Aplicación de ciencia de datos para predecir el éxito de la ejecución de los contratos públicos en Colombia

Javier Andres Arias Sanabria
Código 8971773

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

Director(a)
María Constanza Pabón Burbano

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 1 DE 2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

En este trabajo se desarrolló un modelo de clasificación binaria para predecir el éxito de los procesos de contratación pública en Colombia. El objetivo es categorizar cada proceso en dos estados principales: 'exitoso' (aquellos que han sido "Liquidados") y 'no exitoso' (una agrupación de los estados "Descartado", "Terminado anormalmente después de convocado" y "Terminado sin liquidar").

Este es un problema de gran relevancia para el país, ya que la eficiencia y transparencia en la gestión de los contratos públicos es una de las mayores preocupaciones del sector. La propuesta se fundamenta en el marco del gobierno abierto, que promueve la publicación de información gubernamental para que sea transformada en valor público. La disponibilidad de datos abiertos permite ir más allá de la simple transparencia, habilitando el control ciudadano y la utilización de la ciencia de datos para detectar de forma temprana banderas rojas o indicadores de riesgo de corrupción, ineficiencia o colusión en los procesos.

La información utilizada se obtiene del portal de datos abiertos de Colombia [1], específicamente de la plataforma SECOP I.

El trabajo se desarrolló en varias etapas:

Análisis y Preparación de Datos: Se realizó una exhaustiva ingeniería de características para pre-procesar las variables y transformar los datos brutos en un formato más adecuado para su procesamiento. Esto incluye el uso de Target Encoding para gestionar variables de alta cardinalidad (geográficas y administrativas), capturando la señal de riesgo histórica. Asimismo, se implementó la vectorización del campo de texto "Objeto a Contratar" utilizando la técnica TF-IDF, permitiendo capturar la importancia semántica de los términos en el contexto de la contratación.

Manejo del Desbalance de Clases: Se abordó el desequilibrio de clases, dado que los contratos 'no exitosos' son significativamente menos comunes. Para ello, se optó por un enfoque de balanceo algorítmico mediante el ajuste de pesos de clase (Class Weighting). Esta técnica permite penalizar de mayor forma el error en la clase minoritaria durante el entrenamiento, logrando un modelo sensible al riesgo sin la necesidad de generar datos sintéticos.

Modelado y Comparación: Se implementaron y compararon cinco modelos de clasificación: Regresión Logística, Árbol de Decisión, Naive Bayes, Random Forest y LightGBM (Gradient Boosting), buscando el equilibrio entre precisión predictiva y eficiencia computacional.

Evaluación y Explicabilidad: El rendimiento se evaluó mediante métricas robustas para datos desbalanceados como el AUC-ROC y el F1-Score, priorizando el Recall de la clase minoritaria para minimizar los Falsos Negativos. Además, se integraron técnicas de interpretabilidad para identificar los factores que más influyen en el éxito o el fracaso de la contratación.

1 DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

La ejecución satisfactoria de los contratos en el sector público representa uno de los desafíos administrativos más persistentes del país. Históricamente, la gestión de la contratación estatal ha evidenciado dificultades para garantizar el cumplimiento de los objetos contractuales, una problemática que impacta directamente en la eficiencia del Estado y que tradicionalmente ha carecido de herramientas tecnológicas predictivas para su mitigación.

Para abordar este flagelo, este proyecto propone aprovechar el creciente volumen de información pública disponible en los portales de contratación estatal (SECOP) y en el Portal de Datos Abiertos de Colombia. Mediante la aplicación de técnicas de ciencia de datos, se busca analizar el comportamiento histórico de las compras públicas con el propósito específico de desarrollar un modelo capaz de predecir la probabilidad de éxito de los contratos antes de que estos sean otorgados. De esta manera, se pretende dotar a las entidades de un mecanismo de alerta temprana que apoye la toma de decisiones basada en evidencia.

1.2 FORMULACIÓN DEL PROBLEMA

Por lo anterior, surge la siguiente pregunta:
¿Cómo utilizar la información pública del país para analizarla y predecir si un contrato puede terminar exitosamente o no?

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un modelo que utilice la información sobre contratación pública del país, la analice con técnicas de ciencias de datos y determine si finalizaran de forma exitosa.

2.2 OBJETIVOS ESPECÍFICOS

- Pre-procesar los datos de los portales de información sobre contratación pública del país.
- Desarrollar distintos modelos de ciencia de datos que determinen si un contrato finalizará de forma exitosa o no.
- Evaluar los resultados de los modelos desarrollados.

3 MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

3.1.1 MARCO TEORICO DEL NEGOCIO: PROCESOS DE CONTRATACIÓN PÚBLICA EN COLOMBIA

El sistema de contratación pública en Colombia está regido principalmente por la Ley 80 de 1993, la Ley 1150 de 2007 y el Decreto 1082 de 2015. La gestión y la publicidad de estos procesos se realizan a través del Sistema Electrónico para la Contratación Pública (SECOP), una plataforma administrada por Colombia Compra Eficiente [2]. Los estados de un proceso, reflejados en el SECOP, indican la fase o el resultado de la gestión contractual.

El ciclo de vida de un contrato público se divide en tres fases principales:

Fase precontractual: Incluye la planeación, selección del contratista y la elaboración de documentos. En esta fase, los estados se centran en la convocatoria y la recepción de ofertas.

Fase contractual: Corresponde a la ejecución del objeto del contrato. El estado clave es la celebración.

Fase poscontractual: Se refiere al final del contrato, incluyendo su liquidación y cierre.

A continuación, en la tabla 1, se presenta un resumen de los estados más comunes que se pueden encontrar en el SECOP:

Estado	Significado y Contexto en el Proceso
BORRADOR	Estado inicial y provisional. Indica que el proceso o sus documentos (como el pliego de condiciones) están siendo creados por la entidad estatal, pero aún no han sido publicados oficialmente.
CONVOCADO	Es el primer estado oficial. La entidad estatal ha publicado los documentos del proceso (estudios previos, pliego de condiciones, etc.) para que los interesados puedan consultarlos y presentar sus ofertas.
PUBLICACIÓN PARA MANIFESTACIONES DE INTERÉS	Un estado específico de la fase precontractual en modalidades como el concurso de méritos. La entidad publica su intención de contratar y solicita a los potenciales proponentes que manifiesten su interés.
FINALIZADO EL PLAZO PARA MANIFESTACIONES DE INTERÉS	Marca el fin del período para que los interesados presenten sus manifestaciones. Es el paso siguiente a la publicación y precede a la conformación de la lista corta.
LISTA CORTA	No es un estado generalizado del proceso, pero se refiere a un sub-proceso, común en la contratación de consultores. Indica que la entidad ha preseleccionado un grupo reducido de proponentes con base en sus manifestaciones de interés para que continúen en el proceso.

DESCARTADO	El proponente o su oferta ha sido descalificado del proceso por no cumplir con los requisitos habilitantes (jurídicos, técnicos, financieros, etc.) establecidos en los pliegos de condiciones.
ADJUDICADO	Se ha seleccionado al proponente que presentó la oferta más favorable y objetiva, de acuerdo con los criterios establecidos. Es la culminación de la fase de selección.
CELEBRADO	El contrato ha sido perfeccionado, es decir, firmado por las partes y formalizado según los requisitos legales. Este estado marca el inicio de la fase de ejecución.
LIQUIDADO	El contrato ha finalizado su ejecución y se ha realizado el proceso de liquidación, en el que se define el balance final de derechos y obligaciones entre las partes. Es el estado ideal de un proceso exitoso.
TERMINADO ANORMALMENTE DESPUÉS DE CONVOCADO	El proceso de selección es cancelado por la entidad pública antes de la adjudicación. Esto puede ocurrir por razones de interés público, fuerza mayor o vicios en el proceso.
TERMINADO SIN LIQUIDAR	El contrato ha finalizado su ejecución o ha sido terminado unilateralmente, pero la liquidación no se ha realizado. Esto puede indicar un proceso de cierre pendiente o conflictivo.

Tabla 1. Estados de los procesos de contratación

Se resaltan los tres estados de interés para la definición de la variable objetivo binaria del modelo. Por una parte, el estado **Liquidado** representa el éxito contractual definitivo, indicando que el proceso cumplió con todas las etapas legales y administrativas de cierre [2].

Por otro lado, se agrupan tres estados que denotan fallas en distintas etapas del ciclo de vida público: **Descartado, Terminado anormalmente después de convocado y Terminado sin liquidar**. Si bien los dos primeros corresponden a la fase precontractual y el último a la fase poscontractual, su consolidación en una única clase de "no éxito" se justifica desde la perspectiva de la eficiencia administrativa y el riesgo fiscal [3].

Para la gestión pública, tanto un proceso que no logra adjudicarse como un contrato que finaliza sin el balance de derechos y obligaciones (liquidación), representan una ineficiencia en el uso de recursos del Estado y una interrupción en la entrega de valor público [4]. Esta agrupación permite que el modelo identifique patrones comunes de riesgo que impiden la culminación ideal de la compra pública en cualquiera de sus fases.

La Ciencia de Datos como Herramienta para la Gobernanza y la Transparencia

La gestión pública ha evolucionado en la medida que la tecnología ofrece más herramientas de gestión de datos, y esto ha permitido pasar de un enfoque reactivo o basado en la intuición a uno proactivo fundamentado en la evidencia [4].

La disponibilidad de datos abiertos, un principio promovido a nivel global y formalizado en Colombia por normativas como la Ley 1712 de 2014 [5], ha sido el catalizador de esta

transformación. Inicialmente, el gobierno abierto se limitaba a la simple publicación de información, con la transparencia como fin en sí misma. Sin embargo, en la actualidad, el enfoque se ha desplazado hacia la analítica de datos como una herramienta para generar valor público tangible, habilitando el control ciudadano y la creación de nuevos servicios.

El análisis predictivo, una subdisciplina de la ciencia de datos, permite ir más allá de la mera descripción de fenómenos para revelar patrones, anticipar comportamientos y evaluar el impacto de las políticas públicas. En el contexto de la contratación pública, esto se traduce en la capacidad de detectar tempranamente indicadores de riesgo de corrupción o ineficiencia [6]. Un modelo predictivo puede monitorear miles de procesos de manera eficiente, superando las limitaciones de los métodos de auditoría manuales [7].

Este proyecto se sitúa en el centro de esta evolución, utilizando la información del portal de datos abiertos de Colombia, específicamente el conjunto de datos de SECOP I. La aplicación de un modelo predictivo no solo genera un sistema de alerta temprana, sino que también contribuye al fortalecimiento de las capacidades institucionales al permitir a las entidades estatales concentrar sus recursos limitados en los casos que más lo ameritan, pasando de una revisión general a un monitoreo preventivo y focalizado.

El Problema de la Contratación No Exitosa como un Fenómeno de Riesgo

La identificación de procesos de contratación que no culminan de manera exitosa se conceptualiza como un problema de detección de anomalías y puntuación de riesgo [8]. La variable objetivo, ***Estado del proceso***, no es meramente descriptiva, sino que actúa como una etiqueta que refleja el resultado final de un fenómeno que ha sido sometido a una serie de condiciones. Para este estudio, se han agrupado los estados **Descartado**, **Terminado anormalmente después de convocado** y **Terminado sin liquidar** en una única clase denominada **'no exitoso'**, que representa la materialización de un riesgo en el proceso. Por su parte, el estado **Liquidado representa el resultado exitoso**.

La complejidad de este fenómeno radica en que los riesgos de corrupción o ineficiencia rara vez se manifiestan a través de un único indicador aislado. Más bien, son de naturaleza configuracional. Esto significa que el riesgo se revela a través de combinaciones recurrentes de condiciones institucionales, procedimentales y contextuales. Por ejemplo, una contratación directa no es inherentemente un riesgo, pero si se combina con una cuantía inusualmente alta o con una falta de ofertas de competidores, puede indicar un riesgo de colusión. Un modelo predictivo es la herramienta idónea para descubrir estas combinaciones complejas, superando las limitaciones de los métodos manuales de banderas rojas que se centran en indicadores individuales.

3.2 MARCO TEORICO TÉCNICO: ELEMENTOS ESTADISTICOS Y DE CIENCIAS DE DATOS

3.2.1 *Marco Conceptual del Aprendizaje Automático Supervisado*

La tarea de predecir el Estado del proceso de una contratación se enmarca en el paradigma del

aprendizaje supervisado, dado que se cuenta con etiquetas históricas que el modelo aprende a reconocer [9]. Esto se debe a que el conjunto de datos de SECOP I ya contiene las etiquetas (Liquidado, Descartado, etc.) que el modelo debe aprender a predecir a partir de las variables o características disponibles. Específicamente, este es un problema de clasificación binaria, donde se asigna cada instancia a una de dos categorías posibles: 'exitoso' o 'no exitoso' [10].

El problema de clasificación es fundamental en machine learning y contrasta con la regresión (predicción de un valor numérico, por ejemplo, el precio de un bien). En este caso, la variable dependiente es “Estado del proceso”, mientras que las variables independientes, que servirán para generar las predicciones, son las que seleccionaremos del conjunto de datos detallado del portal de datos abiertos.

3.2.2 *Análisis y Preparación del Conjunto de Datos*

La ingeniería de características (feature engineering), es la etapa donde los datos brutos se transforman en representaciones numéricas que maximizan la capacidad de aprendizaje de los algoritmos [11]. Este proceso constituye el puente entre la información cruda y el modelado predictivo, abordando cuatro pilares fundamentales:

- ***Tratamiento de Datos Faltantes:*** Consiste en gestionar la ausencia de información. Si bien existen técnicas de imputación estadística, la literatura sugiere que cuando la completitud de los datos es alta y la ausencia es aleatoria, la eliminación de registros (listwise deletion) es el método más seguro para evitar introducir sesgos artificiales en la distribución, preservando la pureza de la muestra original [12].
- ***Escalamiento de Variables Numéricas***: Es crucial para algoritmos sensibles a la magnitud de los datos. Técnicas como MinMaxScaler transforman los valores a un rango definido (frecuentemente [0, 1]), asegurando que variables con escalas grandes no dominen injustamente el cálculo de la función de pérdida del modelo [10].
- ***Codificación de Variables Categóricas***: Dado que los modelos matemáticos operan exclusivamente con números, los datos cualitativos deben ser transformados. Este pilar abarca desde representaciones binarias simples hasta mapeos complejos basados en el impacto de la categoría sobre la variable objetivo [9].
- ***Gestión de Ruido y Detección de Anomalías:*** Es fundamental distinguir entre "anomalías de riesgo" (patrones inusuales válidos) y "ruido estocástico" (errores de medición o captura). Según [13], el ruido extremo —como valores fuera de rango lógico o estadístico— debe ser eliminado mediante técnicas de filtrado o podado (trimming) para evitar que distorsione la frontera de decisión del modelo. Una vez saneado el ruido, los algoritmos pueden enfocarse en detectar las verdaderas anomalías que corresponden a comportamientos de riesgo en la contratación***.***

3.2.3 *Ingeniería de Características para Datos Tabulares*

La ingeniería de características permite capturar relaciones complejas mediante la transformación

de variables categóricas.

- ***Codificación de Baja Cardinalidad (One-Hot Encoding)***: Se utiliza para variables nominales donde no existe un orden jerárquico. Transforma cada categoría en una columna binaria independiente. Para evitar la multicolinealidad perfecta (conocida como la trampa de la variable ficticia), se debe omitir una de las columnas resultantes, la cual servirá como categoría de referencia [9].
- ***Codificación de Alta Cardinalidad (Target Encoding)***: Para variables con cientos de categorías, el OHE resulta ineficiente por la explosión dimensional. El Target Encoding reemplaza cada categoría con un valor escalar basado en la media de la variable objetivo para esa categoría. Para prevenir el sobreajuste (overfitting) y el filtrado de información (data leakage), se aplican técnicas de suavizado (smoothing) que combinan la media local con la media global del conjunto de datos [14].
- ***Interacciones de Variables***: La creación de nuevas variables mediante la combinación de características existentes (ej. el producto de una magnitud económica por un tipo de proceso) permite que el modelo identifique configuraciones de riesgo no lineales que las variables individuales no capturan por sí solas [10].

3.2.4 *Tratamiento de Datos de Texto No Estructurados*

Para procesar variables de texto, es necesario convertirlas en vectores numéricos. Este proceso comienza con la limpieza (eliminación de conectores o *stop-words*) y la normalización. Posteriormente, se aplican técnicas de vectorización:

- ***TF-IDF (Term Frequency - Inverse Document Frequency)***: Es un esquema de pesaje que evalúa la importancia de una palabra en un documento específico respecto a todo el corpus. A diferencia de un conteo simple, TF-IDF penaliza palabras muy comunes y resalta términos distintivos que poseen mayor carga semántica y capacidad discriminatoria para el modelo [11].

