesto estatal, lo que subraya la imperante necesidad de mecanismos que garanticen su transparencia y eficiencia (Joković, 2022). En este contexto, el Servicio Nacional de Contratación Pública (SERCOP) inició la publicación de datos abiertos de cada proceso contractual a partir de 2015. El propósito de esta iniciativa era fomentar la rendición de cuentas, permitir el monitoreo ciudadano y facilitar análisis que contribuyeran a optimizar las compras gubernamentales (Benavides et al., 2016), (García et al., 2022).

Sin embargo, a pesar de la disponibilidad de estos datos, la literatura internacional y las guías de organismos como la Red Interamericana de Compras Gubernamentales (RICG) enfatizan que la verdadera utilidad de dichos portales depende críticamente de la calidad, completitud y estandarización de la información publicada. En Ecuador, esta capacidad de aprovechamiento de los datos abiertos del SERCOP aún no se ha verificado de forma sistemática (Red Interamericana de Compras Gubernamentales, 2021).

La realidad ha demostrado que la calidad de estos datos es deficiente, con inconsistencias, campos incompletos y falta de estandarización que limitan severamente su potencial analítico. Esta deficiencia no solo obstaculiza la detección de sobreprecios o la medición de la competitividad, también impide la generación de alertas tempranas de riesgo, elementos cruciales para una gestión pública eficaz y transparente.

El presente estudio aborda directamente esta problemática al realizar un análisis exhaustivo desde la perspectiva de la calidad y la estructura de los datos abiertos del SERCOP, con el fin de establecer variables de riesgo en los diferentes procesos de compras públicas. Al identificar y comprender los patrones inherentes en estos datos, a pesar de sus deficiencias, se pretende contribuir al fomento de la transparencia en los servicios y productos ofertados a través del SERCOP, y sentar las bases para el desarrollo de herramientas predictivas que mitiguen ineficiencias y riesgos de corrupción.

MATERIALES Y MÉTODOS

El presente estudio adopta un enfoque cuantitativo, descriptivo-exploratorio, con énfasis particular en la ingeniería de datos y la analítica de Big Data. La metodología se fundamenta en el marco de trabajo CRISP-DM (Cross-Industry Standard Process for Data Mining), un modelo de procesos estandarizado y robusto para proyectos de minería de datos. Este enfoque se seleccionó por su capacidad para guiar sistemáticamente el descubrimiento de patrones y la extracción de conocimiento significativo en conjuntos de datos complejos, como los de contratación pública. La elección de CRISP-DM, junto con técnicas de reducción de dimensionalidad y agrupamiento no supervisado, se justifica por la naturaleza exploratoria del estudio y la ausencia de etiquetas predefinidas de riesgo en los datos brutos del SERCOP, lo que

hace inviable un enfoque supervisado tradicional en las fases iniciales. (González, 2014); (Joković, 2022).

Fuente y Alcance de los Datos

Durante el proceso investigativo se seleccionó como fuente de datos principal el repositorio oficial "Datos Abiertos de contratación pública del Ecuador en OCDS" del Servicio Nacional de Contratación Pública (SERCOP). Este portal, que recopila datos desde enero de 2015 hasta la fecha del presente estudio (febrero de 2025), constituye una fuente oficial de los procesos contractuales del estado ecuatoriano.

Para asegurar la pertinencia y el alcance del análisis, se seleccionaron datos específicos de las siguientes etapas clave de los procesos de contratación, disponibles en hojas separadas dentro de los datasets originales: "Licitación (tender)", "Premiaciones (Awards)", "Proveedores (AwardSuppliers)" y "Contratos (Contracts)". La unificación de estos registros, utilizando el identificador único OCID (Open Contracting Identifier), resultó en un dataset consolidado que comprende un total de 2,467,699 filas, abarcando así un vasto universo de transacciones.

El universo de datos seleccionado incluye todos los tipos de contratación pública contemplados en la normativa ecuatoriana, lo que permite una visión integral de las prácticas de adquisición. Esto incluye procedimientos como: subasta inversa electrónica (para bienes y servicios normalizados), licitación (para obras o servicios complejos), contratación directa (en casos excepcionales), menor cuantía (para montos reducidos), cotización (para comparar ofertas específicas), concurso público (para servicios intelectuales) y régimen especial (para sectores estratégicos o situaciones específicas). La inclusión de esta diversidad de procedimientos es fundamental para identificar patrones de riesgo transversales y específicos a cada modalidad (Molina, 2023).

Herramientas y Entorno

El lenguaje de programación empleado es Python 3, con las siguientes librerías: pandas, numpy, scikit-learn, nltk, seaborn, matplotlib. El entorno de programación utilizado fue la plataforma de Google Colab Pro, que brinda acceso a procesadores de alta potencia y altas capacidades RAM. Esta elección se basó en la necesidad de procesar eficientemente grandes volúmenes de datos y ejecutar algoritmos computacionalmente intensivos.(González, 2014)

Procedimiento

La etapa de preprocesamiento fue crucial para transformar los datos crudos del SERCOP en un formato adecuado para el análisis, abordando los desafíos de calidad y heterogeneidad inherentes a los datos abiertos.

Pre-procesado de la información

- Se unificó en un solo dataset las hojas “Tender, Award, AwardSuppliers, Contracts” mediante el identificador único OCID para generar un solo dataset.

- Se realiza un EDA y una limpieza general de la información para eliminar filas irrelevantes para efectos de este estudio.
- Se consideraron solamente los procesos de contratación que no pertenezcan a los siguientes grupos:
    - Situaciones de emergencia, estos procesos priorizan la compra directa del bien o servicio requerido de manera inmediata sin concursos.
    - Catálogo electrónico, no requiere concurso y el requerimiento se hace directamente con el vendedor ya que este se encuentra debidamente registrado con su producto por lo que para efectos de este estudio es irrelevante.
    - Regímenes especiales, dada la vinculación de proyectos pasados con el mismo contratante esto le obliga a contratar directamente al mismo proveedor mediante una convocatoria directa lo cual aunque se considera un concurso público no es sujeto de análisis al tener únicamente un solo ganador predefinido.
    - Procesos de estudio de mercado e ínfimas cuantías, aunque son procesos registrados, estos solamente receptan proformas por lo cual aunque existen participantes no se escogen ganadores ni registros adicionales.
- Las variables de interés para este estudio son: Fechas (firma de contrato, inicio, adjudicación, preguntas), descripción del proceso, cantidad participantes, monto del proceso, estado; se descartaron registros que no contengan datos válidos en estos campos.
- El filtro aplicado reduce el dataset a un total de 224109 registros, lo que significa una reducción de un 90%.
- De manera complementaria para este estudio se generaron las siguientes variables calculadas, a partir de las fechas establecidas, con el fin de obtener valor numérico que cuantifique la cantidad de días:
    - Días de duración del proceso (fecha de final de proceso - fecha de inicio de proceso)
    - Días hasta la firma del contrato (fecha de firma de contrato - fecha de adjudicación)

Procesamiento de lenguaje natural sobre la descripción del proceso

La variable textual “descripción” es un párrafo corto que explica de manera breve cuál es el producto o servicio requerido. En esta etapa, se obtuvo un valor numérico basado en palabras clave para cuantificar la riqueza descriptiva. A cada palabra clave se le asignó un peso, lo que permitió calcular la sumatoria de valoración total de la efectividad de la descripción. El proceso se detalló en los siguientes pasos:

Limpieza

- Normalización del texto descriptivo (conversión a minúsculas, eliminación de caracteres especiales).

- Eliminación de referencias a la entidad contratante (para evitar sesgos descriptivos).
- Eliminación de *stopwords* (adverbios, preposiciones, conjunciones, determinantes) y monosílabos menores a 2 letras, con el fin de reducir el ruido y enfocarse en términos con mayor significado.
- Tokenización de las palabras de descripción para permitir un análisis individualizado de cada término.

