



Detección de contratos públicos susceptibles a adiciones.

Oscar Otalora, Camilo Céspedes, Julián Gomez, Juan David Caballero.

Asesor: Juan Fernando Pérez.

**Universidad de los Andes
Faculta de Ingeniería
Maestría en Inteligencia Analítica Para la Toma de Decisiones
Bogotá D.C., Colombia**

2023

Resumen

La contratación pública es un mecanismo crucial en los países ya que permite satisfacer las necesidades de la población y ejecutar planes gubernamentales. Actualmente en Colombia, los procesos de contratación pública están centralizados en una plataforma en línea (SECOP) que pretende asegurar la trazabilidad y la transparencia de la gestión contractual. Sin embargo, en la etapa contractual se generan modificaciones llamadas adiciones (Valor, Tiempo) asociadas a ineficiencias o incluso irregularidades que se traducen en sobrecostos para el gobierno. En este proyecto se compararon 66 modelos de Machine Learning (XGBOOST, Random Forest, Redes Neuronales, Regresiones) que buscan identificar la probabilidad de que los contratos incurran en modificaciones durante la fase contractual, mediante el análisis de datos tanto estructurados como no estructurados disponibles en la fase precontractual. Los resultados muestran que el modelo XGBOOST es el que mejor identifica los contratos con adiciones con una precisión del 74% y recall del 66%

Palabras claves: Contratación Pública, Transparencia, Adiciones, Clasificación, Machine Learning, Modelos predictivos, Redes Neuronales.

Abstract

Public procurement is a crucial mechanism in countries as it allows meeting the population's needs and executing government plans. Currently in Colombia, public procurement processes are centralized on an online platform (SECOP) that aims to ensure traceability and transparency in contract management. However, during the contractual stage, modifications known as "additions" (Value, Time) are generated due to inefficiencies or even irregularities, resulting in cost overruns for the government. This project compared 66 Machine Learning models (XGBOOST, Random Forest, Neural Networks, Regressions) that seek to identify the probability of contracts incurring modifications during the contractual phase by analyzing both structured and unstructured data available in the pre-contractual phase. The results show that the XGBOOST model is the one that best identifies contracts with additions with a precision of 74% and a recall of 66%.

Keywords: Public Procurement, Transparency, Additions, Machine Learning, Predictive Models, Neural Networks.

Tabla de Contenido

1	Introducción	6
1.1	Planteamiento Del Problema	8
1.2	Objetivos.....	8
1.2.1	Objetivo General.....	8
1.2.2	Objetivos Específicos.....	9
2	Marco Teórico.....	9
2.1	Machine Learning.	9
2.2	Procesamiento Del Lenguaje Natural (NLP).....	10
2.3	Tokenización.....	10
2.4	Transformers.	10
2.5	Term Frequency - Inverse Document Frequency (TF-IDF).....	10
2.6	K-Means.	11
2.7	Regresión Logística RL.....	11
2.8	Random Forest (RF).	12
2.9	Support Vector Classifier (SVC).....	12
2.10	Extreme Gradient Boosting (XGBoost).....	12
2.11	Shap Values	12
2.12	Redes Neuronales.	13
3	Metodología.	13
3.1	Extracción De Datos	14
3.2	Preprocesamiento De Datos Y Limpieza.....	14
3.3	Definición Variable Objetivo	16
3.4	Análisis Descriptivo Y Exploratorio.....	20
3.4.1	No Prestación De Servicios.	20
3.4.2	Prestación De Servicios.....	22
3.5	Ingeniería De Características.....	24
3.5.1	Categorizaciones.....	27
3.6	Aplicación Modelos	28
4	Resultados Y Análisis.	30
4.1	Herramienta Dashboard.....	31
5	Conclusiones.	32
6	Trabajo Futuro.	33
7	Referencias.....	33
8	Anexos.....	35

8.1	Anexo 1. Tablas.....	35
8.2	Anexo 2. Figuras.....	45
8.3	Anexo 3. Categorizaciones.	47
8.4	Anexo 4. Ilustración de la Herramienta de cálculo de probabilidad en Dash.	51

Lista de Tablas

Tabla 1.	Tipos de contratos.....	15
Tabla 2.	Participación tipos de adición	17
Tabla 3.	Descripción de los contratos marcados como adición.	19
Tabla 4.	Descripción contratos mal redactados.....	20
Tabla 5.	Tabla de contingencia No Prestación	22
Tabla 6.	Tabla de contingencia Prestación.....	24
Tabla 7.	Identificador Clusters.....	26
Tabla 8.	Nacionalidad Representante legal.....	27
Tabla 9.	Tabla de contingencia de la variable contra la variable objetivo	28
Tabla 10.	Variables Significativas y sus Coeficientes	36
Tabla 11.	Valor del contrato No Prestación	36
Tabla 12.	Recursos Propios No Prestación.....	36
Tabla 13.	Saldo CDP No Prestación	36
Tabla 14.	Distribución Adicion / No adicion por sector No Prestación	37
Tabla 15.	Rama No Prestación	37
Tabla 16.	Tipo de Contrato No Prestación	38
Tabla 17.	Modalidad de Contratación No Prestación	38
Tabla 18.	Entidad Centralizada No Prestación.....	39
Tabla 19.	Destino del gasto No Prestación	39
Tabla 20.	Valor del contrato Prestación	39
Tabla 21.	Recursos Propios Prestación	39
Tabla 22.	Saldo CDP Prestación.....	39
Tabla 23.	Distribución Adicion / No adicion por sector Prestación.....	40
Tabla 24.	Rama Prestación.....	40
Tabla 25.	Tipo de Contrato Prestación.....	40
Tabla 26.	Modalidad de Contratación Prestación.....	41
Tabla 27.	Entidad Centralizada Prestación	41
Tabla 28.	Destino del Gasto Prestación.	41
Tabla 29.	Agrupación etiquetas grupo entidad por expresiones regulares.	42
Tabla 30.	Clasificación Tipo entidad Expresiones regulares.....	43
Tabla 31.	Hiperparametros modelos no prestación servicios	43
Tabla 32.	Hiperparametros modelos prestación servicios.	44
Tabla 33.	Mejores resultados No prestación	44
Tabla 34.	Mejores resultados Prestación.	45
Tabla 35.	Categorización Valor del contrato.....	47
Tabla 36.	Deciles valor del contrato prestación	48