3.2.5 *Abordaje del Desequilibrio de Clases*

En problemas de detección de eventos raros, el desequilibrio de clases (donde la clase de interés es la minoría) puede generar modelos sesgados hacia la clase mayoritaria, incurriendo en el fenómeno conocido como la “paradoja de la exactitud” [15]. Se proponen dos enfoques principales:

a) Estrategias de Remuestreo:

- Sub-muestreo: Reduce la clase mayoritaria para igualar la proporción de la minoritaria, con el riesgo inherente de pérdida de información potencialmente valiosa para el modelo.
- Sobre-muestreo (SMOTE): La técnica Synthetic Minority Over-sampling Technique genera instancias sintéticas de la clase minoritaria mediante la interpolación entre

vecinos cercanos, mejorando la representación estadística sin incurrir en la simple duplicación de registros [16].

b) Aprendizaje Sensible al Costo (Class Weighting):

En lugar de alterar los datos, esta técnica modifica la función de costo del algoritmo. Se asigna un "peso" superior a los errores cometidos en la clase minoritaria durante el entrenamiento. Esto obliga al modelo a prestar más atención a los casos de riesgo, optimizando métricas como el Recall y el AUC-ROC en lugar de la exactitud global [17].

3.2.6 *Modelos de Aprendizaje Basados en Árboles de Decisión y Ensambles*

Para el desarrollo se seleccionaron cinco algoritmos representativos de diferentes paradigmas del aprendizaje automático supervisado:

1. Regresión Logística (Logistic Regression)

Se utiliza como modelo de referencia o baseline debido a su simplicidad y eficiencia en problemas de clasificación binaria. Estima la probabilidad de éxito mediante el uso de la función sigmoide, estableciendo una relación lineal entre las variables de entrada y el logaritmo de las probabilidades de la clase objetivo [10].

Hiperparámetros clave: C (parámetro de regularización inversa para controlar el sobreajuste) y penalty (norma de penalización L1 o L2).

2. Árbol de Decisión (Decision Tree)

Es un modelo no paramétrico que particiona el espacio de los datos en hiperrectángulos mediante reglas lógicas de decisión. Permite identificar las interacciones iniciales entre las variables de riesgo de manera visual e interpretable [9].

Hiperparámetros clave: max_depth (profundidad máxima del árbol) y min_samples_split (mínimo de muestras para dividir un nodo).

3. Naive Bayes

Este modelo probabilístico se basa en el teorema de Bayes. Si bien es popular para conteo de palabras, su variante Gaussian Naive Bayes es la idónea cuando las variables predictoras (como los componentes principales del análisis de texto) tienen valores continuos y se asume una distribución normal [11].

Hiperparámetros clave: var_smoothing (parámetro para añadir estabilidad al cálculo de probabilidades en la variante Gaussiana)..

4. Random Forest (Bosques Aleatorios)

Es un algoritmo de ensamble basado en la técnica de bagging que construye múltiples árboles de decisión de forma independiente. Su fortaleza radica en reducir la varianza y mejorar la robustez frente al ruido presente en los datos [9].

Hiperparámetros clave: n_estimators (número de árboles) y max_features (cantidad de variables consideradas en cada partición).

5. Gradient Boosting (LightGBM)

Es una implementación de Gradient Boosting Decision Tree optimizada para el alto rendimiento y la eficiencia en el uso de memoria. A diferencia de los métodos de ensamble secuenciales tradicionales que expanden los árboles nivel por nivel (level-wise), este algoritmo utiliza un crecimiento por hoja (leaf-wise), seleccionando la hoja con la mayor reducción de pérdida para realizar la partición. Esta característica permite una convergencia más rápida y una precisión superior al procesar grandes volúmenes de datos con características heterogéneas [18].

Hiperparámetros clave:

num_leaves: controla la complejidad del modelo y el número de terminales en el árbol.

learning_rate: determina el impacto de cada nuevo árbol en la corrección de errores de los anteriores.

feature_fraction: parámetro de submuestreo que selecciona un porcentaje de variables al azar en cada iteración para evitar el sobreajuste.

3.3 ANTECEDENTES

La aplicación de técnicas de ciencia de datos para la fiscalización y el control de la contratación pública es un campo emergente que ha ganado tracción global gracias a las políticas de datos abiertos. A continuación, se analizan investigaciones y marcos de referencia que sientan las bases metodológicas para este proyecto.

3.3.1 Anomaly Detection in Public Procurements using the Open Contracting Data Standard

Un referente técnico relevante es el estudio desarrollado por [8], quienes analizaron procesos de contratación en Paraguay bajo el estándar OCDS. En esta investigación, se trabajó con un conjunto de datos que abarcó registros históricos de aproximadamente diez años, procesando miles de transacciones para identificar irregularidades. Mediante el algoritmo Isolation Forest, lograron detectar patrones sospechosos de colusión y sobrecostos, obteniendo una precisión que permitió validar alertas en casos reales de auditoría.

Relación con este proyecto: A diferencia de [8], que se enfocan en la detección de anomalías (casos extraños sin etiqueta previa), este proyecto utiliza aprendizaje supervisado sobre el histórico del SECOP I. Esto permite no solo detectar lo "anómalo", sino predecir específicamente el "fracaso" basado en patrones de resultados ya conocidos, lo que proporciona una herramienta de alerta temprana con un objetivo de negocio más definido.

3.3.2 Using Machine Learning For Anti-Corruption Risk And Compliance.

Desde una perspectiva corporativa y de cumplimiento, el informe de la [7] documenta casos de éxito en empresas globales como Microsoft y AB InBev. El estudio subraya que el despliegue de Inteligencia Artificial Estrecha (ANI) permite reducir sustancialmente los costos de auditoría y mejorar la eficacia de los programas de cumplimiento. Los resultados reportados indican una reducción de hasta el 50% en el tiempo de revisión manual de contratos y una mejora del 30% en la detección de banderas rojas operativas.

Relación con este proyecto: El informe resalta una barrera crítica, como es el alto costo de implementación para organizaciones sin recursos masivos. Este proyecto aborda esa brecha al demostrar que, utilizando herramientas de código abierto (Python, LightGBM) y datos públicos, es posible desarrollar soluciones de alto rendimiento sin requerir infraestructuras de cómputo inalcanzables.

3.3.3 Predicción de ineficiencias en la contratación pública de Bogotá.

Un antecedente fundamental a nivel nacional es el trabajo de [3], quien desarrolló un modelo predictivo utilizando datos de la plataforma SECOP II para identificar contratos con riesgo de presentar anomalías operativas. En esta investigación, se analizó un conjunto de datos compuesto por más de 130,000 registros de contratación del Distrito Capital, abarcando un periodo de ejecución de tres años.

Mediante la implementación de algoritmos de aprendizaje automático, el autor logró una precisión superior al 90% en la detección de ineficiencias específicas, tales como prórrogas injustificadas y sobrecostos en la ejecución. Los resultados demostraron que variables como la cuantía del contrato y la modalidad de selección son predictores críticos de la estabilidad contractual en el contexto colombiano.

Relación con este proyecto: El estudio de [3] valida la importancia de utilizar los datos abiertos como insumo para herramientas prácticas de supervisión gubernamental. Mientras que dicho antecedente se enfoca exclusivamente en la capital (Bogotá) y en predecir anomalías de proceso (prórrogas o costos), esta investigación amplía el alcance analítico a nivel nacional mediante el uso del histórico de SECOP I. Además, este proyecto integra el procesamiento de lenguaje natural (TF-IDF) para el análisis semántico de los objetos contractuales, extendiendo el estudio predictivo hacia el resultado final de la contratación (éxito o fracaso en la liquidación).

4 AMBIENTE DE DESARROLLO Y PLATAFORMA TECNOLÓGICA

Para garantizar la reproducibilidad, eficiencia y escalabilidad del análisis, se ha diseñado un ecosistema de desarrollo basado en herramientas de código abierto líderes en la industria de la Ciencia de Datos.

4.1 Entorno de Desarrollo Integrado (IDE): Visual Studio Code

El desarrollo técnico se centraliza en Visual Studio Code (VS Code). La elección de este entorno se justifica por su arquitectura modular y su capacidad para integrar flujos de trabajo de ingeniería de software con análisis de datos. Entre las capacidades críticas aprovechadas para este proyecto se destacan:

- Integración Nativa de Jupyter: Permite la ejecución de archivos .ipynb, facilitando un enfoque de Programación Literaria, donde el código fuente, las visualizaciones y la documentación técnica coexisten en un único flujo de trabajo interactivo.
- Gestión de Entornos Virtuales: VS Code facilita el aislamiento de dependencias de Python, asegurando que las bibliotecas utilizadas (Pandas, Scikit-Learn, LightGBM) no entren en conflicto con otros recursos del sistema.
- Depuración de Código (Debugging): La capacidad de inspeccionar variables en tiempo real y establecer puntos de interrupción (breakpoints) fue fundamental para la depuración de las etapas complejas de ingeniería de características y vectorización TF-IDF.
- Extensibilidad Técnica: El uso de extensiones específicas como Pylance para el análisis estático de código y Jupyter Keymap para la agilidad en la manipulación de celdas, optimizó los tiempos de desarrollo.

4.2 Lenguaje de Programación y Ecosistema Python

Python fue seleccionado como el lenguaje núcleo debido a su madurez y su ecosistema especializado en aprendizaje automático. Las herramientas clave integradas en el ambiente de desarrollo incluyen:

- Jupyter Notebooks: Utilizados como la unidad principal de experimentación, permitiendo la iteración rápida necesaria en las fases de limpieza y modelado.
- Administración de Paquetes (Pip/Conda): Empleados para la gestión rigurosa de versiones de las librerías, garantizando que los resultados obtenidos puedan ser replicados en diferentes máquinas locales sin discrepancias técnicas.

4.3 Hardware y Recursos de Cómputo

Considerando que el procesamiento se realiza en una estación de trabajo local, el ambiente de desarrollo fue configurado para maximizar el uso de la memoria RAM y el procesamiento multi-núcleo, requisitos críticos para el entrenamiento del modelo LightGBM y la gestión de la matriz de alta dimensionalidad generada por el proceso de TF-IDF.

A continuación, se detallan las especificaciones técnicas del equipo utilizado, lo cual garantiza la reproducibilidad del experimento en hardware de consumo estándar sin depender de clústeres en la nube:

- Procesador: Intel Core i7 (Arquitectura x64).
- Memoria RAM: 16 GB DDR4.
- Almacenamiento: Unidad de Estado Sólido (SSD) para optimizar la lectura/escritura de los archivos Parquet.
- Sistema Operativo: Windows 11.

Eficiencia Operativa y Tiempo de Ejecución Bajo esta configuración de hardware, el tiempo medio total de ejecución del pipeline completo (end-to-end) fue de 48 minutos.

Este cronometraje incluye la ingesta de los 8.6 GB de datos crudos, la limpieza, la ingeniería de características y el entrenamiento secuencial de los cinco modelos.

Esta métrica valida que la solución propuesta es computacionalmente ligera y escalable, permitiendo su despliegue en entidades públicas con recursos tecnológicos limitados.

5 PRE - PROCESAMIENTO DE DATOS

5.1 OBTENCIÓN DE LOS DATOS

Los datos fueron extraídos del portal de datos abiertos de Colombia, específicamente del conjunto "SECOP I - Procesos de Compra Pública" [1]. La muestra obtenida con corte a julio de 2025 consta de 6.129.496 registros y un peso en disco de 8.6 GB.
Esta base de datos abarca la información histórica de contrataciones registradas desde el año 2011 hasta mediados de 2025. Es imperativo destacar que la selección de este marco temporal no obedece a un muestreo arbitrario, sino que representa el universo poblacional completo de los registros históricos disponibles en el Portal de Datos Abiertos para la plataforma SECOP I.
Al utilizar la totalidad de la data existente desde la incepción del sistema hasta la fecha de extracción, se elimina cualquier sesgo de selección temporal, permitiendo que el modelo capture las tendencias estructurales reales de los últimos 14 años en el comportamiento de la contratación estatal en Colombia; es decir, abarca la totalidad del periodo del cual se tiene registro oficial en SECOP I.

Dado que este volumen excede la capacidad de memoria RAM convencional de una estación de trabajo local, se implementó una estrategia de carga por fragmentos (Chunks). Esta técnica permite realizar una lectura iterativa del archivo CSV, procesando bloques de datos (ej. 50.000 filas a la vez) para optimizar el uso de recursos.

Inicialmente, se realiza una carga exploratoria de la totalidad de los registros para mapear los estados presentes en el SECOP I, cuyo conteo se aprecia en la tabla 2. Este paso es fundamental para distinguir aquellos estados que representan un desenlace definitivo del contrato (exitoso o fallido) de aquellos que corresponden a etapas administrativas o procesos aún en curso, permitiendo así la posterior definición de la variable objetivo.

Estado del Proceso	Conteo Absoluto	(%)
TOTALES	**6.129.496**	**100**
CELEBRADO	3.691.287	60,221
LIQUIDADO	1.709.530	27,89
CONVOCADO	432.391	7,054
TERMINADO SIN LIQUIDAR	188.606	3,077
TERMINADO ANORMALMENTE DESPUeS DE CONVOCADO	79.598	1,298
ADJUDICADO	17.943	0,293
DESCARTADO	7.471	0,122
BORRADOR	2.539	0,041
FINALIZADO EL PLAZO PARA MANIFESTACIONES DE INTERes	57	0,001

PUBLICACIoN PARA MANIFESTACIONES DE INTEReS	53	0,001
EXPRESIoN DE INTEReS	15	0
LISTA CORTA	6	0

Tabla 2. Conteo de registros por estado en el cargue inicial

5.2 SELECCIÓN DE ATRIBUTOS Y CARACTERISTICAS

Para la construcción del dataset final, se aplicó un doble criterio de selección: primero, una depuración vertical (selección de atributos) basada en la calidad y relevancia predictiva; y segundo, una depuración horizontal (definición de la población) basada en el ciclo de vida contractual, siguiendo las metodologías estándar de reducción de datos [12].

5.2.1 Selección de Atributos (Variables)

La base de datos original contiene 76 atributos. Para la selección de las variables finales, se realizó un análisis técnico basado en el Diccionario de Datos de SECOP I (ANEXO 1. DICCIONARIO DE DATOS ABIERTOS SECOP I).

Criterios de Descarte de Variables:

- Irrelevancia Predictiva: Variables como UID, NIT de la Entidad, y códigos internos de procesos que funcionan como identificadores únicos sin valor estadístico.
- Fuga de Información (Data Leakage): Se eliminaron variables que se generan después de que se ha iniciado el contrato, como Fecha de Liquidación, Valor Total de Adiciones o Marcación de Adiciones. Incluirlas causaría que el modelo "adivine" el resultado por razones obvias, invalidando su capacidad preventiva.
  Esta exclusión es fundamental para evitar el sobreajuste por fuga de información (data leakage), asegurando que el modelo realice una inferencia ex-ante [9].
- Alta Redundancia: se prefirieron las versiones numéricas (ID) sobre las descripciones textuales. Es importante resaltar que estos identificadores se trataron técnicamente como datos categóricos para evitar que el modelo les asigne una jerarquía numérica inexistente, asegurando así que su representación matemática sea coherente con su naturaleza de etiqueta. Adicionalmente, la variable original “ID Régimen de Contratación” que se había considerado inicialmente como importante para el propósito del proyecto fue excluida tras identificar una redundancia del 100% (correlación perfecta) con “ID Modalidad” en la fuente de datos. Esta decisión previene la multicolinealidad y optimiza la eficiencia del modelo, reduciendo la dimensionalidad sin perder capacidad explicativa [11].

A continuación, en la tabla 3 se presentan las 9 variables finales para trabajar en este proyecto:

Variable	Descripción	Tipo de Dato	Técnica de Preprocesamiento	Justificación Técnica
1. ID Modalidad	Código que identifica la modalidad de selección (ej. Licitación, Directa).	Categórica Nominal	**Codificación One-Hot**	Al no existir un orden jerárquico numérico entre modalidades, se binariza la variable para evitar que el modelo asuma una falsa relación de magnitud.
2. Cuantía Contrato	Valor de firma del contrato en pesos (COP).	Cuantitativa	**Escalamiento Min-Max**	Normaliza los montos al rango [0,1], evitando que las cifras millonarias dominen la función de pérdida y facilitando la convergencia del algoritmo.
3. Plazo de Ejec del Contrato	Valor sobre el cual se determina la duración del contrato	Cuantitativa	**Ingeniería de Características (Fusión)**	Se combina con la variable *Rango* para calcular una nueva variable unificada (duracion_dias) y posteriormente se escala (Min-Max).
4. Rango de Ejec del Contrato	Unidad en la que se define el plazo, pueden ser días, meses o años	Categórica Nominal	**Ingeniería de Características (Fusión)**	Actúa como factor multiplicador para estandarizar la temporalidad. Una vez calculada la duración real en días, esta variable se elimina para evitar redundancia.
5. Objeto a Contratar	Texto libre con detalles del proceso.	Texto no estructurado	**Vectorización TF-IDF**	Transforma las descripciones en vectores numéricos que ponderan la relevancia semántica de palabras clave (ej. "SUMINISTRO"), capturando riesgos ocultos en el texto.
6. Tipo De Contrato	Categoría del servicio (ej. Obra, Consultoría).	Categórica Nominal	**Codificación One-Hot**	Transforma las categorías en variables binarias independientes, permitiendo al modelo aislar el riesgo específico de cada tipología.
7. Orden Entidad	Nivel administrativo (Nacional, Territorial).	Categórica Nominal	**Codificación One-Hot**	Permite identificar patrones de riesgo asociados a la autonomía y jerarquía

				administrativa de la entidad contratante.
8. Municipio Entidad	Ubicación geográfica específica (+1,100 municipios).	Categórica (Alta Cardinalidad)	**Target Encoding (con Suavizado)**	En lugar de generar miles de columnas binarias, se reemplaza el municipio por su **probabilidad histórica de riesgo**, aplicando un suavizado estadístico para evitar sesgos en poblaciones pequeñas.
9. Departamento Entidad	División política superior (32 departamentos).	Categórica Nominal	**Target Encoding (con Suavizado)**	Se aplica la misma lógica de riesgo probabilístico que en municipios, permitiendo al modelo capturar el "Riesgo Regional" como un valor numérico continuo.