Variables calculadas

Para enriquecer el conjunto de datos con información relevante para el clustering, se calcularon las siguientes variables adicionales:

- Cálculo de la frecuencia de aparición de cada palabra.
- Sumatoria de duración de proceso en días.
- Sumatoria categorías por tipo.
- Variable reescalada *amount* (presupuesto) mediante logaritmo, para manejar la asimetría de su distribución y mejorar la estabilidad de los modelos.

Cálculo de cálculo de TF-IDF

Se calculó el valor relativo de cada palabra en función de su frecuencia (TF-IDF), presupuesto y duración. Este enfoque permitió asignar mayor importancia a palabras que son distintivas de procesos con características específicas (alto monto o duración).

- Cálculo de valor relativo de la palabra en función de su frecuencia, presupuesto y duración.
- Eliminación del percentil inferior 1 de 10, ya que representa grupos minoritarios de palabras con poco valor y frecuencia de aparición. El rango de frecuencias resultante fue de 116 hasta 160,137 con un tamaño de diccionario de 2,792 palabras, lo que optimiza la relevancia del vocabulario.
- La fórmula utilizada para el cálculo del valor de palabra en relación al presupuesto y duración del proceso:

Valor = Log(monto medio normalizado) * (duración media normalizada)

Fórmula 1: Valoración relativa de la palabra en el contexto de la descripción.

Cálculo de valoración de la descripción en base al diccionario de valores de palabras

Utilizando el diccionario calculado en el paso anterior, se realizó la sumatoria del valor correspondiente a cada palabra presente en la descripción, obteniendo así un valor total de la descripción de un proceso. Este valor numérico actúa como un indicador de la calidad y especificidad de la descripción.

Reducción de dimensionalidad

La reducción de dimensionalidad es una fase esencial para manejar la complejidad de los datos del SERCOP, que, a pesar de su volumen, presentan redundancia y variables altamente correlacionadas. Esta etapa busca transformar el conjunto de variables originales en un espacio de menor dimensión, preservando la información más relevante para el análisis.

- Se aplicó el Análisis de Componentes Principales (PCA) con dos componentes principales. La elección de PCA se justifica por su eficacia en la reducción de variables numéricas correlacionadas, permitiendo identificar las dimensiones subyacentes que explican la mayor parte de la varianza en los datos de contratación (monto, duración, número de participantes, etc.). Las dos componentes principales fueron suficientes para capturar una proporción significativa de la varianza total, facilitando la visualización y la interpretación de los clústeres en un espacio bidimensional.
- Se realizó un Análisis Factorial (FA) para extraer factores latentes comunes entre las variables. A diferencia de PCA, que se enfoca en la varianza, FA busca identificar la estructura subyacente de las correlaciones entre las variables observadas. Esta técnica es particularmente útil para descubrir constructos teóricos (factores) que no son directamente observables, pero que influyen en las variables de interés (ej., un factor de "Complejidad del Proceso" que agrupe alta duración y bajo número de participantes). La aplicación de FA permitió una interpretación más profunda de las relaciones entre las variables, complementando el análisis de PCA.

Agrupamiento no supervisado

El agrupamiento no supervisado fue el núcleo del análisis de patrones, dada la ausencia de etiquetas de riesgo predefinidas en los datos del SERCOP. Se aplicaron múltiples algoritmos para explorar diferentes estructuras de agrupamiento y asegurar la robustez de los hallazgos.

- **K-Means:** Se utilizó para agrupar las observaciones en diez clústeres, basándose en la distancia euclidiana. La determinación de k=10 se realizó mediante el método del codo (Elbow method), que identifica el punto de inflexión donde añadir más clústeres no produce una mejora significativa en la varianza explicada. Este algoritmo fue aplicado posteriormente a los componentes obtenidos de PCA y FA para mejorar la diferenciación de los clústeres al trabajar en un espacio de menor dimensionalidad.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Se empleó para detectar clústeres de densidad variable y aislar outliers sin requerir un número predefinido de clústeres. DBSCAN es robusto ante la presencia de ruido y permite identificar grupos de forma irregular, lo cual es ventajoso en datos de contratación pública donde pueden existir patrones atípicos que no se ajustan a formas esféricas (como las asumidas por K-Means). Los parámetros *eps* (radio de la vecindad) y *min_samples* (número mínimo de puntos en una vecindad) se ajustaron iterativamente para optimizar la identificación de clústeres significativos.
- **GMM (Gaussian Mixture Models):** Este método se utilizó para estimar clústeres a partir de mezclas de distribuciones gaussianas, permitiendo el solapamiento entre grupos. GMM es más flexible que K-Means, ya que asume que los datos provienen de una combinación de distribuciones gaussianas con diferentes medias y covarianzas. Esta flexibilidad permite

identificar clústeres de formas elípticas y tamaños variados, lo cual es relevante si los grupos de procesos de contratación tienen distribuciones no esféricas.

- **Clustering Jerárquico:** Se construyó un dendrograma de relaciones entre observaciones, utilizando un método aglomerativo (ej., Ward's method) sobre una muestra representativa de los datos. Esta técnica proporciona una visualización de la estructura jerárquica de los clústeres, lo que ayuda a entender cómo se agrupan los procesos en diferentes niveles de similitud y a validar el número óptimo de clústeres de manera exploratoria.

Validación de Clústeres mediante Comparación con Métodos de Clasificación

El enfoque principal se basó en técnicas de clustering no supervisado para explorar los datos de contratación pública del SERCOP. Este método fue elegido estratégicamente por varias razones fundamentales que aprovechan la naturaleza de nuestros datos:

- **Descubrimiento de Patrones Ocultos:** A diferencia de los métodos supervisados que requieren etiquetas predefinidas (ej., clasificar proyectos como "riesgosos" o "normales" de antemano), el clustering no supervisado destaca por descubrir agrupaciones inherentes dentro de los datos. En este proyecto, no partimos de un conjunto de datos previamente etiquetado en cuanto a tipos de proyectos o niveles de riesgo basados en características numéricas. El clustering nos permitió revelar patrones distintos en las características de los proyectos (como monto, duración, número de oferentes y valor descriptivo) que podrían no haber sido inmediatamente obvios.
- **Identificación de Diversos Perfiles de Proyectos:** Los análisis de K-Means y GMM, particularmente cuando se aplicaron después de técnicas de reducción de dimensionalidad como PCA o Análisis Factorial, nos ayudaron a identificar clústeres que representan diferentes tipos de procesos de contratación. Observamos grupos que iban desde procesos 'típicos y competitivos' hasta procesos 'extraordinarios o críticos' con montos muy altos y baja competencia. Esta visión granular de los perfiles de proyectos es un resultado directo de los algoritmos de clustering al encontrar límites naturales en los datos.
- **Manejo de Relaciones Complejas:** Las relaciones entre las variables numéricas pueden ser complejas y no lineales. Los métodos de clustering, especialmente cuando se combinan con técnicas como el Análisis Factorial que identifica factores latentes subyacentes, pueden capturar estas interacciones complejas y agrupar proyectos basándose en su perfil general a través de estas dimensiones.
- **Identificación de Potenciales Anomalías y Valores Atípicos:** Como se observó en el análisis de nuestros clústeres, algunos grupos destacaron debido a valores extremos en ciertas variables (por ejemplo, tiempos de firma de contrato o duraciones de proceso muy largos, montos muy altos con baja competencia). El clustering no supervisado es una herramienta poderosa para señalar tales potenciales anomalías o valores atípicos que justifican una investigación adicional en busca de posibles ineficiencias o riesgos.

- **Base para Análisis Posteriores:** Los clústeres identificados a través de estos métodos pueden servir como base para análisis posteriores. Las etiquetas de clúster generadas pueden ser utilizadas como una nueva variable categórica para explorar relaciones con otras características no numéricas en el dataset, o incluso como una variable objetivo para el aprendizaje supervisado si se desea construir un modelo para predecir a qué clúster podría pertenecer un nuevo proyecto.
- Para la validación interna de los clústeres, se utilizó el **Silhouette Score** para evaluar la cohesión interna y la separación entre los grupos. Además, se empleó una **tabla de contingencia** para comparar la consistencia de las asignaciones entre los diferentes métodos de agrupamiento no supervisado aplicados
- (K-Means vs GMM). Finalmente, se realizó una **inspección exploratoria** de los resultados a través de visualizaciones de *scatter plots* y dendrogramas para una comprensión visual de la estructura de los clústeres.