Lista de Figuras

Ilustración 1. Metodología detección de contratos.....	13
Ilustración 2. Porcentaje de datos nulos por columna	15
Ilustración 3. Conteo por fecha de la firma del contrato.....	16
Ilustración 4. Precisión del modelo Transformers.....	18
Ilustración 5. Proceso identificación de adiciones	18
Ilustración 6. Variables continuas relevantes No Prestación(Valor contrato, Recursos propios, Saldo CDP).....	21
Ilustración 7. Variables continuas relevantes Prestación (Valor contrato, Recursos propios, Saldo CDP).....	23
Ilustración 8. Método del Codo.....	25
Ilustración 9. Nube de palabras presentes en el objeto del contrato Cluster 1.....	26
Ilustración 10. Desempeño de los modelos No Prestación.....	29
Ilustración 11. Desempeño de los modelos Prestación	30
Ilustración 12. Curva ROC No Prestación	30
Ilustración 13. Curva ROC Prestación.....	30
Ilustración 14. Shap Values No Prestación	30
Ilustración 15. Shap Values Prestación.....	30
Ilustración 16. Tipos de contratos.....	45
Ilustración 17. Método del codo selección de número de Clusters	46
Ilustración 18. Score Silhouette.....	46
Ilustración 19. Davis-Bouldin Index	46
Ilustración 20. Nube de palabras presentes en el objeto del contrato Cluster 1,2,3 respectivamente.....	46
Ilustración 21. Distribución recursos Propios por Adición	47
Ilustración 22. Distribución recursos Propios (Alcaldías, Gobernaciones) por Adición.....	47
Ilustración 23. Presupuesto General de la Nación por Adición	48
Ilustración 24. Distribución de recursos propios por Adición Prestación.	49
Ilustración 25. Distribución PGN por adición Prestación.....	49
Ilustración 26. Matriz de confusión No prestación	49
Ilustración 27. Matriz de confusión prestación.....	49
Ilustración 28. Matriz de confusión ajustada No prestación.	50
Ilustración 29. Matriz de confusión ajustada prestación.	50

1 Introducción

La contratación pública es un mecanismo crucial en los Estados, ya que permite satisfacer las necesidades de la población y garantizar la implementación de los planes gubernamentales. Este mecanismo, empleado por el Estado, se convierte en una herramienta esencial para la adquisición de bienes, servicios y obras, facilitando así el cumplimiento de los objetivos y metas del país (OCDE, s.f.).

Existen cinco modalidades de contratación pública en Colombia: contratación directa, licitación pública, selección abreviada, concurso de méritos y mínima cuantía. Estas modalidades se eligen de acuerdo con los requisitos del Estado, los cuales se detallan en un documento denominado pliego de condiciones (Colombia, 2020). Para cargar y supervisar todo el proceso de contratación, se utiliza el Sistema Electrónico para la Contratación Pública, conocido como SECOP. Este sistema se divide en dos partes: SECOP I y SECOP II. La primera es una plataforma exclusiva para la publicidad, donde se realizan las publicaciones de los documentos del proceso. La segunda, SECOP II, es una plataforma transaccional encargada de gestionar los contratos, donde cualquier persona externa puede hacer un seguimiento de estos contratos, contribuyendo así a la transparencia de la información (CompraColombia, 2021).

Los contratos estatales se dividen en tres etapas fundamentales: la precontractual, que engloba la fase de planificación, la invitación a los oferentes y la selección del contratista; la etapa contractual, durante la cual se lleva a cabo la ejecución y cumplimiento del contrato; y finalmente, la etapa postcontractual, que implica la liquidación del contrato. En esta última fase, se realiza un exhaustivo análisis económico, jurídico y técnico de lo ejecutado, cerrando así el ciclo del contrato estatal (Legis, 2021).

En ocasiones, cuando el contrato se encuentra en la etapa contractual, el contratista solicita una adición de un bien o una actividad que no se tuvo en cuenta dentro de la etapa inicial (CompraEficiente, 2022). Las adiciones se pueden generar en valor y prorrogar en tiempo, en este caso no debe superar el 50% del valor inicial del contrato (ServicioCivilDistrital, 2022).

