



Pontificia Universidad
JAVERIANA
Cali

**IDENTIFICACIÓN DE CONTRATISTAS DEL DEPARTAMENTO DE LA GUAJIRA
RELACIONADOS CON PUBLICACIONES DE CORRUPCIÓN**

Nohora Julieta Lara Moreno
Cod 9015509

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director: Mario Julián Mora Cardona

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS

Tabla de Contenido

INTRODUCCIÓN.....	3
1. CONTEXTUALIZACIÓN	5
1.1 DEFINICIÓN DEL PROBLEMA.....	5
1.1.1 PLANTEAMIENTO DEL PROBLEMA	5
1.1.2 FORMULACIÓN DEL PROBLEMA	6
1.2 OBJETIVOS DEL PROYECTO	7
1.2.1 OBJETIVO GENERAL	7
1.2.2 OBJETIVOS ESPECÍFICOS	7
1.3 MARCO DE REFERENCIA	8
1.3.1 MARCO TEÓRICO	8
1.3.2 ANTECEDENTES	12
2. CAPÍTULO I - INTEGRACIÓN DE FUENTES DE DATOS.....	15
3. CAPÍTULO II - ESTRUCTURACIÓN DE DATOS TEXTUALES	41
4. CAPÍTULO III - DETERMINACIÓN Y ELECCIÓN DE VARIABLES PARA CLASIFICACIÓN	44
5. CAPÍTULO IV - EXPERIMENTACIÓN CON DIFERENTES MODELOS DE CLASIFICACIÓN	48
6. CAPÍTULO V - EVALUACIÓN DEL DESEMPEÑO DEL MODELO	50
7. CONCLUSIONES Y TRABAJOS FUTUROS	54
7.1 CONCLUSIONES	55
7.2 TRABAJOS FUTUROS.....	56
8. REFERENCIAS BIBLIOGRÁFICAS	57
9. ANEXOS	59

INTRODUCCIÓN

La contratación pública se define como la *...“forma mediante la cual el Estado suscribe con un persona natural o jurídica un acuerdo para el cumplimiento de sus obligaciones, es esencial para el funcionamiento de los gobiernos locales; las compras públicas garantizan el cumplimiento de los derechos de los ciudadanos ya que cumplen con los objetivos misionales y mejoran la calidad de vida de los ciudadanos, a través de la satisfacción de las necesidades de la población en materia de salud, educación, servicios públicos, entre otros”...*¹ [1] '. Dado lo anterior, es preciso afirmar que la contratación pública, es un eje fundamental para el desarrollo económico y social, principalmente en regiones con enormes desafíos en transparencia y productividad, como el Departamento de La Guajira, Colombia, *...“ donde se han identificado dos (2) modalidades de corrupción: administrativa y política. En la primera, la Radiografía de la Corrupción 2016 – 2020 identificó 37 hechos de corrupción en este departamento. De estos, el 43% corresponde a corrupción administrativa, la cual se da principalmente por medio de irregularidades en la contratación pública. Y, en la segunda, se puede observar que, de los 37 casos de corrupción identificados, el 27% son de corrupción política. Diversos casos de corrupción han involucrado a funcionarios de elección popular, lo que refuerza la larga crisis institucional que ha dado paso a la vulneración de los derechos de los habitantes del departamento.” ...*²[2]

Considerando lo expuesto, la asignación de contratos públicos afronta riesgos relacionados con la malversación de recursos, lo que afecta negativamente el estricto cumplimiento de las condiciones inicialmente acordadas sobre los bienes y servicios adquiridos por el Estado, además genera desconfianza en las instituciones gubernamentales.

Estas afirmaciones son ratificadas también en el documento elaborado en el marco de la Comisión de Expertos Anticorrupción de Fedesarrollo: “La corrupción en Colombia, diagnóstico y recomendaciones para controlarla”³ [3], donde adicionalmente se explora la dependencia de recursos privados para las candidaturas políticas y como consecuencia la recompensa a los donantes a través de la contratación pública, la falta de autonomía frente a intereses políticos y económicos de la Fiscalía, la Contraloría y la Procuraduría, la conformación de redes de corrupción al interior y al exterior de la Rama Judicial, **la falta de transparencia, acceso a información y rendición de cuentas** y **la falta de integración en los sistemas de información y coordinación interinstitucional en información completa y de baja calidad en los registros de SECOP** (Sistema Electrónico de Contratación Pública).

Sumado a lo anterior, otro fundamento para desarrollar este trabajo es tener resultados que a futuro permitan el estudio de las consecuencias de este fenómeno, es decir, la relación entre corrupción y crecimiento económico sostenible, que es un *...“ concepto que abarca el desarrollo*

¹ [Contratación pública y plataformas para acceder a la información](#)

² [Corrupción administrativa y política en La Guajira - Instituto Anticorrupción](#)

³ [La corrupción en Colombia: diagnóstico y recomendaciones para controlarla](#)

económico a largo plazo sin comprometer los recursos y la calidad de vida de las generaciones futuras.” ...⁴ [4].

En este contexto, con este proyecto se inició el análisis de esta problemática a través de la implementación de un modelo de clasificación supervisado que permite identificar adjudicatarios de contratos públicos que presenten patrones o atributos correlacionados con factores de riesgo en la ejecución contractual dentro del sector público y/o con antecedentes de noticias relacionadas con actos de corrupción. Para conseguirlo, se recopilieron datos de SECOP (Sistema Electrónico de Contratación Pública), RUES (Registro Único Empresarial) de la Plataforma Nacional de Datos Abiertos de Colombia, consulta del Boletín de Responsables Fiscales de la Contraloría General de la República (persona que ha sido hallada responsable de un detrimento al patrimonio público y que por lo tanto puede quedar inhabilitado para contratar con el Estado o para ocupar cargos públicos), funcionarios públicos del Departamento Administrativo de la Función Pública, aportantes a campañas políticas de la Comisión Nacional Electoral y noticias disponibles en internet, relacionadas con presuntos casos de corrupción, aplicando técnicas de aprendizaje automático, para luego generar visualizaciones que permiten entender los resultados. Al cruzar datos de contratistas con información pública, reportes de responsables fiscales, listados de aportantes a campañas políticas y medios de comunicación que operan en entornos digitales, se construyó un sistema que detecta patrones y señales de riesgo, con fines académicos.

El resultado es una herramienta analítica que facilita la identificación de contratistas con historial de corrupción y/o patrones relacionados con factores de riesgo en la ejecución contractual. Es la primera etapa de construcción de un mecanismo preventivo para mitigar riesgos asociados al incumplimiento de obligaciones contractuales.

⁴ [La Corrupción y su Impacto en el Crecimiento Económico Sostenible: Análisis de Datos y Big Data para la Identificación de Factores y Generación de un Modelo de Machine Learning](#)

1. CONTEXTUALIZACIÓN

1.1 DEFINICIÓN DEL PROBLEMA

1.1.1 PLANTEAMIENTO DEL PROBLEMA

En contratación pública, el reto frecuente tanto para la ciudadanía como para las entidades y los mismos proponentes es identificar oportunamente a aquellos contratistas con antecedentes de incumplimiento o de irregularidades en la ejecución de antiguos procesos (casos de corrupción) o cuyo comportamiento histórico se encuentre asociado con variables predictoras de riesgo en la ejecución de lo estipulado, lo que puede afectar el cumplimiento de las obligaciones contractuales adquiridas.

Actualmente, ni los registros de contratistas, ni las bases de datos abiertos, permiten conocer la información de antecedentes negativos, como, por ejemplo, publicaciones de noticias concernientes a malos manejos del dinero, malversaciones, sobrecostos, prevaricato, falsificación de pólizas de seguros, entre otros. Lo anterior, debido a que dicha información está distribuida en diferentes fuentes, no está estructurada o no está asociada con los adjudicatarios de nuevos contratos, como consecuencia, las entidades públicas suscriben contratos sin advertir los riesgos relacionados con algunos proveedores.

Dichas fuentes de información (estructuradas y no estructuradas), donde se obtendrían noticias y datos de los contratistas son:

- SECOP I y II - Sistema Electrónico de contratación pública⁵.
- RUES - Registro único Empresarial⁶.
- Noticias disponibles en internet relacionadas con casos de corrupción o presuntos casos de corrupción.
- Consulta de Responsables Fiscales de la Contraloría General de la República⁷.
- Aportantes a Campañas Políticas de la Comisión Nacional Electoral⁸.
- SIGEP - Funcionarios públicos del Departamento Administrativo de la Función Pública⁹.

⁵ <https://colombiacompra.gov.co/compradores/beneficios-del-secop-ii-para-compradores/acceso-al-secop-ii>

⁶ <https://www.rues.org.co/?old=true>

⁷ <https://www.contraloria.gov.co/boletin-pdf>

⁸ [Cuentas Claras en Elecciones](#)

⁹ [¿Qué es? - SIGEP II - Función Pública](#)

Si bien, los procesos contractuales se realizan en todo el territorio nacional, con el fin de desarrollar un modelo más ajustado y aumentar la precisión de los resultados, se trabajará solo una región del país, el departamento de la Guajira que no solo recibe recursos de la Nación, del Sistema General de Participaciones (SGP) sino que además es beneficiario de regalías y que en contraste presenta uno de los mayores porcentajes de personas en situación de pobreza monetaria con un 65.3%, lo que hace que sea una región que reviste de especial cuidado. Adicional a lo anterior, en el departamento de La Guajira, se encuentran casos emblemáticos de corrupción¹⁰ y el nivel de cobertura de servicios básicos es insuficiente¹¹, lo que genera una dependencia directa de los recursos públicos, en especial de los fondos destinados a ayudar a la población vulnerable.

La idea del proyecto es encontrar contratistas con cierto grado de afinidad hacia la corrupción, a partir de modelos de clasificación supervisada con datos etiquetados de fuentes como: Sistema Electrónico de Contratación Pública, el Registro Único Empresarial, el Portal Anticorrupción de Colombia, Consulta de aportantes a campañas políticas y noticias disponibles en internet relacionadas con casos de corrupción (hechos y presuntos hechos).

1.1.2 FORMULACIÓN DEL PROBLEMA

Como ya se ha mencionado, la contratación pública en el Departamento de La Guajira afronta varios desafíos relacionados con la posible participación de contratistas vinculados con actos de corrupción. En este contexto, se requiere formular un problema de investigación que posibilite el análisis de los riesgos asociados a la selección de contratistas y a la ejecución de obras y compras públicas. Dado lo anterior, es indispensable considerar aspectos como por ejemplo la integración y estructuración de diversas fuentes de información. De igual forma, se advierte la necesidad de establecer una metodología de análisis basada en técnicas de ciencia de datos, con el fin de identificar patrones y antecedentes de corrupción a través de un modelo de clasificación supervisada. Con base en este enfoque, se plantean las siguientes preguntas que orientan el desarrollo del estudio.

1. ¿Cómo contrarrestar los riesgos de corrupción y la ejecución de proyectos de contratación pública haciendo énfasis en la posibilidad de que el contratista pueda estar inmerso en actos de corrupción o presente atributos correlacionados con factores de riesgo en la ejecución contractual, en el Departamento de La Guajira?
2. ¿Qué fuentes de datos son necesarias para construir una base estructurada que integre información de contratistas, contratos públicos, registros empresariales, funcionarios públicos, aportantes a campañas políticas, sujetos inhabilitados por se hallados responsables de detrimento al patrimonio público y noticias relacionadas con corrupción?

¹⁰ [Escándalo de carrotanques: esta es la denuncia del secretario de Transparencia de Presidencia contra Olmedo López](#)

¹¹ [Crisis de acceso al agua en La Guajira: comunidades wayuu exigen soluciones urgentes al Estado](#)

3. ¿Cómo se pueden preprocesar y estructurar los datos textuales de las noticias y de los procesos judiciales para que sean útiles en un modelo de clasificación?
4. ¿Cuáles son las variables que se deben tener en cuenta para determinar si un contratista tiene antecedentes de corrupción o está asociado con indicadores que sugieran un mayor nivel de riesgo en la ejecución de contratos públicos?
5. ¿Qué algoritmos de clasificación supervisada son más adecuados para este problema, considerando la naturaleza de los datos?
6. ¿Cómo evaluar el desempeño del modelo en este contexto?

1.2 OBJETIVOS DEL PROYECTO

1.2.1 OBJETIVO GENERAL

Desarrollar un modelo de clasificación supervisado para identificar adjudicatarios de contratos públicos con antecedentes de noticias relacionadas con actos de corrupción, en el Departamento de La Guajira, con el propósito superior de aportar a la disminución del riesgo en los procesos de futuras obligaciones contractuales.

1.2.2 OBJETIVOS ESPECÍFICOS

1. Integrar fuentes de datos de contratistas, contratos públicos, noticias relacionadas con corrupción y consulta de procesos judiciales del departamento de la Guajira, con el propósito de configurar bases de datos estructuradas.
2. Estructurar datos textuales, previo preprocesamiento de contratistas, contratos públicos, noticias relacionadas con corrupción y consulta de procesos judiciales del departamento de la Guajira, con el fin de obtener conjuntos de datos de entrenamiento, pruebas y validación.
3. Determinar y elegir las variables para la clasificación de los contratistas.
4. Experimentar diferentes modelos de clasificación, con el fin de determinar cuál o cuáles se adecuan al contexto de este problema.
5. Evaluar el desempeño del modelo desarrollado.

1.3 MARCO DE REFERENCIA

1.3.1 MARCO TEÓRICO

La transparencia en la información de la contratación pública es un elemento indispensable para garantizar la eficiencia en el uso de los recursos públicos. No obstante, diversas investigaciones han evidenciado la persistencia de prácticas que no se ajustan a la normatividad vigente, afectando la integridad de los procesos contractuales y la confianza ciudadana en las instituciones. En este contexto, el presente marco teórico aborda los conceptos clave que sustentan el desarrollo del proyecto, con el propósito de analizar cómo las herramientas de ciencia de datos pueden contribuir a la identificación de patrones de incumplimiento y posibles riesgos de corrupción, a partir del análisis de información histórica proveniente de múltiples fuentes.

El desarrollo del proyecto se enmarca metodológicamente en el enfoque **CRISP-DM (Cross Industry Standard Process for Data Mining)**, el cual estructura el proceso analítico en fases iterativas que incluyen la comprensión del problema, la comprensión y preparación de los datos, el modelado, la evaluación y la interpretación de los resultados. Este enfoque resulta adecuado para el análisis de fenómenos complejos como la corrupción en la contratación pública, dado que permite integrar fuentes heterogéneas de información, combinar análisis exploratorios con modelos predictivos y mantener una orientación práctica hacia la toma de decisiones informadas.

Causas y dinámicas de la corrupción en Contratación Pública:

1. Fundamentos de la Corrupción en la Contratación Pública

La corrupción en la contratación pública es un fenómeno global que impacta negativamente en el desarrollo económico y social de los países. Se define como el abuso de poder público para beneficio privado, manifestado en irregularidades como sobornos, colusión y fraude en la adjudicación de contratos públicos (*Transparency International, 2022*) [5]. En regiones como La Guajira, estas prácticas han sido reportadas frecuentemente, afectando negativamente la ejecución de proyectos indispensables en sectores como la educación, la salud e infraestructura (*PNUD, 2021*).

2. Datos Abiertos y Contratación Pública

La contratación pública en Colombia se rige bajo el principio de transparencia (Ley 80 de 1993), fomentado mediante plataformas como la del Sistema Electrónico para la Contratación Pública SECOP, que permite el acceso a datos abiertos de procesos contractuales (*Colombia Compra Eficiente*) [6]. Estos datos ofrecen información sobre contratistas, tipos de contrato y sumas adjudicadas, entre otros, que son fundamentales para analizar patrones de conductas en los

procesos de contratación. A pesar de esto, su capacidad es limitada porque no se integra con otras fuentes de información que identifican actuaciones relacionadas con corrupción.

3. Publicaciones de Noticias como Fuente de Información Complementaria

Las noticias son un componente crucial para registrar casos de corrupción. Utilizando técnicas de procesamiento de lenguaje natural (NLP) y minería de texto, se pueden analizar grandes volúmenes de datos no estructurados que provienen de medios de *comunicación* (Manning et al., 2008) [7]. Este planteamiento posibilita la extracción de nombres de personas naturales y jurídicas, correspondencias contextuales y patrones importantes que puedan relacionarse con contratistas públicos.

4. Factores de Riesgo

En el análisis de la contratación pública, es pertinente contemplar algunos **factores de riesgo** que pueden considerarse indicadores de posibles prácticas irregulares o de ineficiencia institucional. Entre otros, sobresale la suscripción de contratos en los últimos meses del año fiscal, situación que puede asociarse con la presión por ejecutar presupuestos antes de su término, lo que puede originar decisiones apresuradas o carentes de la debida planeación. Asimismo, las adiciones en el valor de los contratos constituyen señales de alerta que pueden revelar fallas en la etapa de formulación o empeño en eludir procesos licitatorios (competitivos). Finalmente, la contratación directa o especial, si bien, es legalmente permitida en ciertos casos, también representa un mecanismo susceptible de subjetividad, disminuyendo los niveles de transparencia y competencia que promueven la integridad del gasto público.

5. Transparencia, protección de datos y presunción de inocencia en el análisis de corrupción

La Ley 1712 de 2014, por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional¹², es fundamental en la lucha contra la corrupción en la contratación pública en Colombia. Esta norma determina que cualquier ciudadano pueda acceder a la información relacionada con la gestión de los recursos públicos, los procesos contractuales y las decisiones administrativas, robusteciendo así el control social y la rendición de cuentas. A su vez, la Ley 1581 de 2012 por la cual se dictan disposiciones generales para la protección de datos personales¹³ y su reglamentación mediante el Decreto 1377 de 2013¹⁴ establecen las pautas, derechos y procedimientos para la protección de datos personales, con el fin de garantizar que la información recolectada y tratada por las entidades públicas se maneje con responsabilidad,

¹² [Leyes desde 1992 - Vigencia expresa y control de constitucionalidad \[LEY 1712 2014\]](#)

¹³ [Leyes desde 1992 - Vigencia expresa y control de constitucionalidad \[LEY 1581 2012\]](#)

¹⁴ [Decreto 1377 de 2013 - Gestor Normativo - Función Pública](#)

confidencialidad y respeto por la intimidad de las personas. En el contexto de los análisis de transparencia y riesgo de corrupción, estas normas posibilitan un equilibrio entre la disposición de información y la protección de los derechos individuales.