RESULTADOS Y DISCUSIÓN

La presente sección detalla los hallazgos derivados del análisis de los datos de contratación pública del SERCOP, aplicando las metodologías descritas en la sección anterior. Los resultados se estructuran para ofrecer una comprensión clara del estado inicial de los datos, la correlación entre variables clave, y las agrupaciones identificadas mediante técnicas de aprendizaje no supervisado.

Estado Inicial de los Datos

Tras el proceso de pre-procesamiento y limpieza de la información, se designaron las variables de interés para el estudio. La distribución de estas variables es fundamental para entender la naturaleza del conjunto de datos y se visualiza en la **Figura 1**.

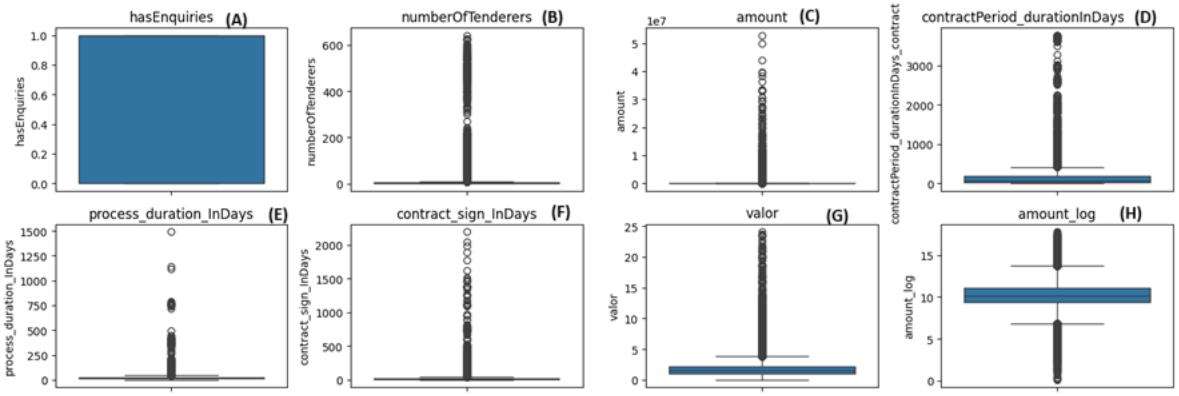
A continuación, se describen las características observadas en cada variable:

1. **Preguntas**: Variable de tipo booleana, que no muestra valores atípicos significativos, indicando una consistencia en su registro.
2. **Cantidad de postulantes**: No presenta valores extremos y se observan dos grupos diferenciados en su distribución.
3. **Presupuesto**: Presenta una distribución orientada hacia valores bajos, lo que sugiere que la mayoría de los procesos de contratación son de montos reducidos.
4. **Duración del contrato**: No se identificaron valores extremos, mostrando una distribución relativamente homogénea.
5. **Duración del proceso en días**: Aunque en el proceso inicial se descartaron algunos valores extremos, la distribución resultante sigue siendo relevante para el análisis.
6. **Firma del contrato en días**: La distribución se orienta a valores bajos, indicando que la mayoría de los contratos se firman en un periodo corto tras la adjudicación.

7. **Valoración de descripción de proceso**: Presenta una distribución homogénea, lo que sugiere una consistencia en la valoración de las descripciones.
8. **Logaritmo de presupuesto**: Tras la transformación logarítmica para manejar la asimetría, se distinguen dos grupos diferenciados, facilitando su interpretación en el modelado.

Figura 1

Diagrama de distribución por variable



Para determinar el grado de correlación lineal entre estas variables de interés, se aplicó la correlación de Pearson. Los resultados se muestran en la **Tabla 1**.

Tabla 1

Correlación de Pearson aplicada a las variables fundamentales

	Hubo Preguntas (bool)	**Duración del contrato días (int)**	**Duración del proceso en días (int)**	**Presupuesto (float)**
Hubo Preguntas (bool)		0.03	0.32	0.23
Duración del contrato días (int)	0.03		0.09	0.21
Duración del proceso en días (int)	0.32	0.09		0.27
Presupuesto(float)	0.23	0.21	0.27	

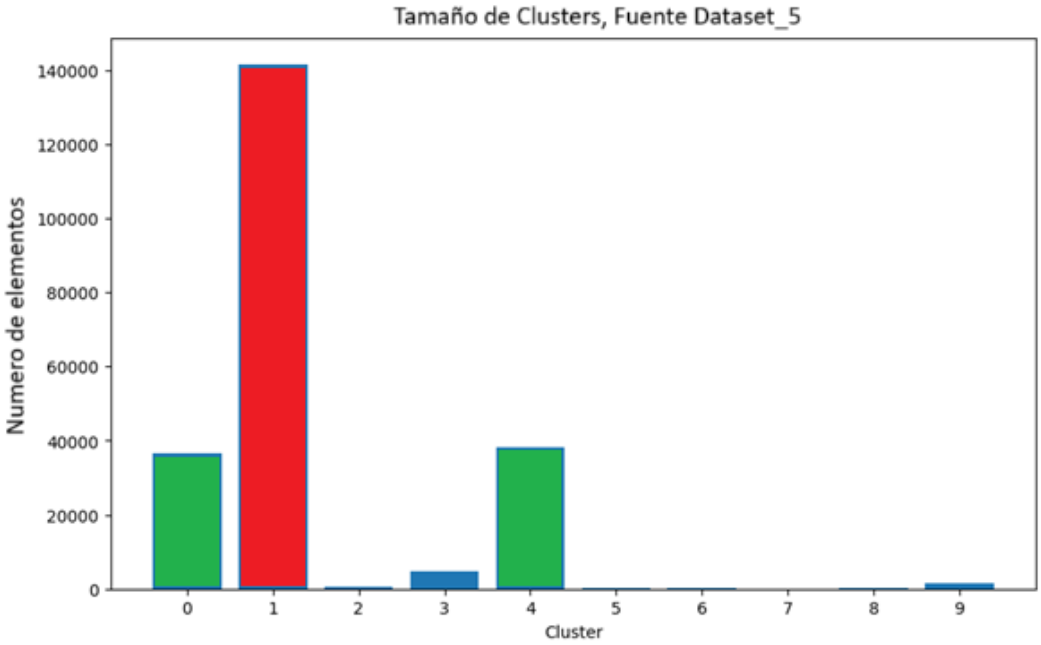
Como se observa en la **Tabla 1**, el nivel de correlación entre las variables es generalmente bajo. Esta baja correlación justifica la necesidad de profundizar en análisis más elaborados, como las técnicas de reducción de dimensionalidad y agrupamiento no supervisado, para descubrir patrones y relaciones no lineales que no son evidentes a través de una simple correlación bivariada.

Agrupamiento (clusters) por Kmeans

Para la segmentación de los datos, se aplicó el algoritmo K-Means, una técnica de agrupamiento no supervisado ampliamente utilizada en análisis exploratorio. El número óptimo de clústeres, determinado mediante el método del codo (Elbow method), indicó la necesidad de establecer 10 clústeres.

Figura 2

Distribución de clusters



La **Figura 2** muestra una clara desigualdad en el tamaño de los clústeres, destacándose el clúster 1 con más de 140,000 elementos. A pesar de que los demás clústeres son significativamente menores en tamaño, su relevancia radica en los patrones específicos que agrupan.2.1. Visualización de Variables en Agrupaciones

A continuación, se presenta la representación de las variables clave por clúster, con el objetivo de caracterizar los grupos identificados.

