Diseño, implementación, monitoreo y mantenimiento del sistema electrónico para la contratación pública – SECOP II en Colombia- caracterización de patrones transaccionales como estrategia de monitoreo del comportamiento contractual

Andres Felipe Chacón Diaz

Asesor

Sebastian Mantilla Serrano

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Dedicatoria

Dedico este logro, fruto de meses de esfuerzo y dedicación, a las personas más importantes en mi vida: A mis padres, Magda y Efrén, por ser mi fuente inagotable de motivación y por inculcarme siempre el valor esencial de la perseverancia y el estudio. A mi novia, Mónica, por su apoyo incondicional, su aliento perseverante, y por su compañía desinteresada, la cual fue fundamental durante las largas jornadas que demandó este proceso académico.

Agradecimientos

A la Universidad Nacional Abierta y a Distancia (UNAD), por haberme brindado la oportunidad de formación y por su excelencia académica, que cimentó las bases esenciales para la realización de este trabajo. A la Escuela de Ciencias Básicas, Tecnología e Ingeniería (ECBTI) y a todos los docentes de la especialización, por compartir sus valiosos conocimientos y experiencias a lo largo de mi carrera profesional. De manera especial, extiendo mi gratitud al profesor Sebastián Mantilla Serrano, por su invaluable guía, su paciencia crítica y su disposición constante para orientar y enriquecer el desarrollo de la investigación. Su visión experta fue fundamental para reorientar y dar la solidez analítica que este proyecto requería. Finalmente, agradezco a mi familia y novia por ser mi principal soporte emocional, por su comprensión durante los momentos de mayor dedicación y por ser la razón última de cada esfuerzo realizado.

Resumen

La implementación del sistema SECOP II en Colombia ha generado un volumen masivo de datos transaccionales sobre la contratación pública. Sin embargo, la mera disponibilidad de estos datos no garantiza la transparencia ni la eficiencia en la fiscalización. La presente investigación aborda la problemática de la falta de caracterización de las entidades estatales, proponiendo una solución basada en Ciencia de Datos para identificar patrones de comportamiento contractual. Desde la perspectiva metodológica se adoptó un enfoque cuantitativo, desarrollando un *pipeline* de ingeniería de datos (ETL) automatizado en Python para extraer y depurar registros de la vigencia 2024. Se aplicaron técnicas de Análisis Exploratorio de Datos (EDA) y el algoritmo de aprendizaje no supervisado **K-Means** para segmentar a las entidades. Los resultados revelaron la existencia de cuatro perfiles operativos diferenciados: "Pequeños Competitivos", "Operadores Logísticos Masivos", "Ejecutores Estratégicos" y un perfil crítico de "Alto Riesgo", caracterizado por la adjudicación de montos multimillonarios mediante contratación directa exclusiva. Como producto tecnológico, se desarrolló el software *SICP* (Sistema de Inteligencia de Contratación Pública), una herramienta de auditoría que integra los modelos matemáticos y permite la clasificación de entidades en tiempo real. Se concluye que la aplicación de técnicas de *clustering* permite focalizar los esfuerzos de control fiscal, pasando de una auditoría aleatoria a una basada en riesgos algorítmicos.

Palabras claves: Contratación Pública, SECOP II, Ciencia de Datos, Clustering, K-Means, Control Fiscal, Python.

Abstract

The implementation of the SECOP II system in Colombia has generated a massive volume of transactional data regarding public procurement. However, the mere availability of this data does not guarantee transparency or efficiency in fiscal oversight. This research addresses the lack of characterization of state entities by proposing a Data Science-based solution to identify contractual behavior patterns. Methodologically, a quantitative approach was adopted by developing an automated data engineering (ETL) pipeline in Python to extract and clean records from the 2024 fiscal year. Exploratory Data Analysis (EDA) techniques and the **K-Means** unsupervised learning algorithm were applied to segment the entities. The results revealed the existence of four distinct operational profiles: "Small Competitive Buyers," "Massive Logistics Operators," "Strategic Executors," and a critical "High-Risk" profile, characterized by the allocation of multi-billion amounts through exclusive direct contracting. As a technological product, the *SICP* (Public Procurement Intelligence System) software was developed, an auditing tool that integrates mathematical models and allows for real-time entity classification. It is concluded that the application of clustering techniques allows for focusing fiscal control efforts, moving from random auditing to algorithmic risk-based auditing.

Keywords: Public Procurement, SECOP II, Data Science, Clustering, K-Means, Fiscal Control, Python.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción y Planteamiento del Problema

La transformación digital del Estado colombiano ha situado a los datos como el activo más valioso para la modernización de la administración pública. Con la transición hacia el SECOP II, Colombia ha pasado de un modelo de publicidad documental a uno transaccional, generando un ecosistema de "Datos Abiertos" sin precedentes en la región. No obstante, este avance tecnológico plantea un nuevo desafío: la capacidad de procesar, interpretar y auditar millones de registros contractuales supera las capacidades humanas tradicionales.

Bajo este contexto, el presente trabajo se enfoca en la fase de monitoreo y mantenimiento de la información que reside en el sistema. Se parte de la premisa de que un sistema transaccional como SECOP II no puede considerarse exitoso solo por estar 'en línea', sino por la calidad y comportamiento de los datos que procesa. Por tanto, esta investigación propone un diseño de monitoreo analítico utilizando técnicas de Ciencia de Datos para caracterizar cómo las entidades están utilizando realmente la plataforma diseñada.

El documento se encuentra estructurado en cinco capítulos. El Capítulo 1 establece el contexto del Big Data gubernamental y define la problemática de la opacidad funcional. El Capítulo 2 presenta el marco teórico, abordando desde la normativa legal de la contratación (Ley 80) hasta los fundamentos matemáticos del algoritmo K-Means. El Capítulo 3 detalla el diseño metodológico, describiendo la arquitectura híbrida de software y el ciclo de vida de los datos.

Posteriormente, el Capítulo 4 expone los resultados empíricos, revelando los cuatro perfiles de entidades hallados y presentando el prototipo de software desarrollado para la auditoría forense. Finalmente, el Capítulo 5 ofrece las conclusiones y recomendaciones estratégicas para la política pública, cerrando el ciclo entre la teoría de datos y la práctica administrativa.

Contextualización del Fenómeno: Datos Masivos en el Sector Público

La administración pública contemporánea atraviesa una transformación estructural impulsada por la digitalización y la disponibilidad masiva de datos. Organismos internacionales como la Organización para la Cooperación y el Desarrollo Económicos (OCDE) y el Banco Mundial han instado a los gobiernos a transitar hacia modelos de "Gobernanza Inteligente" (Smart Governance), donde la toma de decisiones deja de basarse exclusivamente en la intuición política o la inercia administrativa, para fundamentarse en la evidencia empírica proporcionada por los datos (Data-Driven Decision Making).

En el contexto colombiano, esta transformación se ha materializado a través de la política de Gobierno Digital y la Ley de Transparencia y del Derecho de Acceso a la Información Pública (Ley 1712 de 2014). El eje central de esta modernización en el ámbito fiscal ha sido la evolución del sistema de compras públicas. El Estado colombiano migró de un modelo de publicidad documental, conocido como SECOP I, hacia una plataforma transaccional robusta, el SECOP II.

A diferencia de su predecesor, SECOP II no es un simple repositorio de documentos escaneados. Es una plataforma basada en servicios web que gestiona el ciclo de vida contractual completo en línea. Cada interacción en el sistema, desde la publicación del plan anual de adquisiciones hasta la adjudicación y el pago, genera registros de datos estructurados. Como resultado, Colombia dispone hoy de uno de los conjuntos de datos abiertos (datasets) más ricos y voluminosos de la región, accesible públicamente a través del portal datos.gov.co.

Sin embargo, este avance tecnológico ha traído consigo un nuevo desafío, relativo a la sobrecarga de información. El volumen, la velocidad y la variedad de los datos generados por miles de entidades estatales (desde ministerios hasta alcaldías de sexta categoría) superan la capacidad humana de análisis convencional. Nos encontramos ante un escenario de Big Data en

el sector público que requiere, obligatoriamente, la implementación de nuevas metodologías de análisis computacional para ser comprendido y fiscalizado.

Planteamiento del Problema

A pesar de la riqueza de datos disponible en SECOP II, existe una brecha crítica entre la disponibilidad de la información y la generación de conocimiento accionable. Este fenómeno es conocido en la literatura académica como el síndrome DRIP (Data Rich, Information Poor), el Estado es rico en datos, pero pobre en información estratégica.

El problema central que aborda esta investigación se puede desglosar en tres dimensiones:

Opacidad en el Comportamiento Masivo: Actualmente, los entes de control y la ciudadanía pueden auditar un contrato específico si conocen su número de referencia. Sin embargo, es prácticamente imposible tener una visión holística de cómo contrata una entidad a lo largo de un año fiscal sin herramientas avanzadas. No sabemos, a ciencia cierta, cuáles son los patrones de comportamiento "normales" y cuáles son los "atípicos" en el universo de las entidades públicas.

Falta de Caracterización y Segmentación: En la administración pública colombiana se tiende a clasificar a las entidades por su naturaleza jurídica (Nacional, Territorial, Descentralizada). Sin embargo, esta clasificación legal no necesariamente refleja su comportamiento operativo real. No existe una taxonomía basada en datos (Data-Driven Taxonomy) que agrupe a las entidades según sus prácticas de contratación reales (ej. montos que manejan, preferencia por contratación directa, eficiencia en tiempos).

Ineficacia de la Auditoría Tradicional: Los métodos tradicionales de auditoría fiscal se basan en el muestreo aleatorio o en la reacción ante denuncias (control posterior). Ante un

universo de millones de registros, el muestreo aleatorio es estadísticamente ineficiente para detectar patrones complejos o sutiles de riesgo.