Dado lo anterior, es necesario resaltar la presunción de inocencia y el manejo responsable de los señalamientos públicos o mediáticos. Ningún ciudadano debe ser considerado culpable con base en percepciones, publicaciones o indicadores de riesgo, hasta tanto no halla una decisión judicial en firme. Por ello, el score de riesgo o transparencia utilizado en los análisis de contratación pública debe entenderse como un insumo técnico de priorización y alerta temprana, destinado a enfocar esfuerzos de vigilancia de la ejecución del recurso, mas no como una declaración de responsabilidad penal, disciplinaria o fiscal. Su propósito es orientar la toma de decisiones informadas y fortalecer la integridad institucional, sin vulnerar derechos fundamentales ni afectar negativamente la reputación de personas o entidades. Por otro lado, es importante señalar que la publicación de la información de la etapa precontractual, contractual y de liquidación en operaciones contractuales se considera obligatoria, según el artículo 2.2.1.1.1.7.1 del Decreto 1082 de 2015 ¹⁵, mediante el cual se expide el Decreto Único Reglamentario del sector Administrativo de Planeación Nacional, por lo que, la falta de publicación de esta información en SECOP puede considerarse una alerta de corrupción, porque limita la transparencia y el control ciudadano sobre el uso de los recursos públicos.

6. Herramientas computacionales y librerías utilizadas

La implementación de la metodología se desarrolló utilizando el lenguaje de programación Python (versión 3.10.6), apoyado en librerías ampliamente reconocidas en el ámbito de la ciencia de datos. Para la manipulación y análisis de datos estructurados se empleó la librería pandas (versión 2.3.2), mientras que la construcción, entrenamiento y evaluación de los modelos de aprendizaje automático se realizó mediante scikit-learn (versión 1.7.1).

La recolección de información proveniente de fuentes abiertas se apoyó en técnicas de web scraping, utilizando las librerías BeautifulSoup4 (versión 4.13.5) y requests (versión 2.32.5). Para la visualización y análisis gráfico de los resultados se empleó matplotlib (versión 3.10.6). El uso explícito de versiones específicas garantiza la reproducibilidad del análisis y la consistencia de los resultados obtenidos, en concordancia con las buenas prácticas de investigación aplicada en ciencia de datos.

¹⁵ [Decreto 1082 de 2015 Sector Administrativo de Planeación Nacional - Gestor Normativo - Función Pública](#)

Uso de metodologías de ciencia de datos:

1. Algoritmos de clasificación en el contexto de clasificación supervisada

Los algoritmos de clasificación, tales como XGBoost, SVM (Support Vector Machines) y Random Forest, son explorados en el contexto de clasificación supervisada, en diferentes sectores, por ejemplo, en el de la salud, como se describe en el documento “Employing Supervised Classification Algorithms for Diabetes Prediction”¹⁶ [8], aunque se centra en la predicción de la diabetes, esta aplicación también resulta útil para detectar patrones de corrupción.

2. Exploración de enfoques semi-supervisados para manejar datos etiquetados y no etiquetados

Existen documentos en los qu-

e no solamente se analiza cómo algoritmos supervisados como SVM (Support Vector Machines) y Naive Bayes se utilizan en el contexto de clasificación supervisada, sino que además exploran enfoques semi-supervisados para emplear datos etiquetados y no etiquetados, es el caso del estudio denominado “Detection of Fake Online Reviews Using Semi-Supervised and Supervised Models”¹⁷[9], si bien, está enfocado en reseñas, los métodos descritos en este pueden aplicarse a diferentes problemas de clasificación, por ejemplo, la detección de patrones de corrupción.

3. Utilización de algoritmos como SVM (Support Vector Machines), árboles de decisión y Random Forest en el marco de los modelos de clasificación supervisada

El documento “Detection of Financial Fraud Using Machine Learning Algorithms”¹⁸[10], permite observar la aplicación de algoritmos como SVM (Support Vector Machines), árboles de decisión y Random Forest en el marco de los modelos de clasificación supervisada para clasificar transacciones financieras como fraudulentas o no fraudulentas, utilizando diferentes técnicas, entre las que se encuentran: i) Preprocesamiento de datos, ii) selección de características (ejemplo: tiempo entre transacciones, montos inusuales, etc) y iii) evaluación del modelo, utilizando métricas como el F1-score para definir el desempeño del modelo. Dado lo anterior, la relación de este estudio con la detección de patrones de corrupción en contratación pública radica en que con dichas técnicas también se pueden reconocer hechos como adjudicaciones repetidas a determinados contratistas, cifras atípicas, etc.

¹⁶ [EMPLOYING SUPERVISED CLASSIFICATION ALGORITHMS FOR DIABETES PREDICTION](#)

¹⁷ [\(DOC\) Detection of fake online reviews using semi-supervised and supervised learning | JP INFOTECH PROJECTS - Academia.edu](#)

¹⁸ [Detecting Anomalies in Financial Data Using Machine Learning Algorithms](#)

1.3.2 ANTECEDENTES

El uso de herramientas de ciencia de datos para identificar y prevenir la corrupción se está convirtiendo en una estrategia poderosa a nivel mundial. Modelos de clasificación supervisada, como, por ejemplo, los algoritmos de árboles de decisión y redes neuronales han demostrado su capacidad para identificar patrones de comportamiento sospechoso en datos transaccionales (*Hastie et al., 2009*)¹⁹ [11]. Estos modelos se potencian cuando se combinan con análisis de texto, lo que permite integrar datos estructurados y no estructurados en un marco analítico común (*Aggarwal & Zhai, 2012*)²⁰ [12]. Con el fin de asegurar que este proyecto se enmarca en un diálogo académico existente e identificar literatura (académica) relevante que respalde y enriquezca el análisis de esta propuesta, a continuación, se considerarán estudios que utilizan técnicas de ciencias de datos para detectar corrupción en diferentes regiones o países, como, India, Brasil, América Latina y Colombia y algunos estudios que no están enfocados en algún país en particular.

1. Estudio de Minería de Datos para Detectar Corrupción en Contratos Públicos en la India Trabajo: Borge y Singh (2020) [13]

En este estudio se realiza un análisis de patrones de corrupción en la contratación pública en la India, utilizando datos de adjudicaciones y técnicas de minería de datos. Los autores desarrollaron un modelo que combina algoritmos de detección de anomalías con el análisis de redes para identificar vínculos sospechosos entre contratistas y funcionarios públicos.

Proporciona metodologías de análisis de redes y técnicas de detección de desviaciones aplicables al análisis de patrones sospechosos en datos de contratación pública. Es importante mencionar que este proyecto integra el análisis de datos no estructurados provenientes de publicaciones de noticias, con un enfoque textual, con el fin de aumentar el alcance del hallazgo.

Este estudio se diferencia de este proyecto no solamente por el contexto geográfico, sino por el enfoque particular de los datos y las metodologías utilizadas, dado que, mientras que el estudio mencionado utiliza datos de contratación pública como temporalidades, montos, etc (su enfoque se basa en los contratos mismos), los datos para este proyecto no solo provienen de los contratos, sino que también incorporan otras fuentes de información como noticias y consultas de procesos judiciales.

¹⁹ [T. Hastie, R. Tibshirani, J. Friedman the Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition - DocsLib](#)

²⁰ [Mining text data : Aggarwal, Charu C : Free Download, Borrow, and Streaming : Internet Archive](#)

2. Proyecto de Transparencia Pública con Big Data en Brasil [14]

En este estudio desarrollaron sistemas de análisis basados en big data para hacer seguimiento a licitaciones públicas. Implementaron herramientas de minería y visualización de datos para determinar diferencias y generar reportes automatizados.

Este proyecto permite observar cómo la integración de datos abiertos con herramientas de visualización puede fortalecer la transparencia en la contratación pública. Sin embargo, no utiliza modelos de NLP ni noticias como fuente externa.

Este estudio se diferencia de este proyecto debido a que el contexto brasileño tiene una dinámica diferente en el marco de las políticas de transparencia y el uso de datos abiertos, que es precisamente la fuente de información que utilizaron, mientras que en este proyecto no se utilizarán las bases de datos abiertos del Gobierno Nacional porque no hay garantías de que estos datos sean legítimos y actuales. Este proyecto, además involucra técnicas más avanzadas como el procesamiento de lenguaje natural (NLP), minería de datos y modelos de clasificación supervisada.

3. Identificación de Anomalías en Licitaciones Públicas en América Latina [15]

En este estudio utilizaron modelos de machine learning supervisado y no supervisado para detectar anomalías en licitaciones públicas en países latinoamericanos. Los datos utilizados incluían valores de contratos, cronogramas y atributos de los proveedores. La evaluación del modelo reveló un alto desempeño en términos de precisión.

Con este trabajo se demostró el rendimiento de los modelos de machine learning para analizar datos estructurados en la contratación pública, lo que refuerza la selección de técnicas supervisadas en este proyecto. A pesar de lo anterior, no se exploran fuentes como publicaciones de noticias.

Este estudio se diferencia de este proyecto no solamente en el enfoque más localizado (Departamento de La Guajira), sino porque mientras en este documento mencionado se centran en modelos de regresión y clasificación, aplicados a datos estructurados relacionados con contratos o procesos licitatorios, en este proyecto se combinarán estas técnicas con análisis de texto o minería de datos para procesos las publicaciones de noticias y resultados de consultas de la rama judicial.

4. Estudio de Casos de Corrupción en Contratación Pública en Colombia [16]

En este documento se revisaron datos de SECOP II en Colombia para reconocer patrones de riesgo relacionados con actos de corrupción. Aplicaron estadística descriptiva y evaluaron tendencias históricas, sin embargo, no incluyeron modelos predictivos ni otras fuentes de datos.

Este proyecto resalta la relevancia de los datos abiertos como fuente primordial para el análisis de la contratación pública en Colombia.

Sin embargo, mientras que este estudio se concentra en el análisis de datos abiertos a nivel nacional, este proyecto se enfoca en un análisis más específico utilizando datos estructurados y no estructurados para identificar patrones específicos.

5. Procesamiento de Lenguaje Natural para Detectar Indicadores de Corrupción [17]

A través de este documento, presentaron técnicas avanzadas de procesamiento de lenguaje natural (NLP) donde hicieron un análisis de relaciones contextuales respecto de las entidades utilizadas. Estas técnicas han sido empleadas en investigaciones de minería de texto con el objetivo de identificar conexiones entre datos no estructurados.

Este estudio ofrece fundamentos teóricos y prácticos sobre el uso de NLP en la extracción de información fundamental a partir de textos, por ejemplo, nombres de contratistas y relaciones contextuales.

No obstante, se enfoca en métodos de recuperación de información, donde incluyen construcción de índices de texto y representación de documentos y consultas, entre otros; su objetivo es mejorar la eficiencia de los resultados de búsqueda y este proyecto está más centrado en aplicar modelos de clasificación supervisada para identificar patrones específicos de corrupción, utilizando datos estructurados, semi estructurados y no estructurados.

En conclusión, aunque los estudios revisados presentan diferentes enfoques para el análisis de la contratación pública y la detección de corrupción, cada uno de estos documentos contribuyeron con importantes metodologías que complementaron y fortalecieron el enfoque de este proyecto.

6. Enfoques de ciencia de datos aplicados a modelos de clasificación, procesamiento de texto e interpretabilidad [18 - 19]

El desarrollo de modelos analíticos orientados a la identificación de riesgos en la contratación pública requiere el uso de enfoques metodológicos propios de la ciencia de datos, particularmente aquellos relacionados con la clasificación supervisada, el procesamiento de información no estructurada y la interpretación de resultados. Estos enfoques permiten transformar grandes volúmenes de datos heterogéneos en insumos útiles para la toma de decisiones y el monitoreo preventivo de posibles irregularidades.

En el ámbito de la clasificación supervisada, diversos autores destacan su utilidad para identificar patrones complejos en conjuntos de datos donde los eventos de interés son poco frecuentes o presentan un alto nivel de desbalance entre clases. Textos clásicos como *The Elements of Statistical Learning* de Hastie, Tibshirani y Friedman plantean los fundamentos teóricos de modelos como la regresión logística, los árboles de decisión y los métodos de ensamble,

ampliamente utilizados en contextos de predicción y detección de anomalías. Estudios más recientes resaltan que estos modelos resultan especialmente adecuados en escenarios de riesgo, donde la prioridad no es únicamente la precisión global, sino la capacidad de detectar correctamente los casos minoritarios que representan alertas relevantes.

Adicionalmente, la incorporación de técnicas de procesamiento de lenguaje natural (NLP) ha cobrado relevancia en el análisis de información no estructurada, como textos provenientes de noticias, reportes institucionales o publicaciones digitales. De acuerdo con Manning, Raghavan y Schütze, así como con Aggarwal y Zhai, el análisis de texto permite extraer conocimiento a partir de documentos que no siguen una estructura tabular tradicional, ampliando significativamente el espectro de información disponible para los modelos analíticos. En el contexto de este estudio, el uso de NLP posibilita integrar señales provenientes de noticias y registros cualitativos, enriqueciendo la caracterización de los contratistas y fortaleciendo la identificación de posibles antecedentes de riesgo.

Por otra parte, investigaciones aplicadas han demostrado la efectividad de los modelos de aprendizaje automático para la detección de anomalías y comportamientos atípicos en datos financieros y administrativos, evidenciando su potencial para apoyar labores de control y vigilancia. Estos enfoques han sido empleados con éxito en estudios orientados a la detección de fraudes, irregularidades contractuales y prácticas de corrupción, tanto en contextos internacionales como en análisis de datos de contratación pública en América Latina.

Finalmente, un aspecto clave en la aplicación de modelos predictivos en entornos institucionales es la interpretabilidad de los resultados. La literatura reciente enfatiza la necesidad de complementar el desempeño predictivo con mecanismos que permitan comprender cómo y por qué un modelo toma determinadas decisiones. Herramientas de interpretabilidad contribuyen a fortalecer la transparencia, la confianza y la legitimidad del uso de modelos de ciencia de datos, especialmente cuando estos se emplean como apoyo a sistemas de alerta temprana o procesos de supervisión pública. En este sentido, la interpretación de métricas, curvas de desempeño y la importancia relativa de las variables se convierte en un elemento central para la adopción responsable de estos modelos en escenarios reales.

2. CAPÍTULO I - INTEGRACIÓN DE FUENTES DE DATOS

En este capítulo se describe el proceso de integración de las fuentes de información empleadas en el proyecto. La combinación de datos provenientes de distintos sistemas permitió construir una base coherente y útil para el análisis de la contratación pública en el Departamento de La Guajira. Este apartado incluyó la revisión, el análisis exploratorio y estandarización de las bases de datos, con el fin de asegurar su consistencia y posibilitar el estudio de los patrones relacionados al riesgo de incumplimiento contractual.

2.1. Identificar y descargar los datos abiertos de contratación pública disponibles en la plataforma nacional de datos abiertos de Colombia²¹, filtrando información relevante para el departamento de La Guajira, además del Boletín de Responsables Fiscales emitido por la Contraloría General de la República²², teniendo en cuenta lo siguiente:

Tabla No. 1
Repositorio donde están Alojadas las Bases de Datos

No.	NOMBRE	DESCRIPCIÓN	ENLACE	FUENTE
1	SECOP INTEGRADO I y II	Procesos de compra pública que se han registrado en las plataformas SECOP I y II	https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-Integrado/rpmr-utcd/about_data	Plataforma Nacional de Datos Abiertos de Colombia
2	SECOP ADICIONES	Adiciones hechas a los contratos firmados en la plataforma SECOP.	https://www.datos.gov.co/Gastos-Gubernamentales/Adiciones-SECOP/vqec-u7ms/about_data	Plataforma Nacional de Datos Abiertos de Colombia
3	SECOP GARANTÍAS RECHAZADAS	Información de las garantías publicadas en la plataforma SECOP, con estado "rechazada".	https://www.datos.gov.co/Estadisticas-Nacionales/SECOP-II-Garantias/gjp9-cutm/about_data	Plataforma Nacional de Datos Abiertos de Colombia
4	SECOP MULTAS	Registro de las Multas y Sanciones generadas en la plataforma SECOP.	https://www.datos.gov.co/Gastos-Gubernamentales/Multas-y-Sanciones-SECOP-l/4n4q-k399/about_data	Plataforma Nacional de Datos Abiertos de Colombia
5	RUES	Datos básicos del Registro Mercantil que administran las Cámaras de Comercio a nivel Nacional para Personas Naturales, Personas Jurídicas y Entidades Sin Ánimo de Lucro (ESADLs),	https://www.datos.gov.co/Comercio-Industria-y-Turismo/Personas-Naturales-Personas-Juridicas-y-Entidades-c82u-588k/about_data	Plataforma Nacional de Datos Abiertos de Colombia
6	SIBOR	Boletín del Sistema de Información de Responsables Fiscales (personas halladas de ser responsable de un detrimento al patrimonio público), emitido por la Contraloría General de la República.	https://www.contraloria.gov.co/boletin-pdf	Contraloría General de la República
7	SIGEP	Información de los servidores públicos con una vinculación activa en el Sistema de Información y Gestión del Empleo Público	https://www.datos.gov.co/Funcionarios-publicos/Conjunto-servidores-publicos/2jzx-383z/about_data	Plataforma Nacional de Datos Abiertos de Colombia
8	APORTANTES CAMPAÑAS POLÍTICAS	Ingresos de cuentas claras reportadas en los aplicativos del Consejo Nacional Electoral	https://www.datos.gov.co/Estadisticas-Nacionales/Base-ingresos-cuentas-claras-2019/jgra-rz2t/about_data	Plataforma Nacional de Datos Abiertos de Colombia
9	NOTICIAS	Noticias relacionadas con hechos de corrupción (probados o presuntos)	La explicación de enlaces y fuentes se realizará en el numeral 2.2	

²¹ [Datos Abiertos Colombia | Datos Abiertos Colombia](#)

²² [Boletín PDF - Contraloría](#)

Elaborado por: NJLM

- a. **SECOP INTEGRADO I y II:** Información integrada de los procesos de compra pública que se han registrado en las plataformas SECOP I y II, que han finalizado con un contrato, desde 2013 hasta la primera semana del mes de abril del 2025.

Relevancia para el modelo: Los datos de **SECOP I y II** son indispensables porque representan el núcleo del comportamiento contractual del Estado con los diferentes proveedores. Su integración permite analizar el historial de contratación de cada contratista, incluyendo montos adjudicados, entidades contratantes, modalidades de selección y duración de los contratos. Esta información es clave para construir variables que reflejen la magnitud, frecuencia y tipo de relaciones contractuales, lo que ayuda al modelo a identificar patrones inusuales de adjudicación o concentración que pueden asociarse con posibles riesgos de corrupción o favorecimiento.