Visualización de variables en agrupaciones

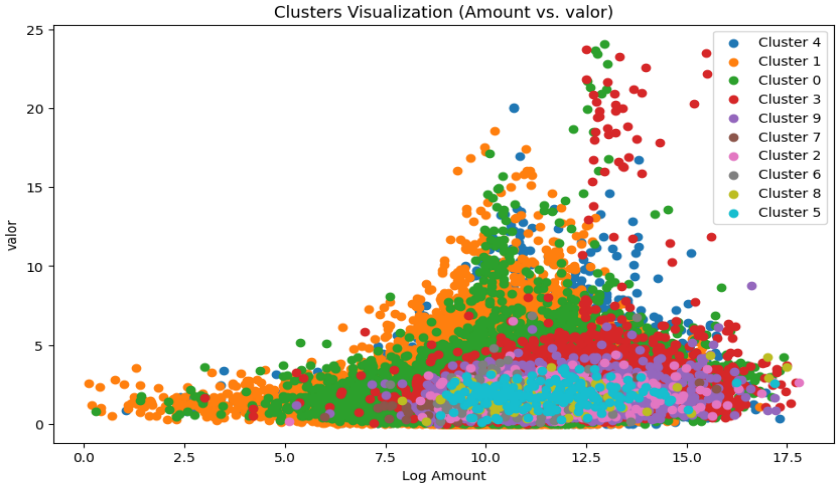
La figura anterior muestra una distribución del número de elementos por cluster tras aplicar el algoritmo K-means con k=10, en el cual podemos observar una clara desigualdad en los clusters, destacándose el cluster 1 con más de 14000 elementos, los demás clusters son menores en tamaño sin embargo son relevantes.

Presupuesto (amount) vs Valor de la descripción (valor)

La **Figura 3** muestra la relación entre el presupuesto del proceso y el valor de su descripción, coloreado por clúster.

Figura 3

Presupuesto (amount) vs Valor de la descripción (valor)



Basado en los clústeres más significativos, se obtienen los siguientes resultados:

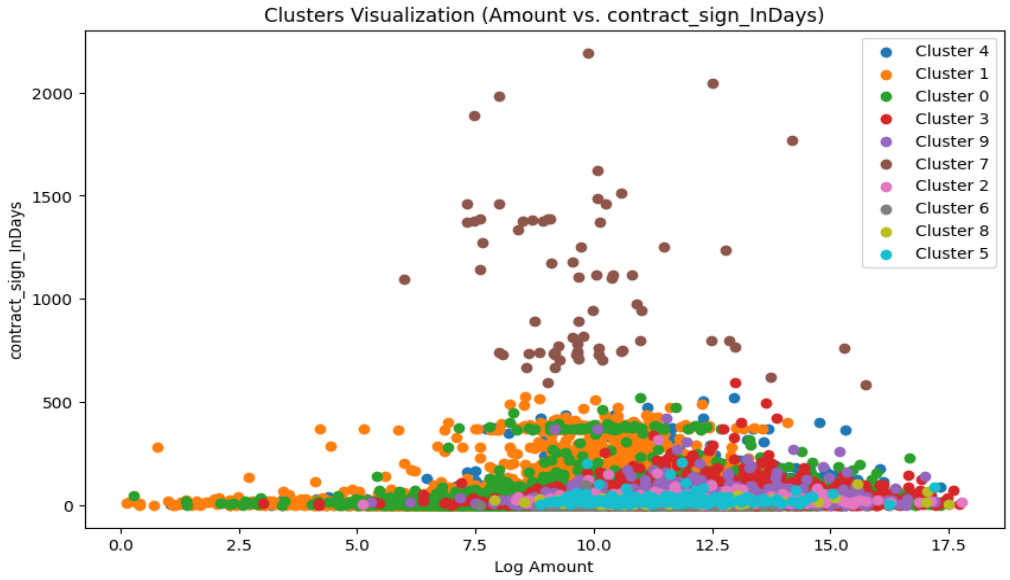
- **CLÚSTER 5**: Caracteriza procesos con una valoración descriptiva baja pero con presupuestos medio-altos. Esto podría indicar descripciones genéricas para contratos económicamente importantes.
- **CLÚSTERES 0 y 3**: Representan procesos con presupuestos medio-altos, donde la valoración de las descripciones es predominantemente baja. Similar al clúster 5, sugiere una posible falta de detalle en la descripción para procesos de relevancia económica.

Presupuesto y Firma del contrato en días

La **Figura 4** ilustra la relación entre el presupuesto y los días hasta la firma del contrato, segmentada por clúster.

Figura 4

Presupuesto y Firma del contrato en días



A partir de esta visualización, se identifican los siguientes patrones relevantes:

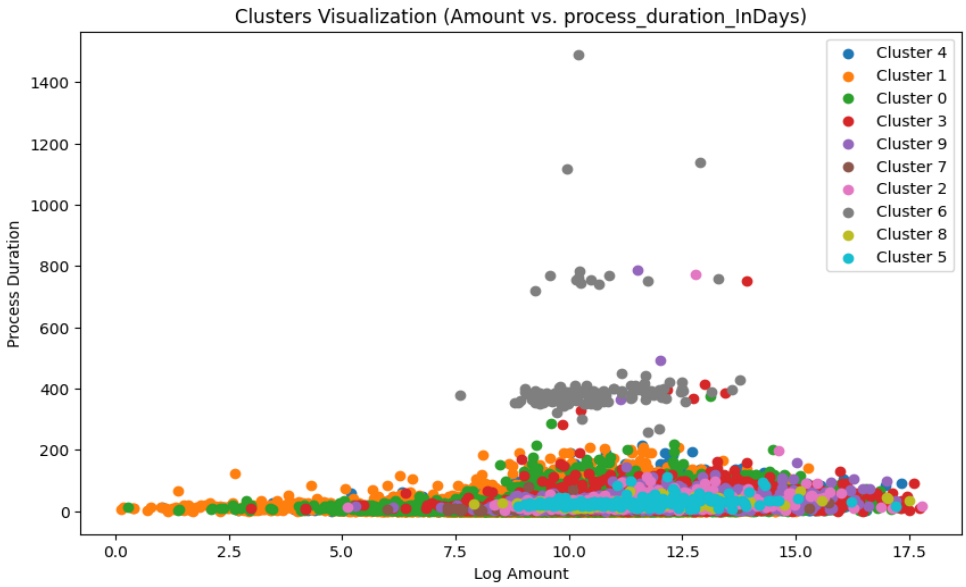
- **CLÚSTER 7**: Agrupa procesos con un presupuesto medio-alto, pero con un tiempo hasta la firma del contrato significativamente más alto que el promedio, lo cual puede ser un indicador de procesos con complejidades o retrasos.
- **CLÚSTER 5**: Se compone de procesos de presupuesto medio-alto con un tiempo de firma de contrato bajo. Esto sugiere procesos ágiles y eficientes, posiblemente debido a descripciones genéricas o falta de controversia.
- **CLÚSTER 1**: Este clúster se observa en procesos cuya duración hasta la firma es menor a 500 días, sin una dependencia clara del monto del contrato.

Presupuesto y Duración del proceso.

La **Figura 5** representa la relación entre el presupuesto y la duración total del proceso en días.