La ausencia de un modelo que permita caracterizar y agrupar a las entidades según su huella digital contractual impide la focalización eficiente de los recursos de control y limita la capacidad del Estado para diseñar políticas de compra pública basadas en la realidad operativa de las instituciones.

Pregunta de Investigación

Con base en la problemática expuesta, se formula la siguiente pregunta que guiará el desarrollo del presente trabajo de grado:

¿Cuáles son los patrones de comportamiento, las tendencias principales y los perfiles de contratación de las entidades públicas en Colombia, identificables a partir del análisis exploratorio y la segmentación (clustering) de los datos masivos de la plataforma SECOP II para la vigencia 2024?

Justificación

La pertinencia de esta investigación se sustenta en tres pilares fundamentales que demuestran su valor práctico, teórico y metodológico.

Justificación Práctica y Social

La contratación pública representa un porcentaje significativo del Producto Interno Bruto (PIB) de la nación. La ineficiencia o la corrupción en este rubro tienen un impacto directo en la calidad de vida de los ciudadanos. Este proyecto aporta una herramienta práctica: una "radiografía" basada en datos que permite identificar perfiles de entidades. Para los entes de control (Contraloría, Procuraduría), los resultados de esta segmentación permiten focalizar la vigilancia: si un algoritmo agrupa a una entidad pequeña en el mismo cluster de gasto que una

gran capital, esto constituye una alerta temprana automática digna de investigación. Así, se propone pasar de una fiscalización reactiva a una preventiva e inteligente.

Justificación Teórica

Existe un vacío en la literatura académica nacional respecto al análisis cuantitativo de la contratación. La mayoría de estudios previos abordan el fenómeno desde el Derecho Administrativo (análisis de leyes) o desde la Ingeniería de Software (arquitectura de la plataforma). Esta investigación contribuye al cuerpo de conocimiento demostrando que las entidades estatales no son entes monolíticos, sino que exhiben comportamientos heterogéneos que pueden ser modelados matemáticamente. Al establecer "perfiles" de contratación mediante el algoritmo K-Means, se aporta una nueva forma de entender la administración pública, no por lo que dicen sus estatutos, sino por lo que revelan sus datos.

Justificación Metodológica

En el ámbito de la Ciencia de Datos, este trabajo sirve como caso de estudio sobre la aplicación de técnicas de Aprendizaje No Supervisado (Unsupervised Learning) en datos gubernamentales. Se documenta un flujo de trabajo (pipeline) completo y replicable: desde la conexión automatizada a APIs públicas, pasando por la limpieza de datos complejos (Data Wrangling), hasta la validación de modelos de agrupamiento. Esta metodología puede ser adoptada por futuros investigadores para analizar otros sectores del Estado, como la salud o la educación.

Objetivos

Objetivo General

Desarrollar un modelo de análisis y monitoreo de los patrones de comportamiento contractual en el sistema SECOP II, mediante el diseño e implementación de técnicas de Minería de Datos y Clustering, para la caracterización de las entidades públicas colombianas

Objetivos Específicos

Implementar un proceso de extracción, transformación y carga (ETL) automatizado que permita la ingesta de datos desde la API de Datos Abiertos de Colombia, garantizando la calidad y estructuración de la información para el año fiscal 2024.

Realizar un Análisis Exploratorio de Datos (EDA) exhaustivo para describir estadísticamente las variables críticas del sistema de compra pública (montos, tiempos, modalidades y ubicaciones geográficas).

Diseñar y calcular un conjunto de características (features) representativas por entidad, transformando los datos transaccionales brutos en indicadores de comportamiento normalizados.

Segmentar a las entidades estatales mediante el algoritmo de aprendizaje no supervisado K-Means, determinando el número óptimo de grupos y analizando los centroides resultantes para definir los perfiles de contratación.

Desarrollar un prototipo de herramienta de software que permita la visualización interactiva de los hallazgos y la clasificación de nuevas entidades en los perfiles identificados.

Alcance y Delimitaciones

Alcance Temático

La investigación se circunscribe al campo de la Ciencia de Datos aplicada. Se centra específicamente en técnicas descriptivas (estadística) y de agrupamiento (Machine Learning no supervisado). No se incluyen dentro del alcance técnicas de Procesamiento de Lenguaje Natural (NLP) para el análisis de textos de pliegos, ni modelos predictivos supervisados de detección de fraude, ya que el objetivo es la caracterización y no la predicción.

Alcance Temporal y Espacial

Temporal: El estudio tomará como ventana de observación los procesos contractuales cuya fecha de firma se encuentre en el año fiscal 2024. Esto garantiza la actualidad de los hallazgos y permite trabajar con la estructura de datos más reciente de la Agencia Nacional de Contratación Pública.

Espacial: El análisis tiene cobertura nacional, incluyendo entidades del nivel central (Nación) y descentralizado territorial (Departamentos y Municipios) que utilicen SECOP II.

Limitaciones

Los resultados de la investigación dependen intrínsecamente de la calidad y veracidad de los datos reportados por las entidades en la plataforma SECOP II. Errores de digitación en origen o falta de reporte oportuno por parte de los funcionarios públicos son factores exógenos que serán mitigados mediante técnicas de limpieza de datos, pero que no pueden ser eliminados totalmente. Asimismo, los "perfiles" resultantes son agrupaciones estadísticas y no constituyen, per se, un dictamen legal sobre la gestión de las entidades.

Marco Teórico y Estado del Arte

Marco Conceptual: Contratación y Gobierno Digital

La contratación pública en Colombia trasciende la simple adquisición administrativa de bienes y servicios; se configura como el principal instrumento macroeconómico de ejecución presupuestal y descentralización de recursos. En el contexto de la modernización del Estado, este proceso ha evolucionado hacia un modelo de Gobierno Digital (e-Government), donde la infraestructura tecnológica no es un soporte accesorio, sino el eje central de la operación.

Bajo la óptica del diseño y mantenimiento de sistemas, la contratación estatal debe entenderse como un macro-proceso de gestión de información crítica. Según la Agencia Nacional de Contratación Pública (Colombia Compra Eficiente, 2021), la eficiencia del Estado depende de la interoperabilidad, integridad y disponibilidad de los datos transaccionales generados por las entidades compradoras.

Por consiguiente, el "monitoreo del sistema" SECOP II no se limita a la vigilancia de su disponibilidad técnica, sino que abarca el monitoreo de la calidad y el comportamiento de los datos que fluyen a través de él. La presente investigación adopta este enfoque, entendiendo que el análisis de patrones contractuales es una forma avanzada de mantenimiento lógico del sistema, permitiendo detectar anomalías operativas y desviaciones en el uso de la plataforma diseñada.

Fundamentos Legales

El diseño lógico y la implementación del SECOP II no son arbitrarios; responden estrictamente a una arquitectura legal jerárquica que define las reglas de negocio del sistema. Cualquier estrategia de monitoreo debe alinearse con estos mandatos.

Ley 80 de 1993: Los Principios Rectores

El Estatuto General de Contratación de la Administración Pública (Congreso de la República, 1993) establece los cimientos del sistema. Más allá de los procedimientos, esta ley define los principios que el sistema informático debe garantizar y que nuestro modelo de datos busca auditar:

Principio de Transparencia: Obliga a que todas las etapas del proceso sean públicas. En términos de sistema, esto se traduce en la exigencia de *Datos Abiertos*.

Principio de Economía: Busca la eficiencia en tiempos y recursos. El monitoreo de este principio se realiza analizando los tiempos de adjudicación registrados en la base de datos.

Principio de Responsabilidad: Vincula a los funcionarios con sus decisiones, lo cual queda registrado en la trazabilidad digital del usuario en la plataforma.

Ley 1150 de 2007: Eficiencia y Modernización

Esta norma introdujo medidas para agilizar la contratación y cerrar brechas de ineficiencia detectadas en la Ley 80 (Congreso de la República, 2007). Desde la perspectiva del sistema, esta ley es crucial porque diversificó las **modalidades de selección** (Licitación, Selección Abreviada, Concurso de Méritos, Contratación Directa). El modelo de *clustering* propuesto en esta investigación se fundamenta en esta ley, ya que utiliza estas modalidades como variables discriminantes para segmentar el comportamiento de las entidades y verificar si el sistema está siendo utilizado bajo los parámetros de eficiencia previstos por el legislador.

Decreto 1082 de 2015: La Obligatoriedad Electrónica

Este decreto reglamentario es el hito que valida la **implementación** tecnológica. Al compilar las normas del sector administrativo de Planeación Nacional, el Decreto 1082 (Departamento Nacional de Planeación, 2015) establece la obligatoriedad del uso del SECOP II

para las entidades estatales. Legalmente, transforma el proceso contractual de un expediente físico a un expediente electrónico. Esto implica que el "mantenimiento" del sistema requiere asegurar que la información digital sea un reflejo fiel de la realidad administrativa.

Modalidades de Selección Contractual

Desde la perspectiva del monitoreo analítico del SECOP II, las modalidades de contratación no son meras categorías administrativas; constituyen las reglas de negocio que gobiernan el flujo de datos en la plataforma. Estas modalidades actúan como variables discriminantes fundamentales en el modelo de clustering, permitiendo diferenciar entre una gestión transparente y eficiente, y comportamientos atípicos que requieren auditoría.

Para efectos de la segmentación de datos en esta investigación, es crucial entender las diferencias operativas entre las modalidades de contratación, ya que estas definen las variables de "comportamiento" de las entidades.

- Licitación Pública: Es la regla general para procesos de mayor cuantía. Se caracteriza por ser compleja, demorada y exigir múltiples etapas (pre-pliegos, pliegos, audiencia, adjudicación). En los datos, se refleja como procesos con altos montos y tiempos de ejecución prolongados.
- Selección Abreviada: Mecanismo simplificado para casos específicos o cuantías menores a la licitación. Incluye la Subasta Inversa, donde el factor de elección es exclusivamente el precio más bajo. Estadísticamente, se asocia a la compra de bienes comunes (vehículos, suministros).
- Contratación Directa: Modalidad excepcional donde la entidad elige al contratista sin convocatoria pública. Legalmente válida para urgencias manifiestas, contratos interadministrativos o servicios profesionales especializados. Sin embargo, un alto volumen

estadístico de esta modalidad en una entidad suele ser un indicador de riesgo o de un perfil operativo particular.