Esta tabla tiene 113.456 filas (contratos) y 22 columnas, así:

Tabla No. 2
Columnas de la Tabla SECOP Integrado

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
1	Nivel Entidad	Nivel de la Entidad pública que desarrolla el proceso de compra, de acuerdo a su cobertura	nivel_entidad	Texto
2	Código Entidad en SECOP	Código de la entidad en la plataforma SEECOP	codigo_entidad_en_secop	Texto
3	Nombre de la Entidad	Nombre de la Entidad pública que desarrolla el proceso de compra pública	nombre_de_la_entidad	Texto
4	NIT de la Entidad	Identificador de la Entidad que desarrolla el proceso de compra pública	nit_de_la_entidad	Texto
5	Departamento Entidad	Departamento de origen de la Entidad en la plataforma SECOP	departamento_entidad	Texto
6	Municipio Entidad	Municipio de origen de la Entidad en la plataforma SECOP	municipio_entidad	Texto
7	Estado del Proceso	Estado del proceso, de acuerdo al flujo que se establece y a la modalidad de adquisición bajo la cual se desarrolla	estado_del_proceso	Texto
8	Modalidad de Contratación	Modo de adquisición, de acuerdo a las características del proceso y del producto o servicio involucrados	modalidad_de_contrataci_n	Texto
9	Objeto del contrato	Descripción detallada del producto o servicio a adquirir, definida al momento de redactar el contrato	objeto_a_contratar	Texto
10	Objeto del proceso	Objeto de contratación registrado en la fase de publicación del proceso	objeto_del_proceso	Texto

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
11	Tipo de Contrato	Tipo de contrato que se firma al finalizar el proceso de compra pública	tipo_de_contrato	Texto
12	Fecha de Firma del Contrato	Fecha en que se firmó el contrato, sea de forma física o digital	fecha_de_firma_del_contrato	Marca de tiempo variable
13	Fecha inicio ejecución	Fecha de Inicio de Ejecución del Contrato	fecha_inicio_ejecuci_n	Marca de tiempo variable
14	Fecha fin ejecución	Fecha definida en el contrato como la fecha de finalización de la ejecución	fecha_fin_ejecuci_n	Marca de tiempo variable
15	ID Contrato	Identificador en la plataforma del contrato firmado	numero_del_contrato	Texto
16	ID Proceso	identificador en la plataforma del proceso desarrollado	numero_de_proceso	Texto
17	Valor Contrato	Valor del contrato firmado	valor_contrato	Número
18	Nom Raz Social Contratista	Nombre del contratista o de la empresa que proveerá los productos o servicios relacionados en el proceso de compra	nom_raz_social_contratista	Texto
19	URL Contrato	URL del proceso en su correspondiente plataforma.	url_contrato	Texto
20	Origen	Plaraforma Origen a donde pertenecen los datos	origen	Texto
21	Tipo Documento Proveedor	Corresponde al tipo de documento del proveedor	tipo_documento_proveedor	Texto
22	Documento Proveedor	Corresponde al número de documento del proveedor	documento_proveedor	Texto

Fuente: [SECOP Integrado | Datos Abiertos Colombia](#)

Elaborado por: NJLM

Se hizo un análisis exploratorio de esta tabla y se encontró que:

- Las entidades que más contratos han suscrito son un hospital de Riohacha, un hospital de Maicao, el departamento (la gobernación) de la Guajira, un hospital del municipio de Fonseca y el SENA regional Guajira.
- Las entidades con mayor número de contratistas son un hospital de Riohacha, el departamento y el municipio de Maicao.
- El tipo de contrato con mayor participación es el de prestación de servicios.
- Las entidades con mayor valor en la contratación son la gobernación, el SENA regional Guajira y el municipio de Fonseca.

- La Contratación objeto de estudio suma 26.5 billones de pesos.

Gráfico No. 1
Entidades del Departamento de la Guajira con
Mayor Número de Contratos



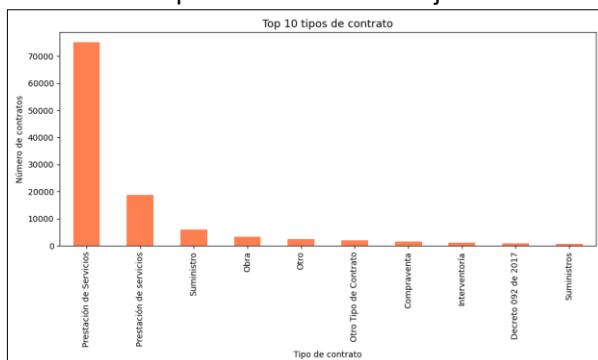
Fuente: [SECOP Integrado | Datos Abiertos Colombia](#)
Elaborado por: NJLM en Visual Code

Gráfico No. 2
Entidades del Departamento de la Guajira con
Mayor Número de Contratistas



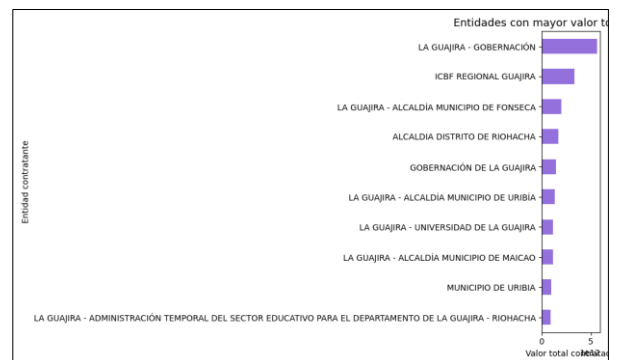
Fuente: [SECOP Integrado | Datos Abiertos Colombia](#)
Elaborado por: NJLM en Visual Code

Gráfico No. 3
Tipos de Contrato – Contratación del
Departamento de la Guajira



Fuente: [SECOP Integrado | Datos Abiertos Colombia](#)
Elaborado por: NJLM en Visual Code

Gráfico No. 4
Entidades del Departamento de la Guajira con Mayor
Valor de Contratación



Fuente: [SECOP Integrado | Datos Abiertos Colombia](#)
Elaborado por: NJLM en Visual Code

- b. **SECOP ADICIONES:** adiciones hechas a los contratos firmados en la plataforma SECOP.

Relevancia para el modelo: La información de Adiciones SECOP aporta una dimensión crítica sobre la ejecución de los contratos, permitiendo identificar incrementos en el valor o tiempo

de estos que podrían revelar sobrecostos o deficiencias en la planeación contractual. Incorporar esta base al modelo permite generar variables que midan la proporción de adiciones respecto al valor inicial, la frecuencia de las modificaciones y su concentración por contratista, brindando señales sobre posibles irregularidades financieras o administrativas. De este modo, las adiciones actúan como un indicador indirecto de riesgo y prácticas contractuales anómalas.

Esta tabla tiene 223.104 filas (contratos) y 5 columnas, así:

Tabla No. 3
Columnas de la Tabla SECOP Adiciones

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
1	Identificador	Identificador único del evento de modificación	Identificador	Texto
2	ID_Contrato	Identificador del contrato al que corresponde la modificación	id_contrato	Texto
3	Tipo	Tipo de modificación, de acuerdo al impacto que tiene sobre el contrato	Tipo	Texto
4	Descripcion	Descripción detallada de la justificación de la adición o modificación	Descripción	Texto
5	FechaRegistro	Fecha en que se hace la modificación	Fecharegistro	Marca de Tiempo Variable

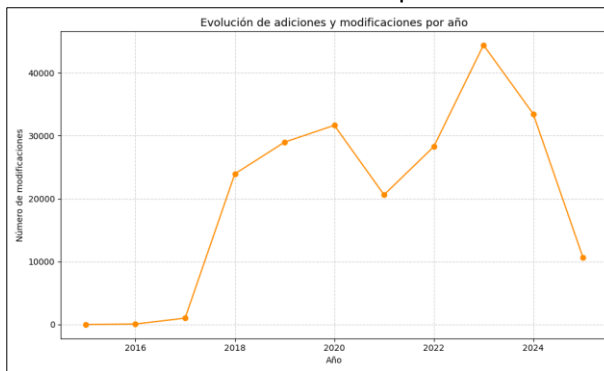
Fuente: [Adiciones SECOP | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

Se hizo un análisis exploratorio de esta tabla y se encontró que:

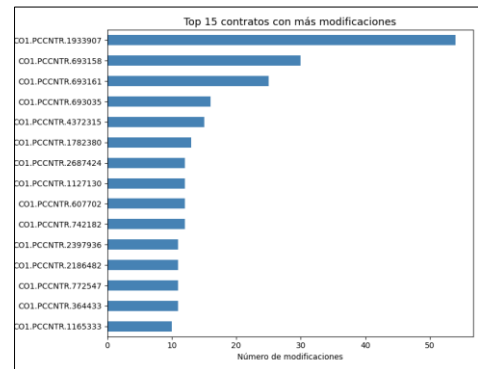
- El mayor número de adiciones de contratos en el departamento de la Guajira no fue durante la pandemia (2020-2021), sino en el año 2023.
- Los contratos con mayor número de adiciones corresponden a una orden de servicio, pólizas de automóviles, pólizas todo riesgo y pólizas de “casco aviación”.

Gráfico No. 5
Evolución de adiciones por año



Fuente: [Adiciones SECOP | Datos Abiertos Colombia](#)
 Elaborado por: NJLM en Visual Code

Gráfico No. 6
Contratos con más adiciones en el departamento de la Guajira



Fuente: [Adiciones SECOP | Datos Abiertos Colombia](#)
 Elaborado por: NJLM en Visual Code

- c. **SECOP GARANTÍAS RECHAZADAS:** información de las garantías publicadas en la plataforma SECOP, con estado “rechazada”.

Relevancia para el modelo: La base de **Garantías Rechazadas SECOP** es relevante porque evidencia comportamientos asociados al cumplimiento formal de los contratistas frente a los requisitos de aseguramiento exigidos por las entidades públicas. Los registros de garantías rechazadas pueden reflejar deficiencias en la gestión contractual, negligencia o incluso intentos de fraude documental. Incorporar esta información permite al modelo considerar la frecuencia y el tipo de rechazos como variables que indican vulnerabilidad o riesgo operativo, fortaleciendo su capacidad para detectar contratistas con comportamientos irregulares.

Esta tabla tiene 802.546 filas (contratos) y 14 columnas, así:

Tabla No. 4
Columnas de la Tabla SECOP Garantías con Estado Rechazada

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
1	ID Contrato	Identificador del contrato, con el que se realiza la garantía.	id_contrato	Texto
2	Aseguradora	Nombre de la Aseguradora, con la que se tramita la garantía.	aseguradora	Texto
3	NumeroPoliza	Numero de póliza, con la que se tramita la garantía.	numeropoliza	Texto
4	Asegurado	Nombre del titular de la póliza, tipo persona natural o jurídica.	asegurado	Texto

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
5	Beneficiario	Nombre de la persona beneficiada en la garantía, tipo natural o jurídica	beneficiario	Texto
6	FechaDeCreacionPoliza	Fecha, en la que se crea la póliza	fechadecreacionpoliza	Texto
7	FechaEnvioPoliza	Fecha, en la que se envía la póliza	fechaenviopoliza	Texto
8	FechaFinPoliza	Fecha, en la que finaliza la póliza	fechafinpoliza	Texto
9	LadoPoliza		ladopoliza	Texto
10	Estado	Estado en la que se encuentra la póliza	estado	Texto
11	TipoPoliza	Objeto con la que se ejecuta la póliza, ejemplo contrato de seguro	tipopoliza	Texto
12	SubTipoPoliza		subtipopoliza	Texto
13	Valor	Valor, de la póliza	valor	Número
14	Fecha de Creacion	Fecha, con la que se crea la garantía en la plataforma	fecha_de_creacion	Texto

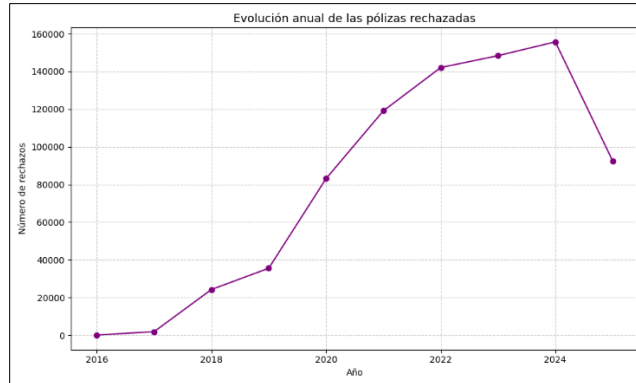
Fuente: [SECOP II - Garantías | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

Se hizo un análisis exploratorio de esta tabla y se encontró que:

- La aseguradora con mayor número de pólizas rechazadas en Seguros del Estado.
- El mayor número de garantías rechazadas se dio en la vigencia 2024.

Gráfico No. 7
Evolución garantías rechazadas en el tiempo



Fuente: [SECOP II - Garantías | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

d. **SECOP MULTAS**: registro de las Multas y Sanciones generadas en la plataforma SECOP.

Relevancia para el modelo: La base de SECOP Multas ofrece una fuente directa de información sobre el historial de cumplimiento de los contratistas frente a las obligaciones documentales. Las multas registradas por las entidades públicas reflejan fallas en el cargue completo de la información. Su integración al modelo permite cuantificar la reincidencia, generando variables que representan la reputación y la confiabilidad del contratista. Estas señales son altamente

predictivas del riesgo de corrupción, al vincular antecedentes de incumplimiento con posibles malas prácticas administrativas.

Esta tabla tiene 1.697 filas y 14 columnas, así:

Tabla No. 5
Columnas de la Tabla SECOP Multas

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
1	Nombre Entidad	Nombre de la Entidad que publica el proceso de compra pública	nombre_entidad	Texto
2	NIT Entidad	NIT de la Entidad que publica el proceso de compra pública	nit_entidad	Texto
3	Nivel	Nivel de cobertura de la Entidad que hace el registro del proceso de compra pública.	nivel	Texto
4	Orden	Orden detallado de la cobertura de la Entidad	orden	Texto
5	Municipio	Municipio en el que está registrado el proveedor al que le aplica la sanción	municipio	Texto
6	Numero de Resolucion	Número de la resolución de la sanción que se aplicará al proveedor	numero_de_resolucion	Texto
7	Documento Contratista	Número de Identificación del Contratista Sancionado	documento_contratista	Texto
8	Nombre Contratista	Nombre del Contratista Sancionado	nombre_contratista	Texto
9	Numero de Contrato	Contrato sobre el cual se aplicará la sanción al contratista	numero_de_contrato	Texto
10	Valor Sancion	Valor de sanción que se aplicará en la resolución actual	valor_sancion	Número
11	Fecha de Publicacion	Fecha en que se registra el incumplimiento y la sanción	fecha_de_publicacion	Marca de tiempo variable
12	Fecha de Firmeza	Fecha en que queda en firme la Multa	fecha_de_firmeza	Marca de tiempo variable
13	Fecha de Cargue	Fecha de cargue de la multa en el SECOP	fecha_de_cargue	Marca de tiempo variable
14	Ruta de Proceso	Ruta del proceso en la plataforma, para consulta directa	ruta_de_proceso	Texto

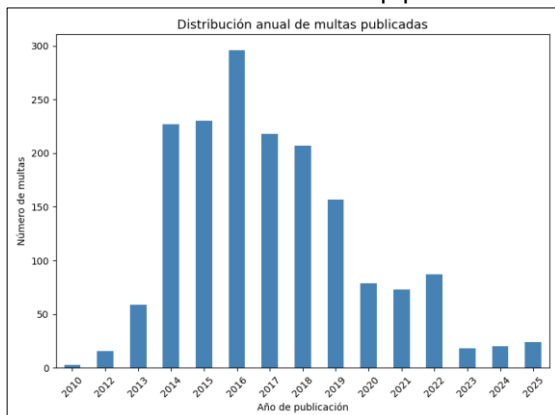
Fuente: [Multas y Sanciones SECOP I | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

Se hizo un análisis exploratorio de esta tabla y se encontró que:

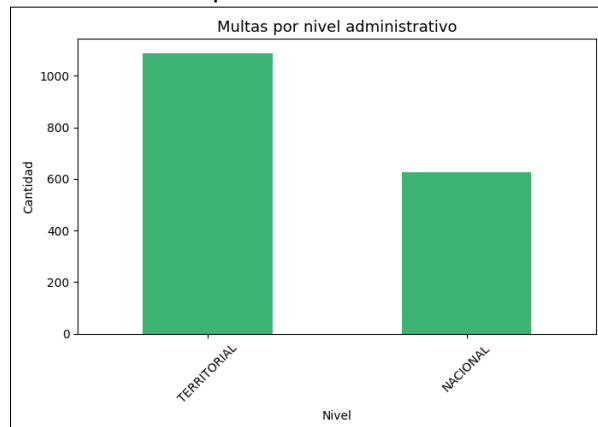
- La vigencia con mayor número de multas fue el 2016.
- Las entidades de orden territorial tienen más multas que las entidades de orden nacional.

Gráfico No. 8
Evolución de Multas Secop por Año



Fuente: [Adiciones SECOP | Datos Abiertos Colombia](#)
 Elaborado por: NJLM en Visual Code

Gráfico No. 9
Multas por Nivel Administrativo



Fuente: [Adiciones SECOP | Datos Abiertos Colombia](#)
 Elaborado por: NJLM en Visual Code

- e. **RUES:** extracción que contiene los datos básicos del Registro Mercantil que administran las Cámaras de Comercio a nivel Nacional para Personas Naturales, Personas Jurídicas y Entidades Sin Ánimo de Lucro (ESADLs), que se encuentran sincronizados en el Registro Único Empresarial y Social (RUES).

Relevancia para el modelo: La base de RUES complementa el modelo con una perspectiva empresarial, al proporcionar información sobre la identidad, estructura y actividad económica de los contratistas. Variables como tipo societario, fecha de constitución, estado de matrícula, CIU y ubicación geográfica permiten caracterizar la formalidad y estabilidad de las empresas. Esta información es relevante para identificar empresas fachada, organizaciones sin trayectoria o entidades inactivas que, sin embargo, participan en procesos contractuales. En el modelo, RUES permite incorporar factores estructurales que explican el contexto organizacional detrás del riesgo de corrupción.