Figura 5

Presupuesto y Duración del proceso



En esta gráfica, los clústeres más significativos revelan:

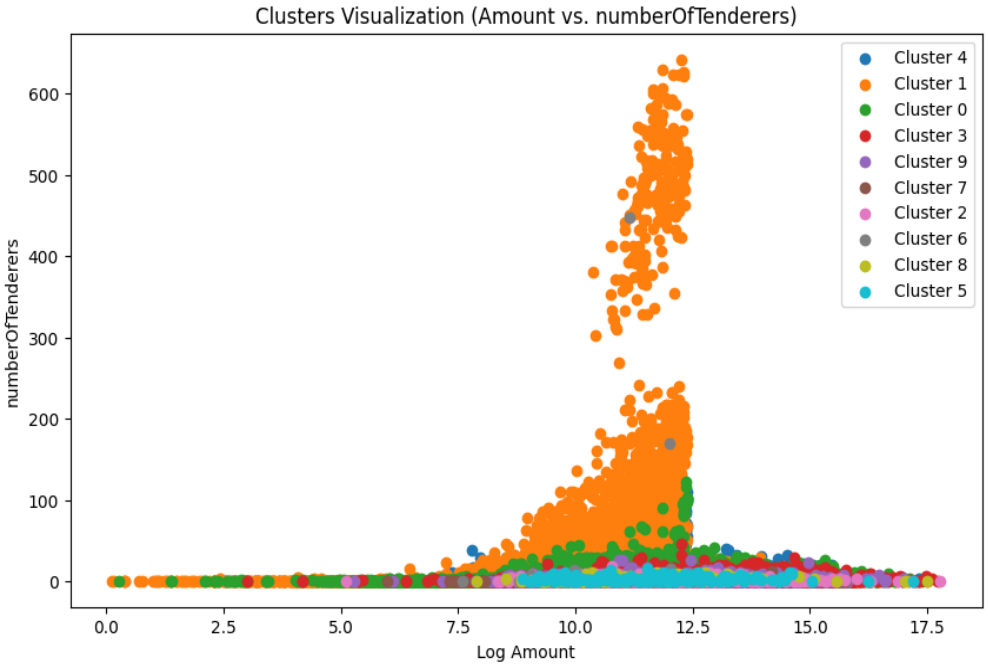
- **CLÚSTER 5**: Reitera la presencia de procesos de presupuesto medio-alto con una duración de proceso baja, sugiriendo una ejecución burocrática ágil, posiblemente facilitada por descripciones genéricas y baja controversia.
- **CLÚSTER 6**: Este clúster agrupa procesos con una duración anormalmente larga, especialmente cuando el presupuesto se encuentra en un rango intermedio (entre 7.5 y 15 en la escala logarítmica). Esto podría indicar problemas o ineficiencias significativas en el proceso.

Presupuesto y Número de participantes

La **Figura 6** muestra la relación entre el presupuesto y la cantidad de participantes en el proceso de contratación.

Figura 6

Presupuesto y Número de participantes



Los patrones observados en esta visualización son:

- **CLÚSTER 1**: Presenta una distribución con dos subgrupos diferenciados por umbrales de presupuesto (entre 7.7 y 12.5 en la escala logarítmica). El primer subgrupo tiene menos de 200 participantes, mientras que el segundo muestra entre 300 y 600 participantes, indicando variabilidad en la competencia para procesos de similar relevancia económica.
- **CLÚSTER 5**: Similar a las observaciones anteriores, este clúster agrupa procesos de presupuesto medio-alto con tiempos de contrato bajos, reforzando la idea de procesos rápidos y, por ende, potencialmente con menor competencia o análisis.

Consideraciones sobre clusters

En la **Tabla 2,** se presentan varias consideraciones sobre los resultados obtenidos de los clusters previamente analizados.

Tabla 2

Descripción extendida sobre las propiedades de los clusters

Cluster	Peso en el conjunto*	Rasgos numéricos dominantes (comparados con la media del dataset)	Lectura operativa / tipo de proceso	Riesgo o acción recomendada
1	≈ 52 %(grupo mayoritario)	• Monto y valor medio-alto • Duraciones bajas • Nº de oferentes muy alto	Procesos competitivos y eficientes	Mantener buenas prácticas; sirven como referencia de desempeño.
4	≈ 14 %	• Montos y duraciones medias • Competencia y valor descriptivo intermedios	Contratación semicompleja (rutinaria pero con cierta variabilidad)	Vigilar tiempos de firma para reducir dispersión.

0	≈ 14 %	• Montos medios-bajos • Duraciones y firmas en plazos normales • Competencia baja-media	Procesos estándar o rutinarios	Automatizar controles; poca alerta de riesgo.
3	≈ 2 %	• Montos moderados • Firma de contrato lenta (> 400 días) • Valor alto	Casos documentados pero lentos	Revisar cuellos de botella administrativos.
9	< 1 %	• Montos altos • Duraciones normales • Competencia y valor altos	Procesos atípicos bien documentados	Buena trazabilidad; buena realización
2	< 0,5 %	• Montos y valor bajos • Pocos oferentes	Procesos poco descriptivos / monótonos	Exigir pliegos más detallados y apertura a más competencia.
5	< 0,1 %	• Montos bajos • Duraciones cortas • Valor bajo	Compras muy simples (baja fiscalización)	Poca relevancia
6	muy residual	• Firma de contrato extremadamente tardía (> 2 000 días)	Procesos con alta eficiencia	Es posible que existan problemas legales de por medio
7	muy residual	• Duración de proceso y firma altas • Montos medios	Procesos críticos o excepcionales	Posiblemente requiera procesos legales
8	muy residual	• Proceso muy prolongado (> 1 400 días)	Procesos estancados / posiblemente	Intervención por posibles irregularidades

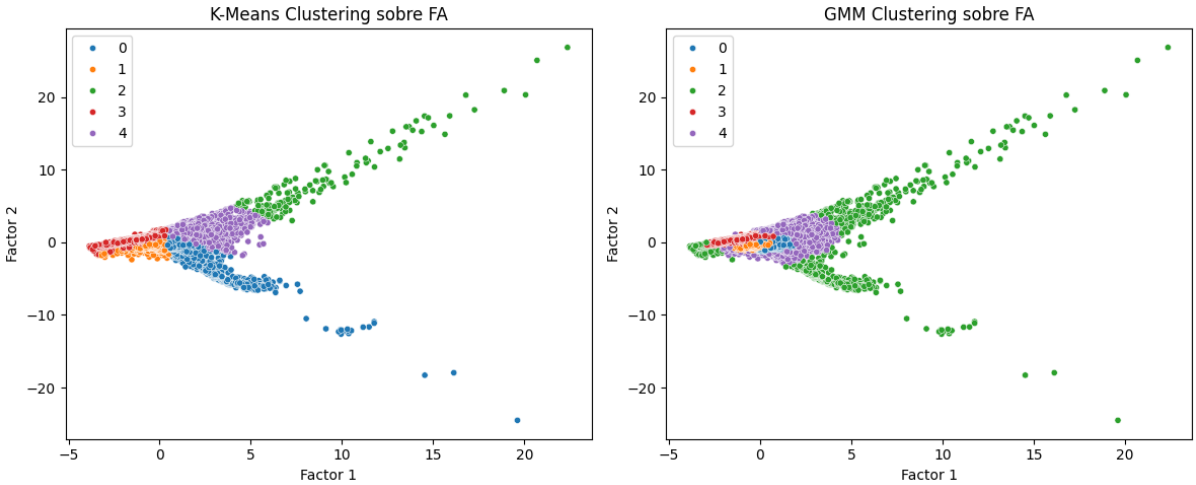
Análisis por Reducción de dimensionalidad con Kmeans

Se realizó un Análisis Factorial (AF) sobre variables numéricas clave para obtener componentes latentes que resumieran la información esencial de los datos. Posteriormente, se aplicaron algoritmos de clustering K-Means y Gaussian Mixture Models (GMM) sobre estos componentes factoriales para identificar agrupaciones significativas en los procesos de contratación. La evaluación de la coherencia interna de los clústeres se realizó mediante el coeficiente de silueta.

La **Figura 7** visualiza la distribución de los clusters generados por K-Means y GMM sobre los componentes del Análisis Factorial.

Figura 7

Clusters con K-Means y GMM



El análisis comparativo de los perfiles de clústeres reveló distintas tipologías de procesos de contratación, incluyendo grupos con alta competencia, procesos de bajo monto, y clústeres que sugieren anomalías en tiempos de proceso o participación. Se evidenció que K-Means generó clústeres más homogéneos y claramente diferenciables, mientras que GMM mostró una mayor propensión a solapamiento, destacando un grupo dominante. Esto resalta la importancia del enfoque de modelado al interpretar agrupamientos no supervisados en datos de contratación pública.