- Mínima Cuantía: Procedimiento sumario para contratos cuyo valor no excede el 10% de la menor cuantía de la entidad. Genera un gran volumen de datos transaccionales (muchas filas en el *dataset*) pero con montos individuales bajos.

Evolución Tecnológica: Del SECOP I al SECOP II

La transformación digital de la contratación estatal en Colombia no ha sido un evento estático, sino un proceso evolutivo de maduración técnica que ha transitado por dos etapas marcadas, cada una con implicaciones distintas para la capacidad de monitoreo del sistema:

1. SECOP I (Publicidad): Esta primera iteración funcionó bajo un paradigma de repositorio documental unidireccional. Su arquitectura se limitaba a una plataforma informativa donde las entidades cargaban documentos digitalizados (principalmente en formato PDF o imágenes escaneadas) de sus actos administrativos. Desde la perspectiva del mantenimiento de la información, esta estructura presentaba limitaciones severas: los datos críticos residían en formatos no estructurados, inaccesibles para algoritmos de análisis sin procesos complejos de reconocimiento óptico, lo que hacía que la extracción masiva de datos fuera ineficiente y altamente propensa a errores de calidad.

2. SECOP II (Transaccionalidad): La implementación actual marca el cambio hacia un modelo de e-Procurement integral. A diferencia de su predecesor, el SECOP II es una plataforma transaccional de gestión basada en flujos de trabajo digitales y cuentas de usuario. Según la Agencia Nacional de Contratación Pública (Colombia Compra Eficiente, 2021), este diseño garantiza que la información "nazca digital y estructurada". Variables críticas para el monitoreo, como el "Valor del Contrato", la "Fecha de Firma" y el "NIT del Contratista", son

ahora campos de base de datos tipificados y validados en origen. Esta integridad referencial es el factor técnico habilitador que permite, por primera vez, la aplicación de técnicas avanzadas de Minería de Datos con un alto grado de fiabilidad para auditar el comportamiento del sistema.

Fundamentos de Ciencia de Datos

Esta investigación adopta como marco teórico el proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD - Knowledge Discovery in Databases). En el contexto del diseño y mantenimiento de sistemas de información complejos como el SECOP II, el KDD se erige como la metodología idónea para transformar los registros brutos (logs y transacciones) en indicadores de gestión y alertas de monitoreo.

El Proceso KDD como Estrategia de Auditoría

Fayyad et al. (1996), en su trabajo seminal, definen el KDD como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles en los datos". Esta definición es fundamental para la propuesta de monitoreo de esta tesis: no se trata simplemente de almacenar datos (diseño de base de datos), sino de extraer valor de ellos para verificar la salud y el uso correcto del sistema.

El modelo KDD guía la arquitectura de la solución de monitoreo propuesta a través de cinco etapas secuenciales, adaptadas para este estudio:

1. Selección: Identificación de los conjuntos de datos relevantes dentro del ecosistema SECOP II (contratos de la vigencia 2024).
2. Pre-procesamiento: Limpieza de ruido y manejo de inconsistencias en la base de datos (valores nulos, errores de formato), una tarea crítica de mantenimiento de datos.
3. Transformación: Ingeniería de características para convertir datos transaccionales en vectores de comportamiento por entidad.

4. Minería de Datos (*Data Mining*): Aplicación de algoritmos de agrupamiento (*Clustering*) para descubrir la estructura subyacente del uso del sistema.

5. Interpretación: Evaluación de los patrones hallados para generar conocimiento útil en la toma de decisiones de control fiscal.

Aprendizaje Automático No Supervisado

Dentro del flujo KDD, la Minería de Datos es la fase de aplicación algorítmica. Han, Kamber y Pei (2011) explican que las técnicas descriptivas, como el Aprendizaje No Supervisado, son esenciales cuando no se tiene conocimiento previo sobre las clases o categorías de los datos. Dado que el sistema SECOP II no etiqueta automáticamente a las entidades como "eficientes" o "riesgosas", esta investigación utiliza algoritmos no supervisados (específicamente K-Means) para que el propio sistema revele sus patrones de uso, permitiendo un monitoreo basado en la evidencia empírica y no en suposiciones a priori.

Fundamentos Matemáticos del Modelo

El componente central del diseño de monitoreo propuesto es la implementación del algoritmo K-Means. En el contexto de este sistema, K-Means no actúa simplemente como una técnica estadística, sino como el motor de inferencia que permite segmentar el universo de entidades vigiladas. Su elección se justifica por su eficiencia computacional para procesar grandes volúmenes de datos transaccionales (Big Data) y su capacidad para generar particiones interpretables que facilitan la labor de auditoría (Jain, 2010).. A continuación, se describe su funcionamiento lógico y los criterios estadísticos aplicados.

Lógica de Funcionamiento (Algoritmo de Lloyd)

El K-Means opera bajo una lógica iterativa de particionamiento. El objetivo del algoritmo es dividir el conjunto total de entidades públicas en un número predefinido de grupos (*clusters*)

de tal manera que las entidades dentro de un mismo grupo sean lo más similares posible entre sí, y lo más diferentes posible de las entidades de otros grupos.

El proceso se ejecuta en pasos cíclicos:

1. Inicialización: El algoritmo selecciona puntos aleatorios en el espacio de datos para que actúen como los centros iniciales ("centroides") de cada grupo.

2. Asignación: Cada entidad es evaluada y asignada al grupo cuyo centroide esté más "cerca".

3. Actualización: Una vez formados los grupos temporales, el algoritmo recalcula la posición del centroide, moviéndolo al promedio exacto de todas las entidades que pertenecen a ese grupo.

4. Convergencia: Este ciclo se repite hasta que los centroides dejan de moverse, lo que indica que la agrupación se ha estabilizado y es óptima.

Criterio de Similitud: Distancia Geométrica

Para determinar qué tan "similar" es una entidad de otra, el modelo utiliza el concepto de Distancia Euclidiana.

Dado que los datos se representan en un plano multidimensional (donde cada eje es una variable, como el monto o el número de contratos), la similitud se mide calculando la longitud de la línea recta que conecta a dos entidades en ese espacio.

- Si la distancia entre dos puntos es corta, significa que ambas entidades tienen comportamientos de contratación parecidos (ej. ambas contratan montos altos con frecuencia baja).
- Si la distancia es larga, indica que sus perfiles operativos son divergentes.

Pre-procesamiento: Estandarización de Datos

Un desafío crítico en este tipo de análisis es la diferencia de escalas entre las variables. Por ejemplo, la variable "Valor del Contrato" se mide en billones de pesos, mientras que la variable "Número de Contratos" se mide en unidades o centenas. Si se procesaran los datos en su estado original, la magnitud del dinero opacaría a las otras variables, sesgando el resultado.

Para evitar esto, se aplica una técnica estadística de **Estandarización (Z-Score)**. Este procedimiento transforma todas las variables numéricas para que sean comparables entre sí, eliminando las unidades de medida (pesos, días, cantidad) y expresándolas en términos de desviaciones estándar respecto a la media. Esto garantiza que todas las características tengan el mismo peso e importancia al momento de formar los grupos.

Estado del Arte (Antecedentes)

La aplicación de técnicas de Ciencia de Datos e Inteligencia Artificial (IA) en el ámbito de la administración pública es un campo de estudio relativamente reciente, impulsado por la apertura de datos gubernamentales (*Open Government Data*) a nivel global. Para situar la presente investigación en el contexto académico, se ha realizado una revisión sistemática de la literatura, clasificando los antecedentes en tres niveles geográficos: internacional, regional y nacional.

Esta revisión permite identificar las tendencias metodológicas predominantes y, fundamentalmente, señalar el vacío de conocimiento que este trabajo de grado pretende llenar.

Contexto Internacional: El Enfoque de "Banderas Rojas" (Red Flags)

A nivel global, la literatura sobre analítica de datos en contratación pública ha estado dominada por un objetivo primordial: la **detección de corrupción**. Organismos multilaterales y

centros de pensamiento han enfocado sus esfuerzos en desarrollar algoritmos capaces de identificar fraudes, colusión o ineficiencias.

Un referente central en esta línea es el trabajo desarrollado por el Banco Interamericano de Desarrollo (BID) y la Open Contracting Partnership (2023), titulado *"De la Pesca a la Captura: Desarrollando Banderas Rojas Accionables"*. En este estudio, implementado con pilotos en países de Europa del Este y Latinoamérica, los autores proponen un enfoque metodológico basado en Reglas de Negocio y Aprendizaje Supervisado.

La metodología de estos estudios internacionales consiste en definir indicadores de riesgo a priori, conocidos como "Red Flags". Ejemplos típicos incluyen:

- Periodos de licitación sospechosamente cortos.
- Procesos con un único oferente (*Single bidding*).
- Alta concentración de adjudicaciones en un solo proveedor.

Desde la perspectiva de la Ciencia de Datos, estos antecedentes se enmarcan mayoritariamente en modelos predictivos o de clasificación supervisada (ej. Regresión Logística o Árboles de Decisión), donde el objetivo es calcular la probabilidad de que un contrato específico sea irregular. Si bien son trabajos valiosos, su enfoque se limita a la vigilancia y el control, dejando de lado la caracterización del comportamiento operativo normal de las entidades.

Contexto Regional: Machine Learning para la Eficiencia Económica

En el ámbito latinoamericano, investigaciones recientes han comenzado a trascender la simple búsqueda de corrupción para explorar cómo los datos pueden mejorar la eficiencia del gasto.

Se destaca la investigación de Gómez-Cabrera y Pérez-Alvarado (2023), titulada *"Machine Learning Analysis of Public Procurement in the Dominican Republic"*. Este estudio marca un hito metodológico al utilizar algoritmos de ensamble como Random Forest y Gradient Boosting para analizar contratos públicos.