Esta tabla tiene 8.627.995 filas y 36 columnas, así:

Tabla No. 6
 Columnas de la Tabla RUES

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
1	codigo_camara	Código de la Cámara de comercio que realiza la actualización de datos.	codigo_camara	Texto
2	camara_comercio	Cámara de comercio que realiza la actualización de datos.	camara_comercio	Texto
3	matricula	Número de Matricula Mercantil	matricula	Texto

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
4	inscripcion_propo nente	Número de Inscripción en el registro de Proponentes	inscripcion_proponente	Texto
5	razon_social	Razón Social	razon_social	Texto
6	primer_apellido	Primer Apellido	primer_apellido	Texto
7	segundo_apellido	Segundo Apellido, otros apellidos	segundo_apellido	Texto
8	primer_nombre	Primer Nombre	primer_nombre	Texto
9	segundo_nombre	Segundo Nombre u otros nombres	segundo_nombre	Texto
10	sigla	Sigla	sigla	Texto
11	codigo_clase_ide ntificacion	Código Clase Identificación	codigo_clase_identificacion	Texto
12	clase_identificacio n	Clase Identificación	clase_identificacion	Texto
13	numero_identifica cion	Número de Identificación	numero_identificacion	Texto
14	nit	Número de Identificación Tributario	nit	Texto
15	digito_verificacion	Digito de Verificación	digito_verificacion	Texto
16	cod_ciiu_act_eco n_pri	Código CIIU actividad económica principal	cod_ciiu_act_econ_pri	Texto
17	cod_ciiu_act_eco n_sec	Código CIIU actividad económica secundaria	cod_ciiu_act_econ_sec	Texto
18	ciiu3	Código actividad económica otras 1	ciiu3	Texto
19	ciiu4	Código actividad económica otras 2	ciiu4	Texto
20	fecha_matricula	Fecha Matricula	fecha_matricula	Texto
21	fecha_renovacion	Fecha de ultima renovación	fecha_renovacion	Texto
22	ultimo_ano_renov ado	Ultimo Año Renovado	ultimo_ano_renovado	Texto
23	fecha_vigencia	Fecha Vigencia	fecha_vigencia	Texto
24	fecha_cancelacio n	Fecha Cancelación	fecha_cancelacion	Texto
25	codigo_tipo_socie dad	Código Tipo sociedad	codigo_tipo_sociedad	Texto
26	tipo_sociedad	Tipo sociedad	tipo_sociedad	Texto
27	codigo_organizaci on_juridica	Código Organización Jurídica	codigo_organizacion_juridica	Texto
28	organizacion_jurid ica	Organización Jurídica	organizacion_juridica	Texto
29	codigo_categoria _matricula	Código Categoría Matricula	codigo_categoria_matricula	Texto
30	categoria_matricu la	Categoría Matricula	categoria_matricula	Texto
31	Representante Legal	Nombre del representante legal de la empresa	representante_legal	Texto
32	segundo_apellido	Segundo Apellido, otros apellidos	segundo_apellido	Texto
33	segundo_nombre	Segundo Nombre u otros nombres	segundo_nombre	Texto
34	sigla	Sigla	sigla	Texto
35	tipo_sociedad	Tipo sociedad	tipo_sociedad	Texto

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
36	ultimo_ano_renovado	Último Año Renovado	ultimo_ano_renovado	Texto

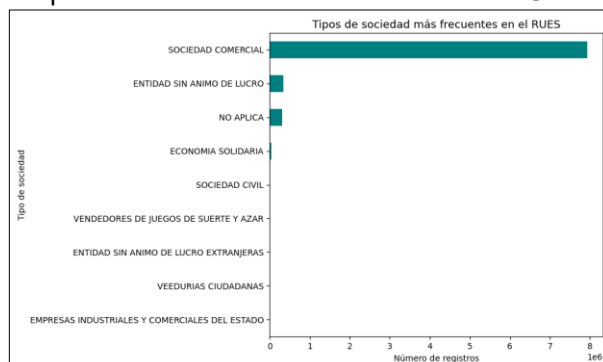
Fuente: [Personas Naturales, Personas Jurídicas y Entidades Sin Animo de Lucro | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

Se hizo un análisis exploratorio de esta tabla y se encontró que:

- Hay 4.820.705 números de identificación comercial con la matrícula cancelada, 3.807.267 con la matrícula activa y 23 con la matrícula inactiva.
- El tipo de sociedad más frecuente en RUES es la sociedad comercial.
- Dentro de la distribución por categoría de matrícula hay más personas naturales que personas jurídicas.

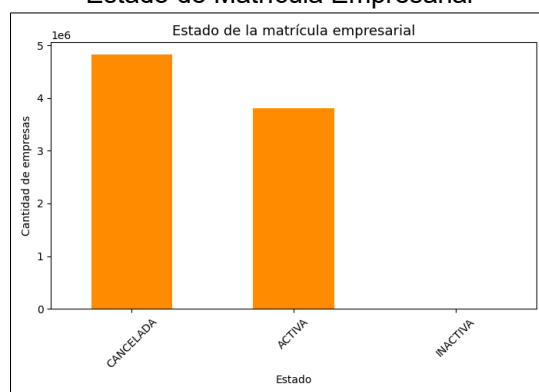
Gráfico No. 10
Tipos de sociedad más frecuentes en el RUES



Fuente: [Personas Naturales, Personas Jurídicas y Entidades Sin Animo de Lucro | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

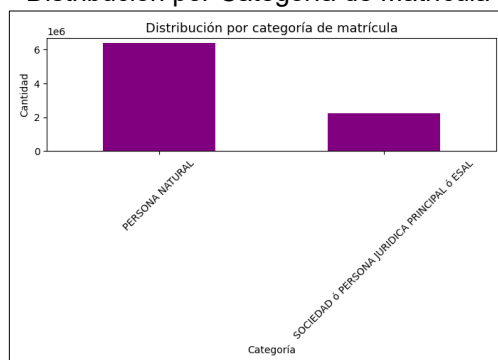
Gráfico No. 11
Estado de Matrícula Empresarial



Fuente: [Personas Naturales, Personas Jurídicas y Entidades Sin Animo de Lucro | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

Gráfico No. 12
Distribución por Categoría de Matrícula



Fuente: [Personas Naturales, Personas Jurídicas y Entidades Sin Animo de Lucro | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

- f. **SIBOR:** boletín del Sistema de Información de Responsables Fiscales, emitido por la Contraloría General de la República.

Relevancia para el modelo: SIBOR es la base que permite identificar a las personas naturales o jurídicas con fallos de responsabilidad fiscal, lo que significa que han sido legalmente sancionados por afectar recursos públicos y que esa situación está registrada oficialmente, lo cual afecta su historial y su capacidad de contratar con el Estado, convirtiéndose en la fuente fundamental para la definición del riesgo de inejecución contractual. Su integración permite etiquetar los casos positivos de riesgo (contratistas sancionados), facilitando el entrenamiento y validación del modelo de clasificación. Además, SIBOR brinda información adicional sobre el tipo de falta, el ente que reporta y la entidad afectada, que puede utilizarse como variables explicativas complementarias. En síntesis, SIBOR proporciona el sustento sancionatorio que permite transformar los datos en un instrumento de predicción del riesgo de corrupción.

Esta tabla tiene 6365 filas y 9 columnas, así:

Tabla No. 7
Columnas de la Tabla SIBOR

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
1	Responsable Fiscal	Nombre del Responsable Fiscal	Responsable Fiscal	Texto
2	Tipo	Tipo de documento de Responsable Fiscal	Tipo	Texto
3	Num_documento	Número de documento de Responsable Fiscal	Num_documento	Texto
4	Entidad Afectada	Entidad que debe ser resarcida por daño al patrimonio público	Entidad Afectada	Texto
5	TR	Tipo de Responsabilidad (individual o solidario)	TR	Texto
6	R	Cantidad de registros por TR del Responsable Fiscal	R	Texto
7	Ente que Reporta	Entidad que reporta la irregularidad	Ente que Reporta	Texto
8	Departamento	Departamento donde ocurrió el hecho que generó la responsabilidad fiscal	Departamento	Texto
9	Municipio	Municipio donde ocurrió el hecho que generó la responsabilidad fiscal	Municipio	Texto

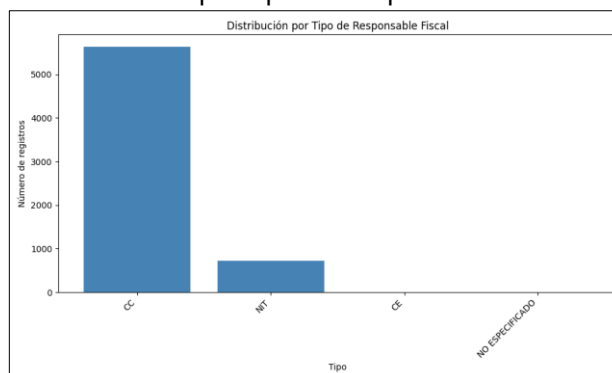
Fuente: <https://www.contraloria.gov.co/boletin-pdf>

Elaborado por: NJLM en Visual Code

Se hizo un análisis exploratorio de esta tabla y se encontró que:

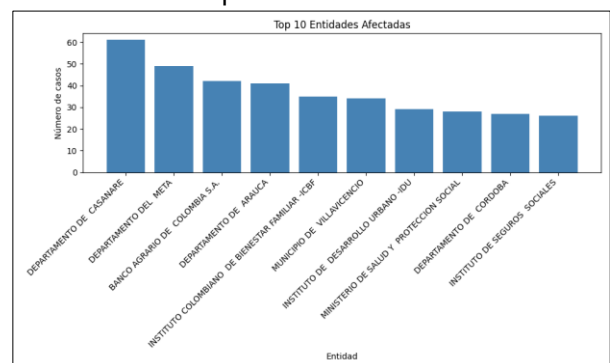
- Hay más responsables fiscales que son personas naturales (más de 5000) que personas jurídicas (menos de 1000).
- Las entidades más afectadas por los responsables fiscales son el departamento del Casanare, el departamento del Meta, el banco agrario, el departamento de Arauca y el Instituto Colombiano de Bienestar Familiar ICBF.
- Hay responsables fiscales con 7 sanciones o más.
- Los departamentos más afectados por los responsables fiscales son Cundinamarca, Antioquia y Bolívar.

Gráfico No. 13
Distribución por Tipo de Responsable Fiscal



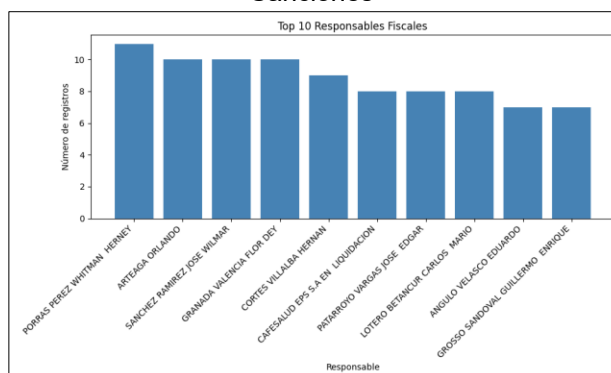
Fuente: <https://www.contraloria.gov.co/boletin-pdf>
Elaborado por: NJLM en Visual Code

Gráfico No. 14
Top 10 de Entidades Afectadas por los Responsables Fiscales



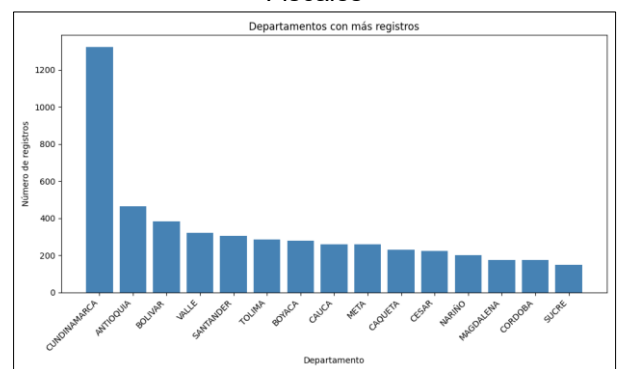
Fuente: <https://www.contraloria.gov.co/boletin-pdf>
Elaborado por: NJLM en Visual Code

Gráfico No. 15
Top 10 de Responsables Fiscales con más Sanciones



Fuente: <https://www.contraloria.gov.co/boletin-pdf>
Elaborado por: NJLM en Visual Code

Gráfico No. 16
Departamentos afectados por los Responsables Fiscales



Fuente: <https://www.contraloria.gov.co/boletin-pdf>
Elaborado por: NJLM en Visual Code

- g. **SIGEP:** información de los servidores públicos con una vinculación activa en el Sistema de Información y Gestión del Empleo Público -SIGEP, de acuerdo con el Título 17 del Decreto 1083 de 2015.

Relevancia para el modelo: La inclusión de SIGEP en el modelo introduce la dimensión institucional y ética del riesgo, al permitir identificar vínculos entre contratistas y funcionarios públicos. Al cruzar información de contratistas con registros de servidores activos, es posible detectar conflictos de interés, nepotismo o relaciones de poder que afectan la transparencia en la contratación. Variables como “empleado público”, “entidad empleadora” o “nivel jerárquico” contribuyen a revelar redes de influencia y comportamientos potencialmente irregulares. Esta fuente fortalece el componente de integridad dentro del modelo predictivo.

Esta tabla tiene 242.474 filas y 23 columnas, así:

Tabla No. 8
Columnas de la Tabla SIGEP

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
1	Número de Identificación	Número que se utiliza para reconocer y diferenciar a una persona, el cual este asignado a un documento oficial	numeroideidentificacion	Texto
2	Nombre	Palabra o conjunto de palabras con las que se designan y se distingue a una persona	nombre	Texto
3	Sexo	Hace referencia a las características fisiológicas y sexuales con las que nacen mujeres y hombres	sexo	Texto
4	ID Departamento de Nacimiento	Número único de identificación y codificación de los departamentos establecidos por el DANE	iddepartamentodenacimiento	Texto
5	Departamento de Nacimiento	División administrativa y política según lo establece la Constitución política de 1991, 33 divisiones, 32 departamentos	departamentodenacimiento	Texto
6	ID Municipio de Nacimiento	Número único de identificación y codificación de los municipios establecidos por el DANE	idmunicipiodenacimiento	Texto
7	Municipio de Nacimiento	Segundo nivel de división administrativa en Colombia, De acuerdo con el Artículo 311 de Constitución de 1991 y la Ley 136 de 2 de junio de 1994.	municipiodenacimiento	Texto
8	Meses de Experiencia Público	Número de meses de experiencia laboral del servidor en entidades del sector público	mesesdeexperienciapublico	Texto
9	Meses de Experiencia Privado	Número de meses de experiencia laboral del servidor en entidades del sector privado	mesesdeexperienciaprivado	Texto
10	Meses de Experiencia Docente	Número de meses de experiencia docente del servidor público	mesesdeexperienciadocente	Texto
11	Meses de Experiencia Independiente	Número de meses de experiencia laboral del servidor público como independiente	mesesdeexperienciaindependiente	Texto
12	Nivel Educativo	Nivel de educación más alto que un servidor ha terminado	niveleducativo	Texto
13	Código Sigep	Número único de identificación de la Entidad.	codigosigep	Texto

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
14	Nombre Entidad	Denominación verbal dada a una Entidad.	nombrentidad	Texto
15	Orden	Campo de acción de la administración pública de acuerdo con el ordenamiento territorial: nacional o territorial.	orden	Texto
16	ID Municipio de Nacimiento	Número único de identificación y codificación de los municipios establecidos por el DANE	idmunicipiodenacimiento	Texto
17	ID Departamento de Nacimiento	Número único de identificación y codificación de los departamentos establecidos por el DANE	iddepartamentodenacimiento	Texto
18	Fecha de vinculación	Fecha en la cual el servidor público inicia su vinculación en la entidad	fecha_de_vinculaci_n	Texto
19	Dependencia Empleo Actual	Área de la entidad en la que se encuentra vinculado el servidor público	dependenciaempleoactual	Texto
20	Departamento de Nacimiento	División administrativa y política según lo establece la Constitución política de 1991, 33 divisiones, 32 departamentos	departamentodenacimiento	Texto
21	Denominación Empleo Actual	Nombre del cargo al que se encuentra vinculado el servidor público que se identifica con un código de varios dígitos de acuerdo con la norma vigente de nomenclatura y clasificación de empleos.	denominacionempleoactual	Texto
22	Codigo Sigep	Número único de identificación de la Entidad.	codigosigep	Texto
23	Asignación Básica Salarial	Salario establecido al servidor público según los decretos emitidos por el gobierno nacional	asignacionbasicasalarial	Texto

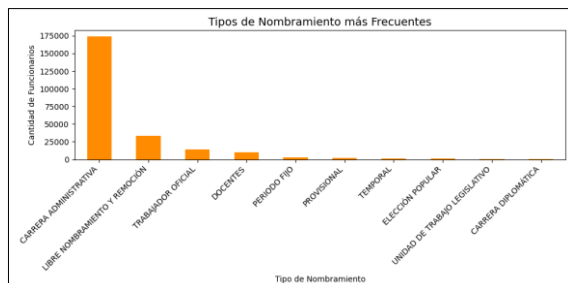
Fuente: [Conjunto servidores públicos | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

Se hizo un análisis exploratorio de esta tabla y se encontró que:

- El tipo de nombramiento más frecuente es el de carrera administrativa.
- El nivel jerárquico más común es el de profesional, seguido por el de asistencial.
- El mayor número de vinculaciones de empleos públicos fue en 2023.

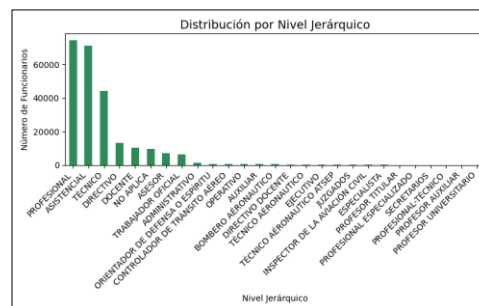
Gráfico No. 16
Tipos de Nombramientos más Frecuentes



Fuente: [Conjunto servidores públicos | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

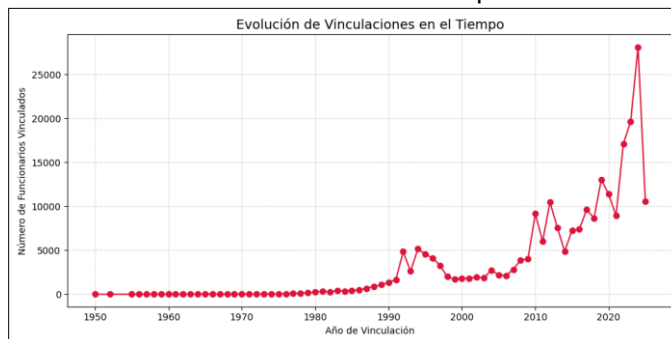
Gráfico No. 17
Tipos de Nombramientos más Frecuentes



Fuente: [Conjunto servidores públicos | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

Gráfico No. 18
Vinculaciones en el Tiempo



Fuente: [Conjunto servidores públicos | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

- h. APORTANTES CAMPAÑAS POLÍTICAS: ingresos de cuentas claras reportadas en los aplicativos del Consejo Nacional Electoral.

Relevancia para el modelo: La variable “aportantes de campañas políticas” se incluye con un enfoque preventivo y objetivo, pues, aunque la financiación electoral es un mecanismo legítimo de participación, puede generar vínculos de interés entre donantes y administradores públicos, lo que amerita análisis desde la perspectiva de transparencia y riesgo. Incorporarla permite explorar si existen patrones de asociación entre apoyo político y adjudicación o ejecución contractual con el Estado, sin asumir causalidad ni presunción de irregularidad, y contribuye a fortalecer sistemas de alerta temprana y monitoreo del riesgo en la contratación pública.

