UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Procesamiento semántico basado en embeddings para clasificar riesgos en compras públicas y generar alertas tempranas.

Mateo Alejandro Fuertes Encalada

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Ciencias de la Computación

Quito, 9 de mayo de 2025

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Procesamiento semántico basado en embeddings para clasificar riesgos en compras públicas y generar alertas tempranas.

Mateo Alejandro Fuertes Encalada

Daniel Riofrío, PhD

Quito, 9 de mayo de 2025

© DERECHOS DE AUTOR

Nombres y apellidos: Mateo Alejandro Fuertes Encalada

Código: 00321987

Cédula de identidad: 1722816616

Lugar y fecha: Quito, 9 de mayo de 2025

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Este proyecto presenta una metodología para el análisis de riesgos en procesos de compras públicas, combinando datos estructurados y no estructurados en la forma de documentos, mediante técnicas de procesamiento semántico. El objetivo es facilitar la identificación de posibles irregularidades a través de modelos de clasificación automatizada y mecanismos de alerta temprana. Se propone un enfoque multimodal que integra embeddings generados por modelos de lenguaje (LLMs) con atributos estructurados extraídos de la base de datos del proyecto Kapak. A partir de esta integración, se aplican modelos de aprendizaje supervisado para estimar niveles de riesgo y técnicas no supervisadas para detectar agrupamientos relevantes en los datos.

Los resultados muestran que los modelos de ensamble alcanzaron los valores AUC-ROC más altos para la clasificación, con un 0.7986 utilizando todos los atributos y de 0.8352 al emplear únicamente atributos estructurados. Además, se desarrolló un sistema de búsqueda semántica para facilitar auditorías automatizadas sobre los documentos procesados. Esta investigación contribuye al fortalecimiento de la transparencia en la gestión pública, proponiendo un marco reproducible y escalable que puede extenderse a otras modalidades de contratación y contextos institucionales.

Palabras clave: contratación pública, aprendizaje automático, embeddings, agrupamiento, clasificación, riesgo, transparencia, lenguaje natural, recuperación semántica.

ABSTRACT

This project presents a methodology for risk analysis in public procurement processes, combining structured and unstructured data in the form of documents, using semantic processing techniques. The objective is to facilitate the identification of potential irregularities using automated classification models and early warning mechanisms. A multimodal approach is proposed, integrating embeddings generated by Large Language Models (LLMs) with structured attributes extracted from the Kapak project database. Based on this integration, supervised learning models are applied to estimate risk levels, while unsupervised techniques are used to detect relevant clusters within the data.

The results show that the ensemble models achieved the highest AUC-ROC values for classification, with 0.7986 using all attributes and 0.8352 using only structured attributes. Additionally, a semantic search system was developed to support automated audits over the processed documents. This research contributes to strengthening transparency in public administration by proposing a reproducible and scalable framework that can be extended to other procurement modalities and institutional contexts.

Keywords: public procurement, machine learning, embeddings, clustering, classification, risk, transparency, natural language, semantic retrieval.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

1.1. Contexto general del problema

Las compras públicas constituyen un pilar fundamental en la gestión gubernamental, ya que representan uno de los principales mecanismos del Estado para asignar recursos, ejecutar políticas públicas, satisfacer necesidades sociales y fomentar el desarrollo económico. En Ecuador, representan una parte significativa del gasto gubernamental, lo que hace imprescindible que estos procesos sean transparentes, eficientes y equitativos (Banco Mundial, 2017). Sin embargo, este objetivo se ve amenazado por prácticas irregulares y por la dificultad de detectar a tiempo señales de alerta relacionadas con riesgos de corrupción.

Uno de los mayores desafíos en este campo es la complejidad de los datos. Los registros generados durante los procesos de contratación pública suelen estar fragmentados, almacenados en distintos formatos (muchos de ellos no estructurados o codificados), y carecen de estandarización. Esto limita la capacidad de instituciones de control y de la ciudadanía para analizar esta información de forma eficiente (Ortiz-Prado et al., 2021). Aunque existen plataformas como el Sistema Oficial de Contratación Pública del Ecuador (SOCE), que han digitalizado estos procedimientos, la supervisión automatizada y en tiempo real sigue siendo limitada (Torres-Berru et al., 2023). En este contexto, el uso de herramientas tecnológicas avanzadas, como el procesamiento de lenguaje natural (NLP) y el análisis de datos, ha demostrado ser efectivo en otros países para detectar irregularidades y fortalecer la fiscalización (Padhi & Mohapatra, 2011; Lyra et al., 2022). En Brasil, por ejemplo, el análisis de redes ha permitido identificar colusión entre proveedores; mientras que en México y otros contextos, el uso de inteligencia artificial ha facilitado la detección temprana de riesgos en contrataciones públicas (Modrušan et al., 2021; Barot, 2023).

En Ecuador, el proyecto Kapak surge como respuesta a esta necesidad. Esta iniciativa integra técnicas de big data, extracción de texto y análisis semántico para estructurar, limpiar y analizar datos de compras públicas con un enfoque en la Subasta Inversa Electrónica (SIE). Su objetivo es contribuir a la transparencia mediante la generación de indicadores y alertas tempranas, facilitando la identificación de posibles irregularidades (Fortuny et al., 2023; Riofrío et al., 2023). Pese a estos avances, aún existe una brecha metodológica para aprovechar al máximo el contenido semántico de los documentos y clasificar el riesgo de forma automatizada y precisa.

1.2. Definición del problema de investigación

A pesar de los avances logrados con el proyecto Kapak y la digitalización del sistema de compras públicas, aún persisten desafíos importantes para lograr un análisis más profundo y efectivo del riesgo de corrupción en estos procesos. Más allá de las limitaciones técnicas propias de la estructura de los datos, se vuelve crucial avanzar hacia metodologías que permitan analizar el contenido semántico de los documentos, como las especificaciones técnicas, las preguntas y respuestas entre actores y las condiciones contractuales.

Hoy en día, gran parte del análisis de riesgo en compras públicas se basa en indicadores estructurados o reglas fijas, lo que, si bien resulta útil, no alcanza a capturar los matices y patrones complejos que pueden revelar señales sutiles de riesgo de corrupción. La capacidad de interpretar y representar de forma automática la información contenida en los documentos es un paso esencial para mejorar la identificación temprana de vulnerabilidades. Incorporar técnicas modernas, como la generación de embeddings a partir de modelos de lenguaje, permitiría representar de manera más rica el contenido textual y habilitar sistemas de clasificación automática de riesgos más precisos.

Fortalecer el análisis semántico de los documentos no solo permitiría detectar riesgos específicos de corrupción, sino también enriquecer el conjunto completo de indicadores disponibles para la supervisión de procesos de contratación pública. Así, se sentarían las bases para el desarrollo de sistemas de alerta temprana más efectivos, escalables y capaces de contribuir significativamente al fortalecimiento de la transparencia y la eficiencia en la gestión pública en Ecuador.

1.3. Hipótesis de trabajo y preguntas de investigación

Hipótesis de trabajo:

Es posible clasificar de manera efectiva el nivel de riesgo de corrupción en procesos de compras públicas mediante el uso de modelos de lenguaje (LLMs) que generen representaciones semánticas (embeddings) de los documentos involucrados, combinadas con información de cada proceso (atributos estructurados), empleando algoritmos de clasificación supervisada y no supervisada.

Preguntas de investigación:

- ¿Qué tan efectivos son los modelos de embeddings generados por LLMs al representar información semántica relevante de los documentos de compras públicas?
- ¿Qué tipo de modelo supervisado logra una mejor clasificación del nivel de riesgo en estos procesos?
- ¿Qué combinación de características y modelos mejora la precisión de las alertas tempranas en un contexto real como el del proyecto Kapak?
- ¿Qué tan aplicable y escalable es la metodología propuesta a otras modalidades de contratación o sistemas similares?

1.4. Objetivos

Objetivo general:

Desarrollar una metodología basada en la generación de embeddings mediante *Large Language Models*, para clasificar procesos de compras públicas según niveles de riesgo de corrupción, con el fin de habilitar mecanismos de alerta temprana que fortalezcan la gestión transparente y eficiente de los recursos públicos.

Objetivos específicos:

- Emplear embeddings generados por *Large Language Models* (LLMs) para representar de manera semántica la información contenida en los documentos relacionados con los procesos de compras públicas.
- Implementar modelos de clasificación supervisada y agrupamiento que, basándose en los embeddings generados y otros atributos de los procesos, permitan categorizar los niveles de riesgo asociados a cada procedimiento.
- Comparar diferentes algoritmos de clasificación y métodos de extracción de información de los documentos, evaluando su impacto en la precisión y efectividad de la clasificación de riesgos.

1.5. Contenido del documento

Este documento está organizado en varios capítulos que permiten comprender de manera estructurada el problema de investigación, la metodología utilizada y los hallazgos obtenidos. En el Capítulo 1, se introduce el contexto general del problema, la relevancia del estudio y los objetivos que guían esta investigación. El Capítulo 2 presenta una revisión del estado del arte, en la que se analizan conceptos clave como embeddings de texto, modelos de

lenguaje grandes (LLMs) y su aplicación en el procesamiento del lenguaje natural, además de métodos de clasificación supervisada y no supervisada. También se examinan estudios previos sobre transparencia en la gestión pública y los vacíos existentes en la literatura. El Capítulo 3 describe la metodología empleada en este estudio, incluyendo el enfoque metodológico, la preparación de los datos, el uso de embeddings generados por LLMs y la comparación de distintos modelos de clasificación. Posteriormente, en el Capítulo 4, se detallan la implementación del modelo y el desarrollo del prototipo, abarcando desde el preprocesamiento de datos hasta la aplicación de técnicas de aprendizaje automático para la categorización de riesgos. En el Capítulo 5, se presentan los resultados obtenidos a partir de la validación experimental de los modelos, comparando su desempeño con diferentes tipos de embeddings y métodos de clasificación. A continuación, el Capítulo 6 discute estos hallazgos en relación con estudios previos, evaluando las fortalezas y limitaciones del enfoque propuesto y sus implicaciones para la detección temprana de riesgos en compras públicas. Finalmente, en el Capítulo 7, se exponen las conclusiones generales del estudio, su contribución al campo del análisis de datos en gestión pública y sus limitaciones. El documento concluye con el Capítulo 8, en el que se sugieren posibles direcciones para futuras investigaciones, incluyendo la integración de modelos más avanzados o la adaptación de esta metodología a otros contextos gubernamentales.

Dado que el análisis de datos en compras públicas involucra conceptos técnicos avanzados, en el siguiente capítulo se presenta una revisión detallada del estado del arte. Allí se explorarán las bases teóricas y los enfoques previos en procesamiento de lenguaje natural, aprendizaje automático y transparencia en la gestión pública, estableciendo el marco conceptual que sustenta esta investigación.

2. ESTADO DEL ARTE

2.1. Introducción al estado del arte

El análisis de antecedentes teóricos y metodológicos es un paso fundamental en cualquier investigación científica, ya que permite contextualizar el problema de estudio, identificar enfoques previos y fundamentar la metodología utilizada. En el caso del análisis de compras públicas mediante técnicas de procesamiento de lenguaje natural (NLP por sus siglas en inglés) y aprendizaje automático, es crucial revisar los avances en la representación semántica de textos, los modelos de clasificación supervisada y no supervisada, así como las aplicaciones previas de estas tecnologías en la detección de riesgos y corrupción en la gestión pública (Jurafsky & Martin, 2023).

Este capítulo tiene como objetivo presentar el marco conceptual y metodológico que sustenta la presente investigación. Se abordará la evolución de los embeddings de texto, desde métodos tradicionales como Word2Vec y GloVe hasta técnicas más avanzadas basadas en modelos de lenguaje grandes (LLMs) como BERT y GPT (Mikolov et al., 2013; Devlin et al., 2019). Asimismo, se analizarán los enfoques de clasificación de riesgos en datos textuales, distinguiendo entre métodos supervisados (e.g., SVM, redes neuronales) y no supervisados (e.g., clustering con K-means o DBSCAN) (Aggarwal & Zhai, 2012). Finalmente, se examinarán estudios previos sobre transparencia y análisis de datos en compras públicas, con especial énfasis en la aplicación de big data y NLP en la detección de irregularidades y el fortalecimiento de la rendición de cuentas (Modrušan et al., 2021).

A lo largo de este capítulo, se establecerá la relación entre los enfoques existentes y la metodología propuesta en este estudio, identificando los vacíos en la literatura que justifican la necesidad de una nueva metodología para la clasificación de riesgos en compras públicas.

2.2. Embeddings de Texto y su Evolución

Los embeddings de texto son representaciones matemáticas en forma de vectores que capturan el significado semántico de palabras, frases o documentos dentro de un espacio continuo de baja dimensión. A diferencia de los enfoques tradicionales basados en representaciones discretas como el modelo de bolsa de palabras (bag-of-words), los embeddings permiten representar relaciones semánticas entre términos, mejorando la capacidad de los algoritmos de NLP para interpretar el contexto en el que aparecen las palabras (Mikolov et al., 2013). Estas representaciones se entrenan utilizando grandes volúmenes de texto y se optimizan para preservar similitudes semánticas, es decir, palabras con significados similares tendrán representaciones vectoriales cercanas en el espacio latente (Jurafsky & Martin, 2023). El uso de embeddings ha sido fundamental en múltiples aplicaciones de NLP, incluyendo la clasificación de texto, la traducción automática y la detección de patrones en datos no estructurados.