Los contratos públicos con adiciones generan un sobrecosto en el presupuesto del gobierno que afecta el desarrollo de las actividades planeadas en función de las necesidades de la población en Colombia, tanto a nivel local como a nivel nacional, por tanto, es fundamental evitar cualquier necesidad adicional.

Hacer uso del análisis de datos y de técnicas de machine learning representan una herramienta valiosa para facilitar la identificación de adiciones en una etapa oportuna que permita gestionar los recursos públicos con mayor eficacia y transparencia, mejorando el proceso y ejecución del presupuesto.

El procedimiento para el entrenamiento discurre de manera típica para el entrenamiento de modelos de Machine Learning, utilizando el poder de los Pipelines que facilitan tareas de preprocesamiento de los datos y puesta a punto para ejecutar los modelos que requieren únicamente matrices numéricas (como las redes neuronales). Se extrae la información de la página de datos abiertos donde se encuentra el SECOP II con todos los contratos del Estado en sus diferentes etapas y sus adiciones (Colombia Compra Eficiente, 2023). Luego, se ejecuta un pre-procesamiento de datos dentro del Pipeline, se tiene la opción de aplicar selección de variables y culmina con el entrenamiento del modelo. En este punto el modelo es iterativo, comparando algoritmos como Regresión Logística, Random Forest, Support Vector Classifier (SVC) y Extreme Gradient Boosting (XGBoost), que, por método de grilla, busca el algoritmo óptimo con sus mejores hiperparámetros. Este proceso se complementa con la implementación de validación cruzada para mejorar sus resultados.

El análisis se da sobre contratos en la ciudad de Bogotá D.C., al tener el mayor porcentaje de contratos del Estado. Luego la exploración de los datos lleva a segmentar el modelamiento para contratos de prestación de servicios y de no prestación. Aquí se presenta un reto para contratos de no prestación, con características de contratos de oferta de servicios personales profesionales. Su diferencia con los demás es que aquellos requieren de otras actividades que no son prestadas por persona nombrada. Se destaca algunos casos específicos de régimen especial que representan aproximadamente el 93% según la Ilustración 16 del [Anexo 2](#) versus los de no prestación de servicios. El tercer paso es definir la variable objetivo, en este caso se define cómo adición aquellos contratos con valor y prórroga al aplicar estos filtros en el tipo de adición o en otros casos se procede a utilizar inferencia con un modelo de clasificación BETO, que es un modelo BERT pero entrenado en español por la Universidad de Chile (Cañete et. al., 2020).

La variable "Objeto del contrato", la cual contiene un texto detallado sobre las tareas específicas que justifican la necesidad del contrato, Mediante el uso de técnicas de procesamiento de lenguaje natural (NLP) y métodos de agrupación, se lleva a cabo una ingeniería de características para extraer información relevante de estas descripciones.

Una vez recopiladas y procesadas todas las características relevantes, se procede a aplicar diversos modelos de aprendizaje automático con el objetivo de clasificar si un contrato es susceptible de tener adiciones, tanto en el caso de contratos de prestación de servicios y similares como en contratos de no prestación. Durante esta fase, se lleva a cabo un proceso de validación cruzada, ajustando distintos hiperparámetros y seleccionando las características más pertinentes para cada modelo. Este proceso revela que el modelo más efectivo para la clasificación en el caso de contratos de no prestación es el XGBoost, el cual logra una precisión del 76% y un Área Bajo la Curva (AUC) del 88%. De manera similar, para los contratos de prestación de servicios, el XGBoost también se destaca como el modelo óptimo, alcanzando una precisión del 85% y un AUC del 92%.

Al analizar los resultados, se identifican cuáles son las variables que impactan en mayor medida que un contrato de cualquiera de las segmentaciones establecidas (con y sin prestación) tenga adición:

- Los contratos con descripciones más detalladas tienen una mayor probabilidad de adición.

- Las entidades que cuentan con recursos propios suelen tener un menor número de adiciones.
- Tener recursos a crédito puede generar un mayor número de adiciones.
- Cuando el valor del contrato es más alto es menos probable que se tengan adiciones.
 - Para prestación de servicio las variables de mayor impacto son similares, lo que se debe resaltar, es que en algunas el impacto positivo o negativo difiere con respecto a no prestación, por ejemplo, los contratos con un mayor valor presentan probabilidades más altas de adición. Sin embargo, para prestación la variable Sector tiene una mayor relevancia a diferencia de no prestación.
- Cuando parte del dinero para la ejecución es de la nación, baja la probabilidad de adición.

Por último, se desarrolla un Dashboard ([Anexo 4](#)) diseñado para que las entidades involucradas puedan ingresar todas las variables precontractuales de un contrato específico. Utilizando el modelo XGBoost previamente entrenado, así como otros modelos para el pre-procesamiento (ej. cluster de objeto del contrato, regex del nombre de las entidades,) se estima el riesgo de tener adiciones futuras en el contrato. Con este indicador las entidades distritales de Bogotá pueden identificar las oportunidades de mejora de sus contratos para reducir la probabilidad de adiciones futuras que sean innecesarias.

En situaciones donde la probabilidad de adición es media o baja, las entidades pueden optar por asignar menos recursos de control de manera proporcional, contribuyendo así a la reducción del número total de adiciones. La implementación eficaz de esta herramienta tiene un potencial significativo de ahorro. Este ahorro no solo optimiza la gestión de recursos, sino que también mejora la eficiencia y la transparencia en el proceso de contratación.