Esta tabla tiene 188.170 filas y 20 columnas, así:

Tabla No. 9
Columnas de la Tabla Aportantes a Campañas Políticas

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
1	CAN_IDENTIFICACION	Número de identificación del candidato	can_identificacion	Número
2	CND_NOMBRE	Entidad a la que está aspirando	cnd_nombre	Texto
3	CLA_NOMBRE	Tipo de entidad a la que está aspirando	cla_nombre	Texto
4	DEP_NOMBRE	Nombre departamento	dep_nombre	Texto
5	MUN_NOMBRE	Nombre municipio	mun_nombre	Texto
6	LOC_NOMBRE	Localización	loc_nombre	Texto
7	ORG_NOMBRE	Organización o Partido Político	org_nombre	Texto
8	NOMBRE_CANDIDATO	Nombre Candidato	nombre_candidato	Texto
9	CCO_ID	ID candidato	cco_id	Número
10	TPE_NOMBRE	Persona Natural o Jurídica	tpe_nombre	Texto
11	NOMBRE_PERSONA	Nombre Aportante	nombre_persona	Texto

No.	Nombre de la columna	Descripción	Nombre del campo	Tipo de Dato
12	ING_VALOR	Valor del aporte	ing_valor	Número
13	TID_NOMBRE	Tipo de documento aportante	tid_nombre	Texto
14	ING_IDENTIFICACION	Número de identificación del aportante	ing_identificacion	Texto
15	ING_ACTA	Acta donde se registra el aporte	ing_acta	Texto
16	TDO_NOMBRE	Tipo de aporte	tdo_nombre	Texto
17	PARTIDO_COALICION	Nombre partido coalición	partido_coalicion	Texto
18	ING_FECHA_COMPROBANTE	Fecha aporte	ing_fecha_comprobante	Marca de Tiempo Variable
19	ING_COMPROBANTE	Número comprobante aporte	ing_comprobante	Texto
20	ING_CONCEPTO	Concepto aporte	ing_concepto	Texto

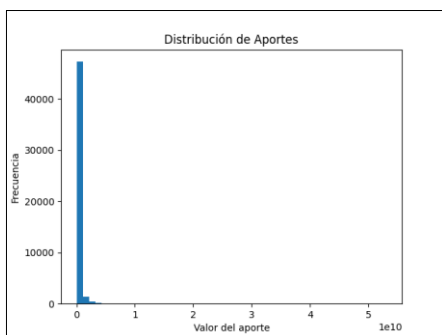
Fuente: [Base ingresos cuentas claras 2019 | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code

Se hizo un análisis exploratorio de esta tabla y se encontró que:

- Los aportes son en cantidad abundantes, pero en valor no muy altos.
- Los mayores aportes los registran sin asociarlos a un partido político.
- Entre los mayores aportantes a las campañas políticas están la fundación grupo argos, Miguel Uribe Londoño, la fundación Bancolombia y el partido Mira de la iglesia ministerial de Jesucristo internacional.

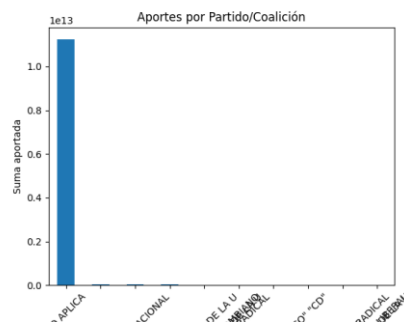
Gráfico No. 19
Distribución de Aportes



Fuente: [Base ingresos cuentas claras 2019 | Datos Abiertos Colombia](#)

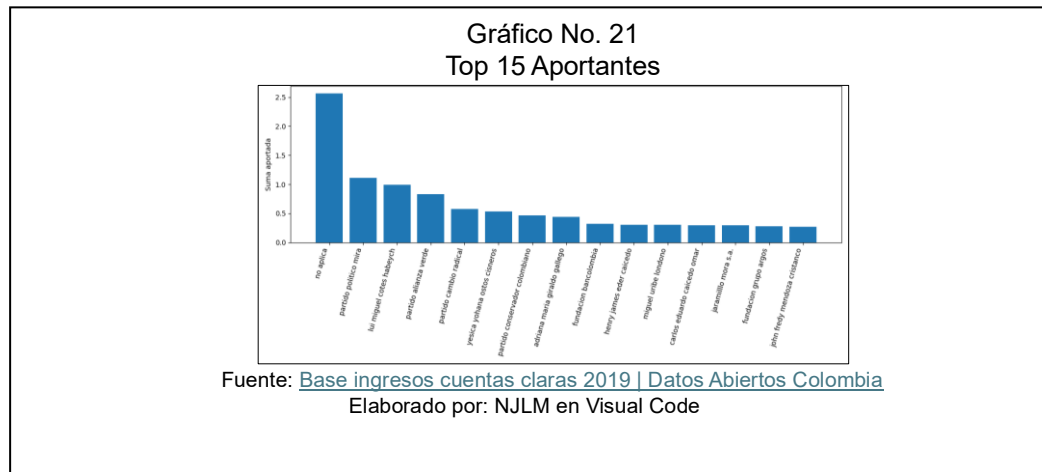
Elaborado por: NJLM en Visual Code

Gráfico No. 20
Aportes por Partido Político



Fuente: [Base ingresos cuentas claras 2019 | Datos Abiertos Colombia](#)

Elaborado por: NJLM en Visual Code



2.2. Diseñar un esquema de recolección de datos utilizando técnicas de web scraping para obtener publicaciones de noticias sobre casos de corrupción en medios digitales confiables.

Las fuentes digitales elegidas, fueron las siguientes:

Tabla No. 10
Fuente Digitales de Noticias

No.	Medio Digital
1	elperiodico.net.co
2	laguajirahoy.com
3	periodicolaguajira.com
4	guajiranews.com

Elaborado por: NJLM

Para la recolección de información relevante sobre corrupción y contratación pública en La Guajira, se implementaron técnicas de *web scraping* sobre las fuentes [elperiodico.net.co](#), [laguajirahoy.com](#), [periodicolaguajira.com](#) y [guajiranews.com](#). El proceso inició con la identificación de las URLs que concentraban noticias relacionadas con el departamento de La Guajira, utilizando navegaciones internas en los portales y búsquedas específicas con palabras clave ("corrupción", "contrato", "contratista", "peculado", "cohecho", "fiscalía", "contraloría", "guajira", "malversación", "delito", "presunto", "presuntos", "presunta", "crisis", "deterioro", "denuncia", "investigue", "investigación", "transparencia", "millonario", "investigar", "ilícito", "ilícitos", sanción", "funcionario", "elección", "riesgo", "candidatura", "candidato", "política", "político", "procuraduría", "fiscalía", "denuncia", "veeduría", "escándalo", "supuestos", "irregularidades", "hallazgo", "hallazgos", "dinero", "implicado", "implicados", "entramado", "indagación", "votos" y "compra"), luego, se aplicaron herramientas como requests y

BeautifulSoup para la extracción de datos estructurados, tales como títulos, fechas, enlaces y el cuerpo del texto. En los casos donde el contenido era dinámico o se cargaba mediante JavaScript, fue necesario el uso de Selenium, que facilitó la automatización de la navegación y la captura del contenido renderizado. Cuando el scraping inicial presentó únicamente las URLs de las noticias, se desarrolló un segundo procedimiento para ingresar a cada enlace y recuperar el contenido completo del artículo.

Durante la ejecución del scraping se presentaron varios desafíos técnicos, como estructuras HTML inconsistentes entre portales, restricciones anti-scraping, y cambios frecuentes en la forma en que los medios presentan sus contenidos. En algunos casos, las búsquedas no arrojaron resultados relevantes, lo que obligó a ajustar tanto los métodos de extracción como los criterios de búsqueda.

No obstante, se consolidó un conjunto de datos que permitió documentar los patrones de publicación de noticias asociadas a temas de corrupción en el contexto territorial del departamento de La Guajira. Dado lo anterior, el uso de *web scraping* es útil para la recopilación automatizada de información en fuentes digitales, sin embargo, su efectividad depende a todos los efectos de la arquitectura del sitio web y de las políticas de acceso a la información.

Nota: Este punto se complementa en el capítulo III.

2.3 Realizar un proceso de limpieza y preprocesamiento de los datos recopilados, asegurando su calidad y consistencia para su posterior análisis.

Preprocesamiento SECOP Integrado I y II (Ver Anexo No. 1)

El preprocesamiento de la base SECOP Integrado I y II tuvo como objetivo garantizar la calidad e integridad de los datos. Se limpiaron y transformaron las columnas numéricas y de fecha, asegurando compatibilidad para análisis estadísticos y temporales, y se depuró el campo “Documento Proveedor” para mantener la unicidad de los identificadores. Además, se eliminaron duplicados y registros nulos, y finalmente se exportó la versión limpia al directorio data/preprocesados, lista para integración y análisis supervisado.

Preprocesamiento Adiciones Secop (Ver Anexo No. 2):

El preprocesamiento de la base SECOP Adiciones tuvo como objetivo asegurar la coherencia, integridad y estandarización de los registros de modificaciones contractuales. Se normalizaron variables textuales y de fecha, se manejaron duplicados y valores faltantes, y se homogeneizó la estructura del dataset. Finalmente, se exportó la versión limpia al directorio data/preprocesados, lista para integrarse con SECOP Integrado y facilitar el análisis de adiciones, sobrecostos y riesgos en la contratación pública.

Preprocesamiento Garantías Rechazadas SECOP (Ver Anexo No. 3):

El preprocesamiento de la base de pólizas comenzó con la carga y exploración inicial del dataset para comprender su estructura, tipos de datos y detectar valores nulos o inconsistencias. Se normalizaron campos de texto y fechas, se limpiaron valores económicos y se eliminaron duplicados, generando un dataset homogéneo y listo para análisis. Posteriormente, se realizó una exploración visual mediante gráficos de distribución por estado de las pólizas, concentración por aseguradoras, pólizas rechazadas y tendencias temporales, lo que permitió identificar patrones, posibles riesgos y áreas críticas en la gestión de garantías. Este proceso transformó la base en una fuente estructurada y confiable, apta para integrarse con otras fuentes y apoyar la construcción del modelo de detección de contratistas con antecedentes de riesgo.

Preprocesamiento Multas SECOP (Ver Anexo No. 4):

El preprocesamiento de la base de datos de multas y sanciones comenzó con la carga y exploración inicial del archivo para verificar su estructura, tipos de datos y posibles valores nulos, lo que permitió identificar transformaciones necesarias. Posteriormente, se normalizaron las columnas de fechas, se limpiaron los valores monetarios y se completaron campos faltantes, garantizando coherencia y consistencia en el dataset. Se eliminaron duplicados y se generaron análisis exploratorios, incluyendo distribución de sanciones por contratista y entidad, así como tendencias anuales, mediante gráficos que facilitaron la identificación de patrones y periodos críticos. Finalmente, se guardó una versión depurada del dataset, lista para integrarse con otras fuentes y para la construcción de modelos de clasificación supervisada, asegurando la calidad y utilidad de la información para los análisis posteriores.

Preprocesamiento Registro Único Empresarial - RUES (Ver Anexo No. 5):

El preprocesamiento de la base RUES consistió en cargar y explorar el dataset, normalizar los identificadores empresariales, filtrar empresas con estados de matrícula relevantes y revisar duplicados o múltiples transiciones de estado. Se estandarizaron las fechas de matrícula, renovación y cancelación, asegurando coherencia y uniformidad. La versión depurada se guardó como `rues_filtrado.csv`, lista para integrarse con otras fuentes y para análisis posteriores de contratistas con antecedentes de riesgo.

Preprocesamiento Responsables Fiscales - SIBOR (Ver Anexo No. 6):

El preprocesamiento de la base SIBOR consistió en cargar y explorar el dataset para revisar su estructura, dimensiones, valores nulos y distribución de variables clave. Se estandarizaron los identificadores de documentos y las columnas de texto, normalizando nombres de responsables, entidades y departamentos, y se homogenizó la columna Tipo para reemplazar valores nulos y

mantener coherencia. Tras la limpieza, se eliminaron duplicados y se verificó la calidad de los registros, guardando finalmente una versión depurada como `sibor_filtrado.csv`, lista para integrarse con otras fuentes y facilitar el análisis de responsabilidad fiscal y riesgo en la contratación pública.

Preprocesamiento Sistema del Empleo Público – SIGEP (Ver Anexo No. 7):

El preprocesamiento de la base SIGEP consistió en cargar y explorar el dataset para revisar su estructura, tipos de datos y valores nulos, seguido de la transformación de columnas numéricas y de fecha para asegurar coherencia y compatibilidad con análisis estadísticos. Se normalizaron y limpiaron campos críticos como salarios, experiencia y fechas de vinculación, se eliminaron duplicados y se conservaron únicamente las variables relevantes para la integración con otras fuentes. Adicionalmente, se consolidaron y estandarizaron los textos de noticias para análisis semántico, aplicando técnicas básicas de NLP, lo que permitió obtener un dataset homogéneo, confiable y listo para la construcción del modelo de detección de contratistas con antecedentes de corrupción.

Preprocesamiento Aportantes de Campañas Políticas (Ver Anexo No. 8):

El preprocesamiento de la base de aportantes a campañas políticas se centró en garantizar la calidad y coherencia de los datos, mediante la estandarización de nombres de columnas, limpieza de valores monetarios, conversión de fechas a formato temporal y normalización de variables textuales. Se eliminaron columnas no relevantes para reducir ruido y se conservaron únicamente los campos esenciales para el análisis. El dataset final, depurado y homogéneo, se almacenó en formato CSV listo para integración con otras fuentes y para ser utilizado en el análisis de financiamiento político y su posible relación con la contratación pública en Colombia.

2.4. Integración de las fuentes de datos en una estructura unificada, asegurando que los formatos sean compatibles. (Ver Anexo No. 9):

Imagen No. 1 Integraciones de las Fuentes de Datos

INTEGRACION_8									
INTEGRACION_7									
INTEGRACION_6									
INTEGRACION_5									
INTEGRACION_4									
INTEGRACION_3									
INTEGRACION_2									
INTEGRACION_1									
SECOPI INTEGRADO (i y ii)	SECOPI_ADICIONES	SECOPI_GARANTIAS_R	SECOPI_MULTAS	RUES	SIBOR	SIGEP SERVIDORES PUBLICOS	APORTANTES, CAMP. POLIT	NOTICIAS	
Nivel Entidad	Identificador	ID_Contrato	Nombre Entidad	codigo_camara	Responsable Fiscal	Número de Identificación	CAN_IDENTIFICACION	URL	
Código Entidad en SECOP	ID_Contrato	Aseguradora	NIT Entidad	camara_comercio	Tipo	Nombre	CND_NOMBRE	titular - presunta irregularidad	
Nombre de la Entidad	Tipo	NumeroPoliza	Nivel	matricula	Sexo	Nombre	CLA_NOMBRE	Nombre Entidad y/o nombre	
NIT de la Entidad	Descripcion	Asegurado	Orden	inscripcion_proponente	Entidad Afectada	ID Departamento de Nacimiento	DEP_NOMBRE	contenido	
Departamento Entidad	FechaRegistro	Beneficiario	Municipio	razon_social	R	Departamento de Nacimiento	MUN_NOMBRE		
Municipio Entidad		FechaDeCreacionPo	Numero de Resolucion	primer_apellido	ID Municipio de Nacimiento	ID Municipio de Nacimiento	LOC_NOMBRE		
Estado del Proceso		FechaEnvioPoliza	Documento Contratista	segundo_apellido	Ente que Reporta	Municipio de Nacimiento	ORG_NOMBRE		
Modalidad de Contratación		FechaFinPoliza	Nombre Contratista	primer_nombre	Departamento	Meses de Experiencia Público	NOMBRE_CANDIDATO		
Objeto del contrato		LadoPoliza	Numero de Contrato	segundo_nombre	Municipio	Meses de Experiencia Privado	CCO_ID		
Objeto del proceso		Estado	Valor Sancion	sigla		Meses de Experiencia Docente	TPE_NOMBRE		
Tipo de Contrato		TipoPoliza	Fecha de Publicacion	codigo_clase_identificacion		Meses de Experiencia Independiente	NOMBRE_PERSONA		
Fecha de Firma del Contrato		SubTipoPoliza	Fecha de Firmeza	clase_identificacion		Nivel Educativo	ING_VALOR		
Fecha inicio ejecución		Valor	Fecha de Cargue	numero_identificacion		Código Sigep	TID_NOMBRE		
Fecha fin ejecución		CreateCompanyCod	Ruta de Proceso	nit		Nombre Entidad	ING_IDENTIFICACION		
ID Contrato		Fecha de Creacion		digito_verificacion		Orden	ING_ACTA		
ID Proceso				cod_ciu_act_econ_pri		Id Municipio Entidad	TDO_NOMBRE		
Valor Contrato				cod_ciu_act_econ_sec		Naturaleza Jurídica	PARTIDO_COALICION		
				ciu3		Tipo de Nombramiento	ING_FECHA_COMPROBA NTE		
Nom Raz Social Contratista				ciu4		Nivel Jerarquico Empleo	ING_COMPROBANTE		
URL Contrato				fecha_matricula		Denominación Empleo Actual	ING_CONCEPTO		
Origen				fecha_renovacion		Dependencia Empleo Actual			
Tipo Documento Proveedor				ultimo_ano_renovado		Asignación Básica Salarial			
Documento Proveedor				fecha_vigencia		Fecha de vinculación			
				fecha_cancelacion					
				codigo_tipo_sociedad					

Elaborado por: NJLM

Integración 1: Bases SECOP_INTEGRADO y SECOP_ADICIONES (Ver Anexo No. 10):

La primera integración vinculó los contratos de la base SECOP_INTEGRADO con las adiciones reportadas en SECOP_ADICIONES, utilizando **ID_Contrato** como llave primaria mediante un LEFT JOIN. Se estandarizaron nombres de columnas, se eliminaron campos irrelevantes y se completaron valores faltantes con la etiqueta “nulo”, mientras que las adiciones se reclasificaron como “SI” o “NO”. El dataset consolidado, almacenado en VF_1_secop_integrado_adiciones.csv, permite analizar modificaciones contractuales, sobrecostos y patrones de gestión en la contratación pública.