Los primeros enfoques para generar embeddings de palabras incluyen Word2Vec, GloVe y FastText, cada uno con características específicas. Word2Vec, propuesto por Mikolov et al. (2013), se basa en redes neuronales superficiales que aprenden representaciones de palabras a partir de su contexto en grandes corpus de texto. Este modelo tiene dos variantes principales: CBOW (Continuous Bag-of-Words), que predice una palabra a partir de su contexto, y Skip-gram, que predice el contexto a partir de una palabra dada. Posteriormente, GloVe (Global Vectors for Word Representation) introdujo un enfoque basado en matrices de co-ocurrencia para capturar relaciones semánticas globales, obteniendo mejores resultados en ciertas tareas de NLP (Pennington et al., 2014). Finalmente, FastText, desarrollado por Facebook AI, mejoró estos modelos al representar palabras como un conjunto de subpalabras (n-grams), lo que le permite manejar vocabularios abiertos y mejorar la representación de palabras morfológicamente complejas (Bojanowski et al., 2017).

Aunque los modelos tradicionales de embeddings lograron avances significativos, tenían una limitación fundamental: cada palabra tenía una única representación fija, sin considerar el contexto en el que aparecía. Para superar esta restricción, surgieron los embeddings contextuales, los cuales permiten que la representación de una palabra varíe dependiendo del contexto en el que se encuentra (Devlin et al., 2019). Modelos basados en transformers, como BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer) y T5 (Text-to-Text Transfer Transformer), utilizan mecanismos de atención que les permiten procesar información en múltiples direcciones y generar embeddings más ricos semánticamente (Brown et al., 2020; Raffel et al., 2020). BERT, por ejemplo, es entrenado bidireccionalmente, lo que le permite entender el significado de una palabra considerando tanto el contexto anterior como el posterior, mientras que GPT se entrena de manera autoregresiva, generando texto de manera secuencial.

Los métodos tradicionales como Word2Vec y GloVe generan embeddings estáticos, lo que significa que una palabra tendrá siempre la misma representación sin importar el contexto en el que aparezca. Esto representa una limitación en tareas que requieren una comprensión más profunda del significado contextual de las palabras, como la resolución de ambigüedades semánticas (Peters et al., 2018). En contraste, los embeddings generados por modelos como BERT y GPT son dinámicos, lo que significa que su representación cambia en función de las palabras circundantes, capturando de manera más precisa las relaciones semánticas dentro de una oración o documento (Tenney et al., 2019). En el contexto del análisis de documentos de compras públicas, los embeddings dinámicos permiten una mejor detección de patrones lingüísticos y una mayor precisión en la identificación de riesgos, lo que los convierte en una herramienta clave en este estudio.

2.3. Modelos de Lenguaje Grandes (LLMs) y su Aplicación en el Procesamiento del Lenguaje Natural (NLP)

Los Large Language Models (LLMs) son modelos de inteligencia artificial diseñados para procesar y generar texto natural con una comprensión profunda del lenguaje. Se basan en arquitecturas avanzadas de redes neuronales profundas, particularmente en transformadores, y son entrenados con grandes volúmenes de datos para aprender patrones lingüísticos complejos (Brown et al., 2020). La relevancia de los LLMs en el campo de NLP radica en su capacidad para realizar tareas como traducción automática, respuesta a preguntas, generación de texto, clasificación de documentos y análisis semántico con un alto grado de precisión (Raffel et al., 2020). A diferencia de modelos previos, los LLMs pueden generar representaciones contextuales más ricas de los textos, lo que los hace fundamentales para tareas avanzadas de análisis de datos, como la clasificación de riesgos en compras públicas.

En la última década, varios modelos han marcado hitos en el campo de NLP. BERT (Bidirectional Encoder Representations from Transformers), introducido por Google, revolucionó la forma en que se generan embeddings al permitir representaciones bidireccionales, es decir, teniendo en cuenta tanto el contexto previo como posterior de una palabra en una oración (Devlin et al., 2019). Por otro lado, GPT (Generative Pre-trained Transformer), desarrollado por OpenAI, introdujo un enfoque basado en aprendizaje autoregresivo, lo que le permite generar texto de manera coherente y contextualizada a partir de un conjunto de palabras de entrada (Radford et al., 2019).

Modelos más recientes han seguido optimizando estas técnicas. LLaMA (Large Language Model Meta AI) es un modelo desarrollado por Meta que optimiza el uso de datos de entrenamiento para lograr un rendimiento eficiente con menos recursos computacionales (Touvron et al., 2023). Mistral, una arquitectura de código abierto, ha sido reconocida por su

capacidad para generar representaciones lingüísticas más compactas y eficientes sin perder precisión en tareas de comprensión del lenguaje (Jiang et al., 2024). Finalmente, DeepSeek v3 ha demostrado mejoras en la capacidad de razonamiento y generación de respuestas más precisas, consolidándose como una opción competitiva en la nueva generación de modelos de lenguaje grande (Zhou et al., 2024).

Una de las principales ventajas de los LLMs es su capacidad para generar embeddings de texto altamente contextualizados, lo que mejora significativamente tareas como clasificación, análisis semántico y detección de anomalías en datos no estructurados (Peters et al., 2018). Sin embargo, la extracción de estos embeddings requiere técnicas especializadas. LLM2Vec es un método emergente que transforma modelos de lenguaje de propósito general en generadores de embeddings eficientes, permitiendo su aplicación en análisis de datos a gran escala sin necesidad de realizar inferencias completas sobre todo el modelo (BehnamGhader et al., 2024). Otras técnicas incluyen el uso de capa de salida de modelos preentrenados, en la que se extraen representaciones de última capa para tareas específicas, y ajuste fino con representaciones intermedias, que permite capturar mejor los patrones lingüísticos específicos de un dominio.

El uso de estos métodos en la presente investigación permite generar embeddings de documentos de compras públicas con mayor precisión semántica, facilitando la identificación de riesgos y patrones de comportamiento en los procesos de contratación estatal.

2.4. Métodos de Clasificación en NLP

La clasificación de texto es una tarea fundamental en el procesamiento de lenguaje natural (NLP), utilizada para asignar etiquetas o categorías a documentos en función de su contenido semántico. Esta tarea se divide en clasificación supervisada y clasificación no

supervisada (clustering), dependiendo de si se dispone de etiquetas previas para el entrenamiento del modelo. En el contexto del análisis de riesgo en compras públicas, estas técnicas permiten detectar irregularidades, identificar patrones de comportamiento en licitaciones y categorizar procesos en función de su nivel de transparencia (Aggarwal & Zhai, 2012).

La clasificación supervisada se basa en modelos entrenados con datos etiquetados, lo que significa que cada entrada del conjunto de entrenamiento tiene una categoría asignada previamente. Estos modelos aprenden patrones en los datos y pueden predecir etiquetas para nuevas observaciones con base en las características extraídas del texto (Sebastiani, 2002). Entre los algoritmos más utilizados en NLP para clasificación supervisada se encuentran:

- *Máquinas de Soporte Vectorial (SVM):* Este método busca encontrar un hiperplano óptimo que separe las clases de manera eficiente en un espacio de alta dimensión. Ha demostrado ser altamente efectivo en tareas de clasificación de texto debido a su capacidad para manejar datos no lineales mediante el uso de kernels.
- *Árboles de Decisión:* Estos modelos dividen los datos en función de características específicas mediante reglas de decisión jerárquicas. Son interpretables y eficientes, aunque pueden ser susceptibles al sobreajuste si no se aplican técnicas de poda adecuadas.
- *Redes Neuronales:* Modelos como los Perceptrones Multicapa (MLP) o arquitecturas más avanzadas como Transformers han logrado grandes avances en clasificación de texto, permitiendo capturar relaciones semánticas complejas entre palabras y frases dentro de los documentos (LeCun et al., 2015).

El uso de clasificación supervisada en análisis de riesgos permite identificar documentos sospechosos o categorizar licitaciones en función de indicadores de transparencia. En el ámbito de las compras públicas, estos modelos pueden entrenarse con datos históricos

para predecir si un proceso de contratación presenta señales de alerta, como falta de competencia, modificación de términos contractuales o adjudicaciones repetitivas a un mismo proveedor (Torres-Berru et al., 2023). Además, técnicas basadas en redes neuronales han sido aplicadas para detectar sesgos en textos de licitaciones, identificando posibles favoritismos o redacciones que favorecen a ciertos oferentes sobre otros (Barot, 2023).

A diferencia de la clasificación supervisada, los métodos de clustering no requieren etiquetas previas y permiten agrupar documentos en función de sus similitudes estructurales y semánticas (Manning et al., 2008). Entre los algoritmos más utilizados en NLP para análisis exploratorio y detección de patrones en datos de compras públicas se encuentran:

- *K-Means:* Es un algoritmo basado en particiones que agrupa datos en k clústeres en función de su proximidad en el espacio vectorial. Se utiliza ampliamente en NLP para organizar documentos según su similitud semántica y ha demostrado ser útil en la detección de grupos de licitaciones con características similares.
- *DBSCAN (Density-Based Spatial Clustering of Applications with Noise): A* diferencia de K-Means, este método detecta clústeres de diferentes formas y tamaños basándose en densidad de datos, permitiendo identificar anomalías y documentos atípicos dentro de un conjunto de datos de contratación pública.
- *Modelos de Clustering Jerárquico:* Agrupan datos de manera progresiva en una estructura de árbol (dendrograma), facilitando la exploración de relaciones entre documentos sin necesidad de predefinir el número de clústeres.

En el contexto de esta investigación, los métodos de clustering permiten analizar grandes volúmenes de documentos de compras públicas y descubrir patrones ocultos que podrían indicar riesgos de corrupción o procesos con características inusuales. La combinación

de enfoques supervisados y no supervisados es clave para obtener una visión integral de los datos y mejorar la precisión en la detección de irregularidades.

2.5. Transparencia y Análisis de Riesgo en Compras Públicas

Para comprender mejor el alcance de este estudio, es necesario definir algunos términos clave relacionados con las compras públicas y las herramientas utilizadas en su análisis. Las compras públicas se refieren al proceso mediante el cual las entidades gubernamentales adquieren bienes, obras y servicios con el objetivo de satisfacer necesidades de interés público. Estos procesos pueden realizarse mediante distintas modalidades, como licitaciones abiertas, contratación directa o subastas electrónicas (Banco Mundial, 2017). Entre estas, la Subasta Inversa Electrónica (SIE) es un mecanismo utilizado en Ecuador para promover la competencia y eficiencia en la adquisición de bienes y servicios, permitiendo que los proveedores reduzcan progresivamente sus ofertas hasta alcanzar el precio más bajo aceptable (SERCOP, 2022).

La transparencia en la gestión pública es un principio fundamental para garantizar la eficiencia en el uso de los recursos estatales, fortalecer la confianza ciudadana y prevenir actos de corrupción. En el ámbito de las compras públicas, la implementación de mecanismos de rendición de cuentas y acceso a datos abiertos permite un mayor escrutinio de los procesos de contratación, reduciendo el riesgo de prácticas ilícitas como el fraude, la colusión y el favoritismo en la adjudicación de contratos (OECD, 2016). Diversos organismos internacionales han promovido iniciativas para mejorar la transparencia en adquisiciones gubernamentales, incluyendo la adopción de estándares de datos abiertos y plataformas de monitoreo ciudadano (Banco Mundial, 2017). En Ecuador, el Sistema Oficial de Contratación Pública (SOCE) ha sido una de las herramientas clave para la digitalización de los procesos de

licitación y contratación estatal, aunque persisten desafíos relacionados con la accesibilidad y el análisis de grandes volúmenes de datos (Fortuny et al., 2023).

Los enfoques tradicionales para la detección de corrupción en compras públicas se han basado en indicadores de riesgo, auditorías manuales y análisis estadísticos. Entre estos métodos, el uso de matrices de indicadores de riesgo ha sido ampliamente adoptado para evaluar la transparencia y eficiencia de los procesos de contratación (Kaufmann et al., 2011). Estas matrices incluyen factores como la cantidad de oferentes en una licitación, la frecuencia de adjudicación a un mismo proveedor y los cambios en los términos contractuales. Sin embargo, estos enfoques presentan limitaciones, ya que dependen de datos estructurados y no permiten el análisis en tiempo real de grandes volúmenes de información (Auriol et al., 2016). En este sentido, han surgido nuevas metodologías basadas en big data y análisis de lenguaje natural, las cuales pueden mejorar la identificación de patrones de riesgo al analizar documentos de contratación y comunicaciones entre actores del proceso.

La aplicación de procesamiento de lenguaje natural (NLP) y aprendizaje automático en compras públicas ha cobrado relevancia en los últimos años como una alternativa para fortalecer la transparencia y automatizar la detección de riesgos. Estudios recientes han utilizado técnicas de minería de texto para analizar licitaciones gubernamentales y detectar señales de manipulación en los términos de referencia, lo que puede indicar sesgos o direccionamiento de contratos hacia proveedores específicos (Torres-Berru & Batista, 2020). Otros trabajos han empleado modelos de clasificación supervisada y no supervisada para identificar irregularidades en procesos de adjudicación, demostrando que estos enfoques pueden mejorar la detección de anomalías en comparación con los métodos tradicionales (Dhurandhar et al., 2015).

En el contexto latinoamericano, iniciativas como SALER (Sistema de Alerta para Licitaciones Estatales con Riesgo) han utilizado técnicas de inteligencia artificial para clasificar procesos de compra en función de su nivel de riesgo, integrando análisis de datos estructurados y no estructurados (Martínez-Plumed et al., 2019). Asimismo, en Brasil se han implementado sistemas basados en análisis de redes, los cuales permiten identificar relaciones sospechosas entre proveedores y funcionarios públicos mediante grafos de contratación (Lyra et al., 2022).

A pesar de los avances en el uso de NLP y machine learning para el análisis de compras públicas, existen diversas limitaciones y vacíos en la literatura. Una de las principales barreras es la falta de acceso a datos abiertos de calidad, ya que en muchos países la información sobre contratos estatales sigue estando fragmentada o restringida (Modrušan et al., 2021). Además, los modelos de análisis actuales enfrentan desafíos en la interpretación de textos legales y documentos contractuales, los cuales suelen contener un lenguaje técnico complejo y ambigüedades semánticas (BehnamGhader et al., 2024).