Segmentación con K-Means sobre FA

Al aplicar K-Means sobre los componentes obtenidos del Análisis Factorial, con un número de 5 clústeres, se logró una identificación de grupos de procesos mejor diferenciados, incluyendo una segmentación más clara de los *outliers*. La **Tabla 3** resume las características principales de estos clústeres.

Tabla 3

Descripción de los clusters con el método K means

cluster	Monto promedio (amount)	Nº Oferentes	Duración del contrato	Valoración (valor)	Comentario general
0	$147K	8.6	179 días	2.10	Contratos medianos con buena competencia
1	$22K	3.9	83 días	1.61	Contratos menores con baja valoración
2	$14.7M	2.7	761 días	2.57	Procesos grandes con poca competencia
3	$32K	1.6	122 días	1.59	Procesos simples y poco participativos
4	$929K	3.4	587 días	2.24	Contratos grandes, larga duración

La **Tabla 2** muestra perfiles de clusters bien definidos: el Clúster 2, por ejemplo, representa procesos de muy alto monto con baja competencia, lo cual podría ser un indicador de riesgo de ineficiencia o incluso colusión. Por otro lado, el Clúster 0 representa procesos medianos con buena competencia, que podrían ser considerados "saludables".

Resultados de GMM sobre FA

El modelo Gaussian Mixture Models (GMM), al ser aplicado sobre los componentes del Análisis Factorial, identificó subgrupos más finos dentro de los clústeres creados por K-Means, revelando patrones anidados pero con una lógica probabilística más flexible. Los principales hallazgos fueron:

- Clústeres similares en cuanto a procesos directos (pocos oferentes y sin consultas) y procesos comunes (bajo monto y buena frecuencia).
- Un clúster adicional con más de 100 oferentes, montos elevados y valoración alta, posiblemente asociado a licitaciones masivas nacionales o internacionales.

Tabla 4

Características de los clústeres identificados por GMM

Cluster	Monto promedio	Nº Oferentes	Duración del contrato	Valoración	Comentario general
0	$188K	5.5	202 días	2.09	Contratos medianos y relativamente comunes
1	$29K	3.8	91 días	1.66	Contratos pequeños
2	$3.87M	117	377 días	3.23	Procesos masivos con altísima competencia
3	$38K	1.6	124 días	1.61	Procesos directos, baja participación
4	$795K	11.3	568 días	2.17	Contratos grandes con buena competencia

La **Tabla 4** complementa la información de K-Means, ofreciendo una perspectiva más granular. El Clúster 2 de GMM, por ejemplo, es particularmente interesante, pues identifica procesos de muy alto monto con una competencia excepcionalmente alta, sugiriendo un perfil de licitaciones de gran envergadura y visibilidad.

Análisis de tabla de contingencia para agrupamientos de K-means y GMM

Para evaluar la consistencia y las diferencias en las asignaciones de los puntos de datos entre los algoritmos K-Means y GMM, se utilizó una tabla de contingencia (tabla de frecuencias cruzadas). Esta tabla resume en forma matricial el número de observaciones en cada combinación de clústeres de ambos métodos. La **Tabla 5** muestra la matriz de contingencia.

Tabla 5

Correlación Cruzada

GMM sobre FA	0	1	2	3	4
Kmeans sobre FA					
0	28600	13896	431	1850	1877
1	2572	80669	23	297	208
2	0	0	161	0	0
3	0	97	90	83078	252
4	2238	0	217	951	6581

Consistencia entre clústeres

Con el objetivo de evaluar la estabilidad y coherencia de los agrupamientos generados, se realizó una comparación cruzada entre los resultados obtenidos mediante los algoritmos K-Means y Gaussian Mixture Models (GMM). Esta comparación permite identificar qué clústeres han sido reconocidos consistentemente por ambos métodos, lo cual puede interpretarse como evidencia de estructuras subyacentes reales en los datos. A continuación, se detallan los principales emparejamientos entre clústeres de ambos modelos y su interpretación:

- K-Means clúster 2 coincide con GMM clúster 2, con un total de 161 observaciones en común; esto indica que ambos métodos identificaron un grupo claro y extremo (probablemente procesos con montos muy altos y poca competencia).
- K-Means clúster 3 se alinea fuertemente con GMM clúster 3, con un total de 83,078 observaciones en común; esto indica que ambos modelos coinciden en identificar procesos pequeños, directos o con baja participación.
- K-Means clúster 1 y GMM clúster 1 también están bastante alineados, con 80,669 observaciones coincidentes; refuerza que ese grupo representa procesos comunes, pequeños, sin características extremas.

Divergencia y mezcla

Además de identificar coincidencias entre clústeres, es importante analizar los casos en los que los algoritmos generan asignaciones divergentes. Esta divergencia puede evidenciar diferencias en la sensibilidad de cada método para captar estructuras internas en los datos. En esta sección se examinan los casos donde un clúster generado por K-Means se distribuye en varios

clústeres de GMM, lo que sugiere una mayor capacidad del modelo de mezclas gaussianas para capturar subgrupos más complejos o con mayor variabilidad interna..

- K-Means clúster 0 se distribuye entre GMM 0, 1, 3 y 4. Esto indica que GMM subdivide más este grupo, es decir, que percibe subestructuras internas que K-Means no distingue.
- K-Means clúster 4 también está disperso, en especial en GMM 4 y 0, esto indica que GMM parece detectar variabilidad interna en los contratos institucionales o de gran escala.

La **Tabla 5** de contingencia revela una coincidencia significativa entre los clústeres 1 y 3 de ambos métodos, mientras que en otros casos como el clúster 0 de K-Means, los elementos se reparten entre varios clústeres de GMM, lo que evidencia una segmentación más refinada por parte del modelo de mezclas gaussianas.

La divergencia en la asignación de algunos clústeres puede atribuirse a la diferencia de supuestos: K-Means segmenta en función de la distancia euclidiana al centroide, mientras que GMM modela distribuciones probabilísticas, permitiendo clústeres de formas elípticas y con diferente varianza. Esta flexibilidad podría explicar la capacidad de GMM para capturar mejor ciertas estructuras latentes del conjunto

Dendrograma Jerárquico por aplicación de cluster Jerárquico

El dendrograma es una representación gráfica de la estructura jerárquica de los datos, derivada del análisis de clustering jerárquico. Su interpretación se basa en los siguientes componentes:

- **Hojas:** Los puntos inferiores del dendrograma, etiquetados con los índices de los puntos de datos individuales (proyectos muestreados), representan cada instancia de dato analizada.
- **Ramas y Nodos:** Las líneas verticales (ramas) conectan las hojas y se unen mediante líneas horizontales en los nodos. Estas uniones ilustran cómo los clústeres más pequeños se fusionan progresivamente para formar clústeres más grandes, revelando la jerarquía de agrupamiento.
- **Altura de las Líneas de Fusión (Distancia):** La altura de cada línea horizontal en el eje y, etiquetada como "Distancia", indica la disimilitud entre los dos clústeres que se fusionaron en ese punto. Una gran distancia vertical entre puntos de fusión sugiere que los clústeres combinados eran relativamente disímiles, mientras que una distancia pequeña indica una alta similitud al momento de la fusión.
- **Corte del Dendrograma (Determinación del Número de Clústeres):** Para obtener un número específico de clústeres, se puede conceptualmente "cortar" el dendrograma a una cierta altura en el eje de distancia. Todas las ramas conectadas por debajo de esta línea de corte pertenecen al mismo clúster. Alternativamente, como se realizó en este estudio, se