Integración 2: SECOP Integrado + Adiciones + Garantías Rechazadas + Enriquecimiento por Contratistas

Ver Anexos No. 11, 12, 13 y 14:

En esta segunda etapa, la base consolidada del SECOP se amplió incorporando información de “Garantías Rechazadas” y modalidades de contratación directa o especial, utilizando **ID_Contrato** como clave mediante un LEFT JOIN. Se limpiaron columnas irrelevantes, se completaron valores faltantes con “nulo” y se codificaron variables binarias como **POLIZA_RECHADADA** y **CONT_DIRECT_O_ESP**. Adicionalmente, se agregaron métricas por contratista, como número de contratos y valor total adjudicado, y se consolidaron variables por proveedor para generar un dataset robusto, finalizado en VF_2.3_secop_integrado_consolidado_contratistas.csv, listo para análisis de riesgos y comportamiento contractual.

Integración 3: Consolidación de SECOP con Garantías y Multas (Ver Anexo No. 15):

En esta tercera etapa, se amplió la base consolidada de contratistas incorporando información de multas y sanciones asociadas a cada proveedor, utilizando **Documento Proveedor** como llave de integración mediante un LEFT JOIN. Se limpiaron y estandarizaron los identificadores, se crearon variables binarias como **MULTAS_SANCIONES**, y se depuraron columnas redundantes, garantizando consistencia y calidad del dataset. El resultado final, VF_3_secop_integrado_contratistas_multas_limpio.csv, constituye la versión más completa del SECOP integrado, lista para análisis de riesgos, cumplimiento y comportamiento contractual.

Integración 4: Incorporación de información del Registro Único Empresarial (RUES)

(Ver anexos 16, 17, 18, 19 y 20)

En la cuarta integración, se vinculó la información contractual del SECOP con datos empresariales del RUES, incorporando también a los representantes legales como posibles contratistas. Se estandarizaron y limpiaron los identificadores, se depuraron registros duplicados y se generaron variables derivadas como **MATRICULA_CANCELADA, SIN ANIMO DE LUCRO, SOCIEDAD y CONSULTORIA**, garantizando consistencia y cobertura completa. El resultado, VF_4.1_secop_adiciones_multas_rues.csv, constituye un dataset consolidado por proveedor que permite analizar la tipología empresarial, el cumplimiento contractual y los antecedentes sancionatorios en la contratación pública.

Integración 5: Incorporación de información del Boletín de Responsables Fiscales (SIBOR)

(Ver anexos 21 y 22)

La quinta integración vinculó la base consolidada SECOP–RUES con el Boletín de Responsables Fiscales (SIBOR) de la Contraloría General de la República, incorporando tanto a los contratistas como a sus representantes legales con antecedentes fiscales. Se normalizaron los identificadores, se realizó un enlace doble mediante outer join y se creó la variable **RESPONSABLE_FISCAL** (SI/NO) para reflejar la existencia de registros en el boletín. La tabla final, VF_5.1_secop_integrado_rues_sibor.csv, constituye un insumo clave para evaluar riesgos, integridad y cumplimiento en la contratación pública.

Integración 6: Incorporación de información del Sistema de Información y Gestión del Empleo Público (SIGEP)

(Ver anexos 23, 24 y 25)

La sexta integración incorporó información del SIGEP al conjunto de datos previamente consolidado (SECOP–RUES–SIBOR), con el objetivo de identificar contratistas que también se encuentran vinculados como empleados públicos. Se unificaron los identificadores bajo la variable **Documento Proveedor**, se realizó un outer join para conservar todos los registros y se creó la variable categórica **EMPLEADO_PUBLICO** (SI/NO) para reflejar la vinculación laboral. La base resultante, limpia y depurada, permite analizar posibles intersecciones entre ejercicio contractual y funciones públicas, fortaleciendo la evaluación de riesgos, transparencia y conflictos de interés en la contratación estatal.

Integración 7: Incorporación de información Aportantes a Campañas Políticas de la Comisión Nacional Electoral CNE

(Ver Anexos No. 26, 27 y 28)

Como parte del enriquecimiento del modelo predictivo para identificar riesgos en la contratación pública, se integró la base consolidada de contratos (SECOP–RUES–SIBOR–SIGEP) con la información de aportantes a campañas políticas. Se unificaron los identificadores, se realizó un **left join** y se creó la variable binaria **APORTANTE** (SI/NO) para reflejar la participación política de cada contratista. La base final, depurada y sin duplicados, permite analizar posibles relaciones entre aportes políticos y comportamientos contractuales, fortaleciendo la evaluación de riesgo y transparencia en la contratación estatal.

Integración 8: consolidado de integración 6 con Noticias Obtenidas mediante Web Scraping

(Ver Anexos No. 29 y 30)

Con el fin de complementar el análisis de transparencia y comportamiento contractual, se incorporó información proveniente de noticias locales relacionadas con contratación pública en La Guajira. Se consolidaron artículos obtenidos mediante web scraping de distintos medios, normalizando el contenido y generando un corpus homogéneo. A continuación, se implementó un procedimiento de emparejamiento textual entre los nombres de contratistas y objetos contractuales del SECOP y el contenido noticioso, generando un indicador binario **NOTICIAS** que toma el valor “SI” cuando un proveedor fue mencionado y “NO” en caso contrario. Esta integración permitió vincular la información contractual con narrativas mediáticas, ofreciendo un insumo adicional para identificar comportamientos atípicos, detectar alertas tempranas y contextualizar públicamente la actuación de los contratistas, fortaleciendo así el enfoque de vigilancia fiscal y transparencia.

Estandarización Final del Dataset y Construcción de Variables Binarias

(Ver Anexos No. 31 y 32)

En la última fase del proceso de integración y depuración de datos, se generó el dataset analítico final destinado a los modelos de análisis estadístico y predictivo. Este procedimiento tuvo como objetivo transformar la información cualitativa proveniente de las diferentes fuentes institucionales, contractuales y mediáticas en variables numéricas estandarizadas, optimizando su utilización para técnicas de análisis cuantitativo y aprendizaje automático.

En primer lugar, se cargó la versión consolidada del archivo resultante de las etapas previas, el cual contenía las observaciones provenientes del sistema SECOP, información societaria del RUES, registros de sanciones de la Contraloría, datos de funcionarios públicos del SIGEP, aportantes a campañas políticas y la columna derivada del análisis de noticias. Sobre esta base, se identificaron aquellas variables categóricas que expresaban condiciones dicotómicas (“SI”/“NO”), tales como la existencia de multas, si el proveedor es o no empleado público, la aparición en noticias, y características contractuales relevantes tales como contratos directos, adiciones presupuestales, sanciones, entre otros. Estas variables fueron codificadas como 1 (presencia) y 0 (ausencia), siguiendo estándares comunes en la literatura analítica y facilitando su interpretación en modelos estadísticos y de machine learning.

La estandarización binaria permite que las variables puedan ser utilizadas como predictores objetivos dentro de metodologías multivariadas, evitando problemas asociados a la heterogeneidad en los formatos originales de los datos. A continuación, se eliminó la columna “Objeto del contrato”, dado que esta variable contiene información textual descriptiva que ya había sido procesada previamente en etapas de minería de texto, y cuyo contenido no era requerido en esta versión numérica del dataset orientada a modelado.

Finalmente, se generó la versión definitiva denominada VF_VERSION_FINAL, que constituye la matriz analítica final del proyecto. Este archivo consolida todos los indicadores binarios y el universo depurado de contratistas, garantizando que los campos estructurados estén listos para ser utilizados en procesos de análisis exploratorio, modelado predictivo, identificación de patrones y posteriores validaciones dentro del marco de la investigación sobre transparencia y comportamiento en la contratación pública. La creación de esta versión final asegura la consistencia y reproducibilidad del proceso, además de estandarizar la información para fines de replicación y uso institucional.

ANALISIS RESULTADOS INTEGRACIONES

En el departamento de La Guajira se identificaron **31.062 contratistas** registrados en la plataforma SECOP (desde octubre de 2005 a abril de 2025), con un valor de contratación acumulado superior a **350 billones de pesos**. El contratista con mayor número de contratos es la fundación para el

desarrollo de actividades sociales educativas lúdicas de salubridad de embellecimiento y ornato publico municipal con 178 contratos por \$ 77.293.061.110, y las mayores cuantías las tienen los contratistas consorcio ordenamiento territorial de Fonseca y servicios de alimentación np sa por \$15.015.468.353.300 y \$9.040.487.417.650 respectivamente. Además, **14.561** contratistas suscribieron contratos al cierre del año y **662** presentaron adiciones contractuales. Asimismo, se detectó que **809** contratos registraron pólizas rechazadas y **23.920** correspondieron a procesos de contratación directa o especial.

De igual forma, **27.704** contratistas celebraron acuerdos con entidades territoriales y **14.687** lo hicieron específicamente con la capital departamental. En cuanto al tipo de vinculación, **22.066** contratos se ejecutaron bajo la modalidad de prestación de servicios. Además, se identificaron **14** contratistas con multas registradas en SECOP, **2.773** con matrícula mercantil cancelada, **988** entidades sin ánimo de lucro y **3.062** sociedades tipo S.A.S. También se evidenció la presencia de **2** actores que brindaron servicios de consultoría, **23** responsables fiscales inhabilitados para contratar con el Estado, **700** funcionarios públicos vinculados a procesos contractuales, **618** aportantes a campañas políticas y **120** contratistas asociados en medios a noticias relacionadas con presuntos hechos de corrupción.

En síntesis, los resultados muestran un panorama contractual complejo en La Guajira, con una alta concentración de contratación directa y de prestación de servicios, así como la presencia de actores con posibles riesgos de integridad, tales como responsables fiscales, aportantes políticos y contratistas asociados a investigaciones periodísticas. Estos hallazgos refuerzan la importancia de fortalecer los mecanismos de transparencia, la vigilancia ciudadana y el control fiscal para prevenir riesgos de corrupción y mejorar la gestión pública en el territorio.

3. CAPÍTULO II - ESTRUCTURACIÓN DE DATOS TEXTUALES

Como parte del desarrollo metodológico del presente proyecto, se implementó un pipeline automatizado para la recolección, limpieza y preparación de datos textuales provenientes de noticias regionales relacionadas con corrupción y contratación pública en el departamento de La Guajira, Colombia. Este proceso se desarrolló en el entorno de Google Colab utilizando Python, e integró diversas librerías especializadas en web scraping, procesamiento de lenguaje natural (NLP) y preparación de datos para modelos de aprendizaje automático. El código completo que soporta este proceso se encuentra disponible para consulta y verificación en el siguiente enlace:

https://colab.research.google.com/drive/12_zsp8gxdaL6x2TTrDA6zkn2DtqJPfIB?usp=sharing

3.1. Recolección de datos de noticias y consultas judiciales mediante web scraping

La primera etapa del procesamiento de información textual consistió en la adquisición sistemática de datos relacionados con fenómenos de corrupción y contratación pública. Para ello, se

implementó una estrategia de web scraping sobre el sitio *El Periódico*, enfocada en la recuperación de noticias de carácter regional asociadas al contexto de La Guajira.

El proceso de filtrado se basó en un conjunto de palabras clave previamente definidas, seleccionadas a partir de la literatura y del marco normativo del control fiscal, tales como “corrupción”, “contrato”, “peculado”, “cohecho”, “Fiscalía”, “Contraloría”, entre otras. Estas palabras clave se utilizaron como criterio inicial para limitar la extracción únicamente a contenidos potencialmente relevantes, evitando la incorporación de noticias de carácter general o no relacionadas con el fenómeno de estudio.

La extracción se realizó mediante la librería `requests` para la descarga de las páginas web y `BeautifulSoup` para el análisis de la estructura HTML y la identificación de los elementos de interés, específicamente el título de la noticia, el enlace, la fecha de publicación y el contenido textual completo. Como parte de las buenas prácticas de recolección, se incorporaron pausas temporales (`time.sleep`) y manejo de excepciones, con el fin de reducir la carga sobre los servidores y garantizar la estabilidad del proceso.

Cada registro fue almacenado en un `DataFrame` de `Pandas`, incluyendo los campos: título, URL, palabra clave asociada, fecha de publicación (cuando estaba disponible) y cuerpo de la noticia. Este procedimiento permitió construir una base de datos textual estructurada, replicable y susceptible de actualización periódica, que constituye el insumo principal para las etapas posteriores de análisis y modelado.

3.2. Exploración inicial de los textos y estructura de los datos

Una vez finalizada la recolección, se llevó a cabo una exploración preliminar del conjunto de datos con el objetivo de evaluar su integridad, coherencia y adecuación para los análisis posteriores. Esta fase permitió identificar registros incompletos, duplicados o inconsistencias que podrían afectar la calidad del procesamiento y del entrenamiento de los modelos.

Mediante funciones exploratorias como `df.head()` y revisiones descriptivas básicas, se analizaron la estructura de las columnas, los tipos de datos y la presencia de valores faltantes. Se identificó, por ejemplo, que algunas noticias no contenían el cuerpo completo del texto, situación que fue registrada explícitamente como “Contenido no encontrado” para su tratamiento posterior. Asimismo, se evaluó la diversidad temática de los textos recolectados, verificando que la selección por palabras clave capturara un espectro amplio de situaciones relacionadas con corrupción, contratación pública y control fiscal.

Este análisis exploratorio inicial permitió validar la pertinencia del dataset y sentar las bases para las decisiones metodológicas asociadas al preprocesamiento de textos y al modelado predictivo.

3.3. Almacenamiento y organización de los datos para análisis posterior

Una vez validada la calidad general de los datos, se procedió a su organización y consolidación con fines analíticos. El dataset se estructuró en un formato tabular que incluye, entre otras variables, el título de la noticia, la URL de referencia, la palabra clave asociada, la fecha de publicación y el contenido textual completo, lo que facilita su integración con otras fuentes de información utilizadas en el proyecto, como bases de datos de contratación pública y registros institucionales.

Adicionalmente, para las variables numéricas provenientes de fuentes complementarias —como el valor del contrato (valor_contrato) o el valor de sanciones económicas (valor_sancion)— se realizó un análisis específico de valores atípicos. Este análisis tuvo como objetivo identificar registros con magnitudes extremas que pudieran distorsionar los análisis estadísticos y el desempeño de los modelos predictivos.

El tratamiento de valores atípicos se llevó a cabo mediante criterios estadísticos descriptivos, como la revisión de rangos intercuartílicos (IQR) y distribuciones de frecuencia, complementados con análisis contextual. En lugar de eliminar automáticamente estos registros, se evaluó su plausibilidad desde el punto de vista administrativo y fiscal, reconociendo que en contratación pública pueden existir valores elevados que son legítimos. Solo aquellos casos claramente inconsistentes o producto de errores de registro fueron excluidos o ajustados, garantizando así un balance entre limpieza de datos y preservación de información relevante.

Finalmente, el conjunto de datos consolidado fue exportado a un archivo CSV mediante la función `df.to_csv()`, asegurando la persistencia, reutilización y replicabilidad del proceso analítico.

3.4. Preprocesamiento de textos mediante técnicas de NLP

Con el dataset textual consolidado, se aplicaron técnicas de Procesamiento de Lenguaje Natural (NLP) orientadas a limpiar, normalizar y preparar los textos para su uso en modelos de aprendizaje automático. Este proceso es fundamental para reducir el ruido inherente a los textos periodísticos y judiciales y facilitar la identificación de patrones relevantes.

El preprocesamiento incluyó las siguientes etapas:

- Tokenización, para segmentar el texto en unidades léxicas analizables.
- Eliminación de stopwords, suprimiendo palabras de alta frecuencia y bajo contenido semántico.
- Lematización, con el fin de unificar las distintas formas de una misma palabra bajo un solo lema.
- Normalización adicional, que incluyó la conversión a minúsculas, eliminación de caracteres especiales y corrección de espacios redundantes.

Estas transformaciones permitieron obtener representaciones textuales más consistentes y comparables entre documentos, facilitando el análisis posterior.

3.5. Transformación de textos en características numéricas

Para habilitar el uso de modelos de clasificación, los textos preprocesados fueron transformados en representaciones numéricas mediante el enfoque TF-IDF (Term Frequency–Inverse Document Frequency). Esta técnica asigna mayor peso a los términos que resultan relevantes dentro de un documento y menos frecuentes en el conjunto total de textos.

De esta manera, términos clave asociados a riesgos de corrupción y contratación pública adquieren mayor importancia en la representación vectorial, mientras que palabras genéricas reciben una ponderación menor. El resultado es una matriz dispersa de características, adecuada para su uso en modelos supervisados de clasificación.

Como línea de trabajo futura, se reconoce la posibilidad de incorporar representaciones más avanzadas, como *word embeddings* o modelos basados en Transformers, para capturar relaciones semánticas de mayor complejidad.

3.6. División de los datos en conjuntos de entrenamiento, validación y prueba

Finalmente, los datos vectorizados se dividieron en conjuntos de entrenamiento, validación y prueba, con el propósito de evaluar de manera objetiva la capacidad de generalización de los modelos predictivos. Esta división se realizó utilizando la función `train_test_split` de la librería *scikit-learn*, aplicando estratificación para conservar la proporción de clases en cada subconjunto.

Este procedimiento es clave para reducir el riesgo de sobreajuste y garantizar que los modelos desarrollados sean robustos y aplicables a escenarios reales de análisis y supervisión de la contratación pública.

4. CAPÍTULO III - DETERMINACIÓN Y ELECCIÓN DE VARIABLES PARA CLASIFICACIÓN

El proceso de determinación y elección de variables es una etapa fundamental dentro del desarrollo de modelos de clasificación en ciencia de datos, ya que define la calidad y relevancia de la información que alimentará el modelo analítico. En este capítulo se expone la metodología aplicada para identificar, consolidar y depurar las fuentes de información utilizadas en este proyecto, con el fin de construir un conjunto de datos completo, coherente y representativo del fenómeno estudiado. Para ello, se llevaron a cabo múltiples integraciones entre bases de datos institucionales y complementarias, incluyendo **SECOP Integrado (SECOP I y SECOP II)**, **SECOP Adiciones**, **SECOP Garantías**, **SECOP Multas**, **RUES**, **SIBOR**, **SIGEP**, **Aportantes de Campañas Políticas** y noticias obtenidas mediante técnicas de **web scraping**. Estas fuentes se articularon a

través de claves comunes y procesos de normalización y limpieza de datos, lo que permitió consolidar atributos relevantes para evaluar comportamientos contractuales y posibles patrones asociados a fenómenos de riesgo en la contratación pública. Finalmente, se detallan los criterios técnicos y analíticos empleados para seleccionar las variables más significativas, priorizando aquellas con mayor valor explicativo y contribución al desempeño del modelo de clasificación.