Tabla 3. Variables seleccionadas

Nota sobre la Normalización de la Variable Temporal: Se identificó que en el SECOP I, la temporalidad de los contratos carece de una unidad de medida estandarizada, fragmentándose en dos campos: una magnitud numérica (Plazo) y un calificador de unidad (Rango: Días, Meses o Años).

Por tanto, la selección de la variable Rango de Ejec del Contrato fue indispensable como variable auxiliar de transformación. Aunque no se ingresa directamente al algoritmo final (para evitar redundancia dimensional), su extracción es pre-requisito para la fase de ingeniería de características, donde permite computar la métrica unificada duracion_dias. Esto garantiza que el modelo interprete correctamente la escala temporal del riesgo, equiparando, por ejemplo, un plazo de '30' en días frente a un '1' en meses.

Este conjunto de 9 variables es técnicamente sólido:

- ✓ Cubre la dimensión Económica (Cuantía).
- ✓ Cubre la dimensión Temporal (Plazo y Rango).
- ✓ Cubre la dimensión Legal (Tipo, Modalidad, Orden).
- ✓ Cubre la dimensión Geográfica (Municipio, Departamento).
- ✓ Cubre la dimensión Semántica (Objeto).

5.2.2 Definición de la Población Objetivo (Filtrado de Registros)

Para garantizar la viabilidad del modelo predictivo y su alineación con el problema de negocio, **se realizó un filtrado necesario** del conjunto de datos original. Esta etapa no responde únicamente

a una mejora en el rendimiento computacional, sino a la exigencia de entrenar el algoritmo exclusivamente con registros que presentan un desenlace administrativo definitivo y comparable, requisito indispensable para el aprendizaje supervisado [9].

En lugar de procesar la totalidad de los 6.129.496 registros, el modelo se enfocó en los contratos que permiten una clasificación binaria clara. Como se observó en la tabla 2, el estado predominante es "CELEBRADO" (60,22%). Sin embargo, este estado indica un contrato en ejecución, careciendo de un resultado final etiquetable. Al excluir estos registros y otros estados preliminares, se redujo el dataset en un 65% aproximadamente , un proceso alineado con las metodologías de selección de datos relevantes y reducción de ruido [12].

Esta delimitación es fundamental por dos razones: primero, permite que las técnicas de ingeniería de características, como el **TF-IDF** para el objeto del contrato y el **Target Encoding** para la ubicación geográfica, operen sobre patrones de éxito real y no sobre procesos inconclusos. Segundo, concentra el aprendizaje del modelo en la distinción crítica de negocio: identificar las variables que separan una ejecución contractual plenamente satisfactoria de una que fracasa en el cumplimiento de sus fines institucionales.

estado_del_proceso	Conteo	Porcentaje	Clase Predictiva
LIQUIDADO	1.709.530	86,11%	**0 (Negativo)**
TERMINADO SIN LIQUIDAR	188.606	9,50%	**1 (Positivo)**
TERMINADO ANORMALMENTE...	79.598	4,01%	**1 (Positivo)**
DESCARTADO	7.471	0,38%	**1 (Positivo)**
TOTAL	**1.985.205**	**100%**	-

Tabla 4. Distribución estados variable objetivo

Como se puede apreciar en la tabla 4, este filtrado inicial redujo el volumen de procesamiento de 6.1 millones de registros a aproximadamente 1.98 millones, enfocando este proyecto exclusivamente en los contratos con desenlace definitivo. Se puede observar que para el propósito de este proyecto que es la predicción de contratos con posibilidad de fracasar, se consideran estados de Positivo = 1 aquellos que fracasan.
No obstante, es imperativo precisar que, bajo la arquitectura de clasificación binaria supervisada, la predicción del fracaso conlleva matemáticamente la predicción del éxito:

$$P(\text{Éxito}) = 1 - P(\text{Fracaso}).$$

Por tanto, el modelo no actúa únicamente como un sistema de alertas restrictivo, sino también como una herramienta de facilitación administrativa. Al segregar con alta precisión los patrones de riesgo, el algoritmo valida implícitamente la idoneidad de los procesos clasificados como seguros, permitiendo a los entes de control agilizar el trámite de la contratación legítima y focalizar los recursos de auditoría humana exclusivamente donde la evidencia estadística sugiere

irregularidades.
El uso de archivos en formato Parquet para almacenar este resultado intermedio garantiza una lectura eficiente en las etapas posteriores de modelado, optimizando el flujo de trabajo de datos [10].

5.2.3 Análisis Exploratorio de Datos (EDA)

Una vez filtrado el conjunto de datos en los procesos con desenlace definitivo, se procedió a realizar un diagnóstico de las variables. Este paso es crítico para comprender la distribución subyacente de los datos y fundamentar las estrategias de transformación matemática que requieren los algoritmos de aprendizaje automático, siguiendo las fases estándar de comprensión de los datos [12].

1. Variables Cualitativas

- *Concentración Geográfica*:

Como se evidencia en la ilustración 1 de los Top 15 Departamentos, la contratación en Colombia no sigue una distribución uniforme. Se observa un comportamiento de "cola larga" (long tail) donde departamentos como Antioquia, Boyacá y Cundinamarca concentran un volumen transaccional masivo, superando incluso a Bogotá D.C. en cantidad de procesos registrados.

Esta disparidad se agudiza al descender al nivel municipal como se puede apreciar en la ilustración 2. La existencia de más de 1,100 categorías territoriales (municipios) genera una alta cardinalidad que hace inviable el uso de una codificación binaria simple (One-Hot Encoding), pues resultaría en una matriz dispersa (sparse matrix) altamente ineficiente. Esta evidencia visual justifica la implementación de Target Encoding, una técnica que permite condensar esta dispersión geográfica reemplazando el nombre del territorio por su probabilidad histórica de riesgo, capturando así la "cultura contractual" de la región sin explotar las dimensiones del modelo [14].

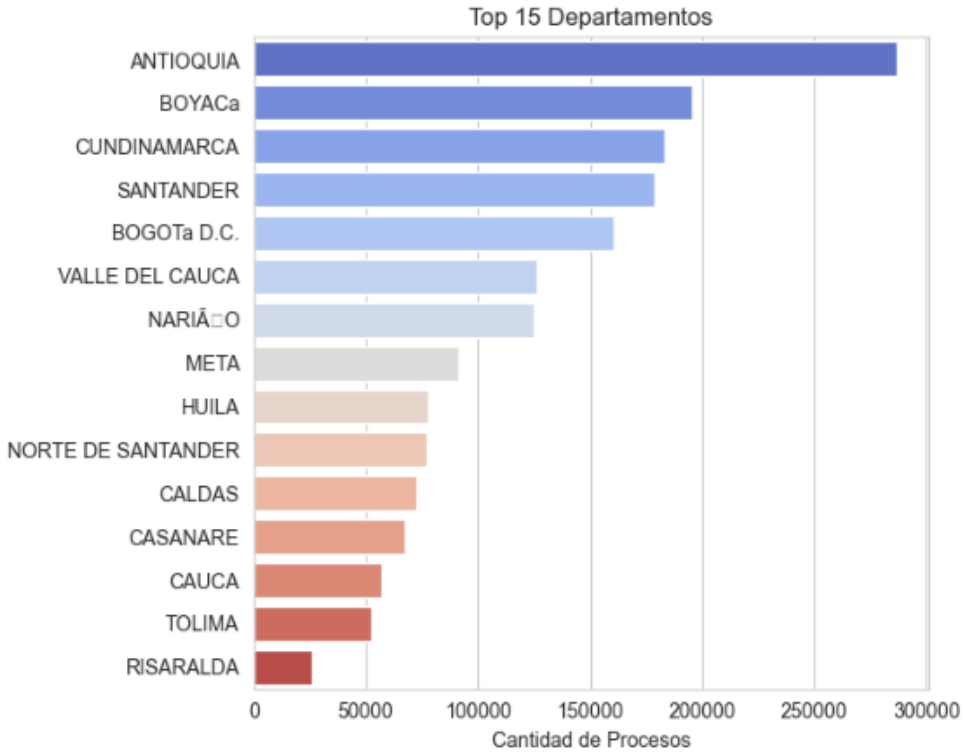

Ilustración 1. Top 15 cantidad de procesos contractuales por departamento

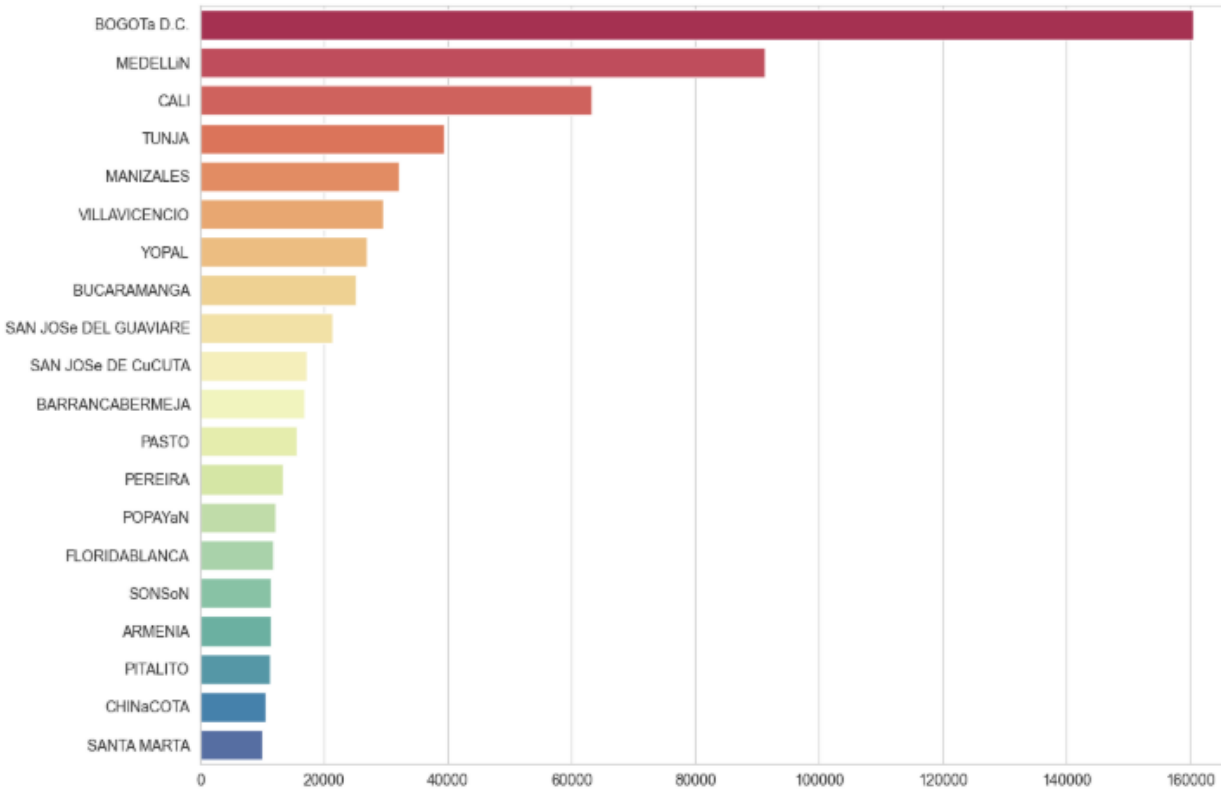

Ilustración 2. Top 20 cantidad de procesos contractuales por municipio

- *Tipología Contractual y Volumen Operativo*

Al observar los Tipos de Contrato, en la ilustración 3, se identifica una hegemonía de la Prestación de Servicios, con más de 1.4 millones de registros. Esta categoría supera por un amplio margen a contratos de mayor complejidad técnica como los de Obra o Suministro. Esto indica que el grueso de la "data" con la que entrenará el modelo corresponde a contratos de personal o apoyo a la gestión, los cuales suelen tener montos menores, pero frecuencias muy altas, configurando un perfil de riesgo basado en el volumen y la gestión administrativa más que en la complejidad ingenieril.

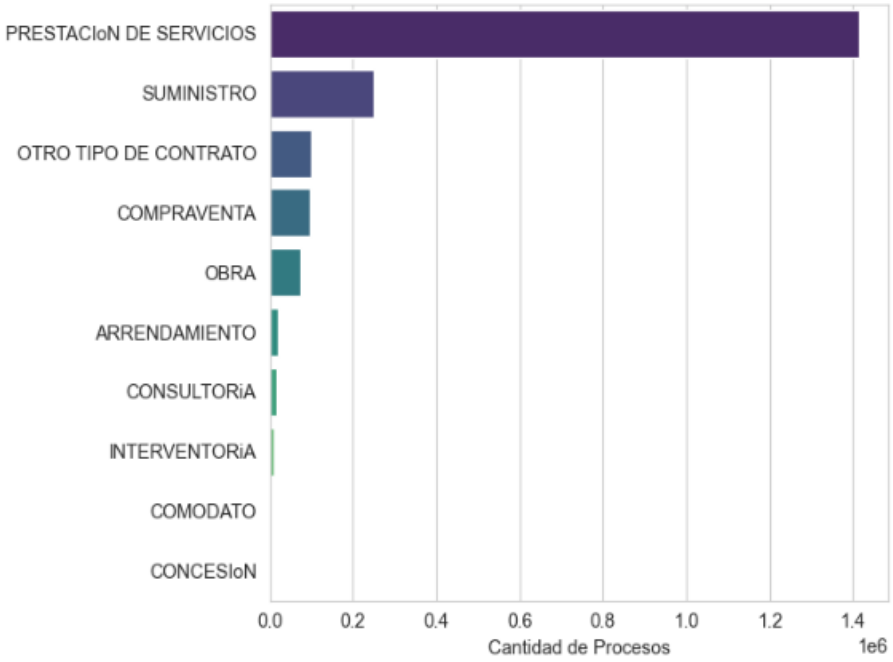

Ilustración 3. Cantidad de procesos contractuales por Tipo de contrato

- *Hegemonía de Modalidades*:

El análisis de las modalidades de selección, que se puede apreciar en la ilustración 4, expone una característica crítica del sistema: la predominancia absoluta de la Contratación Directa y el Régimen Especial.

Estas modalidades, que superan ampliamente a mecanismos competitivos como la Licitación Pública, sugieren escenarios de baja competencia y alta discrecionalidad administrativa. Esto implica que el algoritmo debe aprender a detectar patrones de riesgo intrínsecos en procesos donde la adjudicación es directa, modalidad que ha sido señalada en la literatura como un foco de atención para la detección de irregularidades [6].