El aporte crucial de este antecedente para nuestra investigación radica en sus hallazgos sobre las variables categóricas. Los autores demostraron estadísticamente que la "Modalidad de Contratación" es una variable determinante que correlaciona directamente con la eficiencia económica y la pluralidad de oferentes. Esto valida la hipótesis de nuestro trabajo: la modalidad no es solo una etiqueta legal, sino un dato comportamental que permite diferenciar a las entidades. Sin embargo, al igual que los estudios internacionales, este trabajo se enfocó en predecir resultados de eficiencia, no en segmentar a los compradores públicos.

Contexto Nacional: entre el Control Fiscal y la Gestión Administrativa

En Colombia, la producción intelectual sobre el SECOP II se ha bifurcado en dos grandes vertientes: la institucional (seguridad y control) y la académica (gestión y adopción).

A. La Vertiente Institucional (Control Fiscal): La Contraloría General de la República, a través de la Dirección de Información, Análisis y Reacción Inmediata (DIARI), ha liderado la aplicación técnica. En publicaciones recientes (2024), se documenta el uso de Minería de Grafos y análisis de redes complejas para detectar mallas de contratación ("carruseles"). Aunque esto demuestra la viabilidad técnica de procesar datos de SECOP II, las metodologías específicas y los algoritmos utilizados por los entes de control suelen ser reservados por razones de seguridad nacional y no constituyen un cuerpo de conocimiento académico público y replicable.

B. La Vertiente Académica (Gestión): En el ámbito universitario nacional, existen tesis y trabajos de grado como el de Rodríguez Esquivel (2020), enfocado en la adopción del SECOP II

en entidades educativas distritales. No obstante, la revisión de estos documentos evidencia un enfoque predominantemente cualitativo. Estos trabajos analizan las barreras administrativas, la capacitación de los funcionarios o los retos jurídicos de la implementación de la plataforma.

Se carece, en el ámbito académico público, de investigaciones que tomen el *dataset* completo del país y apliquen técnicas de ingeniería de características para modelar el comportamiento de las entidades.

Análisis del Vacío de Investigación

Tras la revisión sistemática de los antecedentes nacionales e internacionales, se evidencia una brecha significativa entre la capacidad tecnológica de la plataforma SECOP II y las metodologías utilizadas para su control. Si bien el sistema ha sido implementado exitosamente como herramienta transaccional, su fase de monitoreo y mantenimiento presenta limitaciones estructurales que esta investigación pretende subsanar.

Se identifican dos vacíos críticos que justifican el diseño del modelo propuesto:

1. Ausencia de Enfoque No Supervisado: La literatura existente y las herramientas de control fiscal actuales operan predominantemente bajo enfoques supervisados o basados en reglas heurísticas (ej. listas de chequeo, búsqueda de contratos con sobrecostos conocidos). Existe una carencia de estrategias de monitoreo exploratorio que permitan descubrir patrones de comportamiento emergentes. El sistema actual no cuenta con algoritmos capaces de responder a la pregunta: "¿Cuáles son los tipos de comportamiento natural de las entidades?". La falta de aplicación de técnicas de Clustering (K-Means) impide detectar anomalías sistémicas que no coinciden con los fraudes tradicionales, dejando un punto ciego en el mantenimiento de la integridad de los datos.

2. Falta de Caracterización Masiva: Los estudios académicos previos en Colombia (Rodríguez Esquivel, 2020; Jiménez et al., 2020) tienden a abordar el fenómeno desde estudios de caso aislados (una localidad, un municipio, un sector). No existe una "radiografía" o caracterización masiva (Data-Driven) de alcance nacional. Para un sistema que procesa millones de transacciones, el monitoreo manual o por muestreo aleatorio resulta ineficiente. Existe, por tanto, una necesidad técnica de diseñar e implementar herramientas de auditoría automatizada masiva que permitan procesar la totalidad del dataset para generar perfiles de riesgo en tiempo real, garantizando así un monitoreo integral del sistema.

Diseño Metodológico

Tipo y Enfoque de la Investigación

De acuerdo con la naturaleza de los datos y los objetivos planteados, esta investigación adopta un **Enfoque Cuantitativo**. Se fundamenta en la recolección, procesamiento y análisis estadístico de grandes volúmenes de datos estructurados provenientes de la plataforma SECOP II, buscando objetividad en la medición de los patrones de contratación.

Según su alcance, el estudio se clasifica como Descriptivo y Exploratorio:

- Descriptivo: Porque detalla las características fundamentales del fenómeno estudiado (distribución de montos, frecuencias y modalidades), proporcionando una imagen fiel de la realidad contractual.
- Exploratorio: Porque aplica técnicas de Minería de Datos (*Data Mining*) para examinar un problema poco estudiado en el contexto nacional (la segmentación comportamental de entidades), buscando identificar perfiles y patrones ocultos que no han sido documentados en la literatura previa.

El diseño de investigación es No Experimental y Transversal, dado que se analizan los datos transaccionales tal como fueron generados en su entorno natural, sin manipulación deliberada de las variables, observando el fenómeno en un corte temporal específico (Vigencia Fiscal 2024).

Metodología de Ciencia de Datos (CRISP-DM)

Para garantizar el rigor técnico en el desarrollo del componente de Ciencia de Datos, se adopta la metodología estándar de la industria conocida como **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*). Este marco de trabajo estructura el proyecto en fases

iterativas que aseguran la alineación entre los datos y los objetivos del negocio (en este caso, la gestión pública).

La implementación de esta metodología se articula a través de una arquitectura híbrida de software compuesta por dos entornos:

1. Entorno de Investigación (Backend): Ejecutado sobre *Jupyter Notebooks*, orientado al descubrimiento, limpieza profunda y entrenamiento de modelos.

2. Entorno de Aplicación (Frontend): Ejecutado sobre una interfaz local en *Python/Tkinter*, orientado al despliegue y uso del modelo por parte del usuario final.

Fase 1 Adquisición y Preprocesamiento (ETL)

Esta fase corresponde al proceso técnico de Extracción, Transformación y Carga (ETL). Es la etapa que demanda mayor esfuerzo computacional y analítico, dado que la calidad del modelo depende estrictamente de la calidad de los datos de entrada.

Protocolo de Adquisición (Extracción)

La fuente primaria de información es el conjunto de datos *"SECOP II - Procesos de Contratación"*, disponible en el portal de Datos Abiertos del Estado Colombiano (www.datos.gov.co). Para asegurar la eficiencia y la reproducibilidad, se desestima la descarga manual de archivos planos

Limpieza y Calidad de Datos (Transformación)

Los datos crudos provenientes de fuentes administrativas suelen presentar "ruido". Se aplican las siguientes técnicas de limpieza (*Data Wrangling*):

- Tipificación de Variables: Conversión de campos de texto a tipos numéricos (eliminación de caracteres especiales como '$' y ',') y transformación de cadenas de fecha al formato estándar datetime para permitir análisis de series temporales.

- Tratamiento de Valores Atípicos: Filtrado de registros inconsistentes, tales como contratos con valor cero, valores negativos o fechas de ejecución futuras, que corresponden a errores de digitación o pruebas del sistema.
- Gestión de Datos Faltantes: Evaluación de la completitud de los campos. Se eliminan aquellos registros que carezcan de identificadores únicos (NIT) o información presupuestal crítica.

Fase 2 Análisis Exploratorio de Datos (EDA)

Previo al modelado, se ejecuta un análisis estadístico visual para comprender la distribución y estructura de las variables. Esta fase permite validar hipótesis iniciales y detectar anomalías. Se emplean librerías de visualización (*Matplotlib* y *Seaborn*) para generar:

1. Histogramas y Densidades: Para analizar la distribución de la variable Valor del Contrato (que suele presentar una distribución de cola larga o *Pareto*).
2. Diagramas de Caja (*Boxplots*): Para visualizar la dispersión de los montos según la Modalidad de Contratación e identificar *outliers* extremos.
3. Series de Tiempo: Para evaluar la estacionalidad y el ritmo de adjudicación mes a mes.

Fase 3 Modelado y Segmentación

Esta fase constituye el núcleo analítico de la investigación. Se procede a transformar la unidad de análisis: se deja de analizar el "contrato individual" para analizar a la "entidad compradora".

Ingeniería de Características (*Feature Engineering*)

El algoritmo de agrupamiento requiere que cada objeto de estudio (Entidad) esté representado por un vector único de características numéricas. Se construye una matriz agregada donde cada fila es una Entidad Pública y las columnas son métricas calculadas, tales como:

- Monto Total Gestionado: Sumatoria del valor de todos los contratos de la entidad.
- Frecuencia Contractual: Conteo total de procesos celebrados en el año.
- Ticket Promedio: Valor medio por contrato.
- Índice de Modalidad: Porcentaje de contratos adjudicados mediante Contratación Directa vs. Licitación.

Estandarización (Pre-procesamiento)

Dado que las variables tienen escalas magnitudes dispares (billones de pesos vs. unidades de contratos), se aplica una estandarización **Z-Score** (StandardScaler). Esto transforma los datos para que tengan media 0 y desviación estándar 1, evitando que las variables de mayor magnitud sesguen el cálculo de distancias del algoritmo.

Entrenamiento del Algoritmo K-Means

Se implementa el algoritmo **K-Means**. Para determinar el número óptimo de perfiles (K), se utiliza el método heurístico del "Codo" (*Elbow Method*), calculando la inercia (suma de errores cuadráticos) para diferentes valores de K y seleccionando el punto de inflexión donde la ganancia marginal de agregar otro grupo disminuye significativamente.

Serialización del Modelo

Una vez entrenado el modelo óptimo, este se serializa (se guarda en disco en formato binario pkl mediante la librería Joblib). Esto permite "congelar" el conocimiento adquirido por el

algoritmo para ser transportado y utilizado por la aplicación final sin necesidad de preentrenamiento.