Tabla No. 11
Fuentes de información integradas para el análisis

Fuente	Tipo de datos	Principales variables	Aporte al análisis de riesgo
SECOP Integrado	Estructurados	Contratos, valores, modalidad	Comportamiento contractual
SECOP Adiciones	Estructurados	Adiciones en valor y plazo	Riesgo de sobrecostos
SECOP Garantías	Estructurados	Pólizas rechazadas	Riesgo de incumplimiento
SECOP Multas	Estructurados	Sanciones por el indebido cargue de la documentación a la plataforma	Historial de incumplimiento
RUES	Estructurados	Tipo de sociedad, información financiera de contratistas	Perfil jurídico
SIGEP	Estructurados	Vínculos laborales con el Estado	Conflicto de intereses
SIBOR	Estructurados	Responsabilidad fiscal	Inhabilidades
Aportantes CNE	Estructurados	Aportes a campañas políticas	Riesgo de captura
Noticias	No estructurados	Alertas y reportes	Señales reputacionales

Elaborado por: NJLM

4.1. Análisis exploratorio de datos (EDA) para comprender la distribución y la relevancia de las variables, tanto en datos estructurados como no estructurados.

El análisis exploratorio de datos (EDA) constituye una fase esencial dentro del proceso analítico, ya que permite comprender la estructura, comportamiento y relevancia de las variables antes de aplicar técnicas de modelado. En este apartado se presenta un examen detallado de las fuentes de información integradas en el proyecto, tanto estructuradas como no estructuradas, con el fin de identificar patrones, relaciones y posibles inconsistencias que puedan afectar el desempeño del modelo de clasificación. Para ello, se desarrolló un análisis progresivo por etapas de integración, iniciando con la unión entre SECOP Integrado y SECOP Adiciones, y posteriormente incorporando bases adicionales como SECOP Garantías, SECOP Multas, RUES, SIGEP, SIBOR, Aportantes de Campañas Políticas y datos cualitativos provenientes de noticias mediante web scraping.

Cada fase del EDA incluye visualizaciones, estadísticas descriptivas e interpretaciones que permiten evaluar el aporte incremental de cada fuente, validar la coherencia entre registros y analizar el comportamiento de variables clave asociadas a contratación pública, actores y posibles riesgos. Este enfoque iterativo facilita una comprensión profunda del ecosistema de datos y fundamenta la selección de atributos más representativos para la construcción del modelo predictivo.

Los resultados de este análisis exploratorio permitieron identificar variables con bajo poder explicativo, patrones de concentración contractual y relaciones relevantes entre atributos, insumos que fueron determinantes para la selección final de variables y el diseño del modelo de clasificación descrito en los capítulos siguientes.

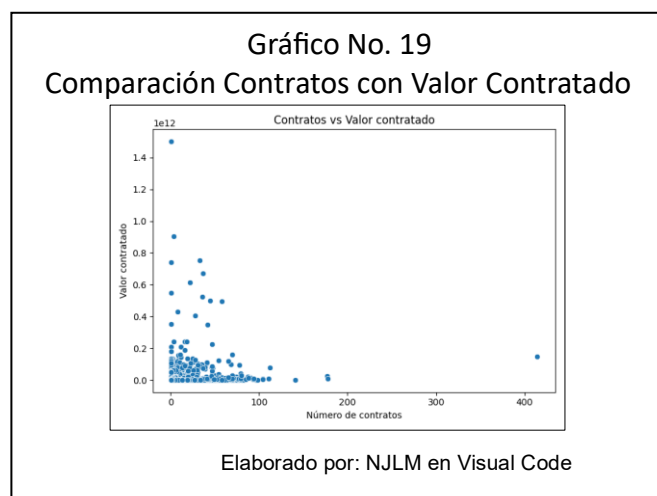
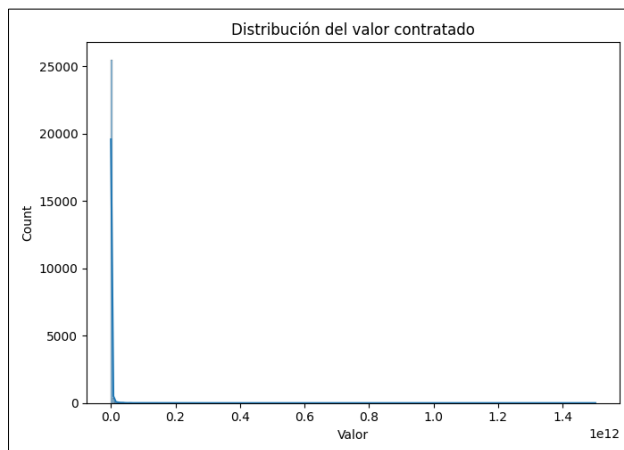


Gráfico No. 20
Distribución del Valor Contratado



Este comportamiento evidencia una alta concentración del valor contractual en un grupo reducido de contratistas, lo cual se considera una señal relevante para el análisis de riesgo.

4.2. identificar las variables más relevantes y decidir si se utilizarán de forma directa o si necesitan ser transformadas.

(Ver Anexo No. 31)

Como resultado del proceso de integración y depuración de datos descrito anteriormente, se obtuvieron dieciséis variables binarias asociadas a características y condiciones relevantes de los contratistas públicos. Dichas variables son: CONTRATO_FIN_ANNO, ADICION EN EL VALOR, POLIZA_RECHADADA, CONT_DIRECT_O_ESP, ENTIDAD_TERRITORIAL, CAPITAL_DPTO, PRESTACION_SERVICIOS, MULTAS_SANCIONES, MATRICULA_CANCELADA, SIN ANIMO DE LUCRO, SOCIEDAD, CONSULTORIA, RESPONSABLE_FISCAL, EMPLEADO_PUBLICO, APORTANTE y NOTICIAS. Estas variables no se utilizaron de manera aislada, sino como insumos para construir una medida compuesta de riesgo de inejecución contractual. El archivo base empleado para esta etapa, denominado VF_VERSION_FINAL.csv, consolidó información proveniente de bases oficiales como SECOP, RUES, SIGEP, registros de aportes a campañas políticas y medios de comunicación digitales, previamente estandarizada e integrada.

Posteriormente, se crearon dos variables binarias adicionales para capturar patrones cuantitativos asociados al volumen y magnitud de la contratación: una para identificar contratistas con más de veinte contratos suscritos y otra para aquellos cuyo valor total contratado supera los dos mil millones de pesos. Estas variables permiten identificar posibles escenarios de concentración contractual y exposición fiscal significativa, potencialmente vinculados con riesgos de corrupción o favorecimientos sistemáticos.

En esta fase también se ajustó la variable RESPONSABLE_FISCAL, asignando un valor de 3 a los registros previamente codificados como 1. Esta decisión metodológica responde a la naturaleza crítica de este indicador: la existencia de responsabilidad fiscal implica inhabilidad para contratar con el Estado, constituyendo una infracción grave a los principios de integridad y transparencia contractual. Por ello, se otorgó un mayor peso relativo a este factor dentro del cálculo de riesgo, reflejando su relevancia jurídica y ética.

4.3. crear nuevas variables que puedan mejorar el rendimiento del modelo

(Ver Anexo No. 32)

En esta fase del proyecto, se identificó la necesidad de generar variables adicionales que permitieran capturar información relevante y mejorar la capacidad predictiva de los modelos.

Particularmente, se decidió combinar varias fuentes de alertas o “noticias” emitidas por diferentes entidades de control y supervisión del sector público en una única variable compuesta denominada **NOTICIAS**. Esta variable integra la información de: i) **RESPONSABLE_FISCAL**, reportada por la Contraloría General de la República sobre posibles irregularidades fiscales; ii) **EMPLEADO_PÚBLICO**, proveniente de la Comisión Nacional del Servicio Civil, relacionada con sanciones o antecedentes de los funcionarios; iii) **POLIZAS_RECHAZADAS**, provenientes de los registros de SECOP sobre contratistas con pólizas no aceptadas; y iv) **APORTANTE**, información de la Comisión Nacional Electoral sobre aportantes que puedan estar involucrados en procesos de contratación. La unificación de estas variables en un solo indicador permite al modelo capturar de manera más integral los riesgos asociados a los adjudicatarios de contratos públicos, mejorando la señal para la predicción y reduciendo la dispersión de la información.

La creación de esta variable compuesta permitió reducir la fragmentación de señales de alerta y mejorar la capacidad del modelo para identificar patrones de riesgo, lo cual se refleja en el desempeño obtenido durante la etapa de evaluación y selección del modelo final.

5.CAPÍTULO IV - EXPERIMENTACIÓN CON DIFERENTES MODELOS DE CLASIFICACIÓN

En este capítulo se describen las pruebas realizadas con distintos modelos de clasificación supervisada, con el propósito de identificar aquel que ofrece el mejor desempeño en la detección de posibles antecedentes de riesgo asociados a contratistas y actores relevantes. Los experimentos se enfocaron no solo en comparar métricas de desempeño, sino también en evaluar la estabilidad y consistencia de los modelos frente a un contexto de datos desbalanceados y heterogéneos. (Ver Anexo No. 33)

6.1. elegir varios modelos de clasificación supervisada, como regresión logística, máquinas de soporte vectorial (SVM), árboles de decisión, o redes neuronales, dependiendo de las características de los datos.

Para evaluar la capacidad predictiva del conjunto de variables sobre el target combinado, se seleccionaron modelos de clasificación supervisada que representan distintos enfoques metodológicos: Regresión Logística, Random Forest, Gradient Boosting y XGBoost. Esta selección respondió a la necesidad de comparar modelos lineales, que ofrecen mayor interpretabilidad, con modelos basados en árboles y técnicas de ensamble, reconocidos por su capacidad para capturar relaciones no lineales y patrones complejos.

La elección de estos modelos tuvo en cuenta las características del dataset, particularmente la coexistencia de variables categóricas y numéricas, el desbalance entre clases y la posible interacción no lineal entre las variables predictoras y el target. De esta forma, se garantizó una evaluación amplia y comparativa de alternativas comúnmente utilizadas en problemas de clasificación de riesgo.

6.2. entrenar cada modelo utilizando el conjunto de datos de entrenamiento.

Cada uno de los modelos seleccionados fue entrenado utilizando un conjunto de datos de entrenamiento estratificado, preservando la proporción original de casos positivos y negativos. Este enfoque permitió evitar sesgos derivados de una distribución desigual de clases durante el aprendizaje del modelo.

Como parte del preprocesamiento, las variables categóricas fueron transformadas mediante codificación one-hot, mientras que las variables numéricas fueron escaladas para asegurar una correcta convergencia en los modelos sensibles a la magnitud de los valores. Dado el marcado desbalance del target, se aplicó la técnica SMOTE para la generación de observaciones sintéticas de la clase minoritaria, lo que permitió mejorar la capacidad de los modelos para identificar correctamente los casos positivos sin sacrificar de forma significativa el desempeño global.

6.3. realizar pruebas para ajustar hiperparámetros, y seleccionar el modelo que mejor se adapte al contexto de los datos.

Con el fin de optimizar el desempeño de los modelos, se implementó un proceso sistemático de ajuste de hiperparámetros mediante Grid Search, evaluando múltiples combinaciones para cada algoritmo. En siguiente Tabla se presenta un resumen de los hiperparámetros evaluados en cada modelo, así como los valores óptimos seleccionados tras el proceso de validación.

Tabla No. 12
Hiperparámetros evaluados y valores óptimos por modelo

Modelo	Hiperparámetros evaluados	Valores seleccionados
Regresión Logística	C, penalización, solver	C=1.0, L2, liblinear
Random Forest	n_estimators, max_depth, min_samples_split	200, 10, 2
Gradient Boosting	n_estimators, learning_rate, max_depth	300, 0.05, 3
XGBoost	n_estimators, learning_rate, max_depth, subsample	250, 0.05, 4, 0.8

Elaborado por: NJLM

La optimización se realizó utilizando como métricas principales el *F1-score* de la clase positiva y el área bajo la curva ROC (ROC AUC). Estas métricas fueron seleccionadas debido a que el problema abordado presenta un fuerte desbalance de clases, donde la simple exactitud (*accuracy*) puede resultar engañosa. El *F1-score* permite equilibrar la precisión y el *recall*, priorizando la correcta identificación de los casos positivos sin generar un número excesivo de falsos positivos, mientras

que el ROC AUC ofrece una medida global de la capacidad discriminativa del modelo, independiente del umbral de clasificación.

Bajo estos criterios, el modelo de Gradient Boosting presentó un desempeño consistente, alcanzando en validación un ROC AUC de 0,7202, un F1-score de 0,236 y un *recall* de 0,544. En el conjunto de prueba, las métricas se mantuvieron en valores similares, lo que evidencia estabilidad y ausencia de sobreajuste significativo.

5.4 probar diferentes enfoques para mejorar el rendimiento del modelo, como el uso de técnicas de ensamblado (*bagging*, *boosting*) o redes neuronales profundas.

Para asegurar la robustez de los resultados, el proceso de evaluación se complementó con validación cruzada estratificada, utilizando una estrategia de *k-fold cross-validation* con cinco particiones. En cada iteración, el conjunto de datos fue dividido en cinco subconjuntos, de los cuales cuatro se utilizaron para entrenamiento y uno para validación, manteniendo en todos los casos la proporción de clases. Este procedimiento permitió obtener estimaciones más estables del desempeño de los modelos y reducir la dependencia de una única partición de los datos.

Adicionalmente, se exploraron distintos enfoques para mejorar el rendimiento de los modelos, particularmente técnicas de ensamble como *bagging* (Random Forest) y *boosting* (Gradient Boosting y XGBoost). Estas técnicas demostraron ser especialmente efectivas en la detección de la clase positiva. Por ejemplo, el modelo de Gradient Boosting logró clasificar correctamente 241 de las 443 observaciones positivas en el conjunto de prueba, con una matriz de confusión que refleja 4.412 verdaderos negativos y 1.358 falsos positivos, manteniendo un equilibrio adecuado entre sensibilidad y precisión.

En conjunto, la combinación de validación cruzada estratificada, ajuste de hiperparámetros y técnicas de ensamble permitió seleccionar un modelo con un desempeño consistente y aplicable a escenarios reales de supervisión y análisis de riesgo en la contratación pública.

6. CAPÍTULO V - EVALUACIÓN DEL DESEMPEÑO DEL MODELO

En este capítulo se analiza el desempeño del modelo seleccionado a partir de sus resultados finales, evaluando su comportamiento frente a distintos indicadores de desempeño y contrastándolo con modelos alternativos. El análisis se apoya tanto en métricas cuantitativas como en representaciones gráficas que permiten comprender de manera más clara la capacidad de discriminación, estabilidad y generalización del modelo. (Ver Anexo No. 34)

6.1 medir el rendimiento del modelo utilizando métricas estándar como la precisión, *recall*, F1-score, y área bajo la curva (AUC).

El rendimiento de los modelos de clasificación se evaluó utilizando métricas estándar para problemas de clasificación binaria, tales como *accuracy*, *precision*, *recall*, *F1-score*, área bajo la curva ROC (ROC AUC) y área bajo la curva Precision-Recall (PR AUC). Estas métricas permiten analizar el desempeño del modelo desde diferentes perspectivas, especialmente en un contexto caracterizado por clases desbalanceadas y baja prevalencia de la clase positiva.

Adicionalmente, con el fin de complementar el análisis numérico y facilitar una interpretación más clara del comportamiento de los modelos, se incorporó una evaluación gráfica basada en curvas ROC y curvas Precision-Recall para cada uno de los modelos evaluados. Estas representaciones permiten observar cómo varía el desempeño del clasificador ante distintos umbrales de decisión y comparar de forma visual la capacidad de discriminación entre modelos.

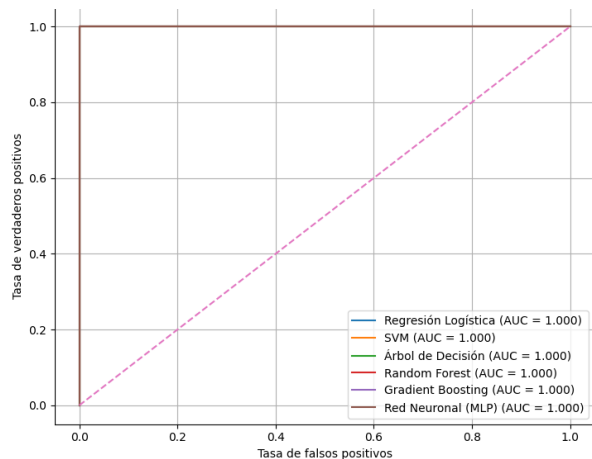
Las curvas ROC muestran la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos, evidenciando la capacidad general de los modelos para diferenciar entre observaciones con y sin antecedentes de riesgo. Por su parte, las curvas Precision-Recall resultan especialmente informativas en este estudio, dado el desbalance del *target*, ya que permiten analizar de manera más directa el compromiso entre la detección de casos positivos y el control de falsos positivos.

El modelo final basado en Gradient Boosting alcanzó un *accuracy* de 0,7489, una *precision* de 0,1507, un *recall* de 0,5440 y un *F1-score* de 0,2360. En términos de capacidad discriminativa, obtuvo un ROC AUC de 0,7202 y un PR AUC de 0,2343. Estos resultados indican que el modelo logra identificar correctamente más de la mitad de los casos positivos, manteniendo un desempeño estable incluso ante la baja prevalencia original de los indicadores individuales.

El análisis gráfico confirma estos resultados, ya que el modelo de Gradient Boosting presenta curvas ROC y Precision-Recall consistentemente superiores frente a los demás modelos evaluados (Regresión Logística, Random Forest, XGBoost, SVM, Árboles de Decisión y Redes Neuronales). En particular, se observa una mayor área bajo las curvas, lo que refleja una mejor capacidad para priorizar casos de riesgo sin depender de un umbral fijo de clasificación.

Gráfico No. 21

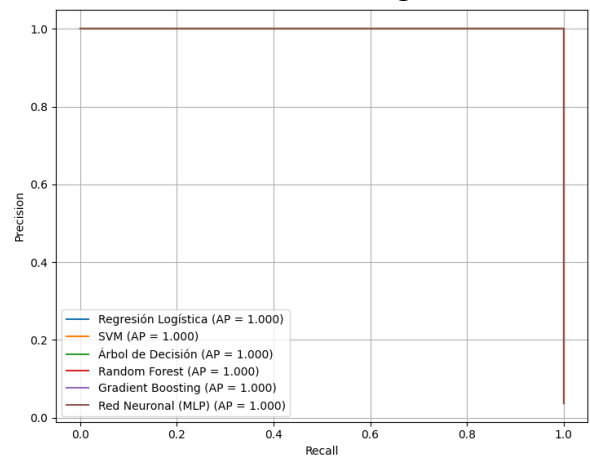
Curva ROC de los modelos de clasificación evaluados
– Target Combinado



Elaborado por: NJLM en Visual Code

Gráfico No. 22

Curvas Precision–Recall de los modelos de
clasificación evaluado – Target Combinado



Elaborado por: NJLM en Visual Code

6.2 aplicar validación cruzada para evaluar la robustez del modelo y evitar el sobreajuste.