1.1 Planteamiento Del Problema

Los contratos públicos con adiciones generan un sobrecosto en el presupuesto del gobierno que afecta el desarrollo de las actividades planeadas en función de las necesidades de la población en Colombia. Si existiera una manera de estimar qué contratos en su etapa precontractual pueden ser propensos a adiciones, podría abrirse la discusión sobre qué factores influyen en esta problemática y cómo mitigarlos en un futuro. Por lo tanto, la pregunta que se busca responder es, ¿Qué contratos tienen una alta probabilidad de tener adiciones con su información disponible durante la etapa precontractual y cuáles son los factores que influyen en ésta?

1.2 Objetivos

1.2.1 Objetivo General

Generar un instrumento para la toma de decisiones donde se identifique la probabilidad de que los contratos incurran en modificaciones durante la fase contractual, mediante el análisis de datos tanto estructurados como no estructurados. Esto con el fin de proporcionar a entidades gubernamentales, entes de control y a cualquier particular, una comprensión detallada de las posibles razones que llevan a estas extensiones contractuales y puedan generar propuestas para mitigar estas ineficiencias en la contratación estatal.

1.2.2 Objetivos Específicos

- Obtener la información del SECOP II y sus adiciones en la plataforma de datos abiertos.
- Generar un análisis exploratorio con la información.
- Generación de características, infiriendo información de textos relacionados con los contratos.
- Modelamiento de la probabilidad de que un contrato tenga adiciones en su etapa contractual.
- Interpretación de las características que argumentan la probabilidad resultante.
- Creación de herramienta que visualice la probabilidad, las características y además el ajuste interactivo de valores de variables, que muestren el impacto estimado en probabilidad.

2 Marco Teórico

A lo largo de esta sección abordamos de manera general los conceptos metodológicos que se aplican en las diferentes etapas del desarrollo de la herramienta.

2.1 Machine Learning.

El Machine Learning (ML) pertenece a la inteligencia artificial y permite a los sistemas aprender patrones a partir de datos y realizar tareas sin una programación explícita. El ML abarca una variedad de enfoques, desde algoritmos clásicos hasta técnicas más avanzadas como el aprendizaje profundo, sus usos abarcan desde clasificación, regresión, hasta clustering y otras tareas (James et al., 2013).

Dentro del ML, se destacan especialmente los modelos supervisados, que aprenden de datos etiquetados para realizar predicciones o clasificaciones en nuevos conjuntos de datos (James et al., 2013). Estos modelos se pueden aplicar para prever la probabilidad de adiciones contractuales y analizar los factores que interfieren en ella.

Existen varios algoritmos supervisados que podrían usarse en la estimación de esta probabilidad. Al ser, en general, un problema de clasificación, se contemplan metodologías como el procesamiento del lenguaje natural (NLP), Regresión Logística (RL), Random Forest (RF), Support Vector Machine (SVM), XGBoost, Redes Neuronales (RN) y algunas otras adicionales que pueden solucionar algunas etapas intermedias.

2.2 Procesamiento Del Lenguaje Natural (NLP).

El Procesamiento de Lenguaje Natural (NLP) hace parte de la inteligencia artificial y se enfoca en la interacción entre las computadoras y el lenguaje humano (Jurafsky & Martin, 2019). Para el propósito del análisis de adiciones en contratación pública en Colombia, el NLP puede ser útil para analizar documentos contractuales, identificar patrones y extraer información relevante.

El NLP abarca una variedad de fases y tareas, desde aspectos básicos como tokenización y análisis morfológico hasta tareas más complejas como la generación de texto y la comprensión del lenguaje natural.

2.3 Tokenización.

La tokenización es un paso importante en el procesamiento de lenguaje natural (NLP) que consiste en dividir un texto en partes más pequeñas llamadas "tokens". Estos tokens pueden ser palabras, frases, o unidades menores, dependiendo del nivel de granularidad deseado. La tokenización es un paso esencial para comprender y analizar el lenguaje humano de manera efectiva (Jurafsky & Martin, 2019).

En NLP, un token es la unidad mínima de procesamiento. La elección de la definición de token afecta directamente la representación del texto y, por lo tanto, la calidad del análisis subsiguiente (Jurafsky & Martin, 2019).

Algunos enfoques de tokenización emplean modelos de ML, como redes neuronales, para aprender patrones. Estos modelos pueden adaptarse mejor a la variabilidad del lenguaje natural.

2.4 Transformers.

En 2017 investigadores de Google publicaron un paper (Ashish Vaswani, 2017) donde dieron a conocer los Transformers, que es una red neuronal que puede aprender del contexto con base en las relaciones que existen entre palabras, esto utilizando técnicas matemáticas que denominaron mecanismo de atención. A diferencia de los modelos anteriores basados en RNNs o LSTMs, los Transformers no procesan los datos secuencialmente, sino que procesan toda la secuencia de entrada a la vez. Este enfoque permite al modelo centrarse en diferentes partes de la secuencia de entrada para realizar una tarea, lo que es especialmente útil para entender el contexto y las relaciones entre palabras en una oración (MERRITT, 2022)