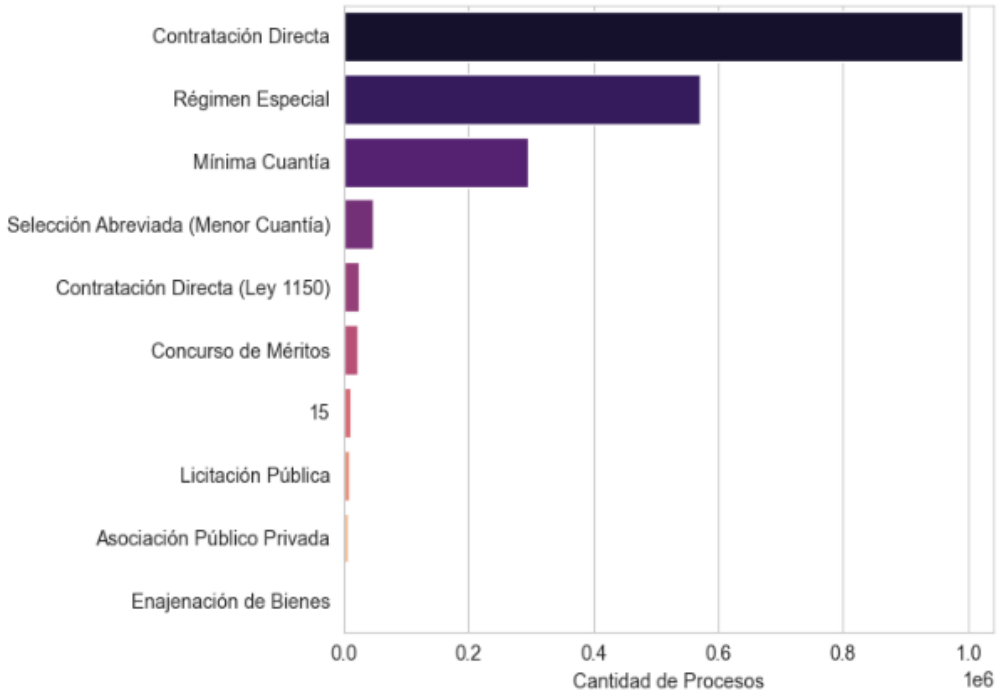

Ilustración 4. Cantidad de procesos contractuales por Modalidad de Contratación

2. Detección de Valores Atípicos y Problemas de Escala (Variables Cuantitativas)
   El diagnóstico visual de las variables numéricas mediante diagramas de caja (Boxplots) revela dos desafíos críticos de calidad de datos que justifican las intervenciones de limpieza posteriores:
   A. Disparidad Extrema en la Cuantía Económica

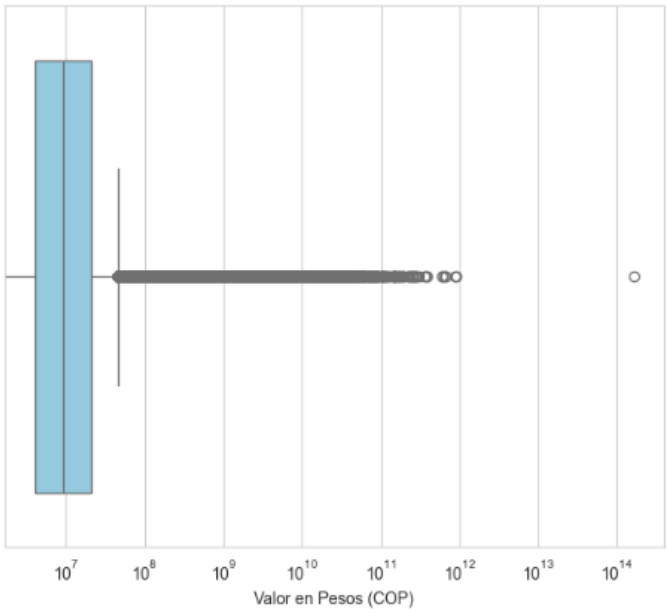

Ilustración 5. Boxplot del Valor de los Contratos en pesos

La distribución de los montos contractuales presentada en la ilustración 5, presenta un sesgo positivo severo que abarca desde los 10^7 (decenas de millones) hasta los 10^{14} (cientos de billones de pesos).

Diagnóstico: La presencia de registros aislados en el extremo derecho del gráfico (círculos lejanos a la caja) indica valores atípicos que distorsionan la escala global. Estos outliers masivos obligan a la aplicación de Recorte Estadístico (Clipping) y Escalamiento Logarítmico; de lo contrario, estos valores dominarían la función de pérdida del modelo, haciendo invisibles los patrones de los contratos de menor cuantía, tal como se recomienda en el análisis de valores atípicos para modelos predictivos [13].

B. Inconsistencia en la Unidad Temporal (Plazo)

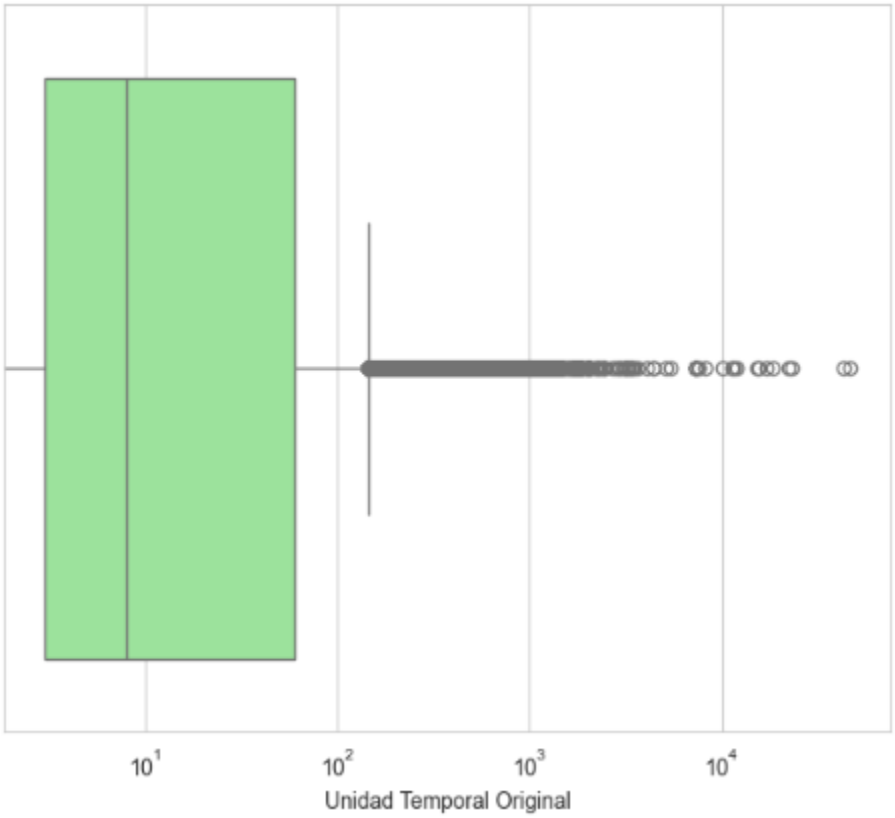

Ilustración 6. Boxplot plazo de ejecución

El diagrama del Plazo de Ejecución, presentado en la ilustración 6, evidencia la naturaleza heterogénea de la variable cruda. Se observa una concentración de datos en valores bajos ($<10^2$), correspondientes probablemente a meses o años) coexistiendo con valores altos ($>10^3$), correspondientes a días).

Justificación de Ingeniería de Características: Esta visualización confirma que la variable numérica "Plazo" por sí sola carece de significado estandarizado. Valida la necesidad operativa de fusionar esta columna con la variable categórica "Rango" (Días/Meses) para computar una nueva métrica unificada (duracion_dias), única forma de hacer comparables los registros antes de entrenar los algoritmos, aplicando técnicas de ingeniería de características para estandarización de datos [9].

5.3 PREPROCESAMIENTO TÉCNICO E INGENIERÍA DE CARACTERÍSTICAS

Con base en los hallazgos del análisis exploratorio (EDA), esta fase metodológica constituye la preparación y enriquecimiento de los datos. El proceso aborda con rigor el aseguramiento de la calidad (gestión de valores ausentes y atípicos) y profundiza en la ingeniería de características (*Feature Engineering*), necesaria para transformar variables categóricas complejas y texto no estructurado en representaciones vectoriales numéricas.

El objetivo central de esta etapa es maximizar el poder predictivo del conjunto de datos que garantice el desempeño óptimo de los algoritmos de aprendizaje supervisado que se implementarán en la fase de modelado, siguiendo las etapas estándar de preprocesamiento de datos [12].

5.3.1 Calidad y Gestión de Nulos

Se realizó una verificación exhaustiva de completitud en las variables seleccionadas.

- **Diagnóstico**: El análisis cuantitativo confirmó una completitud global del 100.0% en el dataset filtrado. No se encontraron celdas vacías en ninguno de los 1,985,205 registros procesados.
- **Justificación**: Este alto nivel de completitud se atribuye a las reglas de validación de la plataforma SECOP I, la cual exige el diligenciamiento obligatorio de campos como *Cuantía, Plazo y Modalidad* como requisito bloqueante para la publicación del proceso.

5.3.2 Consolidación de Variables Temporales (Reducción de Dimensionalidad)

Originalmente, la información de duración estaba dispersa en dos variables: 'Plazo de Ejec del Contrato' (numérica) y 'Rango de Ejec del Contrato' (categórica: Días, Meses, Años). Mantener ambas generaría problemas de redundancia, ya que la duración real es una combinación de las dos.

Para optimizar la estructura de los datos, se procedió a:

1. Calcular la variable unificada duracion_total_dias mediante la estandarización de unidades.
2. Eliminar las variables originales (Plazo y Rango) del conjunto de entrenamiento.

Justificación: Esta estrategia no solo reduce la dimensionalidad del dataset (evitando la creación de columnas binarias innecesarias para los rangos), sino que previene la multicolinealidad, asegurando que el modelo se enfoque exclusivamente en la magnitud temporal del contrato como factor de riesgo, una práctica esencial en la ingeniería de características [9].

5.3.3 Saneamiento de Valores Atípicos (Outliers)

Sobre las variables ya consolidadas, se inspeccionó la distribución estadística, detectando valores máximos inverosímiles producto de errores en la fuente.

- **Evidencia del Hallazgo**: Se identificó, por ejemplo, un contrato de prestación de servicios de transporte en el municipio de Tabio registrado por un valor de **$165.3 billones de pesos** (cifra superior al presupuesto de inversión de grandes ciudades capitales y que sería equivalente aproximadamente al 30% del presupuesto nacional). Este tipo de ruido estadístico distorsiona severamente las escalas de medición.
- **Acciones realizadas:**
  a) Filtro de Consistencia Lógica (Cota Inferior):
  Se identificaron y eliminaron registros con valor de *Cuantía o Duración* iguales a cero. Estos casos constituyen violaciones a la lógica de negocio de la contratación estatal, atribuyéndose a errores.
  b) Tratamiento de Valores Extremos (Cota Superior):
  Tras detectar anomalías severas, se aplicó un recorte (Clipping) basado en el percentil 99.9 (P99.9). Esta técnica estadística permitió excluir el 0.1% superior de la distribución en ambas dimensiones, eliminando el ruido provocado por errores de digitación y garantizando una muestra estadísticamente homogénea para el entrenamiento, tal como se recomienda en el análisis robusto de valores atípicos [13].

5.3.4 División Estratificada del Dataset

Para garantizar una evaluación objetiva de los modelos predictivos y evitar el sobreajuste (overfitting), se procedió a la partición del conjunto de datos procesado en dos subconjuntos disyuntos utilizando un muestreo estratificado.

- Estrategia de División. Se utilizó una proporción de 70/30, considerada óptima dada la gran volumetría del dataset (>1.8 millones de registros):
  - Conjunto de Entrenamiento (Train - 70%): Compuesto por exactamente 1.389.643 muestras (contratos). Este volumen masivo fue utilizado exclusivamente para el aprendizaje de los patrones, el cálculo de los mapas de riesgo (Target Encoding), la vectorización de texto y el ajuste de hiperparámetros de los modelos.
  - Conjunto de Prueba (Test - 30%): Compuesto por 595.562 muestras. Este conjunto fue reservado y aislado estrictamente hasta la fase final de evaluación, permitiendo simular con altísima confianza estadística el comportamiento del modelo frente a medio millón de contratos nuevos no vistos.
- Muestreo Estratificado (Stratified Sampling):

  Dado que el dataset presenta un desbalance de clases (donde los contratos "Positivos" representan la minoría), una división aleatoria simple podría generar subconjuntos donde la clase de interés esté subrepresentada. Para mitigar esto, se aplicó la técnica de estratificación, la cual fuerza a que la proporción original de la variable objetivo (y) se mantenga idéntica tanto

en el conjunto de entrenamiento como en el de prueba, garantizando la representatividad estadística del riesgo en ambas muestras y evitando el sesgo de muestreo [10].

5.3.5 Transformación Matemática de Variables

Para habilitar el procesamiento por parte de los algoritmos de Machine Learning, se aplicaron las siguientes técnicas de codificación y normalización:

- Escalamiento Numérico (Min-Max Scaling):
  Las variables Cuantía y Duración fueron transformadas al rango [0, 1]. Esto es crítico para neutralizar la diferencia de magnitudes (billones de pesos vs. cientos de días) y evitar que la variable de mayor valor numérico domine la función de pérdida del modelo [9].
- Codificación Categórica:
  - Target Encoding: Aplicado a variables de alta cardinalidad (Municipio, Modalidad) para capturar la probabilidad de riesgo asociada a cada categoría sin aumentar la dimensionalidad, siguiendo el esquema propuesto por [14].
  - One-Hot Encoding: Aplicado a variables nominales de baja cardinalidad (Tipo de Contrato, Orden, Departamento), creando variables binarias independientes.
- Procesamiento de Lenguaje Natural (NLP):
  Mediante la técnica TF-IDF (Term Frequency - Inverse Document Frequency), se vectorizó el campo de texto libre "Objeto a Contratar". Para evitar una explosión dimensional y mantener la eficiencia computacional, se configuró el algoritmo para extraer únicamente las 150 características (n-gramas) más frecuentes y con mayor relevancia predictiva, aplicando previamente un proceso de limpieza que incluyó la conversión a minúsculas y la eliminación de palabras vacías (stop words) en español. El vector resultante transformó cada texto en 150 nuevas columnas numéricas continuas, donde cada columna representa el peso específico de términos clave dentro del objeto del contrato (por ejemplo, términos recurrentes asociados a obras, suministros o asesorías) [11].

5.3.6 Arquitectura del Conjunto de Datos Final

Tras la ejecución de todo el flujo de preprocesamiento (selección, limpieza, imputación y transformación matemática), el conjunto de datos final (matriz de características) adquirió su estructura definitiva para la fase de entrenamiento y evaluación.

El dataset consolidado quedó compuesto por una matriz de 1.985.205 filas (correspondientes a los contratos con estado definitivo) y 196 columnas (variables predictoras).

La dimensionalidad horizontal de este conjunto final se compone estructuralmente de dos grandes bloques:

- Las 46 variables tabulares base (escaladas numéricamente y codificadas mediante Target Encoding y One-Hot Encoding), que capturan la dimensión financiera, temporal, geográfica e institucional del contrato.
- El vector resultante de la técnica TF-IDF, el cual integró exactamente 150 características numéricas adicionales correspondientes al análisis semántico del objeto contractual.

Esta consolidación garantiza que los algoritmos de aprendizaje supervisado reciban una matriz puramente numérica, estandarizada, libre de nulos y optimizada para el reconocimiento de patrones de riesgo contractual.

6 DESARROLLO DE MODELOS

6.1 SELECCIÓN Y JUSTIFICACIÓN DE ALGORITMOS

Para abordar el problema de clasificación, se ha seleccionado una taxonomía de algoritmos que abarca desde modelos lineales tradicionales hasta arquitecturas de ensamble de última generación, los cuales se resumen en la tabla 5, y que se han seleccionado por su robustez en tareas de clasificación. Esta selección busca equilibrar la interpretabilidad, la eficiencia computacional en entornos locales y la capacidad predictiva ante datos heterogéneos y desbalanceados.

Algoritmo	Paradigma	Fortalezas Académicas	Limitaciones	Justificación en el Proyecto
Regresión Logística	Lineal / Probabilístico	Alta interpretabilidad; estimación directa de probabilidades; bajo riesgo de sobreajuste.	Asume linealidad entre variables; sensible a valores atípicos.	Funciona como **Baseline** (línea base) para evaluar la ganancia de modelos complejos.
Árbol de Decisión	No paramétrico	Captura relaciones no lineales e interacciones; fácil visualización.	Alta varianza; propenso al sobreajuste (*overfitting*).	Permite identificar las particiones lógicas iniciales de las variables de riesgo.
Naive Bayes	Probabilístico Bayesiano	Extremadamente eficiente; asume independencia condicional (efecto multiplicativo).	La suposición "Naive" de independencia rara vez se cumple.	Óptimo para procesar vectores de alta dimensionalidad provenientes de **TF-IDF**.
Random Forest	Ensamble (**Bagging**)	Reduce la varianza mediante el promedio de múltiples árboles; robusto ante ruido.	Mayor costo en memoria; menor interpretabilidad que un árbol simple.	Modelo de alta confiabilidad que maneja bien la diversidad de variables tabulares.

Gradient Boosting (LightGBM)	Ensamble (**Boosting**)	Minimiza el sesgo y el error mediante optimización secuencial de la función de pérdida.	Requiere sintonización de hiperparámetros; sensible al ruido si no se regulariza.	Representa el estado del arte en **rendimiento predictivo** para datos estructurados.

Tabla 5. Modelos utilizados en el proyecto

6.1.1 *Modelos de Referencia (línea base)*

El uso de modelos de referencia permite establecer un umbral mínimo de rendimiento. La Regresión Logística es fundamental en este sentido, ya que permite observar si el fenómeno puede ser explicado mediante relaciones lineales simples. Por otro lado, los Árboles de Decisión proporcionan un marco de referencia para entender cómo las reglas jerárquicas afectan la clasificación sin la complejidad de un ensamble. Ambos modelos son esenciales para validar que la complejidad adicional de los algoritmos avanzados está justificada por una mejora significativa en las métricas de evaluación.