Otra limitación importante es la capacidad de generalización de los modelos de detección de riesgos, ya que la corrupción en compras públicas adopta múltiples formas y varía según el contexto regulatorio y cultural de cada país (Auriol et al., 2016). Esto resalta la necesidad de desarrollar metodologías adaptativas que combinen técnicas de análisis estructurado y no estructurado, permitiendo la identificación de riesgos de manera más precisa y flexible. La presente investigación busca contribuir a este campo proponiendo un enfoque basado en embeddings contextuales y modelos de clasificación automatizada, con el fin de mejorar la detección de diferentes niveles de riesgo de corrupción en documentos de contratación pública y generar herramientas más eficientes para la supervisión estatal.

2.6. Implementaciones Previas de Análisis de Datos en Compras Públicas

En los últimos años, varios países han implementado sistemas de monitoreo automatizado y análisis de datos para mejorar la supervisión de sus procesos de compras públicas. Uno de los casos más destacados es el sistema PROACT en la Unión Europea, que utiliza aprendizaje automático y minería de texto para identificar irregularidades en licitaciones gubernamentales mediante el análisis de documentos y registros de contratación (Clarke & O'Connor, 2021). Otro ejemplo relevante es Open Contracting Data Standard (OCDS), una iniciativa global impulsada por el Banco Mundial y Open Contracting Partnership, que promueve el uso de estándares abiertos para mejorar la transparencia y accesibilidad de la información en adquisiciones estatales (Banco mundial, 2017).

En América Latina, Brasil ha desarrollado el Sistema de Compras Gubernamentales (ComprasNet), que integra herramientas de análisis de datos para detectar posibles esquemas de colusión y corrupción en licitaciones públicas (Lyra et al., 2022). De manera similar, en México, la plataforma Compranet ha sido utilizada para la supervisión de contratos estatales, aunque aún presenta limitaciones en la explotación de datos no estructurados para análisis avanzados (Martínez-Plumed et al., 2019).

El uso de procesamiento de lenguaje natural (NLP) y big data en la supervisión de compras públicas ha permitido avances significativos en la identificación de patrones de riesgo. Por ejemplo, el sistema SALER (Sistema de Alerta para Licitaciones Estatales con Riesgo) ha aplicado técnicas de machine learning y análisis de texto para clasificar contratos en función de su probabilidad de contener irregularidades (Martínez-Plumed et al., 2019). En Colombia, el Observatorio de Contratación Pública ha integrado modelos de NLP para analizar la redacción de pliegos de licitación y detectar términos sospechosos que podrían estar dirigidos a favorecer a ciertos proveedores (Ariza & Pardo, 2021).

Otro enfoque innovador ha sido el uso de análisis de redes para visualizar relaciones entre empresas contratistas y entidades gubernamentales. En Italia, se ha implementado un sistema basado en grafos de contratación pública, el cual ha permitido detectar redes de colusión en licitaciones mediante algoritmos de detección de comunidades sospechosas (Dhurandhar et al., 2015).

Si bien existen múltiples iniciativas para la supervisión de compras públicas basadas en big data y NLP, muchas de ellas presentan limitaciones en la integración de datos estructurados y no estructurados, así como en la generalización de sus modelos para distintos contextos. La mayoría de los sistemas actuales se enfocan en el análisis de datos estructurados, dejando de lado la riqueza semántica presente en los documentos de licitación y los registros de preguntas y aclaraciones (Modrušan et al., 2021).

El enfoque propuesto en esta tesis busca superar estas limitaciones mediante el uso de embeddings contextuales y modelos de clasificación automatizada para analizar documentos de contratación pública con mayor precisión. A diferencia de iniciativas previas que dependen de reglas predefinidas o análisis de datos tabulares, esta investigación se centra en extraer información semántica de textos utilizando técnicas avanzadas de NLP y aprendizaje profundo. Además, se propone una metodología replicable y adaptable a diferentes entornos gubernamentales, lo que podría facilitar su implementación en países con desafíos similares en la supervisión de adquisiciones estatales.

3. METODOLOGÍA DE INVESTIGACIÓN

3.1. Enfoque metodológico

Este proyecto adopta un enfoque cuantitativo, dado que su objetivo principal es analizar, modelar y clasificar procesos de compras públicas utilizando representaciones numéricas derivadas del lenguaje natural, específicamente mediante embeddings generados por modelos de lenguaje (LLMs). La metodología se basa en la recolección y procesamiento de datos estructurados y no estructurados, a partir de los cuales se desarrollan modelos de clasificación y agrupamiento que permiten detectar patrones asociados a posibles riesgos de corrupción.

La elección del enfoque cuantitativo se justifica porque permite aplicar técnicas estadísticas y de aprendizaje automático para evaluar la efectividad de los modelos implementados. A través de métricas como precisión, recall y F1-score, es posible medir de manera objetiva el desempeño del sistema propuesto y validar su utilidad en contextos reales, como el entorno del proyecto Kapak. Si bien el análisis cualitativo podría ofrecer una visión interpretativa de ciertos patrones de riesgo, el volumen y complejidad de los datos involucrados en la contratación pública hacen que el enfoque cuantitativo sea más adecuado para garantizar una evaluación sistemática, reproducible y automatizada.

3.2. Diseño de la investigación

Este proyecto tiene un diseño exploratorio y experimental, ya que busca evaluar el uso de modelos de lenguaje (LLMs) y técnicas de procesamiento semántico para detectar posibles riesgos de corrupción en procesos de compras públicas. Al tratarse de un campo en el que aún no existe una metodología establecida que utilice embeddings generados por LLMs para este fin específico, el enfoque exploratorio permite indagar sobre su aplicabilidad. Al mismo

tiempo, se plantean pruebas controladas y comparativas entre distintos modelos y configuraciones, lo cual le da un carácter experimental.

La investigación parte de la siguiente hipótesis:

Es posible clasificar de manera efectiva el nivel de riesgo de corrupción en procesos de compras públicas mediante el uso de modelos de lenguaje (LLMs) que generen representaciones semánticas (embeddings) de los documentos involucrados, combinadas con información de cada proceso (atributos estructurados), empleando algoritmos de clasificación supervisada y no supervisada.

A partir de esta hipótesis se derivan las siguientes preguntas de investigación:

- ¿Qué tan representativos son los embeddings generados por LLMs para capturar información semántica relevante en los documentos de contratación pública?
- ¿Qué tipo de modelos (supervisados o no supervisados) permiten una clasificación más precisa de los niveles de riesgo?
- ¿Qué combinación de embeddings y atributos adicionales ofrece mejores resultados en términos de detección de riesgos?

Para validar los resultados, se utilizan criterios de evaluación estándar en el área de aprendizaje automático, especialmente en tareas de clasificación y agrupamiento de texto. Entre ellos se encuentran: precisión (accuracy), recall (sensibilidad), F1-score, AUC-ROC. Estos criterios permiten medir de forma objetiva el desempeño de los modelos y comparar su efectividad en diferentes configuraciones. La evaluación se realizará con datos del proyecto Kapak, lo que garantiza una validación práctica y contextualizada en un entorno real de contratación pública.

3.3. Fuentes de datos

El conjunto de datos utilizado en este estudio proviene del procesamiento de datos de compras públicas en Ecuador durante los años 2020, 2022 y 2023, obtenidos a partir del Sistema Oficial de Contratación Pública (SOCE) y almacenados en la base de datos del proyecto Kapak. Este dataset fue generado mediante un pipeline de limpieza, extracción y consolidación, el cual procesó datos sin estructurar provenientes de documentos contractuales, archivos y registros transaccionales, transformándolos en una estructura lista para análisis.

Tras el procesamiento, el dataset contaba con 73 variables y 5400 registros, representando procesos de contratación pública bajo la modalidad de Subasta Inversa Electrónica (SIE). Las variables se organizan en diferentes tipos de datos:

- Variables numéricas: Incluyen montos económicos (presupuesto referencial, monto adjudicado y monto de contrato), 12 indicadores de riesgo diseñados para evaluar posibles irregularidades en los procesos de contratación (los que aplican a SIE), etc.
- Variables categóricas: Representan información cualitativa como el nombre de la entidad contratante, el estado del proceso, la clasificación del tipo de compra (bienes, servicios o consultoría) y el nombre del proveedor adjudicado.
- Variables booleanas: Reflejan la presencia o ausencia de ciertas condiciones dentro de los procesos de contratación, como la existencia de modificaciones contractuales, el uso de mecanismos de apelación o la publicación de preguntas y aclaraciones.

Además de estos datos estructurados, el dataset incluye documentos en formato PDF y XML obtenidos en el proceso previo descrito. Estos archivos contienen información relacionada con los procesos de contratación, incluyendo términos de referencia, formularios, especificaciones técnicas y otras secciones relevantes. Estos textos son fundamentales para la generación de embeddings y modelos de clasificación.

3.4. Procedimiento experimental

El procedimiento experimental de este proyecto se basa en la comparación entre modelos de aprendizaje supervisado, aplicados al análisis de riesgos en procesos de compras públicas. Esta comparación tiene como objetivo evaluar qué enfoque resulta más efectivo para clasificar los niveles de riesgo utilizando información semántica extraída de documentos mediante técnicas de procesamiento de lenguaje natural.

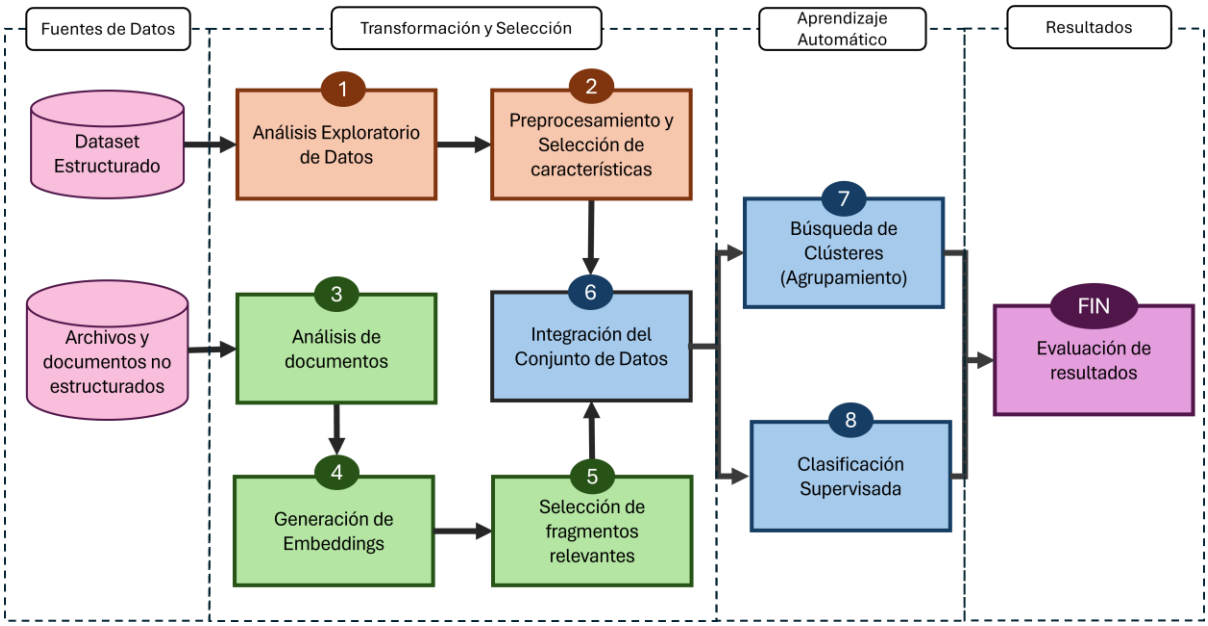

Figura 1. Esquema general del procedimiento experimental

En primer lugar, se lleva a cabo un análisis exploratorio de los datos, seguido por un proceso de limpieza y normalización de los mismos. Una vez organizados, se extraen atributos estructurados relevantes como montos, fechas, tipo de entidad contratante, entre otros. Tal como se muestra en la Figura 1, paralelamente se genera una representación semántica de los documentos utilizando embeddings producidos por modelos de lenguaje (LLMs), lo que permite transformar el contenido textual en vectores numéricos con significado contextual. Para seleccionar documentos clave dentro de cada proceso de contratación, se implementa una

arquitectura basada en Retrieval-Augmented Generation (RAG), que mejora la calidad de la información utilizada por los modelos al enfocar el análisis en los textos más relevantes. Posteriormente, se integra todo el conjunto de datos, combinando atributos estructurados con embeddings semánticos, sobre el cual se aplican las técnicas de aprendizaje automático.

En el enfoque no supervisado, se utilizan algoritmos de agrupamiento como K-Means y DBSCAN, con el objetivo de identificar patrones ocultos o grupos naturales de procesos que presenten similitudes en sus características semánticas y estructurales. En el enfoque supervisado, se implementan modelos como Support Vector Machines (SVM), Random Forest y Gradient Boosting, entrenados con etiquetas de riesgo previamente establecidas (derivadas de reglas del sistema Kapak). La elección de estos métodos responde a tres criterios principales:

- Capacidad para trabajar con representaciones vectoriales complejas, como los embeddings generados por LLMs.
- Rendimiento comprobado en tareas similares de clasificación de texto o detección de anomalías.
- Balance entre interpretabilidad y precisión, ya que uno de los objetivos es que los resultados puedan ser analizados posteriormente por equipos técnicos o de auditoría.

Finalmente, los modelos son evaluados con métricas específicas: para clasificación supervisada se aplican precisión, recall y F1-score; mientras que para agrupamiento se utilizan métricas como silhouette score, calinski-harabasz y análisis visual de los clústeres proyectados mediante técnicas de reducción de dimensionalidad (por ejemplo, t-SNE o PCA). Esta estrategia experimental permite contrastar el desempeño de ambos enfoques y determinar cuál es más adecuado para el contexto de compras públicas, tomando en cuenta tanto la calidad de los resultados como su viabilidad de implementación en entornos reales como el del proyecto Kapak.