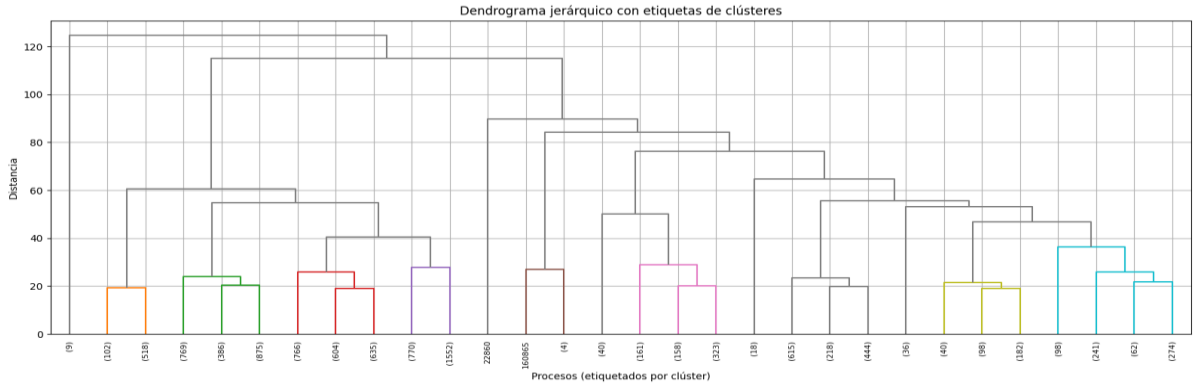
puede especificar el número deseado de clústeres y una función como *fcluster* determina automáticamente la altura de corte necesaria para producir ese número de agrupaciones.

- **Umbral de Color:** El parámetro *color_threshold* utilizado en la generación del dendrograma permite colorear de manera distinta las ramas por debajo de una distancia específica. Esto facilita la visualización de posibles agrupaciones a un nivel de disimilitud determinado, mientras que las ramas por encima de este umbral suelen mostrarse en un color por defecto.

En el contexto de nuestro análisis, la altura de las fusiones en el dendrograma proporciona información visual sobre la disimilitud entre los grupos que se combinan. Aunque un dendrograma con muchos puntos de datos puede parecer complejo, el análisis de las ramas principales y las alturas de fusión ofrece *insights* sobre la estructura jerárquica de los datos y ayuda a validar el número de clústeres seleccionados para análisis posteriores. Los colores distintos por debajo del umbral de color resaltan las agrupaciones potenciales a un nivel de distancia específico.

Figura 8

Cluster Jerárquico



Para explorar la estructura latente de los procesos contractuales, se aplicó un clustering jerárquico aglomerativo utilizando el método de enlace de Ward, sobre una muestra representativa de 10.000 registros. Esta técnica permite construir un árbol de decisiones (dendrograma) que revela las similitudes entre observaciones sin requerir un número de clústeres predefinido.

El dendrograma obtenido muestra la progresiva fusión de observaciones en conglomerados, basándose en la distancia euclidiana multivariada entre registros estandarizados. La visualización truncada en las últimas 30 uniones permite observar con claridad la conformación de varios agrupamientos bien diferenciados.

Se identificaron al menos cinco agrupaciones principales con alturas de fusión significativamente distintas, lo cual sugiere estructura de agrupamiento robusta y heterogénea dentro del conjunto de datos. La existencia de ramas con más de 1.000 observaciones contrastadas con otras de menor densidad (<100 elementos) evidencia la diversidad en patrones contractuales,

posiblemente asociados a diferencias en montos, duración, número de oferentes y tiempos de ejecución.

Esta representación es particularmente útil para interpretar la jerarquía de relaciones entre procesos, ya que uniones a mayor altura indican disimilitud marcada entre clusters, mientras que las uniones a menor distancia reflejan alta similitud interna. Como complemento, se calculó el índice de Silhouette, obteniendo un valor promedio de 0.30, lo que indica una estructura de clusters moderadamente definida y coherente con las observaciones visuales.

A diferencia de métodos como K-Means o GMM, el enfoque jerárquico no requiere supuestos sobre la forma de los grupos ni parámetros iniciales, y proporciona una herramienta valiosa para determinar el número óptimo de clusters en estudios exploratorios. Su aplicación permite sustentar decisiones posteriores sobre segmentación, análisis comparativo o definición de perfiles de riesgo en los procesos de contratación pública.

DISCUSIÓN

Los resultados de este estudio demuestran la viabilidad y el valor de aplicar técnicas de Big Data y aprendizaje no supervisado para analizar los datos de contratación pública del SERCOP, a pesar de sus deficiencias en calidad y completitud. La baja correlación inicial entre las variables, como se mostró en la **Tabla 1**, subraya la insuficiencia de métodos tradicionales y la necesidad de enfoques más complejos como el clustering y la reducción de dimensionalidad para descubrir patrones ocultos.

La caracterización de los clústeres obtenidos mediante K-Means (sección 2.1) proporciona una segmentación granular de los procesos de contratación. Por ejemplo, el Clúster 5, que agrupa procesos de presupuesto medio-alto con valor descriptivo bajo y tiempos de firma rápidos, podría indicar una gestión eficiente, pero también levanta una bandera roja sobre la posible falta de transparencia o detalle en la documentación. En contraste, el Clúster 7 (presupuesto medio-alto, firma de contrato muy alta) o el Clúster 6 (duración de proceso anormalmente larga) sugieren cuellos de botella burocráticos o la presencia de problemas legales subyacentes, lo cual requiere una fiscalización más profunda. Estas observaciones, respaldadas por las **Figuras 3, 4, 5 y 6**, ofrecen *insights* accionables para las autoridades de control.

La integración del Análisis Factorial (AF) con K-Means y GMM (secciones 3 y 4) no solo redujo la complejidad de los datos, sino que mejoró la interpretabilidad de los clústeres. Las **Tablas 2 y 3** detallan cómo cada método reveló perfiles distintos de procesos: desde contratos "saludables" con buena competencia hasta procesos de alto monto con poca participación, lo cual podría indicar ineficiencia o, en el peor de los casos, colusión. La capacidad de GMM para identificar subgrupos con alta competencia en procesos masivos (Clúster 2 de GMM en **Tabla 3**) enriquece la comprensión, destacando la existencia de licitaciones de gran envergadura y visibilidad.

La tabla de contingencia (sección 5, **Tabla 4**) fue crucial para validar la coherencia y las diferencias entre los agrupamientos de K-Means y GMM. Las coincidencias significativas en clústeres extremos (ej., K-Means clúster 2 con GMM clúster 2, o K-Means clúster 3 con GMM clúster 3) refuerzan la robustez de estos hallazgos. Por otro lado, la dispersión de algunos clústeres de K-Means en varios de GMM (ej., K-Means clúster 0 distribuido en GMM 0, 1, 3 y 4) subraya la mayor flexibilidad y capacidad de GMM para capturar estructuras latentes más complejas y sutiles en los datos. Esto demuestra que, si bien K-Means es útil para una segmentación inicial clara, GMM ofrece una visión más rica y matizada de los patrones en los datos. El dendrograma jerárquico (**Figura 8**) complementa este análisis al ofrecer una perspectiva visual de las similitudes y disimilitudes, ayudando a justificar la selección de clústeres y a comprender las relaciones jerárquicas entre los diferentes tipos de procesos de contratación.

En síntesis, este estudio no solo aborda la problemática de la baja calidad de los datos abiertos del SERCOP, sino que también demuestra cómo, a través de una metodología rigurosa basada en Big Data y aprendizaje no supervisado, es posible extraer conocimiento valioso. Los perfiles de clústeres identificados, desde procesos eficientes hasta aquellos con indicadores de posible ineficiencia o riesgo, son un testimonio del potencial de estas herramientas para transformar grandes volúmenes de datos brutos en inteligencia estratégica para la gestión pública. Estos hallazgos sientan las bases para el desarrollo futuro de sistemas de monitoreo y alerta temprana que contribuyan a la transparencia y la eficiencia en la contratación pública ecuatoriana.