6.1.2 *Algoritmos Probabilísticos y Datos de Texto*

Dada la inclusión de variables de texto no estructurado, el algoritmo Naive Bayes se selecciona por su eficiencia histórica en tareas de procesamiento de lenguaje natural (NLP). Al basarse en el Teorema de Bayes, este modelo es capaz de manejar grandes espacios de características (como los generados por TF-IDF) con un costo computacional mínimo, ofreciendo una perspectiva probabilística sobre la influencia de términos específicos en la clase objetivo, siendo un estándar en la minería de texto [11].

6.1.3 *Ensembles: El equilibrio entre Bagging y Boosting*

Los modelos de conjunto (ensembles) combinan múltiples estimadores débiles para construir un predictor robusto. Este proyecto evalúa las dos filosofías principales de ensamble:

a) Bootstrap Aggregating (Bagging): Representado por Random Forest, este método entrena múltiples árboles en paralelo sobre diferentes submuestras de datos. Su fuerza reside en la reducción de la varianza, lo que lo hace muy estable ante fluctuaciones en los datos de entrenamiento, reduciendo el sobreajuste [10].
b) Gradient Boosting: Representado por implementaciones optimizadas como LightGBM [18], este enfoque construye árboles de forma secuencial. Cada nuevo árbol intenta corregir los errores residuales del modelo anterior mediante el descenso de gradiente. Este paradigma es particularmente potente para capturar patrones sutiles y complejos, y permite la integración de pesos de clase para mejorar el rendimiento en conjuntos de datos desbalanceados.

6.1.4 *Importancia de las Variables e Interpretabilidad del Modelo*

En el contexto de la administración pública y la fiscalización, la capacidad predictiva de un

algoritmo debe estar acompañada de explicabilidad. No basta con que el modelo identifique un contrato de alto riesgo; es imperativo comprender qué factores están impulsando esa predicción para que los entes de control puedan tomar decisiones informadas.

La importancia de las Variables (Feature Importance) es el conjunto de técnicas que permiten cuantificar la contribución de cada característica de entrada a las predicciones del modelo. Los mecanismos para calcular esta importancia varían según la arquitectura del algoritmo:

- Modelos Lineales (Regresión Logística): La importancia se deriva directamente de los coeficientes calculados para cada variable. Un coeficiente con mayor magnitud absoluta indica que esa variable tiene un impacto proporcionalmente mayor en la probabilidad de que el contrato sea clasificado como "fracaso", manteniendo las demás constantes.
- Modelos Basados en Árboles (Random Forest y Árboles de Decisión): Utilizan medidas de Impureza de Gini o Entropía. La importancia se calcula observando cuánto disminuye la impureza de los nodos cada vez que se utiliza una variable específica para realizar una partición. Las variables que logran las divisiones más "limpias" y frecuentes en los niveles superiores del árbol se consideran las más importantes, basándose en la reducción de impureza de la información [12].
- Modelos de Gradient Boosting (LightGBM): En estos modelos, la importancia se suele medir mediante dos métricas:
    a. Split (Frecuencia): El número de veces que una variable es seleccionada para realizar una división en todos los árboles del ensamble.
    b. Gain (Ganancia): La contribución total de una variable a la reducción acumulada de la función de pérdida. Esta métrica es especialmente útil para identificar qué variables son las que realmente aportan mayor precisión al modelo.

6.1.5 *Interpretabilidad Global vs. Local*

Metodológicamente, es vital distinguir entre la interpretabilidad global, que permite entender el comportamiento general del modelo sobre todo el conjunto de datos (por ejemplo, determinar que, en términos generales, la ubicación geográfica es el mayor predictor de riesgo), y la interpretabilidad local, que explica por qué el modelo tomó una decisión específica para un contrato individual, siguiendo los principios de Machine Learning Interpretable [15].

Esta capacidad de auditoría algorítmica transforma el modelo de una "caja negra" a una herramienta de transparencia administrativa, permitiendo identificar patrones sistémicos de riesgo y banderas rojas (red flags) que podrían estar asociadas a ineficiencias o irregularidades en la contratación.

6.2 EVALUACIÓN, VALIDACIÓN Y BENCHMARKING

El rendimiento de un modelo de clasificación en entornos con desequilibrio de clases no puede ser evaluado mediante la exactitud (Accuracy), ya que esta métrica tiende a ignorar la clase

minoritaria, sesgando la interpretación del desempeño real del modelo [17]. Para este proyecto, se implementa un marco de evaluación multidimensional basado en la Matriz de Confusión, el cual permite desagregar el rendimiento del algoritmo en función de los aciertos y errores sobre la clase de riesgo. A continuación, la tabla 6 presenta las métricas utilizadas para evaluación de los modelos, seleccionadas por su robustez en problemas de clasificación binaria [9]:

Métrica	Definición Académica	Relevancia Estratégica en el Proyecto
Precision	VP/(VP + FP)	Mide la **confiabilidad de las alertas**. Indica qué proporción de los contratos marcados como "riesgo" realmente lo fueron.
Recall (Sensibilidad)	VP/(VP + FN)	Mide la **capacidad de detección**. Es la métrica crítica del proyecto, ya que indica cuántos contratos de riesgo real fueron capturados por el modelo.
F1-Score	2 * ((Precision * Recall) / (Precision * Recall))	Es la media armónica entre precisión y sensibilidad. Proporciona una medida única del equilibrio del modelo, penalizando valores extremos en cualquiera de las dos métricas.
AUC-ROC	Área bajo la curva de probabilidad	Evalúa la **capacidad de discriminación** del modelo. Indica qué tan bien el algoritmo separa las clases: - Positiva: contratos que fracasaron y - Negativa: contratos que finalizaron exitosamente.

Tabla 6. Métricas para evaluación de los modelos

6.2.1 *Precisión y Sensibilidad (Recall)*

En el control preventivo, existe un compromiso (trade-off) inherente entre la precisión y el recall. Un modelo con alto recall asegura que muy pocos casos de riesgo pasen desapercibidos (pocos Falsos Negativos), aunque esto pueda aumentar los Falsos Positivos (alertas que resultan ser contratos exitosos).

Metodológicamente, este proyecto prioriza la optimización del Recall, bajo la premisa de que el costo social y económico de omitir un contrato con irregularidades es significativamente superior al costo operativo de revisar un contrato que finalmente resulta estar en orden.

6.2.2 *Curva ROC y AUC*

La curva de Característica Operativa del Receptor (ROC) visualiza el rendimiento del clasificador comparando la tasa de verdaderos positivos frente a la tasa de falsos positivos. El AUC (Area Under the Curve) resume esta curva en un valor entre 0 y 1. Un AUC cercano a 1.0 indica un modelo con una capacidad de separación casi perfecta, mientras que un valor de 0.5 sugiere un rendimiento

no mejor que el azar. Esta métrica es la más robusta para el benchmarking (comparativa) entre los cinco algoritmos evaluados, independientemente del umbral de decisión seleccionado [10].

6.2.3 *Validación Cruzada Estratificada (Stratified K-Fold)*

Para garantizar que los resultados obtenidos sean estadísticamente significativos y generalizables a nuevos datos, se emplea la técnica de Validación Cruzada Estratificada. A diferencia de la validación cruzada convencional, la versión estratificada asegura que cada uno de los subconjuntos (folds) mantenga la misma proporción de clases (éxito vs. fracaso) que el conjunto de datos original. Esto es indispensable en este estudio para evitar que algún subconjunto carezca de ejemplos de la clase minoritaria, tal como se recomienda en la validación de modelos con distribución sesgada [12], lo cual sesgaría la estimación del rendimiento y comprometería la fiabilidad del modelo en producción.

Para garantizar una comparación objetiva (benchmarking), se estableció un protocolo de salida estándar para todos los experimentos. Cada modelo es evaluado bajo las mismas condiciones de datos (X_test) y mediante cuatro dimensiones críticas:

1. **AUC-ROC**: Para medir la capacidad global de separación de clases.
2. **F1-Score**: Para balancear la precisión y la sensibilidad en la clase minoritaria.
3. **Matriz de Confusión**:
   Para la presentación de los resultados de clasificación, se adoptó la convención estadística donde las filas representan las clases predichas por el modelo y las columnas representan las clases reales.

   La matriz se desglosa así y se resume en la tabla 7:
   - **Verdadero Positivo (VP)**: El modelo predijo fracaso y el contrato realmente fracasó. Es el éxito del ejercicio de control y prevención, pues identificó correctamente un riesgo.
   - **Falso Positivo (FP)**: El modelo predijo fracaso, pero el contrato fue exitoso (Real Negativo). Representa una "falsa alarma" que podría generar auditorías innecesarias.
   - **Falso Negativo (FN)**: El modelo predijo éxito, pero el contrato realmente fracasó (Real Positivo). Es el error de mayor peligro, pues un contrato con problemas pasó inadvertido.
   - **Verdadero Negativo (VN)**: El modelo predijo éxito y el contrato fue exitoso. Indica que el modelo reconoce correctamente una ejecución contractual normal.

Concepto	Real Fracaso (1)	Real Éxito (0)

Predicho Fracaso (1)	**Verdadero Positivo (VP):** Detección exitosa del riesgo contractual.	**Falso Positivo (FP):** Alerta de riesgo en contrato exitoso (Falsa Alarma).
Predicho Éxito (0)	**Falso Negativo (FN):** Omisión de riesgo en contrato fallido (Punto Ciego).	**Verdadero Negativo (VN):** Identificación correcta de ejecución normal.

Tabla 7. Estructura de la Matriz de Confusión bajo el Enfoque de Control y Prevención

4. **Costo Computacional**: Medido en tiempo de entrenamiento (en segundos).

6.2.4 *Selección de Hiperparámetros y Estrategia de Balanceo*

Una vez definidas las métricas de evaluación, se procedió a determinar la configuración óptima para cada algoritmo. Dado que el conjunto de datos presenta un desbalance de clases significativo (donde los contratos ***No exitosos*** son la minoría) y una alta dimensionalidad tras el preprocesamiento, el uso de las configuraciones por defecto (default) de las librerías estándar resultaba insuficiente.

Para abordar esto, se implementó una estrategia de optimización basada en dos pilares técnicos:

1. **Aprendizaje Sensible al Costo (Cost-Sensitive Learning)**: En lugar de alterar artificialmente los datos con técnicas de sobremuestreo sintético (como SMOTE [16]), se optó por configurar los pesos internos de los algoritmos. Se activaron los parámetros class_weight='balanced' (en Scikit-Learn) e is_unbalance=True (en LightGBM). Esto obliga a la función de pérdida del modelo a penalizar con mayor severidad los errores de clasificación en la clase minoritaria ("Fracaso"), priorizando la sensibilidad del sistema.
2. **Ajuste Experimental (Heuristic Tuning):** A diferencia de una búsqueda exhaustiva automatizada (Grid Search), que implicaría un costo computacional elevado dada la volumetría de los datos, se aplicó una estrategia de ajuste heurístico basado en regularización. Se definieron manualmente los hiperparámetros críticos para controlar la complejidad del modelo, tales como la limitación de la profundidad en los árboles (max_depth), la fuerza de regularización inversa (C) en modelos lineales y el ajuste de pesos por clase (class_weight='balanced'). Esta configuración fue validada empíricamente con el objetivo prioritario de penalizar el sobreajuste (overfitting) y maximizar la sensibilidad frente a la clase minoritaria.

   Si bien este enfoque no garantiza la optimización matemática absoluta que ofrecería una búsqueda exhaustiva, permite obtener un modelo robusto y generalizable con un costo computacional viable para la operación.

La tabla 8 presenta la evidencia de este proceso, contrastando el desempeño de los modelos en su estado base frente a la configuración optimizada seleccionada para este estudio:

Modelo	Escenario	Hiperparámetros (Diferencias Clave)	F1-Score	AUC-ROC
Regresión Logística	Base (Defecto)	class_weight=None, C=1.0	0,3695	0,8914
Regresión Logística	**Optimizado**	**class_weight='balanced', C=0.1**	**0,4767**	**0,8947**
Árbol de Decisión	Base (Defecto)	max_depth=None (Infinito), weight=None	0,7997	0,9056
Árbol de Decisión	**Optimizado**	**max_depth=20, weight='balanced'**	**0,6657**	**0,9638**
Naive Bayes	Base (Defecto)	var_smoothing=1e-9	0,222	0,6234
Naive Bayes	**Optimizado**	**var_smoothing=1e-2 (Estabilizado)**	**0,2356**	**0,7886**
Random Forest	Base (Defecto)	n_est=100, depth=None, weight=None	0,8298	0,9721
Random Forest	**Optimizado**	**n_est=100, depth=20, weight='balanced'**	**0,658**	**0,9603**
LightGBM	Base (Defecto)	leaves=31, lr=0.1, unbalance=False	0,6958	0,9575
LightGBM	**Optimizado**	**leaves=128, lr=0.05, unbalance=True**	**0,7108**	**0,9732**

Tabla 8. Selección de Hiperparámetros y Mejoras de Desempeño

Análisis de la Configuración Seleccionada:

Como se evidencia en la tabla 8, la optimización fue determinante. Evidencia la efectividad de la estrategia de sintonización heurística frente a las configuraciones por defecto de scikit-learn:

1. Recuperación de Modelos Lineales y Probabilísticos:
   En la Regresión Logística, la activación del parámetro class_weight='balanced' y el ajuste de la regularización (C=0.1) incrementaron el F1-Score en más de 10 puntos porcentuales (0.36 → 0.47). Un comportamiento similar se observa en Naive Bayes, donde el suavizado de la varianza (var_smoothing) rescató el modelo de un desempeño casi aleatorio (AUC 0.62) a uno aceptable (AUC 0.78). Esto confirma que los algoritmos base son incapaces de manejar el desbalance severo de estos datos sin intervención experta.

2. Control de Sobreajuste en Modelos de Árbol:
   En el caso del Árbol de Decisión y Random Forest, se observa un fenómeno interesante: el modelo "Base" presenta un F1-Score artificialmente alto (0.79 y 0.82 respectivamente). Sin embargo, esto es síntoma de overfitting (memorización), dado que la profundidad era infinita.Al aplicar la optimización (max_depth=20), el F1 disminuye, pero el AUC-ROC aumenta significativamente en el Árbol de Decisión (0.90 → 0.96). Esto indica que, aunque el modelo optimizado es más conservador al clasificar (menor F1), ordena mejor las probabilidades de riesgo y es más robusto para generalizar ante nuevos datos, cumpliendo el objetivo de regularización.

3. Superioridad de LightGBM:
   El modelo LightGBM Optimizado se presenta como el mejor. A diferencia de los árboles clásicos, el ajuste de hiperparámetros (num_leaves=128, learning_rate=0.05) mejoró simultáneamente tanto la precisión global (F1: 0.71) como la capacidad de discriminación (AUC: 0.973), validando su selección como el modelo final para este proyecto.

Con base en estos hallazgos, todos los resultados presentados a partir de la siguiente sección corresponden a los modelos ejecutados bajo esta configuración optimizada.

6.3 Modelo Regresión Logística (modelo 1)

La Regresión Logística actúa como el punto de partida (modelo base) del estudio. Bajo la configuración optimizada definida previamente, este modelo permite establecer si las variables transformadas tienen una relación lineal directa con la probabilidad de riesgo.

Su implementación se rige por tres criterios técnicos clave:

1. *Manejo del Desbalance*: Al configurar class_weight='balanced', se fuerza al algoritmo a penalizar con mayor severidad los errores en la clase minoritaria (Fracaso), priorizando la sensibilidad, una estrategia estándar para conjuntos de datos desequilibrados [17].
2. *Regularización Fuerte (C=0.1)*: Se estableció explícitamente un parámetro de regularización inversa de 0.1. Esta configuración restringe la magnitud de los coeficientes matemáticos, impidiendo que el modelo se "sobreajuste" al ruido inherente de los datos de contratación y obligándolo a aprender solo los patrones más robustos y generalizables.
3. *Eficiencia Lineal*: Permite validar la hipótesis de que el riesgo contractual puede explicarse, al menos parcialmente, mediante una suma ponderada de factores geográficos, económicos y textuales.

6.3.1 *Evaluación de Dimensiones Globales*

AUC-ROC Score: 0,8947
F1-Score: 0,4767
Costo Computacional: 39,58 segundos

- Eficiencia: El uso de técnicas de reducción de dimensionalidad (Target Encoding) y optimización de solucionadores (liblinear) permitió un tiempo de entrenamiento de solo 39.58 segundos. Esto confirma que el modelo es altamente eficiente y escalable para procesar millones de registros sin requerir infraestructura de cómputo excesiva.