2.5 Term Frequency - Inverse Document Frequency (TF-IDF).

Es una técnica utilizada en procesamiento de lenguaje natural y recuperación de información para evaluar la importancia relativa de una palabra en un documento dentro de un conjunto de documentos o corpus. TF-IDF se utiliza comúnmente en motores de búsqueda,

clasificación de texto, resumen automático y otras aplicaciones de procesamiento de texto. (Rodríguez, 2023)

TF (Frecuencia del Término): Esta parte de TF-IDF mide la frecuencia con la que una palabra específica aparece en un documento. Cuanto más frecuente sea una palabra en un documento, mayor será su valor TF para ese documento. TF se calcula generalmente de la siguiente manera:

$$TF_{t,d} = \frac{\text{Número de veces que la palabra } t \text{ aparece en el documento } d}{\text{Número total de palabras en el documento } d}$$

IDF (Frecuencia Inversa del Documento): Esta parte de TF-IDF mide la importancia relativa de una palabra en todo el conjunto de documentos. Cuanto menos común sea una palabra en el conjunto de documentos, mayor será su valor IDF. IDF se calcula generalmente de la siguiente manera:

$$IDF_t = \log\left(\frac{\text{Número total de documentos}}{\text{Número de documentos que contienen la palabra } t + 1}\right)$$

Finalmente, **TF-IDF** combina la frecuencia del término y la frecuencia inversa del documento, de la siguiente manera:

$$TF - IDF_{t,d} = TF_{t,d} * IDF_{t,d}$$

2.6 K-Means.

Es un método de agrupación que tiene como objetivo dividir las observaciones en grupos en los que cada observación pertenece al grupo cuyo valor medio es el más cercano. El algoritmo comienza seleccionando puntos al azar como centros iniciales de grupos (centroides), luego asigna cada punto de datos al grupo con el centroide más cercano, utilizando la distancia euclidiana. Posteriormente, recalcula los centroides de cada grupo tomando el promedio de todos los puntos asignados a ese grupo. Este proceso de asignación y actualización se repite hasta que los centroides se estabilizan, indicando la convergencia y estabilidad de los grupos. Al final, K-means resulta en una agrupación de los datos en un número limitado de grupos, proporcionando los centroides finales de cada uno. (Ramírez, 2023).

2.7 Regresión Logística RL.

Es un método estadístico para analizar un conjunto de datos en el que hay una o más variables independientes que determinan un resultado. El resultado se mide con una variable dicotómica; en otras palabras, sólo puede tener dos resultados posibles. Este tipo de regresión es útil para situaciones donde el resultado a predecir es categórico, como 'sí/no' o 'éxito/fracaso'. Uno de los aspectos clave de la regresión logística es el uso de la función logística. Esta función convierte una entrada lineal en un valor entre 0 y 1, lo que facilita la interpretación de los resultados como probabilidades. Además, la regresión logística puede

incluir múltiples variables independientes, tanto categóricas como numéricas, lo que la hace muy versátil para el análisis estadístico. (Dalgaard, 2008)

2.8 Random Forest (RF).

Es un algoritmo de aprendizaje conjunto para clasificación, regresión y otras tareas, funciona mediante la creación de múltiples árboles de decisión en el momento del entrenamiento y combina las predicciones de estos árboles para realizar una predicción más precisa que cualquier modelo individual.

Este enfoque tiene varias ventajas, incluyendo un mayor rendimiento en términos de precisión predictiva y una mayor capacidad para manejar conjuntos de datos con grandes dimensiones de entrada. Además, Random Forest es eficaz para estimar la importancia de las variables y tiene la capacidad de modelar interacciones complejas entre variables. (Tan et al., 2019, #)

2.9 Support Vector Classifier (SVC).

Las máquinas de vectores de soporte (SVM) se identifican por hallar un hiperplano que mejor separa las clases de datos en un espacio de características de alta dimensión, el hiperplano se selecciona con una función de optimización que maximiza el margen entre las clases de los datos, este enfoque es práctico al dar una solución robusta y eficiente en los problemas de clasificación, especialmente en situaciones donde la separación entre las clases no es claramente definida. (Hastie et al., 2009, #)

2.10 Extreme Gradient Boosting (XGBoost)

Es una implementación eficiente y escalable del algoritmo de boosting de gradientes. XGBoost se destaca por su capacidad para manejar grandes conjuntos de datos, su velocidad y eficiencia en la ejecución, y por ofrecer mejoras en la precisión de los modelos. El algoritmo utiliza árboles de decisión y optimiza tanto la función de pérdida como la regularización, lo cual ayuda a evitar el sobreajuste. (Tianqi Chen & Carlos Guestrin, 2016).