3.5. Estándares y principios de diseño de ingeniería considerados

Para asegurar la calidad y sostenibilidad del proyecto, se han incorporado principios fundamentales del diseño de ingeniería de software, así como estándares internacionales aplicables a sistemas basados en inteligencia artificial y procesamiento de lenguaje natural. Desde el punto de vista del diseño de software, se prioriza la modularidad, de tal forma que los distintos componentes del sistema puedan desarrollarse y probarse de forma independiente. Esta estructura facilita la mantenibilidad del código y su mejora progresiva. Asimismo, se busca escalabilidad, dado que el sistema debe ser capaz de procesar volúmenes cada vez mayores de datos provenientes de procesos de contratación pública sin perder rendimiento. Para ello, se aplican buenas prácticas de eficiencia computacional y uso adecuado de recursos.

En cuanto a los estándares aplicados, se considera el ISO/IEC 23053:2022, que proporciona una guía estructurada para el diseño e implementación de sistemas de inteligencia artificial basados en aprendizaje automático, abarcando desde la definición del problema hasta el despliegue del modelo. Este estándar permite asegurar que el sistema se construya con una arquitectura robusta y documentada (ISO/IEC, 2022a). Además, se utiliza el ISO/IEC TS 4213:2022, que especifica criterios para la evaluación del desempeño de modelos de clasificación supervisada y no supervisada. Este estándar guía la elección de métricas como precisión, recall, F1-score o silhouette score, garantizando evaluaciones objetivas y reproducibles (ISO/IEC, 2022b).

En cuanto a la reproducibilidad, todo el pipeline del proyecto ha sido documentado, incluyendo las configuraciones utilizadas para entrenar modelos, las fuentes de datos, y los criterios de evaluación. Esto permitirá que futuros investigadores o instituciones puedan replicar la metodología o adaptarla a otros contextos similares.

4. IMPLEMENTACIÓN DE MODELOS Y DESARROLLO DEL PROTOTIPO

4.1. Análisis Exploratorio de Datos

El análisis exploratorio de datos (EDA, por sus siglas en inglés) es una etapa fundamental para la comprensión inicial de las variables disponibles en el conjunto de datos. En esta sección se presentan los primeros pasos realizados para familiarizarse con las características generales de la información, identificar posibles patrones, detectar valores atípicos y tener dirección en el desarrollo de los modelos empleados.

En primer lugar, se cargaron los datos estructurados, con las variables que serán objeto de estudio a lo largo del presente trabajo. A partir de este punto, se llevaron a cabo diferentes análisis descriptivos y gráficos, con el fin de obtener una primera aproximación a la distribución y relaciones entre las variables incluidas. Inicialmente, se visualizó la estructura del conjunto de datos mediante funciones como head() e info(), lo cual permitió verificar el número de registros, las columnas disponibles, los tipos de datos y la presencia de valores nulos. También se analizaron el número total de filas y columnas, así como los nombres de las variables disponibles. Este paso fue clave para detectar posibles problemas de calidad de datos y conocer la dimensionalidad del conjunto.

Valores Faltantes.

Se calculó el porcentaje de valores faltantes por columna, lo que permitió detectar variables con una proporción significativa de datos ausentes. Esta información es crucial para tomar decisiones informadas respecto al manejo de valores nulos, ya sea mediante imputación, eliminación o sustitución por estadísticos representativos. En la tabla 1 se muestra el porcentaje de valores faltantes por columna (en aquellas columnas con valores nulos).

Tabla 1. Porcentaje de valores nulos por columna

Columna	% de valores nulos
sie_ic_ruc	28.21
sie_ic_proveedor	28.21
sie_ic_indicador_01	29.49
sie_ic_indicador_04	0.43
sie_ic_indicador_06	6.17
sie_ic_indicador_09	71.42
sie_ic_indicador_11	71.42
sie_ic_indicador_19	33.90
sie_ic_indicador_22	71.42
sie_ic_indicador_25	74.07
sie_ic_indicador_26	5.22
sie_ic_indicador_27	5.22
sd_funcionario_encargado_del_proceso	0.76
sd_presupuesto_referencial_total_sin_iva	29.49
sd_autoridad_autoridad_ordenadora_de_gasto	7.26
sd_autoridad_maxima_autoridad_institucional	7.26
sd_autoridad_maxima_autoridad_responsable_de_la_gestion_administrativa__financiera	7.26
sd_comision_delegado_del_titular_del_area_requirente	80.63
sd_comision_delegado_del_profesional_afin_al_objeto_de_la_contratacion_designado_por_la_maxima_autoridad	96.01
sd_comision_profesional_afin_al_objeto_de_la_contratacion_designado_por_la_maxima_autoridad	60.26
sd_comision_profesional_designado_por_la_maxima_autoridad,_quien_lo_presidira	56.27
sd_comision_secretario/a	56.27
sd_comision_titular_del_area_requirente	75.64

Análisis de Variables Numéricas.

Se realizó un análisis descriptivo sobre las variables numéricas presentes en el conjunto de datos. Para ello, se utilizaron medidas estadísticas como la media, mediana, desviación estándar, valores mínimos y máximos. Posteriormente, se representaron gráficamente las distribuciones individuales de estas variables mediante histogramas con curvas de densidad, lo que permitió visualizar posibles sesgos, asimetrías y la existencia de valores atípicos. En la figura 2 se muestra la distribución de los valores de la columna considerada como variable objetivo: el indicador compuesto.

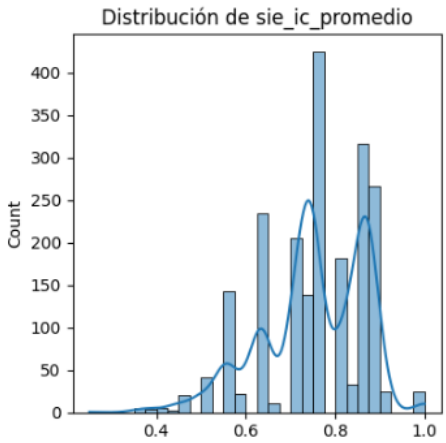

Figura 2. Distribución de la variable “sie_ic_promedio”

Finalmente, se generó una matriz de correlación para evaluar las relaciones lineales entre las variables numéricas seleccionadas. Esta matriz fue visualizada mediante un mapa de calor que permitió identificar asociaciones positivas o negativas entre las variables. Las correlaciones más destacadas pueden ser útiles para detectar redundancia de información o para seleccionar características relevantes en fases posteriores del modelado. Por ello, se construyó una matriz de correlación híbrida capaz de manejar diferentes tipos de datos: numéricos, categóricos y booleanos. Dado que las variables en el dataset poseen distintas naturalezas, se aplicaron métodos específicos para calcular sus correlaciones de manera adecuada:

- *Coeficientes de Pearson y Spearman:* Utilizados para medir la relación entre variables numéricas, permitiendo evaluar correlaciones lineales y no lineales.
- *Prueba de Chi-cuadrado y Cramer’s V:* Aplicadas para analizar la asociación entre variables categóricas y la variable objetivo.
- *Correlación Point-Biserial:* Empleada para determinar la relación entre variables booleanas y variables numéricas.

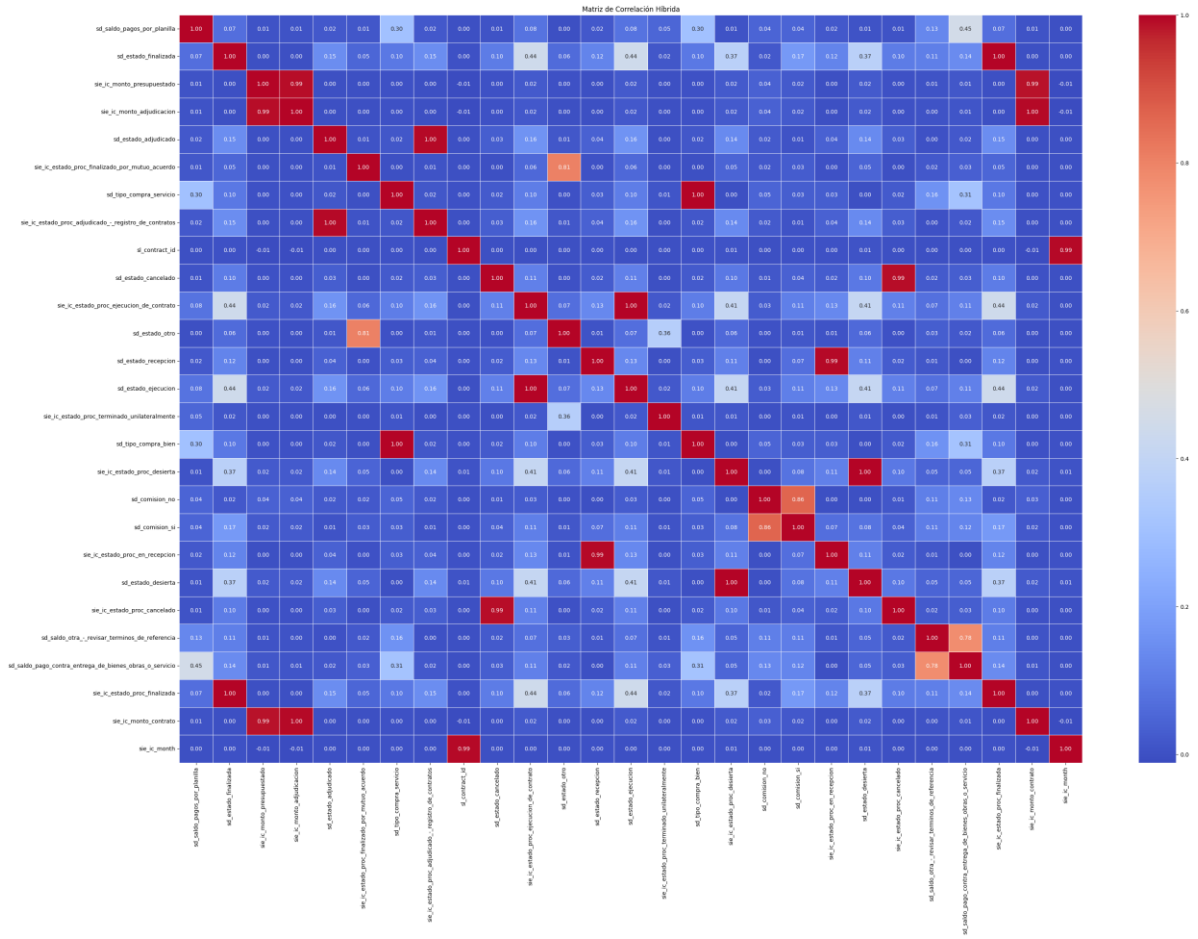

Figura 3. Matriz de correlación híbrida con las 20 variables que mayor correlación tienen con “sie_ic_promedio”

Esta matriz se puede visualizar en la figura 3. A partir de su análisis, se identificaron variables con una relación significativa con sie_ic_promedio, lo que permitió resaltar patrones relevantes en los datos:

- *Variables con correlaciones positivas significativas:* Se identificó que los indicadores 5, 11, 19, 22, 25 y 15 presentaron una asociación positiva con la variable objetivo, lo que sugiere que estos factores pueden estar relacionados con un mayor nivel de riesgo.
- *Variables con correlación negativa:* sie_ic_monto_adjudicacion mostró una correlación negativa con sie_ic_promedio, lo que indica que montos de adjudicación más altos podrían estar asociados con menores niveles de riesgo.

- *Relaciones destacadas:* sie_ic_estado_proc_desierta presentó una fuerte relación con valores altos de sie_ic_promedio, lo que sugiere que procesos declarados desiertos podrían estar vinculados con ciertos patrones de riesgo.

4.2. Procesamiento y limpieza de datos

Una vez realizada una exploración inicial del dataset, se procedió a depurar y limpiar los datos estructurados. El objetivo de este paso fue garantizar la consistencia y completitud de los datos, con el objetivo de hacerlo apto para entrenar diferentes modelos de aprendizaje automático. Para ello, se aplicaron distintas técnicas de imputación y codificación. En primer lugar, se identificaron y gestionaron los valores nulos presentes en las variables seleccionadas. Se empleó el método SimpleImputer de la biblioteca scikit-learn para realizar imputaciones específicas según el tipo de variable: para variables numéricas se utilizó la media o mediana como estrategia de reemplazo, mientras que para variables categóricas se optó por la imputación con la moda o una categoría genérica.

Posteriormente, se codificaron las variables categóricas. Dependiendo de la naturaleza de las variables, se aplicaron dos enfoques: codificación ordinal, cuando existía un orden implícito entre las categorías, y codificación por frecuencia (Frecuency Encoding) para variables sin orden inherente. Este proceso permitió transformar las categorías textuales en variables numéricas compatibles con los modelos de aprendizaje automático. Como resultado de esta fase, se obtuvo una versión limpia y numéricamente codificada del conjunto de datos, libre de valores ausentes y con todas sus variables convertidas a un formato estructurado adecuado para tareas de análisis más avanzadas. En la figura 4, se muestra un flujograma que representa el proceso realizado en esta etapa de imputación y codificación.