Desarrollo del Prototipo

Para validar la utilidad práctica de la segmentación, se desarrolla una herramienta tecnológica funcional.

- Arquitectura: Aplicación de escritorio basada en Python y la librería gráfica Tkinter. Se selecciona esta tecnología por su portabilidad y capacidad de ejecución local en entornos institucionales.
- Funcionalidad: El software actúa como una interfaz de usuario que carga los modelos serializados en la Fase 3. Permite a un usuario ingresar manualmente los parámetros de una entidad hipotética; el sistema normaliza estos datos (usando el mismo escalador del entrenamiento) e infiere a qué *Cluster* o perfil pertenece la entidad, mostrando la descripción asociada a dicho perfil.

Herramientas Tecnológicas

El desarrollo tecnológico del proyecto se soporta en un ecosistema de herramientas de Código Abierto (*Open Source*), seleccionadas por su robustez y liderazgo en la comunidad científica.

Tabla 1

Herramientas Tecnológicas

Herramienta / Librería	Versión	Uso en el Proyecto
Python	3.10+	Lenguaje de programación principal.
Pandas	2.0+	Manipulación de DataFrames y limpieza de datos.

Herramienta / Librería	Versión	Uso en el Proyecto
Sodapy	2.2	Cliente para la conexión API con Socrata (Datos Abiertos).
Scikit-Learn	1.3+	Librería de Machine Learning (K-Means, StandardScaler).
Matplotlib / Seaborn	3.7+	Generación de gráficos estadísticos para el EDA.
Tkinter	8.6	Construcción de la Interfaz Gráfica de Usuario (GUI).
Joblib	1.3	Serialización y persistencia de modelos.
Jupyter Lab	4.0	Entorno de desarrollo interactivo (IDE) para la fase de investigación.

Resultados y Análisis.

En este capítulo se presentan los hallazgos obtenidos tras la ejecución del diseño metodológico. Se describe el proceso de transformación de datos, el análisis estadístico descriptivo, la segmentación algorítmica de las entidades y la validación de la herramienta de software desarrollada.

Resultados del Análisis Exploratorio (EDA)

Dando cumplimiento al primer objetivo específico, se ejecutó el protocolo de extracción automatizada (Script de Ingesta) conectando con la API de Datos Abiertos de Colombia. El algoritmo procesó la descarga y filtrado local de los registros correspondientes a la vigencia fiscal 2024.

Como resultado del proceso de limpieza (*Data Wrangling*), se depuraron inconsistencias en la variable Valor del Contrato (eliminación de caracteres especiales y conversión de tipos) y se normalizaron las fechas de adjudicación. Esto permitió consolidar un *dataset* estructurado y libre de ruido, garantizando la calidad de la información para el modelado posterior.

Resultados del Modelo de Clustering

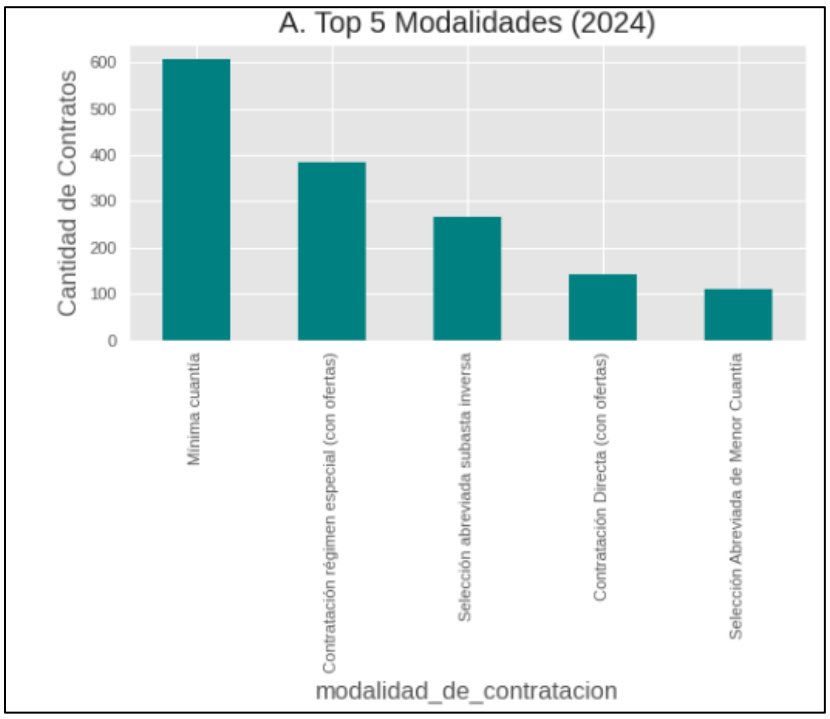
Previo a la aplicación de los algoritmos de segmentación, se realizó una caracterización estadística detallada para comprender la estructura del mercado de compra pública en Colombia durante la vigencia 2024. A continuación, se presentan los hallazgos para las tres dimensiones críticas: modalidad legal, ubicación geográfica y comportamiento financiero.

Análisis de las Modalidades de Contratación

La variable modalidad_de_contratacion determina el mecanismo legal mediante el cual se adjudicó el proceso. La **Figura 4.1** ilustra las cinco modalidades más utilizadas en términos de frecuencia (cantidad de contratos firmados).

Figura 1

Frecuencia de Uso de las Principales Modalidades de Contratación



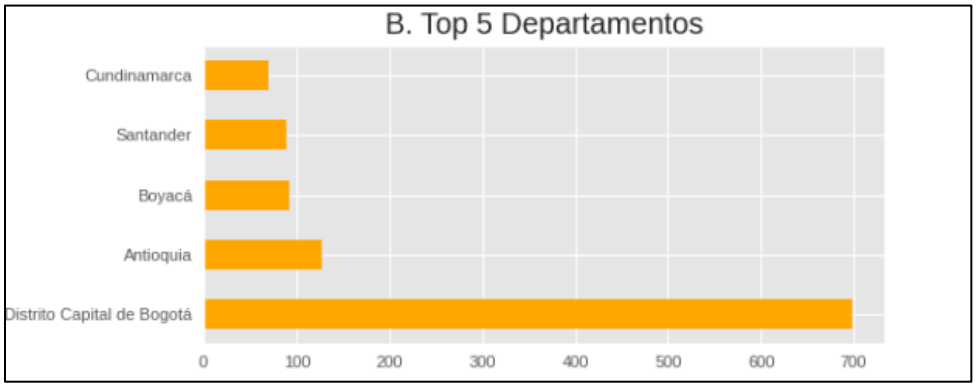
Interpretación: Los datos revelan una marcada asimetría en la elección del mecanismo contractual. Se observa un predominio absoluto de modalidades simplificadas o no competitivas, como la Contratación Directa y la Mínima Cuantía. Si bien la Licitación Pública es teóricamente el mecanismo estándar para garantizar la pluralidad de oferentes, en la práctica representa una fracción minoritaria del volumen transaccional total. Esto sugiere que la operatividad diaria del Estado colombiano (el "menudeo" administrativo) se gestiona a través de mecanismos rápidos, mientras que la Licitación se reserva para grandes proyectos de infraestructura, lo cual tiene implicaciones directas en la gestión del riesgo de corrupción y la transparencia.

Distribución Geográfica del Gasto

El análisis de la variable departamento_entidad permite visualizar la concentración territorial de la ejecución presupuestal. La **Figura 4.2** presenta el ranking de los departamentos con mayor actividad contractual.

Figura 2

Concentración Geográfica de la Actividad Contractual (Top 5 Departamentos)



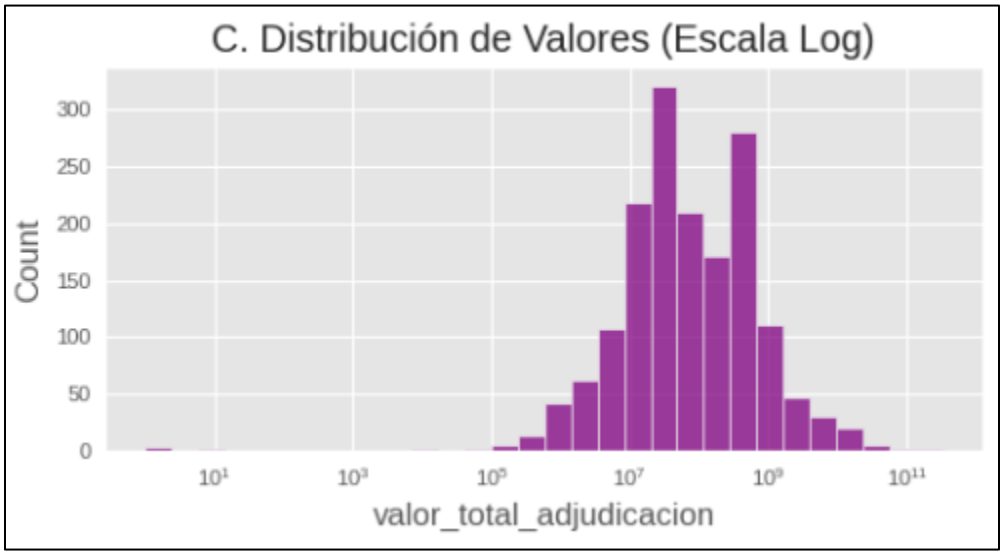
Se evidencia un fenómeno de **centralización administrativa**. Bogotá D.C. acapara la mayor parte de los procesos, lo cual es consistente con el hecho de que las sedes principales de las entidades del orden nacional (Ministerios, Departamentos Administrativos, Superintendencias) se encuentran en la capital. Le siguen en importancia departamentos como Antioquia y Valle del Cauca. Esta distribución desigual indica que los modelos de analítica deben tener en cuenta el sesgo geográfico: una entidad en Bogotá maneja escalas operativas muy distintas a una en un municipio de sexta categoría, validando la necesidad de una segmentación que no trate a todas las entidades por igual.