2.11 Shap Values

Los valores SHAP, derivados de la teoría de juegos cooperativos, son una técnica avanzada en el aprendizaje automático para la interpretación de modelos. Basándose en los valores de Shapley, proporcionan una explicación detallada de cómo cada característica de un conjunto de datos contribuye a la predicción de un modelo para una instancia específica. Estos valores son particularmente útiles por su capacidad para ofrecer explicaciones a nivel de instancia, lo que permite entender las decisiones individuales tomadas por un modelo. La suma de los valores SHAP de todas las características, junto con el valor base del modelo, iguala la predicción del modelo, ofreciendo así un desglose aditivo y coherente. Se destacan por su aplicabilidad a una amplia gama de modelos, desde árboles de decisión hasta redes

neuronales profundas, y se pueden visualizar a través de varios métodos, facilitando la interpretación y comprensión de modelos complejos. En esencia, los valores SHAP son una herramienta valiosa para desmitificar las decisiones de los modelos de aprendizaje automático, promoviendo una mayor transparencia y confianza en las predicciones basadas en datos. (Molnar, 2023)

2.12 Redes Neuronales.

Es un enfoque del aprendizaje automático inspirado en la forma en que funciona el cerebro del humano, estas redes consisten en capas de nodos que están interconectadas similar a como son las neuronas del cerebro. Cada neurona recibe entradas, las procesa y pasa su salida a otras neuronas. El aprendizaje se realiza ajustando los pesos de las conexiones entre las neuronas, lo cual se hace típicamente a través de un proceso conocido como retro propagación y optimización mediante algoritmos como el descenso de gradiente. Este enfoque permite a las redes neuronales aprender a realizar tareas complejas, como reconocimiento de imágenes, procesamiento del lenguaje natural y tareas de regresión o clasificación. (Nielsen, 2015)

3 Metodología.

En este apartado se aborda todo el tratamiento metodológico utilizado para el desarrollo del instrumento que estima la probabilidad de que un contrato pueda ser susceptible de adiciones.

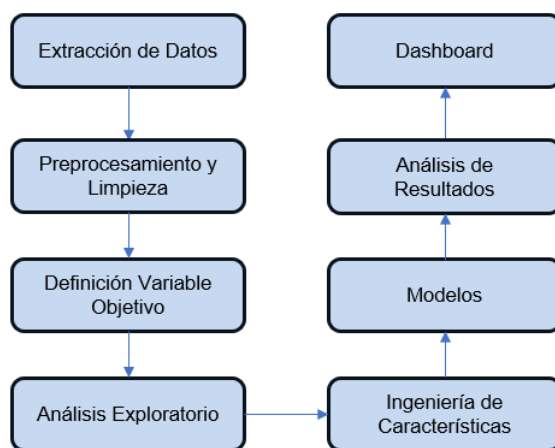


Ilustración 1. Metodología detección de contratos

La metodología inicia con la extracción de datos de los contratos públicos y sus adiciones desde el portal Datos Abiertos (Colombia Compra Eficiente, 2023). Seguidamente, se efectúa un preprocesamiento y limpieza de esta información, incluyendo tareas como la diferenciación entre contratos de prestación de servicios y no prestación de servicios, y la definición de variables precontractuales. Posteriormente, se define la variable objetivo, identificando aquellos contratos que experimentan adiciones en valor y prorrogas en tiempo. El siguiente paso implica un análisis exploratorio de variables relevantes, seguido de la aplicación de ingeniería de características basada en las descripciones de los contratos. Con la tabla

maestra preparada, se ejecutan diversos modelos para contratos de prestación y no prestación, con el fin de predecir la probabilidad de adiciones en valor y prorrogas durante la fase precontractual. Finalmente, se analizan los resultados de los modelos para discernir el impacto de distintas variables en la variable objetivo. Este análisis se integra en una herramienta interactiva Plotly-Dash-HTML (ver [Anexo 4](#)) que muestra la probabilidad de adición de un contrato y permite el ajuste interactivo de valores para evaluar su influencia en la probabilidad estimada.

Cada uno de los anteriores pasos se detallan a lo largo del documento.

3.1 Extracción De Datos

En la extracción de datos se descargaron dos bases de datos desde la plataforma Datos Abiertos del SECOP II-Contratos Electrónicos y SECOP II-Adiciones (Colombia Compra Eficiente, 2023), estos datos son administrados por la agencia nacional de contratación pública Colombia Compra Eficiente que es una Entidad descentralizada de la rama ejecutiva del orden nacional, dentro de sus funciones están principalmente la formulación de políticas, planes y programas buscando optimizar la oferta y demanda en el mercado, y la coordinación con otras Entidades públicas para el cumplimiento de sus objetivos.

El SECOP II es la nueva versión del SECOP (Sistema Electrónico de Contratación Pública) para pasar de la simple publicidad a una plataforma transaccional que permite a Compradores y Proveedores realizar el Proceso de Contratación en línea. Desde su cuenta, las Entidades Estatales (Compradores) pueden crear y adjudicar Procesos de Contratación, registrar y hacer seguimiento a la ejecución contractual. Los Proveedores también pueden tener su propia cuenta, encontrar oportunidades de negocio, hacer seguimiento a los procesos y enviar observaciones y ofertas.

En el ámbito del SECOP II, existen dos bases de datos fundamentales denominadas “Contratos” y “Contratos electrónicos”. Para el desarrollo del proyecto, se ha optado por utilizar la base de “Contratos electrónicos”. Esta elección se debe a que la base “Contratos” presenta una mayor cantidad de campos vacíos, lo cual podría complicar el análisis. Además, nuestra variable objetivo del proyecto se encuentra dentro de la base antes mencionada SECOP II-Adiciones, esta base contiene información detallada sobre los distintos tipos de adiciones aplicadas a los contratos públicos. Más adelante en el proyecto, nos enfocaremos en identificar y clasificar estas adiciones, distinguiendo entre aquellas que corresponden a incrementos en valor y aquellas que implican prórrogas de tiempo.