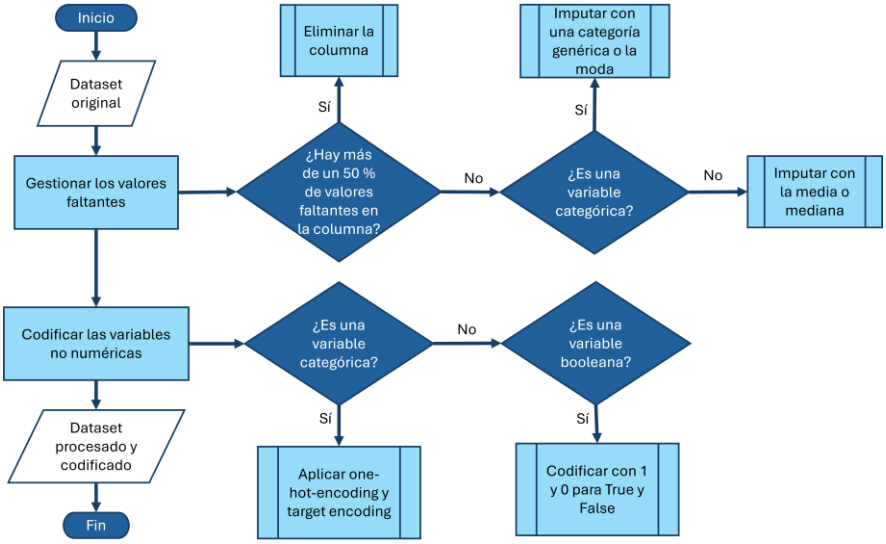

Figura 4. Flujograma del procesamiento y limpieza de datos

4.3. Extracción y selección de atributos estructurados

Una vez finalizada la etapa de preprocesamiento, se procedió a la selección de los atributos estructurados más relevantes para la tarea de modelado supervisado. Esta selección se basó en criterios estadísticos orientados a identificar las variables con mayor capacidad explicativa respecto a la variable objetivo. Para ello, se utilizó el método SelectKBest con la función de evaluación f_classif, que calcula la puntuación F para cada variable en función de su capacidad de discriminar entre clases. Este enfoque permitió conservar únicamente aquellos atributos que mostraban una asociación estadísticamente significativa con la variable objetivo.

Durante esta etapa también se analizaron las correlaciones entre variables para evitar la presencia de atributos redundantes o altamente correlacionados entre sí, lo cual podría generar problemas de multicolinealidad[1] en etapas posteriores. El resultado de esta fase fue un subconjunto de variables seleccionadas que conserva la mayor parte de la información relevante del conjunto original, al tiempo que reduce la dimensionalidad y mejora la eficiencia del proceso de aprendizaje automático.

[1] La multicolinealidad ocurre cuando dos o más variables independientes en un modelo estadístico están altamente correlacionadas, dificultando la estimación precisa de sus efectos individuales (Rodó, 2019).

4.4. Generación de vectores de representación (embeddings) con LLMs

En esta fase del trabajo se generaron representaciones vectoriales de los documentos asociados a cada proceso de contratación. El objetivo de este paso fue transformar el contenido textual no estructurado en una forma numérica interpretable por algoritmos de aprendizaje automático, permitiendo así incorporar información semántica proveniente de los textos a los modelos analíticos posteriores. Como entrada se tomó archivos PDF y XML correspondientes a los documentos descargados por cada proceso. Se utilizó el modelo text-embedding-3-small provisto por OpenAI, un modelo de lenguaje optimizado para la generación de embeddings semánticos.

El procedimiento general constó de los siguientes pasos:

1. *Cargado y extracción del contenido textual:* Se recorrieron las carpetas asociadas a cada sl_id, extrayendo el contenido de archivos PDF mediante pdfplumber y PyMuPDF, así como de archivos XML usando ElementTree.

2. *Filtrado y limpieza del contenido:* Se descartaron documentos cuyo contenido no superaba un umbral mínimo de 20 palabras, y se aplicaron expresiones regulares para eliminar caracteres innecesarios, encabezados irrelevantes u otros elementos no textuales.

3. *Segmentación:* El texto fue segmentado en fragmentos de 300 palabras, sin solapamiento entre ellos. Esta operación fue crítica para asegurar el procesamiento completo de documentos largos.

4. *Generación de embeddings:* Finalmente, se enviaron los textos al endpoint de OpenAI para la generación de vectores de embeddings. Los resultados se almacenaron en un archivo .parquet, el cual contiene, para cada sl_id, una lista de vectores que representan el contenido semántico de sus documentos asociados.

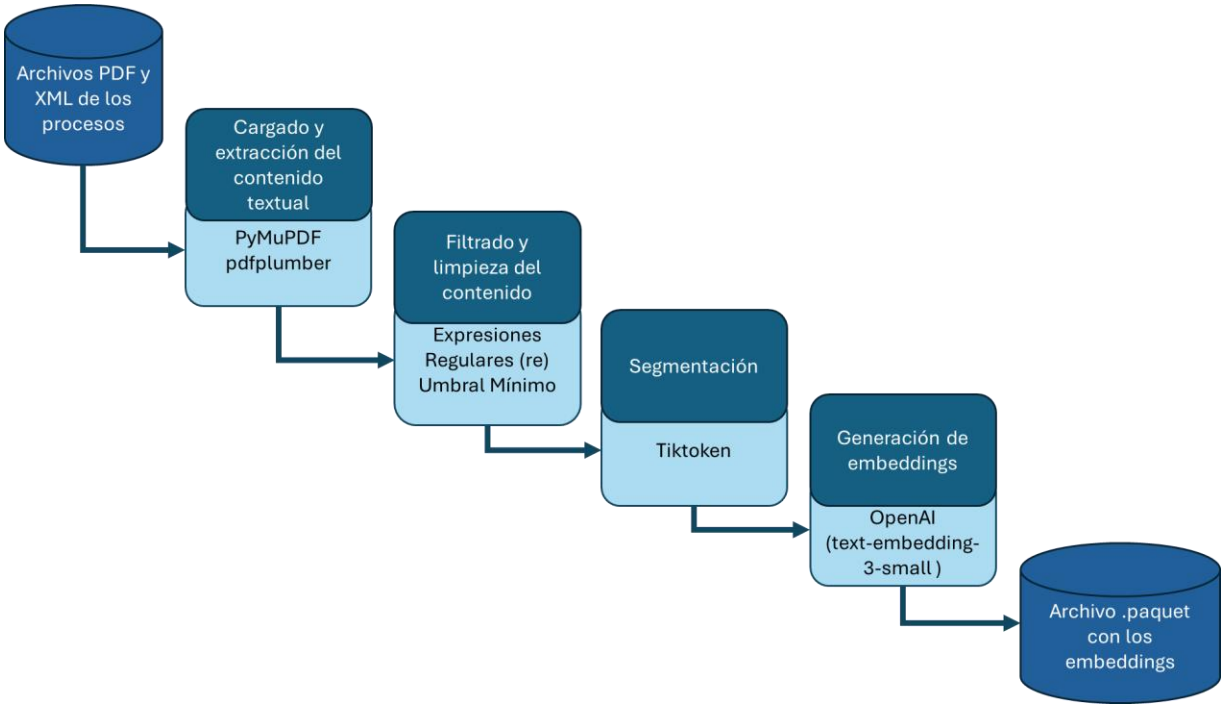

Figura 5. Flujograma de la generación de embeddings

En la figura 5 se muestra el flujo del procesamiento de los documentos para obtener las representaciones vectoriales. Esta etapa representa un puente crucial entre los datos no estructurados (documentos legales y administrativos) y las herramientas de análisis automatizado. Al generar embeddings de alta calidad, se posibilita el uso de técnicas avanzadas como clustering, reducción de dimensionalidad o clasificación basada en similitud semántica.

4.5. Segmentación, evaluación y selección de documentos relevantes

Dado que los modelos de generación de embeddings proporcionados por OpenAI, como `text-embedding-3-small`, imponen un límite de 8192 tokens por consulta, se hizo necesario diseñar una estrategia que permitiera representar documentos extensos sin perder su esencia semántica. Muchos de los procesos superan ampliamente este umbral, por lo que no pueden ser enviados directamente al modelo en una sola pieza. Frente a esta limitación técnica, la solución fue fragmentar los textos en bloques manejables, calcular sus embeddings

individuales y luego seleccionar aquellos fragmentos más representativos para componer una versión reducida y relevante del documento completo.

Además, se observó que una gran proporción de los documentos siguen estructuras repetitivas o utilizan plantillas comunes, lo cual introduce redundancia y dificulta la diferenciación semántica entre ellos. En estos casos, procesar todo el texto de forma indiscriminada podría resultar en embeddings muy similares, incluso si los documentos se refieren a procesos diferentes. Por ello, se incorporó un mecanismo de selección de fragmentos basado en la diversidad semántica interna, priorizando aquellos bloques de texto que contienen información menos redundante o más distintiva dentro de cada documento. Además, se comparó la información de los archivos con el objeto de compra de cada proceso, para buscar información relacionada con cada proceso en particular. Esta selección busca reducir el ruido informativo, mejorar la calidad del embedding final y facilitar tareas posteriores como la clasificación, la detección de anomalías o el análisis de riesgos.

4.5.1. Obtención de Embeddings finales a partir de fragmentos seleccionados

Inicialmente, cada documento es preprocesado para limitar su longitud a un máximo definido de palabras. A continuación, se divide en fragmentos consecutivos de tamaño fijo (por ejemplo, 300 palabras). Cada fragmento es convertido en un vector numérico (embedding) mediante un LLM (por ejemplo, text-embedding-3-small). Estos modelos transforman textos en vectores de alta dimensión donde la cercanía entre vectores refleja similitudes semánticas. Estos modelos se basan en redes neuronales profundas entrenadas sobre grandes corpus de datos textuales, utilizando técnicas como transformers y mecanismos de atención para capturar relaciones contextuales entre palabras.

Una vez obtenidos los embeddings de todos los fragmentos, se aplica un proceso de selección para identificar aquellos fragmentos que aportan información diversa dentro del documento. Para esto se calcula la matriz de similitud coseno entre todos los embeddings del mismo documento. La similitud coseno permite evaluar qué tan parecidos son dos fragmentos entre sí, sin importar su magnitud, considerando únicamente la orientación de los vectores en el espacio. Esta operación se realiza de forma eficiente utilizando FAISS (Facebook AI Similarity Search), una biblioteca optimizada para realizar búsquedas y comparaciones vectoriales a gran escala.

FAISS funciona como un motor de búsqueda de vectores, especializado en calcular distancias o similitudes en espacios de alta dimensión. En este caso, se emplea el índice IndexFlatIP, que calcula el producto interno (inner product) entre vectores. Para que este producto represente efectivamente la similitud coseno, se aplica previamente una normalización L2 a los vectores, de modo que todos tengan magnitud unitaria. Arquitectónicamente, FAISS está diseñado en C++ con enlaces a Python, y aprovecha instrucciones SIMD y GPU para acelerar operaciones de búsqueda. Internamente, permite almacenar grandes volúmenes de vectores en estructuras planas o jerárquicas, e indexarlas para consultas rápidas mediante técnicas como clustering, inverted files o quantization (aunque estas no se usan en IndexFlatIP).

Tras calcular la matriz de similitud, se determina para cada fragmento su similitud promedio con el resto del documento. Se descartaron aquellos fragmentos cuyo promedio de similitud con los demás fragmentos superaba un umbral de 0.85, criterio que típicamente corresponde a plantillas, documentos legales estándar o contenido poco específico. Si ningún conjunto cumple con esta condición, al menos se selecciona el fragmento con menor similitud media. Después de esta primera depuración, se realizó una nueva selección basada en el objeto

de compra de cada proceso. Para ello, se generó un embedding específico del objeto de compra y se calcularon las similitudes entre este embedding y los embeddings de los fragmentos remanentes de cada contrato. Los fragmentos fueron priorizados de acuerdo con esta similitud, y se seleccionaron aquellos que, en conjunto, no excedieran un límite máximo de 8192 tokens por proceso de compra. Este proceso buscó asegurar que el contenido seleccionado fuera tanto representativo como pertinente al propósito particular del proceso de compra.

Una vez seleccionados los fragmentos relevantes, se mapearon nuevamente al texto original, conservando el orden en el que aparecían dentro de los documentos fuente. Los fragmentos seleccionados fueron concatenados y, sobre el texto resultante, se generó un embedding final que resume el contenido más representativo del proceso.

Adicionalmente, para efectos comparativos, se implementaron otras dos estrategias de generación de embeddings finales:

- Promedio de los embeddings individuales de cada fragmento seleccionado.
- Suma de los embeddings individuales, buscando conservar la fuerza de los vectores en lugar de normalizarlos.

Estas variantes permitieron evaluar el impacto que distintas estrategias de agregación de información tienen en la calidad de las representaciones finales de los procesos de contratación.

El procedimiento completo fue desarrollado a través de la clase FragmentSelector, la cual implementa la división, generación de embeddings, filtrado por similitud, selección por objeto de compra, generación de embeddings finales, y visualización de matrices de similitud. Este proceso se muestra en la figura 6.