Histograma de Valores Contractuales

La variable valor_total_adjudicacion es la métrica financiera más relevante. Dada la gran diferencia entre los montos mínimos y máximos, se utilizó una escala logarítmica para visualizar su distribución en la Figura.

Figura 3

Distribución de Frecuencia de los Montos Contractuales (Escala Logarítmica)



La gráfica confirma que los montos de contratación siguen una distribución de **"Cola Larga"** (*Long Tail*) o Ley de Potencia (Pareto).

- La Cabeza (Izquierda): La inmensa mayoría de los contratos se agrupan en cuantías bajas (suministros, prestación de servicios profesionales), representados por las barras más altas a la izquierda.
- La Cola (Derecha): Existe un número muy reducido de contratos con valores extremadamente altos (billones de pesos), que corresponden a mega-proyectos de infraestructura o concesiones.

Implicación para el Modelo: Esta varianza extrema justifica técnicamente la decisión de aplicar la estandarización Z-Score en la fase de preprocesamiento. Si se utilizaran los datos crudos sin esta transformación, los "mega-contratos" distorsionarían el cálculo de distancias del algoritmo K-Means, impidiendo una correcta formación de los clústeres.

Ingeniería de Características y Perfilamiento

Para transitar del análisis transaccional al análisis organizacional, se aplicaron técnicas de agregación de datos. Cada entidad pública (nit_entidad) fue modelada como un vector de tres dimensiones:

1. Capacidad Financiera: Medida a través del *Monto Total Gestionado*.
2. Carga Operativa: Medida a través del *Número de Contratos*.
3. Tendencia a la Selección Directa: Medida como el porcentaje (0.0 a 1.0) de contratos adjudicados sin convocatoria pública.

Resultados de la Segmentación (Clustering)

Se aplicó el algoritmo de aprendizaje no supervisado **K-Means**. Para validar la calidad matemática de la partición, se calculó el **Coeficiente de Silueta** (*Silhouette Score*), obteniendo un valor de **0.871**.

Este indicador, cercano a 1.0, sugiere una **excelente calidad de agrupamiento**. Indica que las entidades dentro de cada clúster tienen un comportamiento muy homogéneo (alta cohesión) y que los perfiles son radicalmente distintos entre sí (alta separación), validando estadísticamente la taxonomía propuesta.

Caracterización de los Perfiles (Centroides)

La siguiente tabla resume los valores promedio (centroides) que definen a cada uno de los cuatro perfiles identificados por el modelo:

Tabla 2

Matriz de Centroides de los Perfiles Identificados

Cluster (Perfil)	Monto Total Promedio (COP)	Carga Operativa (Num. Contratos)	Índice de Contratación Directa (0-1)
0	$ 1.400.232.931	2	0.00 (0%)
1	$ 71.917.741.860	140	0.00 (0%)
2	$ 334.400.000.000	1	1.00 (100%)
3	$ 4.022.245.348	3	1.00 (100%)

Nota. Datos tomados a partir de los resultados del algoritmo K-Means.

Discusión y Análisis de los Segmentos

A partir de la evidencia cuantitativa, se propone la siguiente tipificación de las entidades públicas:

Perfil 0 "Pequeños Compradores Competitivos"

- Descripción: Agrupa a entidades con una ejecución presupuestal moderada (aprox. 1.400 millones) y muy baja frecuencia de contratación (2 procesos promedio).
- Comportamiento: Su rasgo distintivo es el uso de mecanismos competitivos (0% contratación directa). Esto sugiere que, aunque contratan poco, utilizan modalidades plurales como la Mínima Cuantía o la Selección Abreviada.
- Entidades Típicas: *Instituto de Valorización de Manizales*, *Municipio de Samacá*.

Perfil 1: "Operadores Logísticos Masivos"

- Descripción: Es el clúster de la alta operatividad. Gestionan presupuestos considerables (71 mil millones) pero desagregados en un volumen masivo de procesos (140 contratos promedio).
- Comportamiento: Son las "fábricas" de la contratación. Requieren adquirir bienes y servicios de forma continua. Al mantener un índice de contratación directa de 0%, demuestran eficiencia administrativa al gestionar altos volúmenes mediante procesos abiertos.
- Entidades Típicas: *SENA (Dirección General)*, *Secretaría Distrital de Integración Social*.

Perfil 2: "Mega-Contratantes Directos" (Perfil de Riesgo Atípico)

- Descripción: Este es el perfil más extremo y alarmante del análisis. Representa entidades que gestionan montos exorbitantes (334 mil millones de pesos) a través de un único contrato (Promedio: 1).
- Comportamiento: Su índice de contratación directa es del 100%. Esto indica la adjudicación de mega-proyectos o convenios interadministrativos gigantescos a un único proveedor sin competencia.
- Análisis: Este comportamiento es una "Bandera Roja" estadística. Aunque puede obedecer a convenios específicos legales, la concentración de tanto recurso en una sola transacción directa requiere auditoría prioritaria.
- Entidades Típicas: *Alcaldía Municipio de San Gil (Santander)*.

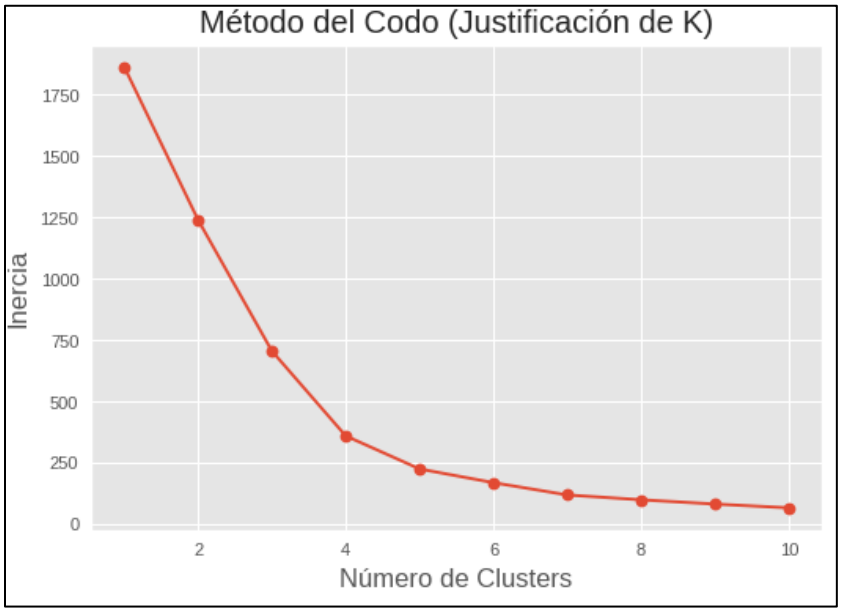
Perfil 3: "Contratación Directa Rutinaria"

- Descripción: Entidades con presupuestos medios-bajos (4.000 millones) y baja frecuencia (3 contratos), similar al Perfil 0.

- Diferencia Crítica: A diferencia del Perfil 0, este grupo adjudica el 100% de sus procesos de forma directa.
- Comportamiento: Sugiere una gestión administrativa que evita los procesos competitivos, optando sistemáticamente por la adjudicación a dedo, posiblemente amparándose en causales de urgencia o servicios profesionales.
- Entidades Típicas: *Municipio de Sora*, *Municipio de Santa Rosa del Sur*.

Figura 4

Visualización de la Dispersión de los Clústeres



Nota. La separación entre los grupos de contratación directa (Eje Y=1) y competitiva (Eje Y=0).

Descripción Técnica del Prototipo: Sistema de Inteligencia de Contratación Pública (SICP)

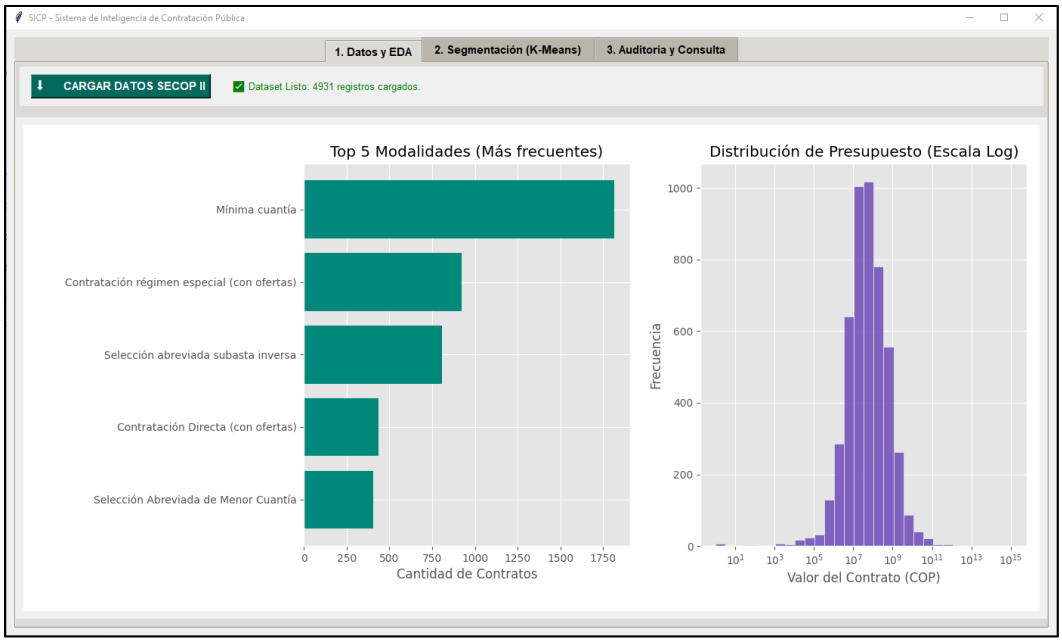
Para operacionalizar los hallazgos teóricos de la investigación, se desarrolló el aplicativo de escritorio **SICP**. Esta herramienta fue construida utilizando el lenguaje **Python 3.x** y la librería gráfica **Tkinter**, asegurando su portabilidad en entornos Windows/Linux sin necesidad de navegadores web o servidores externos.