3.2 Preprocesamiento De Datos Y Limpieza

El primer paso que se realizó fue identificar la completitud de los datos en toda la base del SECOP II, esta base cuenta con 71 columnas descubriendo que 4 columnas cuentan con más del 80% de sus datos como nulos, por lo tanto, se considera eliminar estas columnas, como se muestra en el siguiente gráfico.

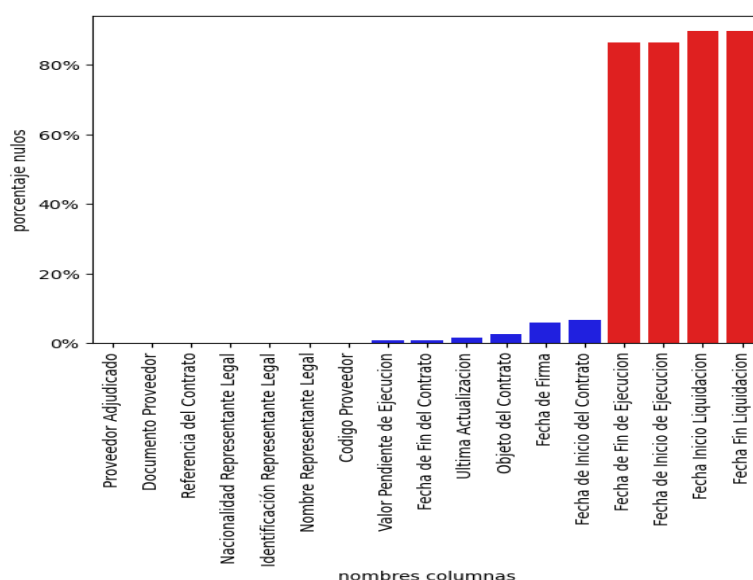


Ilustración 2. Porcentaje de datos nulos por columna

La secretaria distrital de gobierno recomienda a las entidades distritales diligenciar la información de contratación en su totalidad para promover la transparencia en la gestión pública según la ley 1712 de 2014 (SECRETARÍA DISTRITAL DE GOBIERNO, 2014). Con el objetivo de centrar nuestro análisis en Bogotá, que representa el 40% del total de contratos en el país, adoptamos un enfoque sistemático en tres etapas. Inicialmente, filtramos los datos para incluir solo aquellos contratos del Distrito Capital de Bogotá. Posteriormente, para no considerar entidades nacionales con sede en Bogotá, aplicamos un segundo filtro 'Territorial' en la columna 'Orden'. A pesar de esto, algunas entidades departamentales que no corresponden aún permanecían en la base de datos. Por lo tanto, se implementa un filtro final basado en la columna 'Nombre de Entidad', eliminando todas aquellas entidades que incluyen la palabra 'CUNDINAMA' en su nombre, dado que esta palabra es comúnmente presente en las denominaciones de las entidades departamentales.

El siguiente paso consiste en aislar los contratos de prestación de servicios, considerando que este tipo representa el mecanismo más comúnmente utilizado en las entidades. Dada la naturaleza específica de estos contratos y sus posibles complicaciones adicionales, es crucial prestarles atención detallada. Como se observa en la Tabla 1, constituyen el 88% del total de los contratos, como se muestra a continuación.

Tipo de Contrato	% Participación
Prestación de servicios	88.5%
Decreto ley 92 2017	5,6%
Otros	5,9%

Tabla 1. Tipos de contratos

Hemos categorizado las bases de datos en dos tipos: una para contratos de “prestación de servicios” y otra denominada “no prestación de servicios”. Esta nomenclatura facilitará su identificación a lo largo del documento.

Al analizar la fecha de firma en no prestación de servicios, se observan anomalías notables, como se muestra en la ilustración 3, "Conteo por fecha de la firma del contrato". En tres fechas específicas, se registra un número inusualmente alto de más de 400 contratos firmados. Si bien esto es común en contratos de prestación de servicios, especialmente para la contratación de personal en nuevas sedes de subredes, colegios o hospitales, en la base no prestación de servicios no se esperarían estos picos.