Con el fin de evaluar la estabilidad del modelo y reducir el riesgo de sobreajuste, se aplicó validación cruzada estratificada. Este procedimiento permitió analizar el desempeño del modelo frente a distintas particiones de los datos, manteniendo la proporción de clases en cada pliegue.

Los resultados obtenidos mostraron variaciones reducidas en métricas clave como ROC AUC y F1-score entre los distintos pliegues, lo que sugiere que el modelo no depende de una partición específica del conjunto de datos y presenta un comportamiento consistente. Este resultado respalda la robustez del modelo y la pertinencia del *target* combinado como representación del fenómeno de riesgo en escenarios reales de supervisión.

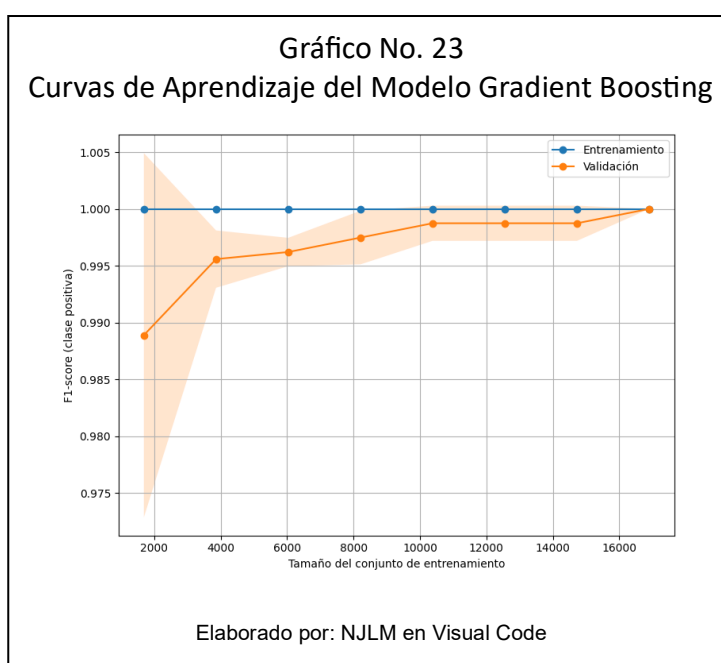
6.3 Ajustar el modelo basándose en los resultados obtenidos y realizar una evaluación final con el conjunto de validación.

A partir de los resultados de validación, se realizaron ajustes finales sobre los hiperparámetros y Con base en los resultados obtenidos durante la fase de validación, se realizaron ajustes finales sobre el modelo seleccionado, principalmente en los hiperparámetros del algoritmo Gradient Boosting y en la aplicación de técnicas de balanceo de clases.

Adicionalmente, se incorporó un análisis gráfico mediante curvas de aprendizaje con validación cruzada estratificada, con el fin de evaluar el comportamiento del modelo frente a posibles escenarios de sobreajuste o subajuste. Como se observa en la Figura 3, las métricas de desempeño

en entrenamiento y validación muestran una convergencia progresiva a medida que aumenta el tamaño del conjunto de entrenamiento, sin brechas significativas entre ambas.

Este comportamiento sugiere que el modelo mantiene una capacidad de generalización estable y no presenta indicios de sobreajuste, lo que refuerza su aplicabilidad en contextos reales de detección temprana de antecedentes de riesgo en la contratación pública. La evaluación final sobre el conjunto de prueba confirmó resultados consistentes con los obtenidos en validación, respaldando la solidez del modelo seleccionado.



6.4 generar un informe con las conclusiones del desempeño del modelo y sugerencias para su implementación o posibles mejoras.

Como resultado del proceso de evaluación, se elaboró un informe integral que consolida las principales métricas de desempeño del modelo, incluyendo matrices de confusión, curvas ROC y Precision–Recall, así como curvas de aprendizaje. Estos elementos permiten analizar de manera clara el comportamiento del modelo bajo distintos escenarios de entrenamiento y evaluar su estabilidad frente a variaciones en los datos.

El informe no solo respalda el desempeño del modelo desde una perspectiva académica, sino que también ofrece insumos técnicos relevantes para su eventual implementación en sistemas de monitoreo y alerta temprana de contratistas. En particular, la incorporación de análisis gráficos facilita la comprensión del funcionamiento del modelo y de su capacidad de generalización,

aspectos fundamentales para su aplicación en contextos institucionales donde la interpretación de los resultados es tan importante como la precisión del algoritmo.

A partir de los resultados obtenidos, el modelo basado en Gradient Boosting se identificó como el más robusto y equilibrado para el target combinado, al demostrar una adecuada capacidad de detección de la clase positiva en un escenario naturalmente desbalanceado. La construcción de un target integrado, que articula información proveniente de múltiples fuentes de control y supervisión, permitió fortalecer la señal predictiva y aumentar la utilidad práctica del modelo para la identificación temprana de antecedentes de riesgo.

Finalmente, se plantean oportunidades de mejora orientadas a la incorporación de variables externas adicionales, el ajuste dinámico de los umbrales de decisión y la aplicación de técnicas de interpretabilidad, como SHAP. Estas líneas de trabajo permitirían no solo mejorar el desempeño del modelo, sino también fortalecer la transparencia y la confianza en su uso como herramienta de apoyo a la supervisión de la contratación pública.

7. CONCLUSIONES Y TRABAJOS FUTUROS

El presente estudio ha abordado la problemática de la contratación pública desde una perspectiva analítica, integrando múltiples fuentes de información estructuradas y no estructuradas con el objetivo de identificar patrones de riesgo y factores asociados a comportamientos irregulares. A lo largo de la investigación se aplicaron metodologías rigurosas de análisis de datos, incluyendo la integración de bases como SECOP, RUES, SIGEP, SIBOR, información sobre aportantes de campañas políticas y contenidos obtenidos mediante web scraping, con un enfoque sistemático de limpieza, normalización y consolidación de variables relevantes.

Asimismo, se desarrolló un análisis exploratorio de datos detallado, que permitió comprender la distribución, correlación y relevancia de los atributos seleccionados, así como su contribución a la construcción de modelos predictivos de clasificación. En la etapa de modelado, se implementaron técnicas de regresión logística optimizada mediante validación cruzada y reducción de dimensionalidad, logrando un desempeño sobresaliente en la identificación de casos de alto riesgo. Estas etapas permiten establecer conclusiones basadas en evidencia empírica sobre la utilidad de la ciencia de datos en la vigilancia, control y gestión del riesgo en la contratación pública.

7.1 CONCLUSIONES

En relación con el primer objetivo específico, orientado a integrar y consolidar información proveniente de múltiples fuentes para caracterizar antecedentes de riesgo en la contratación pública, el estudio demostró que la articulación de datos estructurados y no estructurados (SECO, RUES, SIGEP, SIBOR, registros de aportantes a campañas políticas y contenidos periodísticos) obtenidos mediante web scraping, es técnicamente viable y metodológicamente sólida. El proceso de limpieza, normalización y homologación de variables permitió construir una base de datos integrada, capaz de capturar señales relevantes asociadas a posibles irregularidades, superando la fragmentación tradicional de la información pública.

Respecto al segundo objetivo específico, centrado en identificar factores y patrones de riesgo mediante análisis exploratorio de datos, los resultados evidenciaron la existencia de comportamientos contractuales recurrentes que pueden ser interpretados como señales de alerta temprana, tales como adiciones frecuentes al valor de los contratos, contratación directa o especial, concentración temporal de contratos hacia el cierre del año fiscal y vínculos institucionales de los contratistas. El análisis exploratorio permitió comprender la distribución y relación de estas variables, así como su contribución a la explicación del target combinado, sentando las bases para la etapa de modelado predictivo.

En cumplimiento del tercer objetivo específico, orientado a desarrollar y evaluar modelos de clasificación supervisada para la identificación de casos de riesgo, se comprobó que los modelos basados en técnicas de ensamble, particularmente Gradient Boosting, ofrecen un mejor equilibrio entre capacidad de detección de la clase positiva y control de errores en la clase mayoritaria. La aplicación de técnicas de balanceo como SMOTE y esquemas de validación cruzada estratificada permitió enfrentar de manera adecuada el desbalance inherente al problema, garantizando resultados consistentes y una capacidad de generalización aceptable en conjuntos de datos no vistos.

Finalmente, en relación con el cuarto objetivo específico, enfocado en evaluar la aplicabilidad práctica del modelo como herramienta de apoyo a la supervisión y control de la contratación pública, el estudio concluye que los resultados obtenidos respaldan su uso como un mecanismo de priorización y alerta temprana. La inclusión de métricas estándar, matrices de confusión, curvas ROC, Precision–Recall y curvas de aprendizaje proporciona insumos claros para la interpretación del desempeño del modelo y la toma de decisiones informadas. No obstante, se resalta que el modelo no constituye un instrumento sancionatorio, sino un apoyo técnico para orientar acciones de vigilancia, respetando la presunción de inocencia y el marco normativo vigente.

En conjunto, las conclusiones confirman que el cumplimiento de los objetivos específicos permitió validar la pertinencia de la ciencia de datos como herramienta para fortalecer la transparencia, el

control y la gestión del riesgo en la contratación pública, ofreciendo una base técnica sólida para su posible implementación y futuras mejoras en contextos institucionales.

7.2 TRABAJOS FUTUROS

A partir de los hallazgos obtenidos y del análisis exhaustivo del desempeño del modelo, se identifican oportunidades significativas para ampliar la investigación y profundizar en el entendimiento de los fenómenos asociados al riesgo y la corrupción en la contratación pública. Estos desarrollos futuros permitirán no solo mejorar la capacidad predictiva de los modelos, sino también explorar el impacto económico y social de la corrupción, así como su relación con el crecimiento económico sostenible y la gobernanza institucional.

1. Expansión del análisis a escenarios dinámicos y heterogéneos: Se propone incorporar datos adicionales con dimensiones temporales y geográficas que permitan evaluar el comportamiento del modelo en contextos más complejos, mejorando su robustez y capacidad predictiva.
2. Exploración de modelos avanzados: Se plantea evaluar técnicas de ensamblado, boosting y redes neuronales profundas para capturar patrones no lineales complejos y fortalecer la capacidad del sistema de identificar riesgos emergentes en la contratación pública.
3. Análisis del impacto económico y social de la corrupción: En investigaciones futuras, se espera explorar la relación entre los patrones de corrupción detectados y su influencia sobre el crecimiento económico sostenible, considerando variables macroeconómicas, desarrollo regional y bienestar social. Este enfoque permitirá comprender no solo la presencia de riesgos, sino sus consecuencias sistémicas sobre la economía y la sociedad.
4. Integración con políticas públicas y estrategias de control fiscal: Los modelos predictivos y los indicadores derivados del análisis de datos pueden servir como soporte para la formulación de políticas de prevención, fortalecimiento institucional y diseño de estrategias de control proactivo, promoviendo transparencia y eficiencia en la gestión pública.
5. Investigación interdisciplinaria: Se recomienda fomentar colaboraciones entre ciencias de datos, economía, derecho y ciencias sociales, con el fin de generar un enfoque integral que permita comprender las causas, efectos y posibles intervenciones ante los fenómenos de corrupción, potenciando el impacto de las decisiones basadas en evidencia.
6. Desarrollo de un observatorio estratégico: A futuro, se proyecta la creación de un observatorio de riesgo que combine modelos predictivos, análisis de tendencias y visualización de datos, con el objetivo de brindar soporte a la toma de decisiones y facilitar la evaluación continua de riesgos en contratación pública, contribuyendo al fortalecimiento de la transparencia institucional.

8. REFERENCIAS BIBLIOGRÁFICAS

- [1] Transparencia por Colombia, *Contratación pública: plataformas para acceder a la información*. [En línea]. Disponible en: https://transparenciacolombia.org.co/Documentos/Publicaciones/cuidado-paz/caqueta/Contratacion_p%C3%BAblica_plataformas_acceder_informacion.pdf. [Accedido: 30-nov-2024].
- [2] Centro de Estudios Anticorrupción, *Corrupción administrativa y política en La Guajira*. [En línea]. Disponible en: <https://www.estudiosanticorrupcion.org/corrupcion-administrativa-y-politica-en-la-guajira/>. [Accedido: 30-nov-2024].
- [3] Kahn, F., & Saavedra, A. (2021). *Co_Eco Diciembre 2021*. Fundación para la Educación Superior y el Desarrollo. Recuperado de [https://repository.fedesarrollo.org.co/bitstream/handle/11445/4213/Co_Eco Diciembre 2021 Kahn y Saavedra.pdf?sequence=1&isAllowed=y](https://repository.fedesarrollo.org.co/bitstream/handle/11445/4213/Co_Eco_Diciembre_2021_Kahn_y_Saavedra.pdf?sequence=1&isAllowed=y)
- [4] U. Compensar, "Repositorio CRAI," *Repositorio CRAI - Corporación Universitaria Compensar*. [En línea]. Disponible en: <https://repositoriocrai.ucompensar.edu.co/entities/publication/b8671856-0317-41eb-b902-3d1b45591723>. [Accedido: 30-nov-2024].
- [5] Transparency International, "Corruption Perceptions Index 2022," Transparency International, 2022. [En línea]. Disponible en: <https://www.transparency.org/en/cpi>. [Accedido: 25-nov-2024].
- [6] Colombia Compra Eficiente, "Datos Abiertos de Contratación Pública," Colombia Compra Eficiente, 2023. [En línea]. Disponible en: <https://www.colombiacompra.gov.co/>. [Accedido: 25-nov-2024].
- [7] C. D. Manning, P. Raghavan y H. Schütze, *Introduction to Information Retrieval*. Cambridge, Reino Unido: Cambridge University Press, 2008.
- [8] B. Wang, "Employing Supervised Classification Algorithms for Diabetes Prediction," *Master's Thesis*, California State University San Marcos, San Marcos, CA, USA, Apr. 2023. [En línea]. Disponible en: <https://scholarworks.calstate.edu/downloads/v405sh95f>.

- [9] R. Hassan y M. R. Islam, "Detection of fake online reviews using semi-supervised and supervised learning," en *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox's Bazar, Bangladesh, 7-9 Feb. 2019, pp. 1-5. [En línea]. Disponible en: [\(99+\) Detection of fake online reviews using semi-supervised and supervised learning | JP INFOTECH PROJECTS - Academia.edu](#).
- [10] A. Bakumenko y A. Elragal, "Detecting anomalies in financial data using machine learning algorithms," *Systems*, vol. 10, no. 5, p. 130, 2022. [En línea]. Disponible en: <https://doi.org/10.3390/systems10050130>.
- [11] T. Hastie, R. Tibshirani y J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2ª ed. Nueva York, EE. UU.: Springer, 2009. [En línea]. Disponible en: [T. Hastie, R. Tibshirani, J. Friedman the Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition - DocsLib](#).
- [12] C. C. Aggarwal y C. Zhai, *Mining Text Data*. Nueva York, EE. UU.: Springer, 2012. [En línea]. Disponible en: [Mining text data : Aggarwal, Charu C : Free Download, Borrow, and Streaming : Internet Archive](#).
- [13] V. Borge y R. Singh, "Analyzing Public Procurement Data for Corruption Detection: A Case Study from India," *Journal of Data Science*, vol. 18, no. 3, pp. 345–360, 2020.
- [14] Instituto de Transparencia Pública, "Open Data for Transparency: Public Procurement Analysis in Brazil," 2020.
- [15] A. Rodríguez, J. Martínez y P. López, "Machine Learning Models for Public Procurement Analysis in Latin America," *Data Science in Governance*, vol. 9, no. 2, pp. 112–130, 2021.
- [16] C. Restrepo, M. Gómez y L. Vargas, "Open Data and Corruption Monitoring in Colombia's Public Procurement," *Latin American Journal of Data Governance*, vol. 5, no. 1, pp. 87–105, 2023.
- [17] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- [18] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" *Explaining the Predictions of Any Classifier*.
- [19] Molnar, C. (2022). *Interpretable Machine Learning*.

9. ANEXOS

Los anexos que respaldan y complementan la presente investigación se encuentran organizados en una carpeta adjunta al proyecto, la cual contiene los archivos correspondientes a cada análisis, integración de datos y documentación metodológica. Para facilitar su identificación y consulta, a continuación, se presenta una tabla que relaciona el número de cada anexo con el nombre del archivo correspondiente, permitiendo al lector acceder de manera estructurada a los materiales complementarios que apoyan los resultados y procedimientos descritos en los capítulos anteriores.

Dichos anexos se encuentran en el siguiente enlace: [ANEXOS PROYECTO](#)

Tabla No. 13
Listado de Anexos

NUMERO DE ANEXO	NOMBRE DEL ARCHIVO
1	exploración_Secop_integrado.py
2	exploración_adiciones.py
3	exploración_garantias_rechazadas.py
4	exploracion_multas_secop.py
5	exploración_rues.py
6	exploración_sibor.py
7	exploración_sigep.py
8	exploración_aportantes_cam_polit.py
9	integracion_3_py
10	VF_1_secop_integrado_adiciones.csv
11	VF_2_secop_integrado_adiciones_garantias.csv
12	VF_2.1_secop_integrado_adiciones_garantias_enriquecido.csv
13	VF_2.2_secop_integrado_consolidado_contratistas.csv
14	VF_2.3_secop_integrado_consolidado_contratistas.csv
15	VF_3_secop_integrado_contratistas_multas_limpio.csv
16	VF_4_secop_mas_rues.csv
17	VF_4.1_secop_mas_rues_representantes.csv
18	VF_4.2_secop_contratistas_y_representantes.csv
19	VF_4.3_secop_adiciones_multas_rues.csv
20	VF_4.4_secop_adiciones_multas_rues.csv
21	VF_5_secop_integrado_rues_sibor.csv
22	VF_5.1_secop_integrado_rues_sibor.csv

NUMERO DE ANEXO	NOMBRE DEL ARCHIVO
23	VF_6_secop_integrado_rues_sibor_sigep.csv
24	VF_6.1_secop_integrado_rues_sibor_sigep_limpio.csv
25	VF_6.2_secop_integrado_rues_sibor_sigep_limpio.csv
26	VF_7_secop_integrado_rues_sibor_sigep_aportantes.csv
27	VF_7.1_secop_integrado_rues_sibor_sigep_aportantes.csv
28	VF_7.2_secop_integrado_rues_sibor_sigep_aportantes.csv
29	VF_8_secop_noticias_doblemerge.csv
30	VF_8.1_secop_noticias_resumen.csv
31	VF_VERSION_FINAL.csv
32	ANTECEDENTE_RIESGO.py
33	modelos.py
34	informe_desempeno_NOTICIAS.txt

Elaborado por: NJLM

Esta organización asegura la trazabilidad de la información y facilita la revisión de los datos y procedimientos utilizados a lo largo del proyecto aplicado.