- *Discriminación (AUC):* Con un 0,8947, el modelo demuestra una capacidad aceptable para separar las clases, validando que las variables seleccionadas contienen una fuerte señal predictiva independientemente de la linealidad del algoritmo.
- *Balance (F1): El F1-Score de 0,4767 refleja una mejora respecto a la configuración base (ver Tabla 8), demostrando que la estrategia de balanceo de pesos fue efectiva para evitar que el modelo ignorara la clase minoritaria.*

6.3.2 *Análisis de la Matriz de Confusión (Estructura de Decisión)*

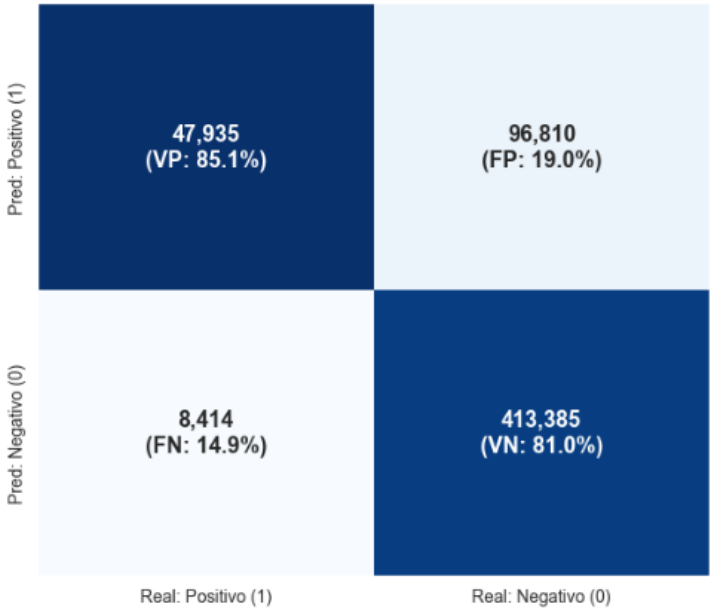

Ilustración 7. Matriz de confusión modelo Regresión Logística

- **Sensibilidad (85,07%)**: El modelo logró detectar el 85.07% de los contratos que realmente fracasaron. Este es un indicador excelente para un modelo lineal, cumpliendo el objetivo primario de "alerta temprana" al capturar la gran mayoría de los riesgos.
- **Especificidad / Eficiencia Operativa (81,02%)**: El modelo reconoció correctamente el 81.02% de los contratos exitosos. Esto indica que la gran mayoría de la contratación normal fluye sin ser obstaculizada por alertas falsas.
- **Tasa de Falsa Alarma (18,98%)**: El modelo clasificó erróneamente como riesgo al 18.98% de los contratos exitosos. Esta cifra sugiere que, aunque el modelo es sensible, su estructura lineal le impide distinguir sutilezas en casi 1 de cada 5 contratos sanos, generando un volumen considerable de alertas para auditoría.
- **Riesgo Omitido (14,93%)**: Únicamente el 14.93% de los fracasos reales no fueron alertados.

6.3.3 *Interpretación y Banderas Rojas*

La Regresión Logística Optimizada confirma que los datos tienen calidad predictiva y el modelo es computacionalmente ligero. Sin embargo, una tasa de falsa alarma cercana al 19% sugiere que la frontera de decisión lineal es demasiado rígida para la complejidad de la contratación pública.

Se concluye que, para reducir las falsas alarmas sin sacrificar la alta sensibilidad lograda (85%), es necesario transitar hacia modelos no lineales capaces de capturar interacciones más complejas.

6.4 Modelo Árbol de Decisión (modelo 2)

A diferencia de la rigidez lineal del modelo anterior, el Árbol de Decisión permite capturar relaciones no lineales y jerárquicas mediante la partición recursiva del espacio de datos. Este enfoque es particularmente útil en el contexto del SECOP I, donde la normativa de contratación sigue reglas lógicas condicionales.

Para su implementación, se aplicó la configuración optimizada definida en la fase experimental, regida por tres parámetros clave:

1. Control de Profundidad (max_depth=20): Se limitó el crecimiento vertical del árbol para evitar que el algoritmo memorizara el ruido de los datos (overfitting), mejorando la capacidad de generalización del modelo [10].
2. Robustez Estadística (min_samples_leaf=50): Se exigió un mínimo de 50 contratos por hoja para tomar una decisión, garantizando que las reglas generadas no se basen en casos aislados o anécdotas.
3. Pesos Balanceados (class_weight='balanced'): Se mantuvo la prioridad sobre la clase minoritaria para asegurar una alta tasa de detección de fraudes.

6.4.2 Evaluación de Dimensiones Globales

AUC-ROC Score: 0,9638
F1-Score: 0,6657
Costo Computacional: 100,57 segundos

- *Eficiencia*: El tiempo de entrenamiento fue de 100,57 segundos. Aunque representa un incremento respecto a la Regresión Logística (aprox. 2.5 veces más lento), este costo es despreciable en un entorno real considerando la mejora masiva en la capacidad predictiva.
- *Discriminación (AUC):* El AUC subió drásticamente de 0,89 a 0,9638. Este incremento valida la hipótesis de que el riesgo en la contratación pública no sigue una línea recta, sino que depende de interacciones complejas (ej. Municipio X es riesgoso solo si el Contrato es de Obra) que el árbol logró decodificar exitosamente.

- *Balance (F1):* El F1-Score alcanzó 0,6657, superando ampliamente el 0.47 del modelo lineal. Esto indica que el árbol logra un equilibrio mucho más sano entre precisión y recuperación.

6.4.3 *Análisis de la Matriz de Confusión (Estructura de Decisión)*

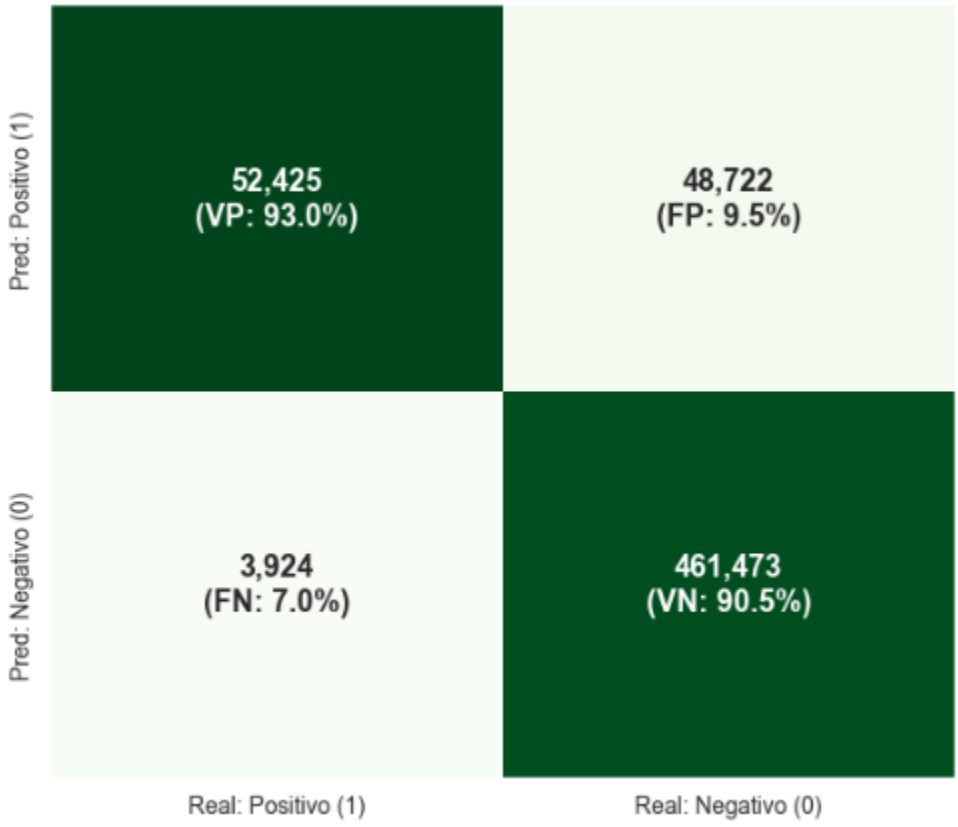

Ilustración 8. Matriz de confusión modelo Árbol de Decisión

- **Sensibilidad (93,04%)**: El modelo detectó el 93,04% de los contratos que realmente fracasaron. La capacidad de detección de riesgo es sobresaliente, superando incluso a la regresión logística y dejando escapar muy pocos casos problemáticos.
- **Especificidad / Eficiencia Operativa (90,45%)**: Se identificó correctamente el 90.45% de los contratos exitosos. Este es un salto cualitativo crítico: el árbol entiende mucho mejor qué hace a un contrato "seguro", reduciendo la fricción operativa.
- **Tasa de Falsa Alarma (9,55%)**: Se logró romper la barrera del 10%. Solo el 9.55% de los contratos exitosos fueron marcados erróneamente (frente al ~19% del modelo anterior). Esto significa que las auditorías innecesarias se redujeron prácticamente a la mitad.
- **Riesgo Omitido (6,96%)**: Únicamente el 6,96% de los fracasos reales no fueron alertados, consolidando al modelo como una herramienta confiable de seguridad preventiva.

6.4.4 *Interpretación y Banderas Rojas*

El Árbol de Decisión Optimizado valida contundentemente la hipótesis de no linealidad. Al aplicar reglas de decisión jerárquicas (limitadas a 20 niveles de profundidad), el sistema logró "limpiar" gran parte del ruido que confundía al modelo lineal.

Sin embargo, aunque bajar las falsas alarmas al 9.55% es un logro técnico notable, en un universo de millones de contratos, ese porcentaje aún podría representar una carga operativa considerable. Esto sugiere que, para perfeccionar la precisión ("limar" ese último 9% de error), se requiere el uso de técnicas de ensamble (Ensemble Learning) que promedien el conocimiento de múltiples árboles, como Random Forest o LightGBM.

6.5 Modelo Naive Bayes (modelo 3)

Tras evaluar los modelos lineales y las reglas jerárquicas, el tercer experimento introduce el paradigma probabilístico. Se implementó el algoritmo Gaussian Naive Bayes, alineado con la naturaleza continua de los datos obtenidos tras el escalado.

Para esta ejecución, se aplicó la configuración optimizada definida previamente (var_smoothing=1e-2). Este parámetro añade una pequeña corrección a la varianza para mejorar la estabilidad numérica del cálculo, especialmente útil para manejar el desbalance de clases y evitar que probabilidades cercanas a cero anulen la predicción.

A diferencia de los modelos anteriores, este algoritmo asume independencia condicional entre las variables predictoras (ej. asume que la Cuantía no tiene relación con el Tipo de Contrato), una simplificación teórica conocida como la hipótesis de independencia ingenua [11]. Su inclusión busca establecer una línea base de "eficiencia pura" y verificar si un enfoque estadístico simple es suficiente.

6.5.1 Evaluación de Dimensiones Globales

AUC-ROC Score: 0,7886
F1-Score: 0,2356
Costo Computacional: 11,62 segundos

- *Eficiencia*: El modelo confirmó su reputación de velocidad, entrenándose en apenas 11,62 segundos. Es, con diferencia, el algoritmo más rápido del estudio, lo que demuestra su ventaja teórica en entornos de Big Data.
- *Discriminación (AUC):* El AUC alcanzó un 0,7886. Si bien es una mejora respecto a la configuración por defecto (gracias al suavizado de varianza), sigue estando muy por debajo del Árbol de Decisión (0.96). Esto evidencia la limitación estructural del algoritmo:

al ignorar las correlaciones entre variables, pierde capacidad para separar con precisión los casos complejos.

- *Balance (F1):* El puntaje de 0,2356 es inaceptablemente bajo. Este valor revela el colapso de la precisión: el modelo "dispara a todo lo que se mueve", detectando fraudes a costa de generar un ruido ensordecedor.

6.5.2 *Análisis de la Matriz de Confusión (Estructura de Decisión)*

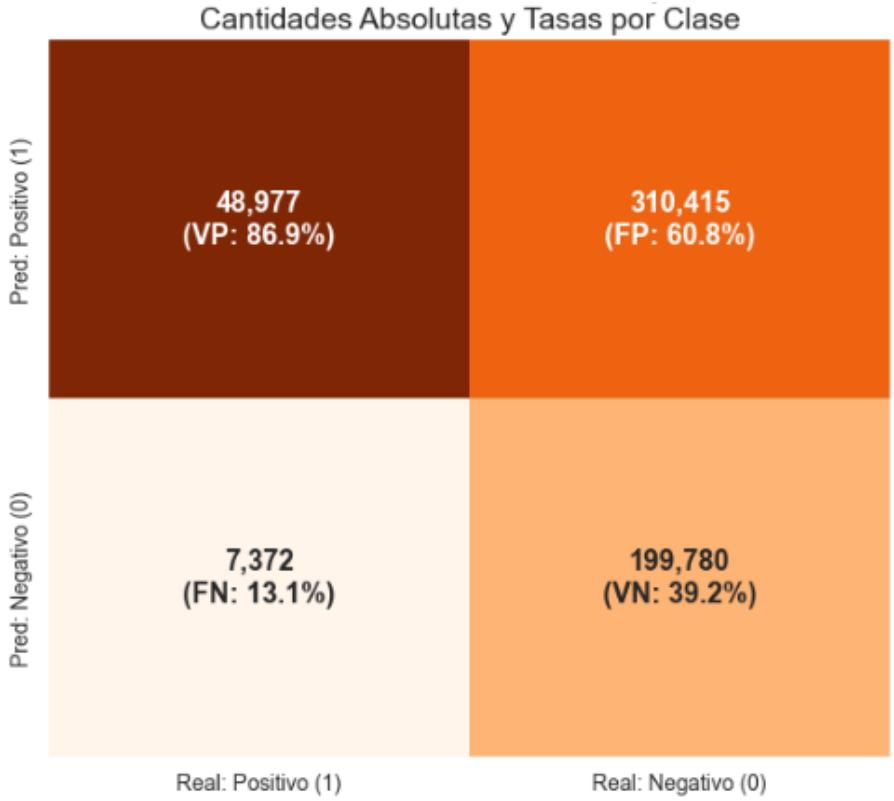

Ilustración 9. Matriz de confusión modelo Naive Bayes

- **Sensibilidad (86,92%)**: El modelo mantiene una alta capacidad de detección, capturando el 86,92% de los fracasos. Sin embargo, esta cifra es engañosa si no se contextualiza con la tasa de error.
- **Especificidad / Eficiencia Operativa (39,16%)**: Este es el punto de quiebre. El modelo solo logró identificar correctamente el 39,16% de los contratos sanos. La mayoría de la contratación legítima fue marcada como sospechosa.
- **Tasa de Falsa Alarma (60,84%)**: El modelo clasificó erróneamente al 60,84% de los contratos exitosos como riesgosos. Esto generaría una carga operativa insostenible, obligando a auditar más de la mitad de toda la contratación pública sin justificación real.
- **Riesgo Omitido (13,08%)**: El 13.08% de los riesgos reales pasaron desapercibidos.

6.5.3 *Interpretación y Banderas Rojas*

El desempeño de Naive Bayes ilustra claramente el peligro de asumir independencia entre variables en un ecosistema tan interconectado como la contratación pública. Si bien la sensibilidad rozó el 87%, la especificidad colapsó al 39%, indicando que el algoritmo sobreestimó masivamente la probabilidad de fracaso, etiquetando erróneamente a la mayoría de los contratos que terminaran exitosamente como riesgosos.
La tasa de falsa alarma superior al 60% invalida al modelo para uso práctico, ya que el "ruido" generado haría imposible cualquier auditoría eficiente. El experimento confirma que las variables del SECOP (como cuantía, modalidad y objeto) poseen correlaciones fuertes que requieren algoritmos capaces de modelar interacciones complejas y no lineales.

Este experimento es importante para este proyecto porque permite concluir empíricamente que:

- La fuerza probabilística aislada de las palabras (TF-IDF) no es suficiente para predecir el riesgo.
- El fenómeno del fracaso contractual es multivariado y jerárquico, lo que explica por qué el Árbol de Decisión fue tan superior.
- Se justifica plenamente el paso hacia modelos de ensamble (Random Forest), que combinan la potencia de los árboles con una mayor estabilidad estadística.

6.6 Modelo Random Forest (modelo 4)

Superando la inestabilidad inherente de un único árbol, el modelo Random Forest implementa la técnica de ensamble tipo Bagging (Bootstrap Aggregating). Tal como se definió en el marco teórico, este algoritmo construye múltiples estimadores independientes e introduce aleatoriedad en la selección de variables para descorrelacionar los errores individuales , reduciendo la varianza del modelo final [12].

Para garantizar la robustez del modelo, se definió una arquitectura basada en tres hiperparámetros estructurales clave:

1. Votación Mayoritaria (n_estimators=100): Se instanciaron 100 árboles de decisión. La predicción final no depende de un solo evaluador, sino del consenso democrático de los 100 estimadores, lo que permite filtrar el ruido estadístico y reducir la varianza.
2. Generalización Controlada (max_depth=20): A diferencia de la implementación estándar que permite árboles infinitos, se limitó la profundidad de cada estimador a 20 niveles. Esto obliga a cada árbol a aprender patrones generales en lugar de memorizar excepciones (overfitting).
3. Sensibilidad al Costo (class_weight='balanced'): Se instruyó al algoritmo para ajustar los pesos en el cálculo de impureza, penalizando con mayor severidad los errores en la clase minoritaria (Fracaso) para mantener alta la sensibilidad del sistema.