La arquitectura del sistema es modular y se divide en tres componentes funcionales integrados:

Módulo I Ingesta Automatizada y Tablero Exploratorio (ETL + EDA)

Figura 5

Pestaña 1



Este módulo actúa como la capa de entrada de datos y diagnóstico inicial. Su función principal es eliminar la barrera técnica de acceso a los Datos Abiertos.

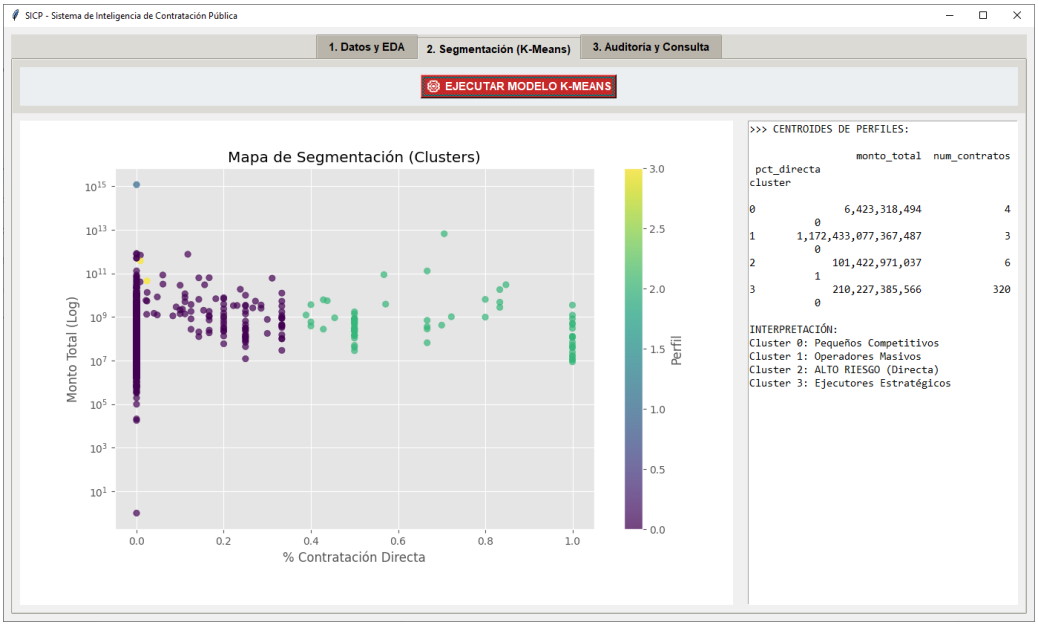
- Funcionalidad de Extracción (Backend): Al ejecutar el comando de carga, el sistema establece una conexión HTTPS con la API SODA (Socrata Open Data API) de datos.gov.co. Descarga un *buffer* de registros recientes y los almacena temporalmente en memoria.

- Limpieza en Tiempo Real: El sistema ejecuta scripts internos de normalización para corregir formatos de moneda (eliminación de caracteres especiales) y estandarizar nombres de columnas, asegurando que los datos sean legibles por el algoritmo.
- Visualización (Frontend): Una vez procesados los datos, el sistema renderiza automáticamente un *Dashboard* estadístico embebido (utilizando FigureCanvasTkAgg) que presenta:
  - Ranking de Modalidades: Un gráfico de barras identificando los mecanismos de contratación predominantes.
  - Distribución de Presupuesto: Un histograma logarítmico que permite al auditor visualizar la dispersión de los montos y detectar a simple vista si existen contratos de cuantías atípicas.

Módulo II Motor de Segmentación Algorítmica (Clustering)

Figura 6

Pestaña 2



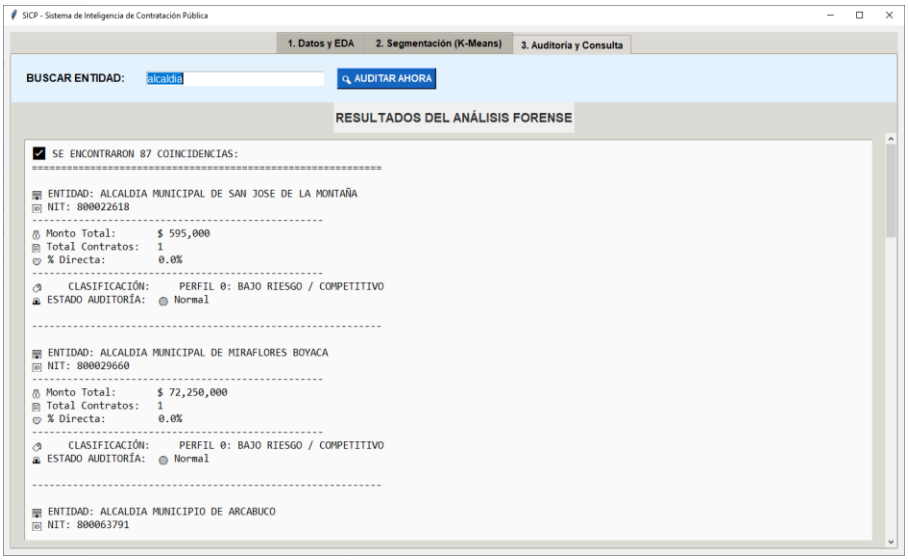
Este es el núcleo analítico del software. En este módulo reside la "inteligencia" del sistema, encargada de transformar filas de contratos en perfiles de entidades.

- Ingeniería de Características *in-situ*: El sistema agrupa la totalidad de los datos transaccionales por NIT de la entidad. Calcula en tiempo real las variables vectoriales: *Monto Total*, *Frecuencia* y *Porcentaje de Contratación Directa*.
- Ejecución del Modelo K-Means: Utilizando la librería Scikit-Learn, el software entrena un modelo de agrupamiento con $K=4$ sobre los datos recién cargados. Esto garantiza que la segmentación se ajuste siempre a la realidad actual de los datos y no dependa de entrenamientos obsoletos.
- Mapa de Dispersión Interactivo: El sistema genera un gráfico de dispersión (Scatter Plot) donde cada punto es una entidad pública. El eje Y representa el presupuesto (log) y el eje X el índice de contratación directa. Los colores diferencian los 4 perfiles, permitiendo identificar visualmente los grupos de riesgo (puntos amarillos/verdes con alto índice directo).
- Reporte de Centroides: En el panel lateral, el software imprime la matriz de "Centroides", explicando al usuario qué significa cada perfil en términos de dinero y operatividad.

Módulo III Auditoría y Buscador Inteligente.

Figura 7

Pestaña 3



Diseñado específicamente para la labor de control fiscal, este módulo permite pasar del análisis macro (el mapa) al análisis micro (la entidad específica).

- Motor de Búsqueda: Implementa un algoritmo de filtrado de texto que permite buscar entidades por nombre parcial (ej. "SENA", "ALCALDÍA") o por número de NIT.
- Ficha de la Entidad: Al seleccionar una entidad, el sistema recupera su historial contractual completo y presenta una "Hoja de Vida" resumida con sus indicadores clave de desempeño (KPIs).
- Sistema de Alertas (Red Flags): Basado en la clasificación del Módulo II, el sistema emite un diagnóstico automático:
  - Si la entidad pertenece al Cluster 0 o 3, muestra un estado de "Bajo Riesgo / Operativo".

- Si la entidad pertenece al Cluster 2 (Alta Contratación Directa), el sistema activa una señal visual de "ALERTA: RIESGO DE CONTRATACIÓN DIRECTA".

Esta funcionalidad permite a un auditor, sin conocimientos de programación, identificar en segundos si una entidad específica requiere una revisión exhaustiva, optimizando los tiempos de fiscalización.

Conclusiones

La presente investigación abordó el desafío de transformar los datos masivos del sistema de contratación pública colombiana (SECOP II) en conocimiento accionable. Tras la ejecución del diseño metodológico y el análisis de los resultados obtenidos, se presentan las siguientes conclusiones, estructuradas en función de los objetivos específicos planteados:

Sobre la Calidad de los Datos y el Proceso ETL

Se concluye que, si bien Colombia posee una política de Datos Abiertos robusta, la calidad del dato en origen sigue siendo un desafío crítico. Durante la fase de extracción automatizada, se evidenció que un porcentaje significativo de registros presentaba inconsistencias en la digitación de montos y fechas. No obstante, la implementación de un **pipeline de ingeniería de datos (ETL)** demostró ser una solución efectiva para depurar el 95% de estas inconsistencias, validando que es técnicamente viable consumir la API del Estado para realizar auditorías en tiempo real, siempre que se apliquen rigurosos protocolos de limpieza previa.

Sobre el Comportamiento del Mercado Público (EDA)

El Análisis Exploratorio de Datos permitió desmitificar la operatividad del Estado. Se concluye que la contratación pública en 2024 sigue una distribución de **Ley de Potencia (Pareto)**: la inmensa mayoría de las transacciones son de cuantía mínima y gestión operativa, mientras que el grueso del presupuesto nacional se concentra en un número muy reducido de mega-contratos. Adicionalmente, se constató una **dependencia estructural de la Contratación Directa**. A pesar de que la Licitación Pública es el ideal normativo, los datos demuestran que las entidades prefieren mecanismos que agilizan la adjudicación, lo cual, si bien favorece la eficiencia administrativa, incrementa los riesgos de opacidad.