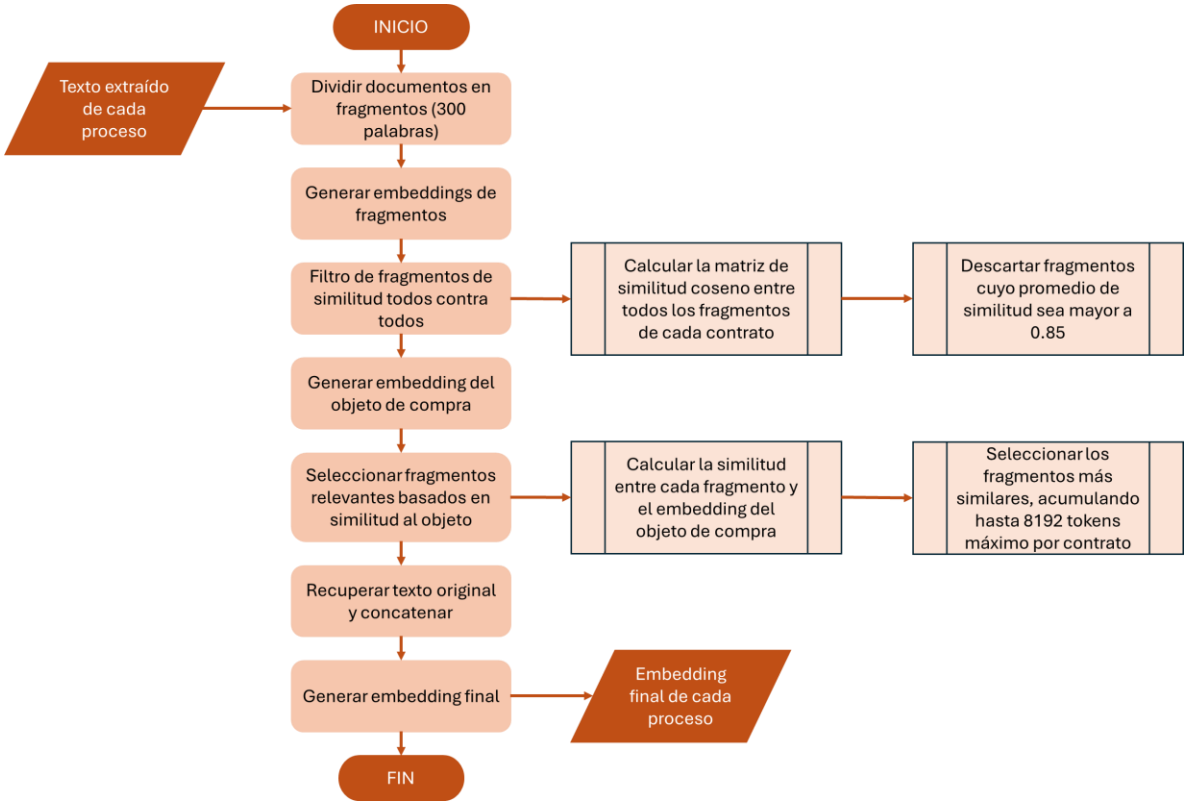

Figura 6: Proceso de selección de información relevante de los archivos y generación de embeddings

4.5.2. Consultas y recuperación semántica

Una vez construido el índice vectorial, se diseñó un sistema de consulta basado en el paradigma *Retrieval-Augmented Generation* (RAG), el cual combina mecanismos de recuperación de información con generación automática de texto. Este enfoque permite formular preguntas o enunciados y obtener respuestas generadas por un modelo de lenguaje a partir del contenido recuperado. En el flujo de trabajo principal desarrollado en este estudio, se empleó exclusivamente la etapa de recuperación de información mediante búsqueda semántica, sin incorporar la generación de lenguaje natural. Los componentes de este sistema se muestran en la tabla 2.

Tabla 2. Componentes del módulo de consultas y recuperación semántica.

No.	Componente	Descripción
1	Tokenización y embeding de la pregunta	La consulta textual es convertida en un vector de embeddings utilizando el mismo modelo de OpenAI empleado en la indexación.
2	Búsqueda en la base FAISS	Se ejecuta una búsqueda de vecinos más cercanos usando el vector de la consulta, recuperando los fragmentos más similares.
3	Presentación de resultados	Se genera una respuesta mediante un LLM de OpenAI a partir de un query realizado.

4.6. Integración del conjunto de datos final

Una vez completadas las etapas de preprocesamiento, generación de embeddings y selección de fragmentos relevantes, se procedió a la integración de la información proveniente de fuentes estructuradas y no estructuradas en un único conjunto de datos consolidado. Este dataset final constituye la base sobre la cual se desarrollarán los modelos de aprendizaje automático y análisis posteriores. El propósito de esta fase fue unificar, mediante una clave común, los vectores semánticos generados a partir de los documentos (embeddings) y los atributos estructurados codificados de cada proceso de contratación. La columna sl_id fue utilizada como identificador único para realizar esta fusión. Una vez realizada esta integración, se efectuaron comprobaciones adicionales para eliminar duplicados y asegurar que todos los registros contaran con información tanto semántica como estructurada.

Después de consolidar el dataset, se procedió a guardar distintas versiones del mismo. Primero, se almacenó una versión sin etiquetas, pensada para tareas de clustering no supervisado. Luego, se generaron múltiples datasets etiquetados según el método de obtención de los embeddings finales (selección por similitud promedio, suma o promedio). Las etiquetas de riesgo se asignaron en función de un indicador compuesto, y se clasificaron en dos categorías: bajo riesgo y riesgo alto. La determinación del umbral óptimo para esta

clasificación fue realizada en dos etapas. En primer lugar, se aplicó un análisis de varianza (ANOVA) para evaluar cómo se distribuían los valores del indicador respecto a distintas divisiones posibles del umbral. Este análisis permitió identificar el punto de corte que maximiza la varianza entre grupos, reflejado en el mayor valor de la estadística F. Como se observa en la figura 6, el valor óptimo de umbral obtenido fue 0.56, correspondiente al pico más pronunciado de la curva de F-score promedio.

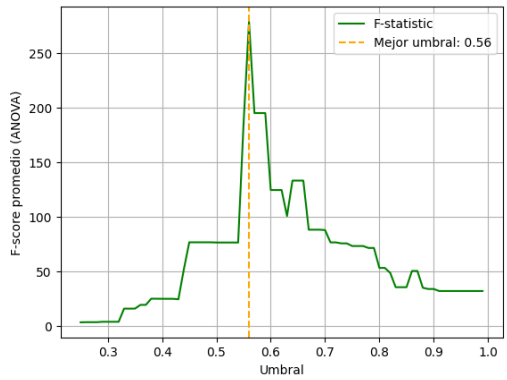

Figura 7: Definición del umbral óptimo para etiquetar con varianza intergrupal

En una segunda fase, se optimizó el umbral mediante un enfoque supervisado, ajustando modelos de clasificación binaria y evaluando su desempeño con la métrica F1-score sobre distintas divisiones del indicador. El gráfico resultante muestra que el mejor rendimiento del modelo también se alcanzó con un umbral cercano a 0.55, validando la elección previa (Figura 7). Estos análisis combinados permitieron definir con mayor precisión un punto de corte robusto para la asignación de etiquetas en el conjunto de datos.

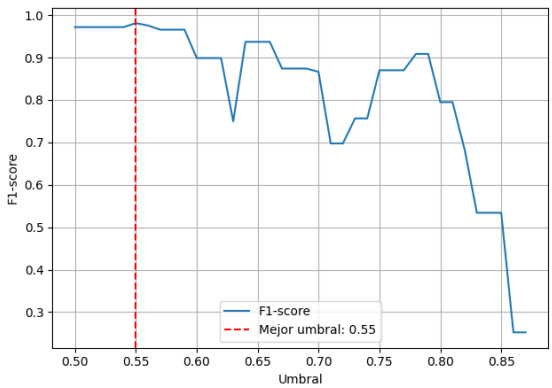

Figura 8: Optimización del umbral mediante análisis de F1-Score de clasificación rápida

Para construir el conjunto de datos final utilizado en los experimentos, se aplicó undersampling aleatorio con el objetivo de equilibrar las clases, garantizando que ambas categorías de riesgo tuvieran la misma cantidad de observaciones. Antes de proceder con esta técnica, se examinó la distribución del indicador compuesto de riesgo para los tres años considerados en el análisis (2020, 2022 y 2023). Los resultados mostraron que las distribuciones eran suficientemente similares como para justificar el uso de undersampling sin introducir sesgos relevantes (Ver figura 9). Esta verificación se complementó mediante un análisis estadístico: la Wasserstein distance entre los años comparados fue baja (0.0115 entre 2020 y 2022; 0.0228 entre 2020 y 2023; 0.0129 entre 2022 y 2023), indicando una alta similitud entre las distribuciones. Adicionalmente, se realizaron pruebas de Kolmogorov-Smirnov (KS), que arrojaron estadísticos moderadamente bajos (0.0545, 0.1168 y 0.0648 respectivamente) aunque, debido al gran tamaño de las muestras, los p-valores resultaron bajos. En conjunto, estos análisis proporcionan evidencia suficiente para considerar que el procedimiento de undersampling es adecuado en este contexto. Así, el dataset final cuenta con 465 registros por cada categoría, para un total de 930 filas.

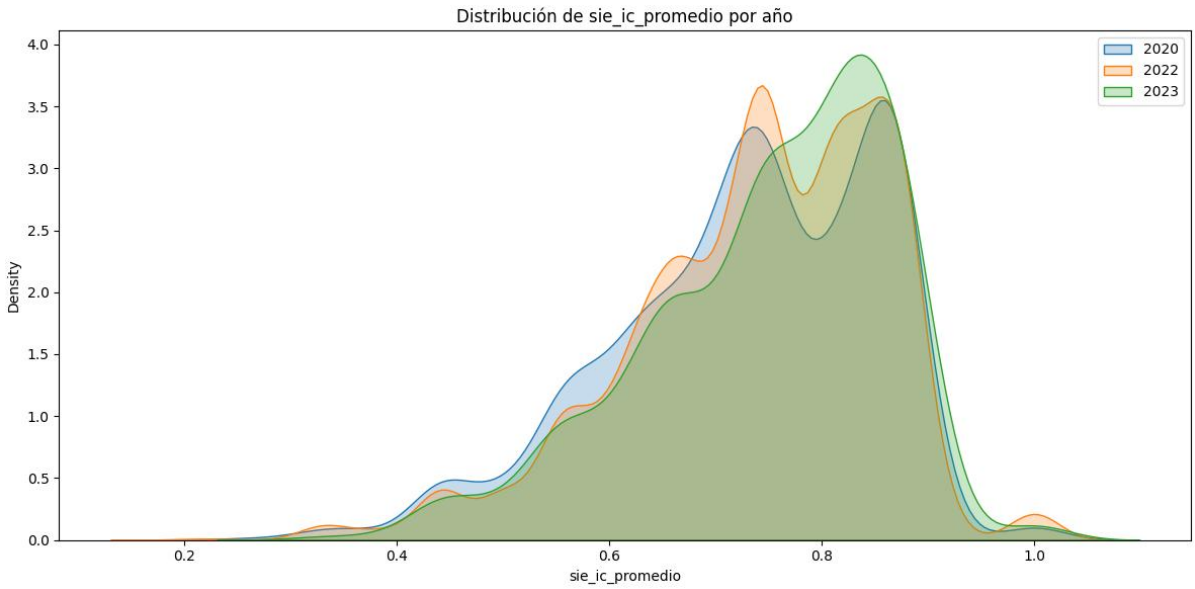

Figura 9: Distribución del indicador compuesto por cada año

4.7. Análisis de clústeres sobre el conjunto de datos

El análisis de clústeres constituye una técnica no supervisada utilizada para identificar patrones de agrupamiento dentro de un conjunto de datos. Esta técnica se aplicó con el objetivo de explorar estructuras latentes en los procesos contractuales, a partir de la combinación de variables estructuradas y representaciones semánticas obtenidas previamente. Para facilitar la visualización e interpretación de los grupos, se aplicó una transformación estándar de los datos mediante StandardScaler, con el fin de normalizar las escalas de las variables. Se experimentó con distintos algoritmos de agrupamiento, para comparar su comportamiento y resultados:

- ***K-Means y MiniBatchKMeans:*** Algoritmos basados en la partición de los datos según centroides. Se probaron con diferentes valores de k (número de clústeres).
- ***DBSCAN:*** Algoritmo basado en la densidad, que permite detectar grupos de distinta forma y tamaño, así como identificar puntos atípicos (ruido).

Para evaluar la calidad de los agrupamientos generados se utilizaron métricas internas, presentadas en la tabla 3.

Tabla 3. Métricas utilizadas en el análisis del proceso de agrupamiento.

No.	Métrica	Descripción	Objetivo
1	Silhouette Score	Mide la coherencia de los clústeres, comparando la distancia intra-clúster con la distancia al clúster más cercano.	Cercano a 1
2	Calinski-Harabasz	Evalúa la separación entre los clústeres con base en la dispersión interna y externa.	Alto
3	Davies-Bouldin	Estima la media de la relación entre la dispersión intra-clúster y la separación entre clústeres.	Cercano a 0

Estos clústeres pueden representar categorías funcionales, estructuras de riesgo, niveles de complejidad o grupos semánticamente similares, y resultan útiles para futuras tareas de auditoría, análisis exploratorio o modelado supervisado.

4.8. Implementación de modelos de clasificación supervisada

Con el objetivo de predecir la categoría de riesgo asociada a cada proceso contractual, se implementaron distintos modelos de clasificación supervisada. Estos modelos se entrenaron utilizando el conjunto de datos final consolidado y etiquetado, lo que permitió evaluar su capacidad para aprender patrones a partir de características tanto estructuradas como semánticas. Previo al entrenamiento, se realizó una separación del conjunto de datos en subconjuntos de entrenamiento y prueba utilizando la función train_test_split, con el fin de evaluar el rendimiento de los modelos en datos no vistos.

4.8.1. Modelos implementados.

Se exploraron diversos algoritmos de clasificación supervisada, abarcando enfoques tanto lineales como no lineales, estos se muestran en la tabla 4.

Tabla 4. Modelos de clasificación supervisada utilizados

No.	Modelo	Descripción
1	Random Forest	Algoritmo basado en ensamblado de árboles de decisión, robusto frente a ruido y eficaz en la captura de interacciones no lineales.
2	Gradient Boosting	Variante eficiente del método de gradient boosting, optimizada para grandes conjuntos de datos y capaz de manejar valores faltantes de forma nativa.
3	Support Vector Classifier (SVC)	Modelo basado en hiperplanos de separación que maximiza el margen entre clases. Requiere normalización de datos y selección adecuada de kernel.
4	Red neuronal multicapa	Red neuronal multicapa capaz de modelar relaciones complejas entre variables mediante capas ocultas y funciones de activación.

4.8.2. Evaluación de desempeño.

Para cada modelo se calcularon las principales métricas de evaluación, destacando especialmente:

- *AUC-ROC:* área bajo la curva ROC, utilizada como métrica principal para medir la capacidad discriminativa de los modelos entre las categorías de riesgo.
- *F1-score:* media armónica entre precisión y recall, particularmente útil en escenarios donde podría haber cierto desbalance entre clases.
- *Precisión y Recall*: métricas individuales por clase que reflejan, respectivamente, la exactitud y la cobertura de las predicciones.