6.6.2 Evaluación de Dimensiones Globales

AUC-ROC Score: 0,9603
F1-Score: 0,6580
Costo Computacional: 313,35 segundos

- *Eficiencia*: El tiempo de entrenamiento ascendió a 313.35 segundos, triplicando el costo del árbol individual. Este aumento es el "precio" de la estabilidad: calcular 100 árboles independientes requiere significativamente más recursos, convirtiéndolo en el modelo más exigente hasta el momento.
- *Discriminación (AUC):* Con un 0,9603, el resultado es prácticamente idéntico al del árbol único (0.9638). Este hallazgo técnico es crucial: indica que la señal predictiva capturada por el primer modelo era robusta y no producto del azar. El ensamble validó la calidad de las reglas descubiertas, pero no logró extraer información adicional significativa, sugiriendo que se ha alcanzado el límite de lo que el enfoque de partición (Bagging) puede ofrecer.
- *Balance (F1):* El valor de 0.6580 confirma la estabilidad. A diferencia de las pruebas iniciales donde el ensamble degradaba la precisión, la optimización de profundidad (max_depth=20) permitió mantener la calidad de la predicción, evitando que la votación mayoritaria diluyera los aciertos.

6.6.3 *Análisis de la Matriz de Confusión (Estructura de Decisión)*

Ilustración 10. Matriz de confusión modelo Random Forest

- **Sensibilidad (90,93%)**: El modelo detectó el 90.93% de los casos de fracaso. Aunque hubo una leve disminución respecto al árbol único (93%), esto se interpreta como una corrección positiva: el consenso de 100 árboles filtró algunos casos limítrofes, haciendo la detección ligeramente más conservadora pero estadísticamente más confiable.
- **Especificidad / Eficiencia Operativa (90,56%)**: El modelo clasificó correctamente al 90.56% de los contratos exitosos. La capacidad de distinguir la normalidad se mantiene sólida y consistente con el modelo anterior.
- **Tasa de Falsa Alarma (9,44%)**: Se logró mantener las falsas alarmas por debajo del umbral psicológico del 10% (9.44%). Esto desmiente la hipótesis de que el ensamble generaría más ruido; por el contrario, validó las "zonas de seguridad" identificadas previamente.
- **Riesgo Omitido (9,07%)**: El 9,07% de los contratos que fracasaron no fueron detectados, con lo cual el modelo tiene un rango de seguridad aceptable.

6.6.4 *Interpretación y Banderas Rojas*

El Random Forest arroja una conclusión definitiva e interesante sobre la relación costo-beneficio.

Los resultados demuestran que el Árbol de Decisión Optimizado (Modelo 2) ya había capturado la mayor parte de la estructura de riesgo latente en los datos. La implementación de Random Forest funcionó como una herramienta de validación científica: confirmó que el alto desempeño (AUC ~0.96) es real y reproducible, pero lo hizo a un costo computacional tres veces mayor sin aportar un incremento marginal en la precisión.

La persistencia de un 9% de falsas alarmas inamovibles sugiere que la técnica de Bagging (promediar árboles paralelos) ha llegado a su techo técnico. Para "limar" ese último porcentaje de error y optimizar el tiempo de cómputo, es necesario cambiar de paradigma hacia el aprendizaje secuencial correctivo (Boosting), justificando la implementación final de LightGBM.

6.7 Modelo LightGBM (modelo 5)

Para cerrar la fase de desarrollo, se implementó LightGBM, configurado específicamente para maximizar la precisión predictiva sin sacrificar la velocidad operativa. A diferencia de los modelos anteriores, se utilizó la configuración optimizada validada en la Sección 6.2.4, regida por tres pilares técnicos:

1. Complejidad Estructural (num_leaves=128): Aprovechando la estrategia de crecimiento por hojas (Leaf-wise), se permitió una alta granularidad en el árbol. Esto capacita al modelo para capturar reglas de negocio profundas y específicas (ej. cruces entre municipio y objeto contractual) que algoritmos más simples omiten, tal como se describe en la arquitectura original de LightGBM [18].

2. Aprendizaje Secuencial Lento (learning_rate=0.05 con 500 iteraciones): Se redujo la tasa de aprendizaje para que el modelo corrigiera los errores de manera progresiva y meticulosa, evitando la convergencia prematura en mínimos locales.
3. Manejo Nativo del Desbalance (is_unbalance=True): Se activó el re-balanceo interno de la función de pérdida, asegurando que la detección de la clase minoritaria (Fracaso) tuviera prioridad matemática sobre la clase mayoritaria.

6.7.2 Evaluación de Dimensiones Globales

AUC-ROC Score: 0,9732
F1-Score: 0,7108
Costo Computacional: 60,68 segundos

- *Eficiencia*: El modelo convergió en apenas 60,68 segundos. A pesar de su alta complejidad interna (500 estimadores y 128 hojas), resultó ser más veloz que el Árbol de Decisión (100s) y cinco veces más rápido que el Random Forest (313s), validando la superioridad del algoritmo Leaf-wise para procesar grandes volúmenes de datos en tiempo récord.
- *Discriminación (AUC):* Con un puntaje de 0,9732, este modelo establece el estándar más alto de todos. Superar la barrera del 0,97 confirma que el enfoque de Boosting (aprendizaje correctivo) logró refinar las fronteras de decisión que los modelos anteriores dejaron borrosas.
- *Balance (F1):* El F1-Score ascendió a 0,7108. Es el único modelo que logró romper el techo del 0.70, indicando el mejor equilibrio matemático entre detectar fraudes y no generar alertas basura.

6.7.3 Análisis de la Matriz de Confusión (Estructura de Decisión)

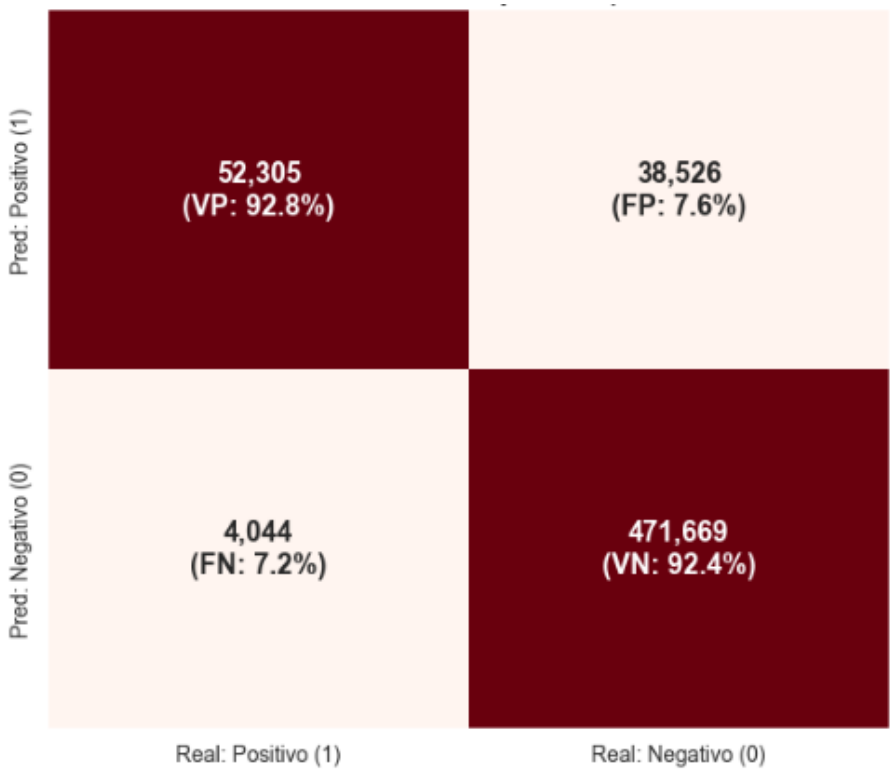

Ilustración 11. Matriz de confusión modelo LightGBM

- **Sensibilidad (92,82%)**: El modelo presenta una capacidad de detección robusta, identificando el 92.82% de los contratos fallidos. Esto lo sitúa al mismo nivel de sensibilidad que el Árbol de Decisión, asegurando que la gran mayoría de los casos de riesgo sean capturados.
- **Especificidad / Eficiencia Operativa (92,45%)**: Este indicador representa el mayor avance del estudio. Se clasificó correctamente el 92.45% de los contratos exitosos, superando el desempeño del Random Forest (90.5%). Esto implica una reducción sustancial en la carga operativa.
- **Tasa de Falsa Alarma (7,55%)**: La tasa de falsos positivos descendió al 7,55%, el valor más bajo de toda la batería experimental. A diferencia de la Regresión Logística (~19%) o el Random Forest (~9.5%), esta configuración logra filtrar eficazmente el ruido, otorgando mayor credibilidad a cada alerta generada.
- **Riesgo Omitido (7,18%)**: Solo el 7,18% de los riesgos reales no fueron alertados. Este margen de error se mantiene controlado y es consistente con los mejores resultados de sensibilidad obtenidos.

6.7.4 *Interpretación y Banderas Rojas*

La configuración optimizada de LightGBM demuestra que es posible reconciliar la alta sensibilidad con la precisión operativa. Al combinar la profundidad de los árboles (num_leaves=128) con el manejo nativo del desbalance (is_unbalance=True), el algoritmo logró corregir los errores de generalización que penalizaban a los modelos anteriores.

El resultado es un sistema que detecta casi el 93% de los riesgos, reduciendo las falsas alarmas a un mínimo del 7.55%.

Este desempeño convierte a LightGBM en el modelo ganador de este proyecto. Ofrece la mejor sinergia entre potencia predictiva (AUC 0,973) y viabilidad técnica: su capacidad para procesar esta complejidad en solo 60 segundos lo habilita para despliegues en entornos reales de auditoría continua, donde la velocidad y la minimización del ruido administrativo son requisitos críticos.

7 SELECCIÓN DE LOS MEJORES MODELOS

Al contrastar los cinco modelos evaluados tras la optimización de hiperparámetros, se evidencia una evolución clara en la capacidad de predicción. La Regresión Logística y Naive Bayes, aunque eficientes en tiempo, mostraron limitaciones estructurales para capturar la complejidad del riesgo contractual, presentando métricas de equilibrio (F1-Score) inferiores a 0,50.

La transición hacia modelos basados en árboles representó un salto cualitativo, con todos los algoritmos de esta familia (Árbol, RF y LightGBM) superando el umbral de 0,96 en AUC. No obstante, es el algoritmo LightGBM el que se posiciona como la solución definitiva por tres razones fundamentales:

- Precisión (Menor Ruido): Logró reducir la tasa de falsas alarmas al 7,55%, el menor de todos. Esto es crítico, pues supera la barrera del 8% que ni el Árbol de Decisión (9.55%) ni el Random Forest (9.44%) pudieron romper. Para una entidad de control, esto significa auditar menos contratos exitosos.
- Balance Sensibilidad/Precisión: Aunque el Árbol de Decisión tuvo una sensibilidad marginalmente superior (93.04% vs 92.82%), lo hizo a costo de mayor "ruido". LightGBM ofrece el mejor equilibrio global, evidenciado en el F1-Score más alto del proyecto (0,7108).
- Escalabilidad y Velocidad: Su costo computacional (60,68 segundos) demuestra una eficiencia muy superior a los métodos de ensamble tradicionales. Es cinco veces más rápido que el Random Forest (313 segundos), lo que facilita su implementación en un entorno de producción real con actualizaciones diarias o procesamiento de millones de registros.

El resumen de las cifras comparativas de los modelos optimizados se presenta en la tabla 9:

Modelo	AUC-ROC	F1-Score	% Detección (Sensibilidad)	% Falsa Alarma (FP Rate)	% Riesgo Omitido (FN Rate)	Tiempo (s)
LightGBM	0,9732	0,7108	92,82%	7,55%	7,18%	60,68
Árbol de Decisión	0,9638	0,6657	93,04%	9,55%	6,96%	100,57
Random Forest	0,9603	0,658	90,93%	9,44%	9,07%	313,35
Regresión Logística	0,8947	0,4767	85,07%	18,98%	14,93%	39,58
Naive Bayes	0,7886	0,2356	86,92%	60,84%	13,08%	11,62

Tabla 9. Tabla comparativa de las métricas de los modelos

7.1 Análisis e Importancia de Variables (LightGBM)

Identificado el modelo LightGBM como la arquitectura óptima, se procedió a extraer la importancia de las variables utilizando la métrica de Ganancia (Gain).

A diferencia de la importancia basada en frecuencia (Split), que simplemente contabiliza cuántas veces se utiliza una variable para dividir un nodo, la Ganancia mide la calidad de dichas divisiones. Matemáticamente, representa la reducción total acumulada de la función de pérdida (o impureza, como la Entropía o Gini) aportada por la variable a lo largo de todos los árboles del ensamble.

Esta distinción es fundamental en este estudio debido a la presencia de variables de alta cardinalidad (como el Municipio). Una métrica de frecuencia podría sesgar la importancia hacia variables con muchas categorías simplemente porque ofrecen más oportunidades de corte. La Ganancia, en cambio, penaliza estas divisiones triviales y resalta únicamente aquellas características que reducen drásticamente la incertidumbre del sistema, proporcionando una 'anatomía del riesgo' más fiel a la realidad del fenómeno de corrupción e ineficiencia [12].

Los resultados presentados en la ilustración 12 revelan la "anatomía del riesgo" en la contratación pública:

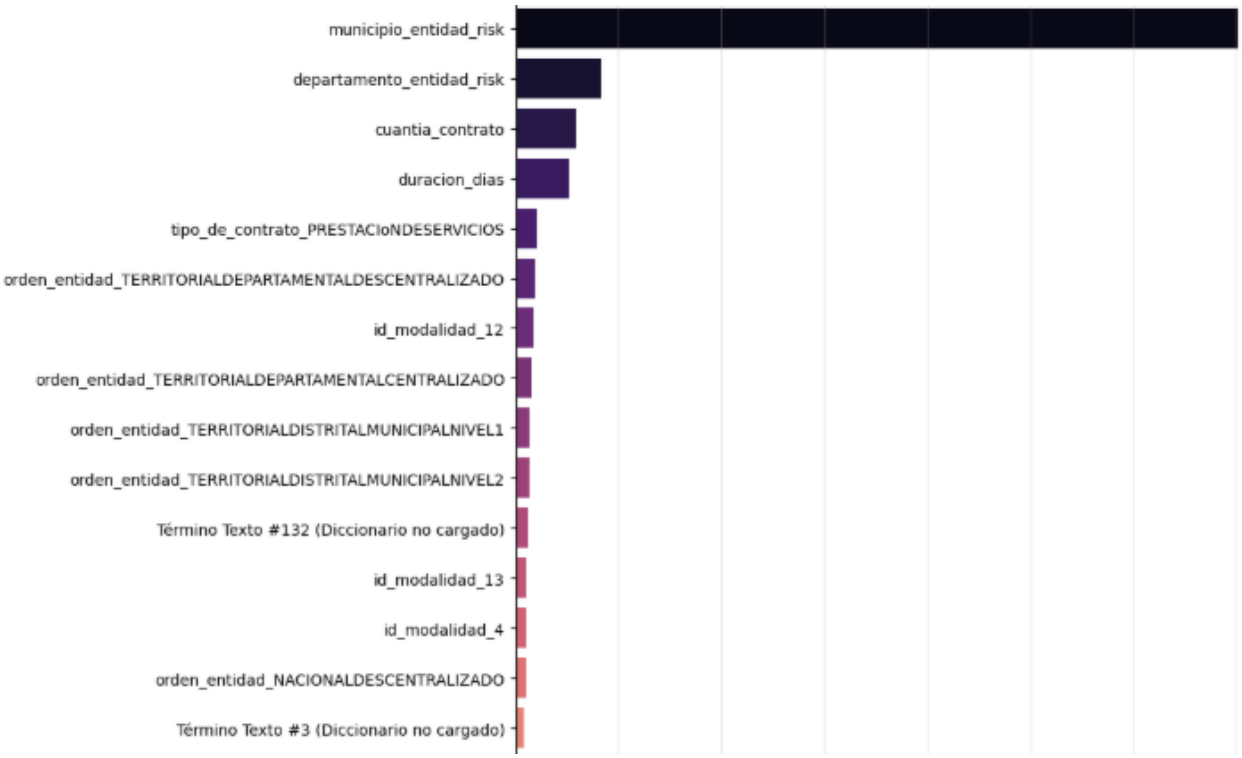

Ilustración 12. Top 10 variables predictoras modelo LightGBM

1. Hegemonía del Riesgo Geográfico (Target Encoding)
El modelo confirma de manera contundente que la ubicación es el predictor más fuerte.

- municipio_entidad_risk (14.0M): Esta variable es, por un orden de magnitud, la más determinante. Al utilizar Target Encoding, el modelo aprendió que el historial de comportamiento de un municipio es el mejor predictor de su futuro, validando la eficacia de esta técnica en variables de alta cardinalidad [14].
- departamento_entidad_risk (1.6M): Refuerza lo anterior a nivel regional. Esto sugiere que el riesgo no es un evento aleatorio, sino sistémico y cultural: depende drásticamente de las prácticas administrativas instaladas en la entidad territorial específica , un patrón coincidente con los estudios sobre riesgos de corrupción en la región [6].