CONCLUSIONES

Esta investigación demostró la aplicabilidad y utilidad de técnicas de análisis no supervisado en el contexto de los datos abiertos de contratación pública del Ecuador (SERCOP, 2015–2025). Inicialmente, se exploraron técnicas normales de regresión, pero no fue posible obtener correlaciones significativas debido a la naturaleza dispersa y no curada de los datos. Por esta razón, se optó por un enfoque metodológico basado en el paradigma de Big Data, integrando procesos de preprocesamiento, reducción de dimensionalidad y segmentación automática (Pita, 2021). Esto permitió estructurar y analizar una base de datos de gran volumen, logrando así identificar patrones y obtener insights relevantes que no fueron posibles con los métodos tradicionales de regresión.

El **Análisis Factorial (FA)** permitió transformar el conjunto de variables con correlaciones mínimas en representaciones latentes más manejables, capturando la varianza del sistema sin sacrificar información clave. Esta reducción dimensional fue crucial para facilitar la posterior aplicación de algoritmos de agrupamiento como **K-Means**, **Gaussian Mixture Models (GMM)** y **clustering jerárquico**, los cuales, en conjunto, ofrecieron una visión integral y multiescala de la estructura interna de los datos.

El algoritmo **K-Means** generó agrupaciones compactas y bien diferenciadas, útiles para establecer tipologías claras de procesos contractuales. En cambio, **GMM** reveló estructuras más complejas y superpuestas, capaces de identificar subgrupos no capturados por métodos basados únicamente en distancia euclidiana. Por su parte, el **clustering jerárquico** permitió analizar la similitud entre procesos desde una perspectiva evolutiva, sin necesidad de predefinir el número de grupos, lo que aportó profundidad al análisis exploratorio.

Entre los hallazgos más relevantes se identificaron clusteres asociados a:

- Contratos masivos con alto índice de participación,
- Procesos aparentemente direccionados con un solo oferente y sin consultas,
- Contratos de largo plazo, cuya información es ambigua o no justifica la duración.
- Procesos estándar de bajo monto, con diferentes características que promueven su rápida ejecución.

La **tabla de contingencia entre K-Means y GMM** fue fundamental para evaluar la coherencia entre los modelos, revelando una alta concordancia en clústeres extremos y divergencia en zonas más heterogéneas. Este hallazgo refuerza la validez del enfoque aplicado y subraya la riqueza estructural del conjunto de datos.

En resumen, este trabajo resalta cómo las técnicas de aprendizaje no supervisado, integradas en un flujo analítico estructurado, pueden ofrecer un valor significativo en contextos gubernamentales con datos voluminosos y heterogéneos. La combinación de reducción de dimensionalidad, agrupamiento y validación cruzada transforma datos administrativos en conocimiento útil, replicable y escalable para la eficiencia, transparencia y fiscalización inteligente en el sector público.

Recomendaciones

A partir de los hallazgos de este estudio, que evidencian la deficiencia de la calidad de los datos abiertos del SERCOP y su impacto en la capacidad de detectar riesgos (como los clústeres de procesos prolongados o con descripciones ambiguas), se plantean las siguientes recomendaciones con un enfoque en su implementación práctica en el contexto del Sistema Nacional de Contratación Pública:

Mejora de la Calidad y Completitud de los Datos Publicados

- **Acción Propuesta:** El SERCOP debe liderar un proyecto de mejora continua de la calidad de datos, enfocándose en la **estandarización y obligatoriedad de registro de campos clave** y la **inclusión de metadatos** esenciales.

Plan de Implementación

- **Corto Plazo (6-12 meses):** Revisión y actualización de los formularios de carga de información en el Sistema Oficial de Contratación del Estado (SOCE) para hacer obligatorios campos críticos como "unidad de medida", "cantidad", "justificación de la duración del contrato para contratos de largo plazo" y "categorías de

bienes/servicios más específicas". Implementar validaciones de datos en tiempo real en la plataforma para prevenir la carga de valores nulos o inconsistentes en estas variables.

- **Mediano Plazo (1-2 años):** Desarrollo de un **módulo de metadatos** dentro del portal de datos abiertos que proporcione información sobre la frescura de los datos (fechas de actualización), porcentaje de valores nulos por campo y alertas sobre *outliers* detectados automáticamente.
- **Ejemplo Concreto (vinculado a tus clústeres):** Para clústeres identificados con **"contratos de largo plazo cuya información es ambigua o no justifica la duración"**, la obligatoriedad de un campo de justificación detallada permitiría a futuros análisis discriminar entre procesos legítimamente extensos y aquellos que podrían ocultar ineficiencias o riesgos, facilitando la labor de la Contraloría General del Estado al proveer datos más claros para su fiscalización

Para facilitar el acceso y la reutilización de la información, se sugiere consolidar la API oficial para que proporcione datasets limpios y actualizados, con estructuras homogéneas a lo largo del tiempo, así como promover la interoperabilidad con otras bases de datos públicas relevantes. Estas acciones permitirán que futuros estudios se realicen con mayor precisión y sin pérdidas significativas de registros durante la etapa de preparación de datos.

Impacto Social y Económico Esperado: La implementación de estas acciones derivadas de estudios de esta índole no solo reducirá las ineficiencias y los riesgos de corrupción en la contratación pública, liberando recursos que pueden ser destinados a mejorar servicios esenciales como salud, educación e infraestructura (impacto social), sino que también fomentará un ambiente de competencia justa y transparente que atraerá inversión y promoverá el crecimiento de empresas honestas en el país (impacto económico). Una gestión pública más transparente y eficiente, respaldada por datos de calidad y herramientas analíticas, fortalecerá la confianza ciudadana en las instituciones y contribuirá al desarrollo sostenible de Ecuador."

REFERENCIAS

Banco Mundial. (s. f.). Proyectos y programas de adquisiciones. https://www.worldbank.org/en/programs/project-procurement

Barrera Borbor, K. J. (2021). Análisis del Sistema Nacional de Contratación del Estado (SOCE) actual y posibles alternativas para el mejoramiento de la compra pública [Tesis de maestría, UCSG]. http://repositorio.ucsg.edu.ec/handle/3317/16850

Benavides, J. L., M'Causland Sánchez, M. C., Flórez Salazar, C., & Roca, M. E. (2016). Las compras públicas en América Latina y el Caribe y en los proyectos financiados por el BID: Un estudio normativo comparado. Banco Interamericano de Desarrollo.

García, M., Rodríguez, V., Ballesteros, P., Love, P. E. D., & Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. Automation in Construction, 133, 104047. https://riunet.upv.es/entities/publication/ffa52850-ffa8-4995-b71b-df93befdb105

González, J. (2014). La duración de los procedimientos de licitación pública en Ecuador: Análisis y propuestas para su mejora [Tesis de licenciatura, Universidad Pontificia Comillas]. https://repositorio.comillas.edu/xmlui/handle/11531/4873

Joković, S. (2022). Competition and integrity in public procurement (2.ª ed.). Notion Press.

Molina, M., Acaro, X., Molina, M., Quinoñez, M., Alvarez, G., & Fernandez, J. (2023). Application of explainable artificial intelligence to analyze basic features of a tender. En Proceedings of the International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME 2023) (pp. 1-6). IEEE. https://doi.org/10.1109/ICECCME57830.2023.10253063

Nai, R., Meo, R., Morina, G., & Pasteris, P. (2023). Public tenders, complaints, machine learning and recommender systems: A case study in public administration. Computer Law & Security Review, 51, 105887. https://doi.org/10.1016/j.clsr.2023.105887

Pita, C. (2021). Proyecto de sistema de recomendación de filtrado colaborativo basado en machine learning. Revista PGI, 8, 48-51. https://ojs.umsa.bo/ojs/index.php/inf_fcpn_pgi/article/view/46

Red Interamericana de Compras Gubernamentales. (2021, 9 de diciembre). Guía para la identificación de riesgos de corrupción en contratación pública, utilizando la ciencia de datos. Organización de los Estados Americanos; Banco de Desarrollo de América Latina. https://ricg.org/es/publicaciones/lanzamiento-guia-para-la-identificacion-de-riesgos-de-corrupcion-en-contratacion-publica-utilizando-ciencia-de-datos/