El entrenamiento de los modelos se llevó a cabo utilizando un grid search para encontrar de forma estadística la mejor combinación de hiperparámetros. Posteriormente, se realizó una validación cruzada k-fold con k=10 para evaluar la robustez y generalización de los modelos de manera más rigurosa. Adicionalmente, se generaron informes de clasificación y, cuando fue pertinente, matrices de confusión para interpretar los patrones de error más comunes entre las clases.

5. EVALUACIÓN DE RESULTADOS

5.1. Resultados generales de los modelos implementados

Los modelos desarrollados a lo largo del presente trabajo evidencian la viabilidad de aplicar técnicas de aprendizaje automático, tanto supervisadas como no supervisadas, al análisis de procesos contractuales a partir de datos estructurados y no estructurados. El enfoque integral adoptado permitió comprender mejor el contenido de los documentos y su relación con los indicadores de riesgo calculados dentro del proyecto Kapak. La metodología empleada facilitó un análisis más profundo de la información disponible, fortaleciendo la capacidad de interpretar los documentos contractuales en el contexto de las señales de riesgo que se buscan detectar.

Los resultados del análisis de clústeres revelan que los modelos que emplearon dos grupos (k=2) obtuvieron los mejores desempeños generales según las métricas de evaluación interna. En particular, el algoritmo MiniBatchKMeans con k=2 alcanzó el mayor Silhouette Score (0.136) y un valor Davies-Bouldin (2.376) más bajo que sus contrapartes con mayor número de clústeres, lo que sugiere una adecuada cohesión interna y separación entre los grupos. Asimismo, el modelo KMeans con k=2 presentó resultados muy similares, con una ligera ventaja en el índice de Calinski-Harabasz (133.1), indicador que también favorece la partición en dos clústeres. A medida que se incrementa el número de clústeres (k=3 o k=4), las métricas tienden a deteriorarse: el Silhouette Score disminuye y los índices Davies-Bouldin aumentan, lo que denota una menor claridad en la estructura de los grupos. En conjunto, estos resultados sugieren que el conjunto de datos presenta una segmentación natural más clara cuando se agrupa en dos clústeres, siendo este el valor óptimo de k bajo los criterios aplicados. Todos estos resultados se resumen en la tabla 5.

Tabla 5. Mejores modelos de clustering

Modelo	# Clusters	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
MiniBatchKMeans	2	0.136456	132.588164	2.376296
MiniBatchKMeans	3	0.135729	82.097952	3.191962
KMeans	2	0.133711	133.117611	2.341603
KMeans	3	0.133020	90.768012	3.141648
KMeans	4	0.101021	73.499451	2.955494

En cuanto al modelado supervisado, se observó un mejor desempeño al utilizar la información estructurada de los procesos, logrando resultados aceptables en términos de clasificación utilizando técnicas como árboles de decisión y ensambles. La combinación de atributos estructurados con representaciones semánticas derivadas de los documentos permitió enriquecer los modelos predictivos, aunque el aporte de los documentos fue más evidente en el entendimiento cualitativo que en una mejora cuantitativa sustancial del rendimiento. A pesar de estas limitaciones, el enfoque supervisado demostró que es posible construir modelos capaces de capturar patrones relevantes asociados al riesgo de corrupción, proporcionando una base sólida para futuras mejoras metodológicas.

5.2. Comparación de desempeño entre representaciones y clasificadores

Los modelos de clasificación implementados fueron evaluados en función de su rendimiento sobre el conjunto de prueba, utilizando métricas estándar como AUC-ROC, precision, recall y F1-score. Además, se documentaron los mejores hiperparámetros encontrados a través de búsqueda en rejilla (GridSearchCV), lo cual permitió optimizar el comportamiento de cada clasificador. Posteriormente, se aplicó validación cruzada K-fold con $k = 10$ para estimar de manera robusta el rendimiento de los modelos. En la tabla 6, se reportan

los promedios y desviaciones estándar (±) de cada métrica luego de aplicar la validación cruzada. Se pueden identificar varios hallazgos clave que aportan evidencia relevante sobre el comportamiento de los distintos modelos de clasificación utilizados, así como sobre el valor predictivo de las diferentes representaciones del dataset (atributos estructurados, embeddings, y su combinación).

Tabla 6: Rendimiento promedio de los modelos de clasificación supervisada tras validación cruzada (k = 10)

Dataset	Modelo	AUC-ROC	Precision	Recall	F1-Score
Solo atributos	Red Neuronal	0.83 ± 0.05	0.75 ± 0.04	0.79 ± 0.08	0.76 ± 0.05
	SVM	0.81 ± 0.05	0.72 ± 0.03	0.80 ± 0.07	0.76 ± 0.04
	Random Forest	0.82 ± 0.05	0.73 ± 0.06	0.82 ± 0.09	0.77 ± 0.06
	Gradient Boosting	0.84 ± 0.05	0.74 ± 0.05	0.81 ± 0.08	0.77 ± 0.05
Atributos + embeddings (suma)	Red Neuronal	0.76 ± 0.06	0.70 ± 0.04	0.71 ± 0.07	0.70 ± 0.05
	SVM	0.80 ± 0.06	0.74 ± 0.05	0.70 ± 0.06	0.72 ± 0.05
	Random Forest	0.73 ± 0.06	0.70 ± 0.06	0.63 ± 0.07	0.67 ± 0.06
	Gradient Boosting	0.79 ± 0.05	0.74 ± 0.05	0.68 ± 0.06	0.71 ± 0.05
Atributos + embeddings (selección todos contra todos)	Red Neuronal	0.71 ± 0.05	0.65 ± 0.04	0.65 ± 0.05	0.65 ± 0.04
	SVM	0.70 ± 0.06	0.65 ± 0.05	0.67 ± 0.06	0.66 ± 0.05
	Random Forest	0.74 ± 0.06	0.71 ± 0.06	0.60 ± 0.08	0.65 ± 0.07
	Gradient Boosting	0.78 ± 0.06	0.75 ± 0.05	0.67 ± 0.07	0.71 ± 0.06
Atributos + embeddings (selección con objeto de compra)	Red Neuronal	0.71 ± 0.06	0.66 ± 0.05	0.65 ± 0.06	0.65 ± 0.05
	SVM	0.69 ± 0.06	0.65 ± 0.06	0.64 ± 0.08	0.64 ± 0.07
	Random Forest	0.73 ± 0.05	0.70 ± 0.04	0.62 ± 0.08	0.66 ± 0.06
	Gradient Boosting	0.80 ± 0.05	0.76 ± 0.04	0.67 ± 0.07	0.71 ± 0.05

El análisis de los modelos entrenados únicamente con atributos estructurados muestra que se obtuvo el mejor desempeño general. Los valores de AUC-ROC oscilaron entre 0.8133 (SVM) y 0.8352 (Gradient Boosting), con F1-Scores relativamente consistentes, todos por

encima de 0.75. En particular, Gradient Boosting logró el mayor AUC-ROC (0.8352) y un F1-Score de 0.7686, evidenciando su capacidad para capturar las relaciones estructurales de los datos de manera efectiva. Random Forest también presentó un rendimiento competitivo, ligeramente inferior pero consistente. La Red Neuronal y SVM mostraron un desempeño razonable, aunque algo por debajo en precisión y estabilidad.

Cuando se incorporaron embeddings generados por suma al conjunto de atributos, se observó un descenso generalizado en el rendimiento de todos los modelos. Los AUC-ROC se redujeron, situándose entre 0.7346 (Random Forest) y 0.7984 (SVM). Asimismo, los F1-Scores disminuyeron en comparación con el dataset de solo atributos. Esto sugiere que la simple suma de embeddings no aportó una representación semántica suficientemente diferenciadora y, en algunos casos, pudo haber introducido ruido, afectando la capacidad de clasificación de los modelos.

Utilizando embeddings generados a partir de la selección por filtrado de fragmentos (todos contra todos), se mantuvo una tendencia similar de desempeño moderado. El AUC-ROC máximo alcanzado fue de 0.7813 (Gradient Boosting), mientras que el F1-Score más alto fue de 0.7085, también obtenido por este modelo. Aunque hubo una ligera mejora frente a la simple suma de embeddings en algunos casos, los resultados siguieron estando por debajo de los obtenidos utilizando solo atributos. Esto evidencia que el filtrado basado únicamente en similitud no fue suficiente para capturar la totalidad del contenido relevante para la clasificación.

Al emplear la selección de fragmentos basada en la similitud con el objeto de compra, los resultados mejoraron respecto a la selección por similitud general, aunque aún no alcanzaron el nivel de desempeño del dataset de solo atributos. Gradient Boosting nuevamente destacó con un AUC-ROC de 0.7986 y un F1-Score de 0.7070. La inclusión del objeto de

compra como criterio de relevancia permitió preservar fragmentos más significativos para la tarea, aportando mejoras modestas, pero consistentes, en la capacidad de predicción de riesgo.

En conjunto, los resultados muestran que la información estructurada de los procesos es, por sí sola, muy informativa para la predicción del nivel de riesgo, superando consistentemente las combinaciones que incluían representaciones semánticas. La figura 8 muestra una comparación del F1-Score y ACU-ROC para todos los algoritmos. Si bien la integración de embeddings permitió explorar nuevas dimensiones del contenido documental, los métodos utilizados para su agregación y selección aún presentan limitaciones para capturar plenamente la complejidad del problema. No obstante, la metodología implementada permitió obtener importantes aprendizajes sobre la estructura de los documentos, la relación entre los textos y los indicadores, y proporcionó una base sólida para el diseño de futuros métodos de procesamiento semántico más eficaces.

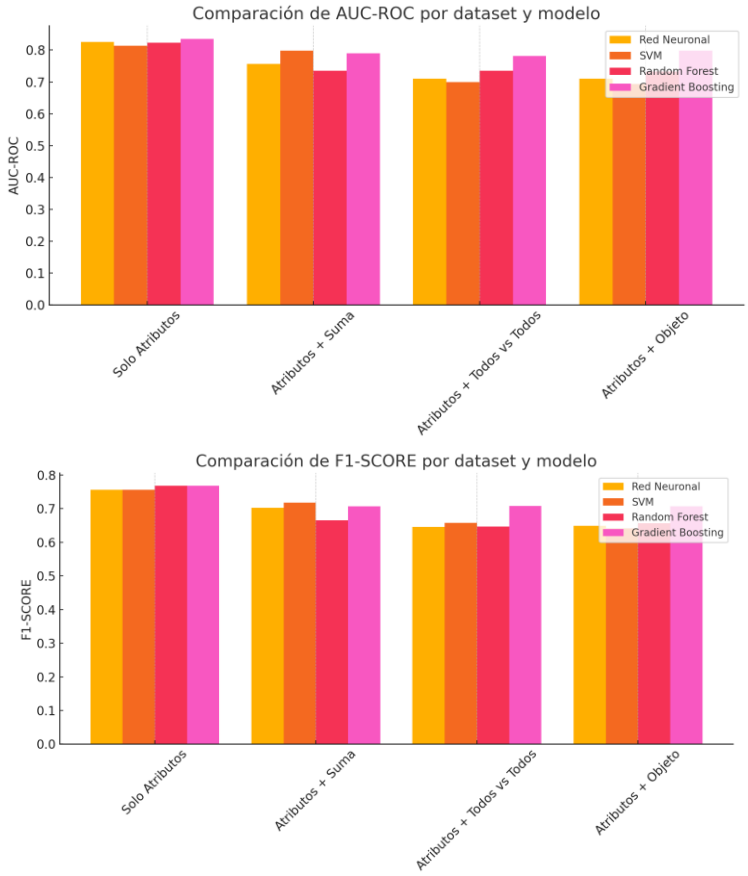

Figura 8. Comparación de métricas obtenidas para los modelos de clasificación

5.3. Análisis de métricas de evaluación

La evaluación objetiva del rendimiento de los modelos implementados se realizó mediante un conjunto de métricas que permitieron valorar tanto la calidad de los agrupamientos no supervisados como la eficacia de las predicciones en el caso de los modelos supervisados. La elección y análisis de estas métricas fueron fundamentales para garantizar la validez de los resultados y evitar interpretaciones parciales o sesgadas. Al evaluar los modelos desarrollados, se observan diferencias relevantes entre los valores de AUC-ROC y F1-Score obtenidos en los distintos experimentos. Conceptualmente, estas métricas capturan aspectos complementarios de la calidad de un modelo, por lo que su interpretación conjunta es fundamental.

El AUC-ROC mide la capacidad de un modelo para discriminar entre clases, evaluando su rendimiento en todos los posibles umbrales de decisión. Un valor alto de AUC-ROC, como los observados principalmente en los modelos entrenados con atributos estructurados (superiores a 0.8 en varios casos), indica que los modelos tienen una buena capacidad para distinguir procesos de bajo y alto riesgo, incluso antes de elegir un umbral de corte para hacer predicciones. Es decir, aunque el modelo pueda cometer errores puntuales en la clasificación final (por ejemplo, etiquetar mal algunos procesos), su estructura interna de puntuaciones refleja correctamente el orden de riesgo.

El F1-Score refleja el balance entre precisión (qué tan confiables son las predicciones positivas) y recall (qué proporción de los verdaderos positivos es capturada). A diferencia del AUC-ROC, el F1-Score depende directamente de un umbral de clasificación específico. Los valores de F1-Score obtenidos, aunque aceptables (en torno a 0.70-0.76 en los mejores casos), son consistentemente menores que los AUC-ROC. Esto indica que, aunque el modelo discrimina bien en términos relativos, cuando se fija un umbral estándar (por ejemplo, 0.55), la clasificación no logra capturar de manera equilibrada todas las instancias relevantes.

6. DISCUSIÓN

El análisis de las métricas de evaluación muestra tendencias claras sobre el desempeño de los modelos desarrollados en este trabajo. En general, los modelos entrenados únicamente con atributos estructurados obtuvieron los mejores resultados tanto en AUC-ROC como en F1-Score, superando a los modelos que integraban embeddings derivados de documentos.