2. La Escala del Proyecto (Dinero y Tiempo)
En segundo nivel de influencia aparecen las variables estructurales:

- cuantia_contrato: El valor económico actúa como una señal de alerta inmediata.
- duracion_dias: El plazo de ejecución es igualmente crítico. Esto valida la hipótesis de que los contratos que se desvían de la norma en magnitud (muy costosos o muy largos, o sospechosamente cortos para su monto) activan patrones de riesgo específicos en el algoritmo.

3. Naturaleza Institucional y Modalidad
Una novedad importante en este modelo final es la alta relevancia del Orden de la Entidad (ej. orden_entidad_TERRITORIALDEPARTAMENTALDESCENTRALIZADO). El algoritmo detectó que el nivel de centralización y la naturaleza jurídica de la entidad (si es una Alcaldía, una Gobernación o un ente descentralizado) influyen significativamente en la probabilidad de éxito o fracaso, probablemente debido a diferencias en capacidad de gestión y vigilancia. Paralelamente, la variable tipo_de_contrato_PRESTACIoNDESERVICIOS se mantiene en el Top 5, confirmando que esta figura contractual específica es un foco crítico de atención.

4. Huellas Textuales (Minería de Texto)
La presencia de variables como Término Texto #132 (derivada del TF-IDF) en el ranking superior valida la utilidad del procesamiento de lenguaje natural. Aunque el peso es menor comparado con la geografía, el modelo logró aislar palabras clave específicas dentro del objeto contractual que actúan como "banderas rojas" semánticas, demostrando que lo que se escribe en el contrato también predice su desenlace y que el análisis semántico es una herramienta viable para la auditoría forense [7].

Conclusión del Análisis
LightGBM ha revelado que el riesgo en el SECOP I tiene una estructura jerárquica clara: Dónde se contrata (Geografía) > Quién contrata (Naturaleza Institucional) > Qué se contrata (Cuantía y Tipo).

Esta jerarquía sugiere que una estrategia de control preventivo eficiente debe priorizar la focalización geográfica e institucional. El modelo no solo predice fraudes, sino que señala que el problema de la corrupción/ineficiencia está fuertemente arraigado en clusters territoriales específicos que el algoritmo ha logrado mapear con precisión.

7.2 Valor Agregado de la Ciencia de Datos frente al Análisis Estadístico Tradicional

Si bien las técnicas estadísticas estándar (como el análisis de correlación bivariado o las regresiones clásicas) son metodológicamente válidas para identificar la relevancia de ciertas variables aisladas (por ejemplo, confirmar empíricamente que la ubicación geográfica o la cuantía están asociadas al fracaso contractual), la implementación de técnicas avanzadas de Ciencia de Datos (CD) y Aprendizaje Automático aporta un valor diferencial y superlativo al proyecto en tres dimensiones críticas:

1. Captura de Interacciones No Lineales y Complejas: Los métodos estadísticos tradicionales suelen evaluar el peso de las variables de forma aislada o asumiendo relaciones lineales estrictas. Por el contrario, los algoritmos basados en árboles de decisión secuenciales (como LightGBM) tienen la capacidad matemática de descubrir interacciones multivariadas profundas de forma automatizada. El modelo no solo identifica que el "Municipio" y el "Tipo de Contrato" son importantes, sino que mapea reglas condicionales complejas (ej., un contrato es altamente riesgoso en el Municipio X, pero únicamente si es de Prestación de

Servicios, su duración es menor a 30 días y la cuantía supera cierto umbral). Estas sutilezas sistémicas son matemáticamente indetectables para los enfoques estándar.

2. Procesamiento de Datos No Estructurados e Hiperdimensionalidad: El análisis tradicional se ve limitado al uso de variables numéricas o categóricas simples. El uso de la Ciencia de Datos permitió trascender estas limitaciones al incorporar técnicas de Procesamiento de Lenguaje Natural (NLP). La extracción de 150 características matemáticas a partir del texto libre del "Objeto a contratar" mediante TF-IDF, sumado al manejo de la altísima cardinalidad geográfica (más de 1.100 municipios) a través del Target Encoding, representa un tratamiento de datos de alta dimensionalidad (196 variables finales) que superaría la capacidad de convergencia y análisis de las herramientas convencionales.

3. Transición de la Explicación Histórica a la Predicción Preventiva (Escalabilidad): El mayor valor agregado radica en el objetivo final de la herramienta. Mientras que las técnicas estándar tienen un enfoque explicativo e inferencial (entender el pasado o hacer auditoría ex-post), el modelo de Machine Learning desarrollado actúa como un artefacto predictivo ex-ante. La Ciencia de Datos permite operacionalizar los hallazgos en un sistema automatizado capaz de ingerir millones de registros nuevos, aplicar el conocimiento aprendido y emitir alertas tempranas de riesgo en cuestión de segundos, garantizando una especificidad superior al 92% que optimiza los recursos de control fiscal y reduce las falsas alarmas.

8 CONCLUSIONES Y TRABAJOS FUTUROS

8.1 CONCLUSIONES

1. El algoritmo LightGBM se consolidó como la solución tecnológica más adecuada para cumplir el objetivo general del proyecto. Este modelo demostró ser la arquitectura más competente para determinar si un contrato finalizará de forma exitosa, alcanzando un AUC-ROC de 0,9732. Este hallazgo constituye el insumo técnico principal para que las entidades de control puedan transitar de modelos de auditoría reactivos a preventivos.

2. Se identificó y realizó el preprocesamiento adecuado que aseguró la calidad predictiva necesaria para este tipo de herramientas:

   - Integridad Temporal: Se implementó un filtrado estricto de data leakage, eliminando variables posteriores a la firma, garantizando que el modelo simule un escenario real de predicción ex-ante.
   - Transformación de Variables: El uso estratégico de Target Encoding para las variables geográficas y de TF-IDF para el texto no estructurado del objeto contractual fue determinante. Estas técnicas permitieron convertir datos categóricos de alta cardinalidad y texto libre en señales numéricas de alto valor predictivo.

3. Se desarrollaron cinco modelos que abarcan diferentes paradigmas algorítmicos:
   - Modelos Lineales: Regresión Logística (Baseline).
   - Modelos Basados en Árboles (Single): Árbol de Decisión.
   - Modelos de Conjunto (Ensemble): Random Forest y LightGBM.
   - Modelos Probabilísticos: Naive Bayes.

   El desarrollo de este portafolio evidenció que la complejidad no lineal es indispensable para este problema. Mientras que modelos como Naive Bayes y Regresión Logística mostraron limitaciones severas (F1 < 0.50), los modelos basados en árboles (Árbol de Decisión, Random Forest y LightGBM) superaron consistentemente el umbral de 0.96 en AUC, confirmando que las reglas de decisión jerárquicas son la mejor aproximación matemática al riesgo contractual.

4. Tras evaluar los resultados, se seleccionó objetivamente a LightGBM como el modelo óptimo, superando a sus competidores en las tres dimensiones críticas de negocio:
   - Sensibilidad (Detección): Logró identificar el 92.82% de los contratos en riesgo, manteniendo una capacidad de detección robusta y estable.

- Especificidad (Reducción de Ruido): Su mayor aporte fue la reducción de la Tasa de Falsas Alarmas al 7.55% (el mínimo histórico del estudio). Aunque el Random Forest optimizado mejoró su desempeño (bajando al 9.44%), LightGBM sigue siendo superior filtrando alertas incorrectas, lo que optimiza los recursos de auditoría humana.
- Velocidad y Eficiencia: Con un tiempo de entrenamiento de 60.68 segundos, resultó ser cinco veces más rápido que el Random Forest (313 segundos). Esto demuestra la viabilidad técnica de LightGBM para procesar grandes volúmenes de datos históricos en entornos productivos, donde el Random Forest sería costoso computacionalmente.

5. El análisis de importancia de features del modelo LightGBM genera una hoja de ruta clara para la gestión de riesgos:

    1. **Enfoque en el Riesgo Geográfico (Máxima Prioridad):**
    La evidencia es concluyente: los factores de localización (municipio y departamento) son los principales impulsores del fracaso contractual. Se recomienda la creación de un Índice de Alerta Geográfica que clasifique automáticamente los contratos provenientes de ubicaciones con bajas tasas históricas de éxito como Riesgo Alto, exigiendo una debida diligencia intensificada desde la fase de planeación.

    2. **Naturaleza Institucional**: Se identificó que el Orden de la Entidad (Nacional vs. Territorial/Descentralizado) es un predictor clave, sugiriendo diferencias estructurales en la capacidad de gestión entre niveles de gobierno.

    3. **Protocolo para la cuantía y modalidad**: También en un segundo nivel se encuentran la cuantia_contrato y la duracion_dias. Los contratos que se desvían de la norma económica (muy costosos o atípicamente baratos) configuran perfiles de alerta inmediata.

8.2 TRABAJOS FUTUROS

1. Extensión del Alcance Predictivo (Clasificación Multi-Clase)

Sugerencia: Desarrollar un modelo de clasificación multi-clase para predecir no solo si el contrato fracasará, sino cómo fracasará.

Clases de Salida: [Liquidado Exitoso], [Terminado Anormalmente], [Descargado/Anulado], [Terminado Sin Liquidar].

Justificación: El modelo actual solo distingue éxito vs. fracaso. Para las entidades sería de ayuda tener más “pistas” si el fracaso se debe a una terminación por incumplimiento (TERMINADO ANORMALMENTE) —que implica litigio y penalidades— o a un simple descarte administrativo (DESCARTADO). Predecir el tipo específico de riesgo permite asignar recursos legales y de gestión de manera más precisa.

2. Mayor Interpretabilidad a Nivel Local (SHAP Values)

Sugerencia: Implementar técnicas de explicabilidad post-hoc como los SHAP (SHapley Additive exPlanations) Values en el modelo LightGBM.

Justificación: El análisis actual identifica las variables más importantes a nivel global (ej. municipio_entidad). El siguiente paso es la interpretabilidad local. Con SHAP, se podría generar una "tarjeta de puntuación de riesgo" por cada nuevo contrato, indicando: "Este contrato tiene una probabilidad del 8% de fracaso debido a que el Plazo de Ejecución es un 30% más largo que el promedio histórico para ese Municipio". Esto daría a los gestores contractuales de información accionable en tiempo real.

3. Ingeniería de Features Temporales y de Monitoreo

Sugerencia: Enriquecer el set de datos con nuevas features dinámicas que cambian durante el ciclo de vida del contrato.

Variables a Explorar:

Variables de Licitación: Número de oferentes, desviación de la oferta ganadora respecto al presupuesto oficial, tiempo entre la publicación y la adjudicación.

Features de Riesgo de Contratista: Historial de incumplimiento previo del contratista (si esa información está disponible y es viable de usar).

Justificación: Las features actuales son estáticas (conocidas al inicio). Las variables dinámicas,

como el comportamiento atípico durante la licitación, podrían actuar como señales de advertencia tempranas incluso más poderosas que las variables geográficas y administrativas, mejorando la robustez del modelo.

4. Detección de Colusión y Patrones Atípicos

Sugerencia: Aplicar modelos de detección de anomalías no supervisada (como Isolation Forest o Autoencoders) sobre los datos de contratación.

Justificación: Un contrato que finaliza con éxito, pero con características altamente inusuales (ej. un sobrecosto no justificado, una licitación con un único oferente que gana por un margen mínimo y en un municipio de alto riesgo), podría ser un indicio de colusión o fraude. Un modelo de detección de anomalías puede identificar estos patrones de "éxito sospechoso" que el modelo predictivo tradicional ignora.

5. Optimización Presupuestaria y Asignación de Recursos

Sugerencia: Transformar la salida predictiva del modelo en una herramienta de optimización de costos.

Justificación: El modelo ya identifica los FN (contratos que posiblemente tendrán dificultades). El trabajo futuro consistiría en calcular el costo esperado de fracaso (pérdidas, litigios) para cada FN y luego asignar los recursos de supervisión y gestión al conjunto de contratos que maximicen el ahorro potencial por cada unidad de esfuerzo invertida. Esto convierte el modelo de riesgo en una herramienta de gestión de portafolio y eficiencia presupuestaria.

6. Escalabilidad y Reproducibilidad en Otros Ecosistemas Contractuales

Sugerencia: Extrapolar la arquitectura del pipeline de Machine Learning (procesamiento NLP, codificación de variables y el algoritmo LightGBM) hacia plataformas más modernas como SECOP II, o hacia la gestión de proveedores en el sector privado.

Justificación: El marco metodológico desarrollado es agnóstico a los datos específicos de SECOP I. Dado que los pilares predictivos identificados (cuantías, plazos, historiales geográficos y objetos contractuales) son universales en cualquier relación contractual, este modelo puede ser reentrenado fácilmente para predecir riesgos en otros dominios. Esto incluye la auditoría preventiva de subsidios, la adjudicación de licencias o la evaluación de la cadena de suministro corporativa, demostrando que la solución construida es altamente transferible y reproducible para cualquier ecosistema que requiera vigilancia contractual continua.

REFERENCIAS BIBLIOGRÁFICAS

[1] Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia, "Portal de Datos Abiertos de Colombia," Datos Abiertos. [En línea]. Disponible en: https://www.datos.gov.co/Estad-sticas-Nacionales/SECOP-I-Procesos-de-Compra-P-blica/f789-7hwg/about_data. [Accedido: 09-ene-2026].

[2] Colombia Compra Eficiente, "Portal de contratación pública de Colombia: SECOP," Agencia Nacional de Contratación Pública. [En línea]. Disponible en: https://www.colombiacompra.gov.co/secop/secop-i. [Accedido: 09-ene-2026].

[3] S. Rodríguez Arévalo, "Predicción de ineficiencias en la contratación pública de Bogotá," Tesis de Maestría, Universidad del Rosario, Bogotá, Colombia, 2020.

[4] J. E. Pérez Lara, "Analítica de datos y gobierno abierto: hacia una gestión pública basada en evidencia," Gestión y Análisis de Políticas Públicas, no. 36, 2025.

[5] Congreso de la República de Colombia, "Ley 1712 de 2014: Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional," Diario Oficial No. 49.084, 2014.

[6] Organización de los Estados Americanos (OEA) y CAF, "Guía para la identificación de riesgos de corrupción en contratación pública, utilizando la ciencia de datos," Secretaría General de la OEA, Washington, D.C., 2021.

[7] Coalition for Integrity, "Using machine learning for anti-corruption risk and compliance," 2021. [En línea]. Disponible en: https://www.coalitionforintegrity.org/wp-content/uploads/2021/04/Using-Machine-Learning-for-Anti-Corruption-Risk-and-Compliance.pdf. [Accedido: 09-ene-2026].

[8] M. Kehler, J. Paciello, and J. Pane, "Anomaly Detection in Public Procurements using the Open Contracting Data Standard," Universidad Nacional de Asunción, Paraguay, 2021.

[9] A. C. Müller and S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists. Sebastopol, CA: O'Reilly Media, 2017.

[10] A. Géron, Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow: Conceptos, herramientas y técnicas para conseguir sistemas inteligentes, 2nd ed. Madrid: Anaya Multimedia, 2020.

[11] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA: Morgan Kaufmann, 2016.

[12] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA: Morgan Kaufmann, 2011.

[13] C. C. Aggarwal, Outlier Analysis, 2nd ed. New York, NY: Springer, 2017.

[14] D. Micci-Barreca, "A preprocessing scheme for high-cardinality categorical attributes in classification designs," ACM SIGKDD Explorations Newsletter, vol. 3, no. 1, pp. 27-32, 2001.

[15] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Victoria, BC: Leanpub, 2020.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[17] G. Lemaître, F. Nogueira, and C. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," Journal of Machine Learning Research, vol. 18, no. 17, pp. 1-5, 2017

[18] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, vol. 30, 2017.

ANEXOS

ANEXO 2. DICCIONARIO DE DATOS ABIERTOS SECOP I

Descripción: El presente anexo hace referencia al documento oficial de metadatos proporcionado por la Agencia Nacional de Contratación Pública, el cual describe la estructura de las variables contenidas en el conjunto de datos SECOP_I_Procesos_de_Compra_Publica.

Fuente: Agencia Nacional de Contratación Pública - Colombia Compra Eficiente. (2022). Diccionario de Datos Abiertos SECOP I - Procesos de Compra Pública. Ministerio de Tecnologías de la Información y las Comunicaciones.

Disponibilidad: El documento se encuentra disponible para consulta pública en el siguiente enlace: https://www.datos.gov.co/Estad-sticas-Nacionales/SECOP-I-Procesos-de-Compra-P-blica/f789-7hwg/about_data

Nota: Teniendo en cuenta que este diccionario puede ser actualizado periódicamente, una copia de este archivo en formato PDF con la versión vigente a julio de 2025 se incluye en la carpeta de material suplementario de este proyecto bajo el nombre:
“Anexo 1. Diccionario_de_Datos_Abiertos_SECOP I.pdf”