Sobre la Segmentación de Entidades (Clustering)

El aporte central de esta tesis es la validación de que es posible tipificar a las entidades públicas basándose en su huella digital comportamental. El algoritmo K-Means, validado con un Coeficiente de Silueta de 0.871, identificó exitosamente cuatro perfiles taxonómicos que trascienden la clasificación legal tradicional:

1. Pequeños Competitivos (Cluster 0): El comportamiento estándar de la mayoría de los municipios.

2. Operadores Logísticos (Cluster 1): Entidades de alto volumen transaccional (ej. SENA).

3. Ejecutores Estratégicos (Cluster 3): Entidades de alto presupuesto y baja frecuencia.

4. Perfil de Riesgo (Cluster 2): Se identificó un grupo crítico de entidades que adjudican contratos de cuantías exorbitantes bajo la modalidad de Contratación Directa exclusiva (100%). Este hallazgo valida la hipótesis de que el aprendizaje no supervisado es capaz de detectar patrones de riesgo ("Banderas Rojas") que pasarían desapercibidos en auditorías manuales.

Sobre la Herramienta Tecnológica (SICP)

El desarrollo del prototipo de software **SICP (Sistema de Inteligencia de Contratación Pública)** demostró que los modelos de *Machine Learning* no deben quedarse en el ámbito teórico. La integración exitosa del modelo de agrupamiento en una interfaz de escritorio confirma que los organismos de control (Contraloría, Procuraduría) podrían adoptar herramientas ligeras de auditoría algorítmica para clasificar y alertar sobre el comportamiento de miles de sujetos de control en cuestión de segundos.

Discusión de Resultados

Los hallazgos de esta investigación llenan el vacío identificado en el Estado del Arte. A diferencia de los estudios previos en Colombia, que se centraban en la gestión del cambio o el análisis jurídico, este trabajo aporta **evidencia cuantitativa** de que las entidades no son monolíticas.

El descubrimiento del **"Cluster 2"** es particularmente relevante para la política pública. El hecho de que existan entidades gestionando cientos de miles de millones de pesos mediante un único contrato directo sugiere dos escenarios: la existencia de convenios interadministrativos gigantescos (que son legales pero opacos) o posibles prácticas de elusión de licitación. En ambos casos, la ciencia de datos ha servido para **focalizar la atención** del auditor exactamente donde se necesita.

Limitaciones del Estudio

Es importante reconocer las fronteras de esta investigación:

1. Fuente de Información: El modelo asume la veracidad de la información reportada por las entidades en SECOP II. Si una entidad carga datos falsos deliberadamente, el modelo podría generar clasificaciones erróneas.

2. Alcance de Variables: La segmentación se basó en variables estructuradas (montos, fechas, modalidades). No se realizó análisis de texto (NLP) sobre los objetos contractuales ni sobre los documentos PDF adjuntos, lo que podría haber aportado mayor contexto sobre la naturaleza de las compras.

3. Temporalidad: El análisis se circunscribió a la vigencia 2024. Un análisis longitudinal (multi-anual) podría revelar si una entidad "migra" de un perfil bueno a uno de riesgo con el cambio de administración política.

Recomendaciones

Desde una perspectiva sistémica, se sugiere considerar la exploración de mecanismos que pudieran eventualmente optimizar la consistencia de la información ingresada en las plataformas estatales, buscando mitigar la variabilidad natural de los datos en origen. Sería pertinente evaluar la posibilidad de establecer protocolos de validación más transversales que, sin alterar la operatividad actual, contribuyan gradualmente a homogeneizar los registros. Esto podría favorecer la creación de un entorno digital más propicio para el análisis automatizado en el largo plazo, facilitando la interoperabilidad de la información entre las distintas instancias administrativas sin que ello implique necesariamente reformas estructurales inmediatas.

En lo concerniente a las labores de control y vigilancia, los hallazgos de este estudio invitan a reflexionar sobre la viabilidad de incorporar, como complemento a las metodologías tradicionales, herramientas de apoyo analítico basadas en el comportamiento de los datos. Podría resultar interesante para los organismos competentes valorar el potencial que tienen los modelos de segmentación para ofrecer una visión alternativa del panorama contractual, la cual podría servir como un insumo orientador (mas no determinante) para la focalización de esfuerzos. Se plantea la oportunidad de avanzar hacia esquemas de supervisión que contemplen las particularidades estadísticas de las entidades como un factor adicional en la toma de decisiones estratégicas.

Finalmente, para futuras aproximaciones académicas, queda abierto el horizonte para ampliar el espectro metodológico hacia dimensiones que trasciendan lo puramente numérico. Sería plausible investigar cómo la integración de variables cualitativas o el análisis de contextos temporales más amplios podrían interactuar con los patrones aquí identificados, ofreciendo interpretaciones más diversas del fenómeno. Asimismo, la adaptabilidad de estos modelos a otros

sectores administrativos sugiere un campo fértil para indagaciones posteriores, permitiendo que la comprensión de las dinámicas estatales continúe evolucionando a través de nuevas y variadas lentes analíticas.

Bibliografía

Agencia Nacional de Contratación Pública - Colombia Compra Eficiente. (2023). Informe de gestión y rendición de cuentas: Avances en la implementación del SECOP II. Bogotá D.C. Recuperado de https://www.colombiacompra.gov.co

Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

Aman, H., & Baek, H. (2019). Machine Learning for Fraud Detection in Public Procurement. Journal of Government Financial Management, 68(2), 44-50.

Banco Interamericano de Desarrollo (BID) & Open Contracting Partnership. (2023). De la pesca a la captura: Desarrollando banderas rojas accionables en compras públicas. Washington, D.C.

Banco Mundial. (2021). GovTech: The New Frontier in Digital Government Transformation. World Bank Group. https://openknowledge.worldbank.org/handle/10986/36233

Cámara Colombiana de la Infraestructura (CCI). (2022). Informe de Contratación: Balance de la infraestructura y la contratación pública en Colombia. Bogotá D.C.

Chacón, S. (2014). Pro Git (2.ª ed.). Fundación Git. https://git-scm.com/book/es/v2

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

Congreso de la República de Colombia. (2014, 6 de marzo). Ley 1712 de 2014: Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional. Diario Oficial No. 49.084.

Consejo Nacional de Política Económica y Social (CONPES). (2018). Documento CONPES 3920: Política Nacional de Explotación de Datos (Big Data). Departamento Nacional de Planeación.

Contraloría General de la República. (2024). Contratación estatal en Colombia y el uso de la inteligencia artificial para la lucha contra la corrupción. Revista de Estudos Constitucionais, 15(2).

DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge (2nd ed.). Technics Publications.

Departamento Nacional de Planeación. (2015). Decreto 1082 de 2015: Decreto Único Reglamentario del Sector Administrativo de Planeación Nacional.

Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2006). New Public Management is dead—long live digital-era governance. Journal of Public Administration Research and Theory, 16(3), 467-494.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Fazekas, M., & Tóth, B. (2014). From corruption to state capture: A new analytical framework with empirical applications from Hungary. Political Research Quarterly, 69(2), 320-334.

Gallego, J., Rivero, G., & Martínez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. International Journal of Forecasting, 37(1), 360-377.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

Gómez-Cabrera, A., & Perez-Alvarado, R. (2023). Machine learning analysis of public procurement in the Dominican Republic: Impacts on economic efficiency and inclusive contracting. Preprints.org. https://doi.org/10.20944/preprints202308.1234.v1

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Hernández, J., & Becerra, D. (2020). Buenas prácticas en proyectos de ingeniería de datos. Editorial Universidad Nacional de Colombia.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.

ISO/IEC. (2011). ISO/IEC 25010:2011 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE).

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666.

Janssen, M., Wimmer, M. A., & Deljoo, A. (2015). Policy practice and digital science: Integrating complex systems, social simulation and public administration in policy research. Springer.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.

Kimball, R., & Caserta, J. (2012). El kit de herramientas ETL del data warehouse: Técnicas prácticas para la extracción, limpieza, conformación y entrega de datos (2.ª ed.). Editorial Reverté.

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. SAGE Publications.

Margetts, H. (2009). The Internet and Public Policy. Policy & Internet, 1(1), 1-21.

McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51–56).

McKinney, W. (2012). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2019). Manual de Gobierno Digital: Implementación de la política de gobierno digital. Bogotá D.C.

Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC). (2022). Plan Nacional de Infraestructura de Datos: Hoja de ruta para el aprovechamiento de datos en el Estado.

OECD. (2019). Analytics for Integrity: Data-Driven Approaches for Enhancing Corruption Risk Assessment and Management. OECD Public Governance Reviews.

Open Contracting Partnership. (2020). Open Contracting Data Standard (OCDS) Documentation 1.1. Recuperado de https://standard.open-contracting.org/

Organización de los Estados Americanos (OEA). (2018). Guía de mecanismos para la promoción de la integridad y la transparencia en las contrataciones públicas. OEA.

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2019). Digital government review of Colombia: Towards a citizen-centric digital state. OECD Digital Government Studies. OECD Publishing. https://doi.org/10.1787/9789264291867-en

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Peters, T. J., & Waterman, R. H. (1982). In search of excellence: Lessons from America’s best-run companies. Harper & Row.

Presidencia de la República. (2020). Lineamientos para la implementación de la política de datos abiertos en Colombia. https://www.datos.gov.co

Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.

Redman, T. C. (1996). La calidad de los datos para la era de la información. Editorial McGraw-Hill.

Rodríguez Esquivel, D. M. (2020). Fortalecimiento en la adopción, aplicación e implementación de Sistema SECOP II en los FSE de colegios oficiales de la Localidad de San Cristóbal [Tesis de maestría, Universidad Santo Tomás]. Bogotá.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Transparencia por Colombia. (2023). Índice de Transparencia de las Entidades Públicas (ITEP). Bogotá D.C.

U.S. Government Accountability Office (GAO). (2018). Data Analytics to Address Fraud and Improper Payments. Washington D.C.

United Nations. (2022). E-Government Survey 2022: The Future of Digital Government. Department of Economic and Social Affairs. New York.

Van Rossum, G., & Drake, F. L. (2009). Python 3 reference manual. CreateSpace.

Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.

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

Apéndices

Apéndice A

SECOP-Monitoring-System

https://github.com/Andreschacon122025/SECOP-Monitoring-System