Desde un punto de vista conceptual, los valores de AUC-ROC observados en los modelos de solo atributos (superiores a 0.8 en varios casos) indican que existe una capacidad robusta para ordenar los procesos en función de su nivel de riesgo, aun sin un umbral de decisión fijo. Esto sugiere que la información estructurada capturada en los registros de procesos contractuales como montos, número de oferentes o adjudicaciones, contiene signos relevantes para diferenciar entre procesos de bajo y alto riesgo. Sin embargo, al analizar los F1-Score, se observa que, aunque razonables (en el rango de 0.75 a 0.77 para los mejores modelos), los valores son inferiores a los AUC-ROC. Esta diferencia muestra que, si bien los modelos capturan correctamente la tendencia de riesgo, la clasificación binaria en un punto de corte específico es más complicada.

En contraste, la incorporación de embeddings semánticos, ya sea mediante suma o mediante selección basada en similitud, no logró mejorar los resultados de manera sustancial. De hecho, en muchos casos, el desempeño empeoró respecto al uso exclusivo de atributos estructurados. Esto sugiere que, en su forma actual de implementación, los embeddings no añadieron suficiente información diferenciadora o, alternativamente, introdujeron ruido que complicó el proceso de clasificación. Una posible explicación es que la representación semántica obtenida a partir de fragmentos de documentos, aunque permite entender el contenido textual, no siempre se traduce en información predictiva directa para la clasificación de riesgos. Los textos de los documentos pueden contener información redundante, formal o

genérica (por ejemplo, cláusulas legales o plantillas de contratación) que no aportan señales claras sobre irregularidades. Otro elemento a considerar es que el modelo de generación de embeddings utilizado no fue específicamente entrenado para el dominio de compras públicas o riesgo contractual, lo cual podría haber limitado su capacidad de capturar matices sutiles que sí son relevantes para el problema en cuestión.

No obstante, el uso de embeddings aportó beneficios indirectos: permitió comprender mejor la relación entre los documentos y los indicadores de riesgo utilizados en el proyecto Kapak. El análisis semántico contribuyó a identificar qué fragmentos de documentos tienden a ser más relevantes en los procesos de riesgo alto, ofreciendo una vía de exploración importante para futuras estrategias de selección de información o para sistemas de alerta temprana basados en contenido textual.

En conjunto, estos resultados resaltan que, si bien el procesamiento semántico de documentos abre nuevas posibilidades para enriquecer los modelos de riesgo, su impacto efectivo depende críticamente de la calidad, relevancia y forma de integración de los embeddings en el flujo de análisis. El reto no radica únicamente en generar representaciones vectoriales, sino en asegurar que éstas aporten información complementaria y específica que mejore el poder predictivo de los sistemas.

7. CONCLUSIONES

Este trabajo realizó una exploración de la construcción de una metodología efectiva para la clasificación del riesgo en procesos de compras públicas mediante el uso combinado de datos estructurados y representaciones semánticas generadas a partir de modelos de lenguaje de gran escala (LLMs). La hipótesis planteada se abordó desde múltiples enfoques, lo que permitió evaluar no solo la eficacia de los modelos de embeddings como representación de contenido documental, sino también su interacción con atributos estructurados obtenidos mediante el procesamiento realizado dentro del proyecto Kapak.

A lo largo de este trabajo se pudo comprobar que los datos estructurados que se extraen de los procesos de contratación pública sí permiten, en cierta medida, identificar procesos con mayor o menor nivel de riesgo. Los modelos de Machine Learning, especialmente los basados en árboles de decisión, lograron capturar bien la información que estaba reflejada en el indicador compuesto, lo que demuestra que los atributos como montos, estados y número de oferentes tienen una relación clara con las señales de riesgo que se buscan detectar.

Por otro lado, al analizar los documentos asociados a los procesos (como pliegos y especificaciones técnicas) se encontró que, al menos con el enfoque usado, no existe una relación fuerte entre el contenido textual de los archivos y los niveles de riesgo medidos por el indicador. A pesar de aplicar técnicas modernas de embeddings para representar el contenido de los documentos, los resultados no mejoraron en comparación con usar solo los atributos estructurados. Esto puede explicarse porque muchos de los textos disponibles son plantillas legales o documentos muy estandarizados que no reflejan diferencias reales en el riesgo de cada contratación.

Un aporte importante de este proyecto fue el desarrollo de un algoritmo que limpia y selecciona fragmentos relevantes de los documentos. Primero, se eliminaron las secciones que

eran demasiado parecidas entre sí y después se priorizaron los fragmentos que tenían mayor relación con el objeto de compra de cada proceso. Aunque este enfoque todavía tiene margen de mejora, permitió acercarnos a una forma más inteligente de trabajar con documentos largos y variados.

En cuanto a los modelos de clasificación, se concluye que Gradient Boosting fue sistemáticamente el más eficaz, tanto en escenarios con atributos estructurados como en combinaciones con embeddings. Esta consistencia sugiere que los modelos basados en árboles y boosting están especialmente bien adaptados para aprovechar la riqueza y diversidad de características en este tipo de problemas.

La realización de este trabajo implicó diversas dificultades, tanto técnicas como metodológicas. Una de las principales fue el procesamiento de documentos extensos con contenido redundante o de baja variabilidad semántica, lo cual afectó inicialmente la calidad de los embeddings generados. Fue necesario desarrollar e implementar estrategias de truncamiento, segmentación y selección informativa para superar esta barrera. Otra dificultad relevante estuvo asociada a la necesidad de etiquetar los datos de forma consistente, lo cual requirió una calibración del umbral de riesgo sobre el indicador compuesto. Este proceso implicó un balance entre criterios estadísticos (como ANOVA) y desempeño empírico en modelos supervisados, que debió iterarse cuidadosamente. Finalmente, la integración entre datos estructurados y semánticos presentó desafíos adicionales de alineación, validación de llaves primarias y detección de duplicados, que requirieron una atención técnica para garantizar la consistencia del dataset final.

En definitiva, este trabajo permitió explorar y validar un enfoque integral para el análisis automatizado de riesgo en procesos de compras públicas, combinando capacidades semánticas de modelos de lenguaje con la información de cada proceso.

8. TRABAJO FUTURO

Los resultados obtenidos en este estudio abren diversas líneas de investigación para profundizar en el análisis de riesgo de corrupción en procesos de contratación pública mediante técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático.

Una posible dirección es el desarrollo de métodos que utilicen embeddings para identificar si los documentos asociados a los procesos de contratación presentan indicios de haber sido dirigidos hacia una entidad específica. Esto podría lograrse mediante la detección de patrones lingüísticos o estructuras textuales que favorezcan a ciertos proveedores, lo cual sería indicativo de prácticas de favoritismo o concursos dirigidos (Bonina, 2023). La implementación de modelos de NLP en este contexto permitiría automatizar la identificación de tales patrones, facilitando la detección temprana de posibles irregularidades.

Asimismo, se propone la construcción de indicadores de riesgo de corrupción basados en los embeddings generados a partir de los documentos de los procesos de contratación. Estos indicadores podrían complementar los existentes, que suelen centrarse en datos estructurados, proporcionando una visión más completa del riesgo al incorporar información semántica. La integración de estos nuevos indicadores en modelos predictivos podría mejorar la precisión en la identificación de procesos con alto riesgo de corrupción (CAF, 2024).

Finalmente, se considera necesario depurar la metodología propuesta, centrándose en cómo hacer más robustos los pasos realizados. En este sentido, por ejemplo, sería importante especializar el uso de embeddings, de modo que no capturen únicamente información general de los documentos, sino que se alineen con patrones relacionados a prácticas relacionadas con el riesgo de corrupción. Trabajar en la definición más precisa del dominio de conocimiento al momento de procesar los documentos podría permitir que las representaciones semánticas sean realmente útiles para construir indicadores de riesgo más enfocados y efectivos.

REFERENCIAS BIBLIOGRÁFICAS

Aggarwal, C. C., & Zhai, C. (2012). *A survey of text classification algorithms*. In C. C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 163-222). Springer.

Ariza, D., & Pardo, A. (2021). *Natural Language Processing for Government Transparency: The Case of Public Procurement in Colombia*. *Government Information Quarterly, 38*(4), 101567.

Banco Mundial. (2017). *Las adquisiciones públicas transparentes y eficientes son clave para el desarrollo*. Banco Mundial.

Barot, R. (2023). *Detecting Collusion in Public Procurement: A Comparative Study of Machine Learning Models* [Master's thesis, Concordia University, Montreal, Canada].

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., & Reddy, S. (2024). *LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders*. *COLM 2024*. https://arxiv.org/abs/2404.05961

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information*. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Bonina, N. (2023). *Inteligencia Artificial y Contrataciones Públicas*. Fiscalía General del Estado. Recuperado de https://fiscalia.chubut.gov.ar/wp-content/uploads/2023/05/BONINA-IA-Contrataciones-publicas.pdf

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language Models are Few-Shot Learners*. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

CAF. (2024). *Rol de los datos y las innovaciones digitales en la integridad de las contrataciones públicas*. Recuperado de https://www.caf.com/es/capacitaciones/rol-de-los-datos-y-las-innovaciones-digitales-en-la-integridad-de-las-contrataciones-publicas/

Clarke, J., & O'Connor, R. (2021). *Automated Detection of Fraud in Public Procurement: The Case of PROACT*. *European Journal of Public Administration, 23*(2), 145-167.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186. Disponible en: https://aclanthology.org/N19-1423/

Dhurandhar, A., Graves, B., Ravi, R., Maniachari, G., & Ettl, M. (2015). *Big data system for analyzing risky procurement entities*. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1741–1750.

Fortuny, M., Guerrero, E., Riofrío, D., & Simon, F. (2023). *Towards smart citizen control in public procurement: Ecuador's case study*. *2023 Ninth International Conference on eDemocracy & eGovernment (ICEDEG)*, 1–6.

ISO/IEC. (2022a). *ISO/IEC 23053:2022: Framework for Artificial Intelligence (AI) systems using Machine Learning (ML)*. International Organization for Standardization.

ISO/IEC. (2022b). *ISO/IEC TS 4213:2022: Artificial intelligence — Assessment of machine learning classification performance*. International Organization for Standardization.

ISO/IEC. (2020). *ISO/IEC TR 24028:2020: Information technology — Artificial intelligence — Overview of trustworthiness in AI*. International Organization for Standardization.

Jiang, Z., Wu, L., & Liang, P. (2024). *Mistral: A Scalable and Lightweight Transformer Model for NLP*. *NeurIPS 2024*.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed.). MIT Press.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2011). *The Worldwide Governance Indicators: Methodology and Analytical Issues*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *Nature, 521*(7553), 436-444.

Lyra, M. S., Damásio, B., Pinheiro, F. L., & Bacao, F. (2022). *Fraud, corruption, and collusion in public procurement activities: A systematic literature review on data-driven methods*. *Applied Network Science, 7*, 83. https://doi.org/10.1007/s41109-022-00523-6

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Martínez-Plumed, F., Casamayor, J. C., Ferri, C., Gómez, J. A., & Vendrell Vidal, E. (2019). *SALER: A Data Science Solution to Detect and Prevent Corruption in Public Administration*. *ECML PKDD 2018 Workshops*, 103–117. https://doi.org/10.1007/978-3-030-13453-2_9

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. *arXiv preprint arXiv:1301.3781*. Disponible en: https://doi.org/10.48550/arXiv.1301.3781

Modrušan, N., Mrsic, L., & Rabuzin, K. (2021). *Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies*. *International Journal of Industrial Engineering and Management, 12*(2).

OECD. (2016). *Public Procurement for Innovation: Good Practices and Strategies*. OECD Publishing.

Ortiz-Prado, E., Fernández-Naranjo, R., Torres-Berru, Y., Lowe, R., & Torres, I. (2021). *Exceptional prices of medical and other supplies during the COVID-19 pandemic in Ecuador*. *The American Journal of Tropical Medicine and Hygiene, 105*(1), 81–87.

Padhi, S. S., & Mohapatra, P. K. J. (2011). *Detection of collusion in government procurement auctions*. *Journal of Purchasing & Supply Management*, *17*(4), 207–221. https://doi.org/10.1016/j.pursup.2011.03.001

Pennington, J., Socher, R., & Manning, C. (2014). *GloVe: Global Vectors for Word Representation*. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep Contextualized Word Representations*. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2227–2237.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. *OpenAI Technical Report*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. *Journal of Machine Learning Research*, *21*(140), 1–67.

Rodó, P. (2019). *Multicolinealidad. Economipedia*. Disponible en: https://economipedia.com/definiciones/multicolinealidad.html

Sebastiani, F. (2002). *Machine learning in automated text categorization*. *ACM Computing Surveys*, *34*(1), 1-47.

Servicio Nacional de Contratación Pública (SERCOP). (2022). *Guía de Subasta Inversa Electrónica*. Quito, Ecuador.

Tenney, I., Das, D., & Pavlick, E. (2019). *BERT Rediscovers the Classical NLP Pipeline*. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601.

Torres-Berru, Y., & Batista, J. (2020). *Artificial Intelligence Techniques to Detect and Prevent Corruption in Procurement: A Systematic Literature Review*. *Information Technology and Systems*, 255–265. Disponible en: https://doi.org/10.1007/978-3-030-42520-3_21

Torres-Berru, Y., López-Batista, V. F., & Conde Zhingre, L. (2023). *A Data Mining Approach to Detecting Bias and Favoritism in Public Procurement*. *Intelligent Automation & Soft Computing*, *36*(3), 3501–3516.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., ... & Scialom, T. (2023). *LLaMA: Open and Efficient Foundation Language Models*. *arXiv preprint arXiv:2302.13971*.

Zhou, Y., Li, R., & Han, Q. (2024). *DeepSeek v3: Advancements in Large-Scale Language Modeling for Scientific Applications*. *ACL 2024*.