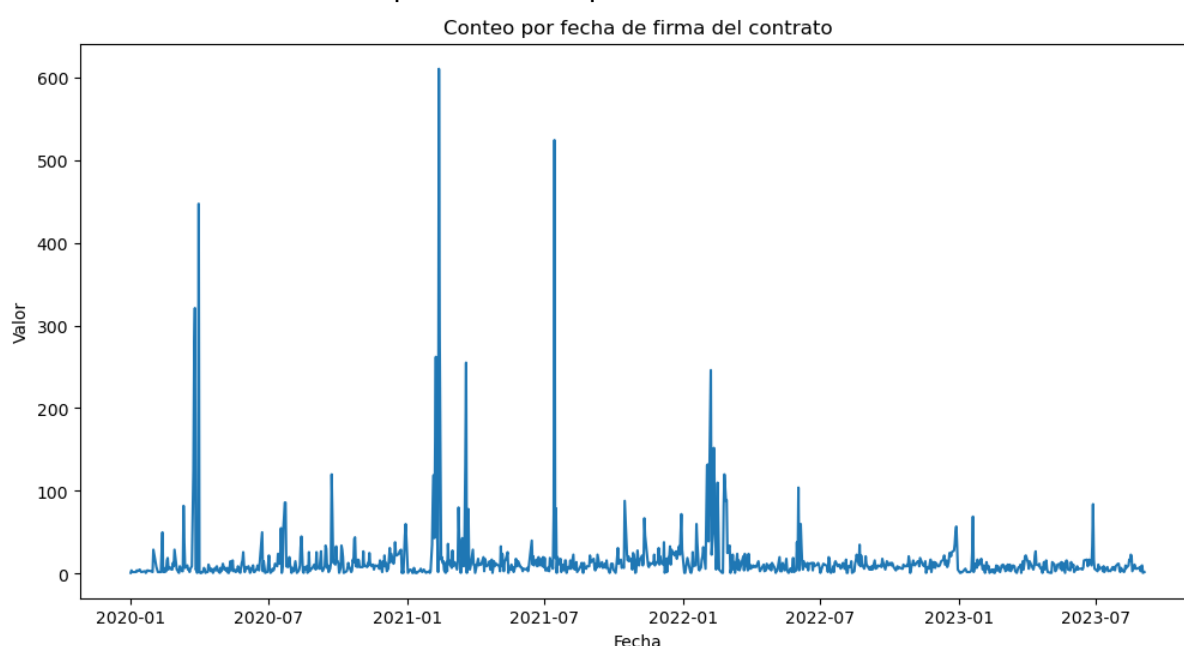


Ilustración 3. Conteo por fecha de la firma del contrato

Una revisión detallada reveló que estos contratos pertenecen a la modalidad de régimen especial, frecuente en entidades públicas y sujeta a legislación específica para garantizar un control y supervisión más estrictos. Es importante destacar que, aunque estos contratos son técnicamente de régimen especial, algunos presentan características similares a los contratos de prestación de servicios. Esto es evidente en aquellos contratos que involucran profesiones que son equiparables a la prestación de un servicio, lo que justifica su reubicación en esta categoría para una evaluación más coherente.

Para conocer si estos contratos están enfocados en profesiones específicas o servicios profesionales personales, se requiere una larga lista de profesiones. Se esperaba, con apoyo de webscrapping, mantener una lista actualizada de profesiones, pero no fue posible encontrar una fuente sin restricciones. Se implementó una regla básica utilizando una lista de 200 profesiones comunes en entidades públicas, generada por ChatGPT y complementada manualmente. Los contratos cuya "Descripción del proceso" coincidiera con estas profesiones se excluyeron de no prestación y se clasificaron como de prestación de servicios. Así, se identificaron un total de 22.017 contratos que fueron reasignados para fines de análisis. Estos, originalmente en el segmento de no prestación, serán trasladados a la base de datos de prestación.

3.3 Definición Variable Objetivo

Para la variable objetivo se usará el archivo SECOP II – Adiciones (Colombia Compra Eficiente, 2023), en esta base se realizó un proceso de identificación de adiciones de valor y

prórroga que se dividió en tres partes. En la primera, se realiza un filtro en la columna “Tipo de adición” donde se toman las etiquetas que correspondan a “Adición en el valor” y “Extensión” que según en la Tabla 2 representa menos del 5% del total de adiciones.

Tipo de Adición	% Participación
Modificación General	57,26%
Terminación	16,84%
Conclusión	13,91%
Adición en el valor	4,47%
No definido	2,71%
Suspensión	1,71%
Cesión	1,60%
Reactivación	1,48%
Extensión	0,01%

Tabla 2. Participación tipos de adición

En la segunda parte, la etiqueta "Modificación general" es la más común y abarca diversas causas de adición, donde algunas son de tipo valor y prórroga. Estas se pueden identificar con la columna “Descripción de la adición”, donde se puede utilizar reglas básicas, como buscar términos clave "adición en valor" y "prórroga" que filtran de 792.510 contratos a 352.373 sin identificar.

En la tercera parte del análisis, se usan herramientas de NLP (Procesamiento de lenguaje natural) para la clasificación de contratos, en específico se usó un modelo Transformers denominado BETO que en teoría es un modelo BERT pre-entrenado en español para la clasificación de textos y otras tareas del lenguaje (Cañete et. al., 2020).

Se procede a aplicar Transformers al tener algunos problemas al usar reglas básicas en su totalidad, ya que existen algunas descripciones donde sus palabras no dicen explícitamente que son adición de tiempo o valor, pero su contexto si lo dice como por ejemplo la siguiente descripción:

“SE REQUIERE CONTINUAR CON LA PRESTACIÓN DEL SERVICIO”

En ningún momento se puede encontrar palabras relacionadas con la fecha o prórroga, aunque para un humano es fácil inferir que según esta descripción la continuación del contrato se interpreta como una prórroga, para una máquina es más complejo.

Se toma como variable objetivo las “adiciones” que se filtraron en la Tabla 2 y las que se hallaron con las reglas básicas de términos clave. Las “no adiciones” son aquellas que se identifican en la Tabla 2 con la etiqueta “Terminación”, “Conclusión”, “No definido”, “Suspensión”, “Cesión” y “Reactivación”. Se separa aleatoriamente en un grupo de entrenamiento (80%) y un grupo testeo (20%) el conjunto de datos con la descripción de la adición y su etiqueta mencionado previamente. A continuación, realizamos un proceso de tokenización del texto de la descripción, este proceso consiste en tomar un modelo pre-entrenado y dividir el texto en subpalabras lo cual ayuda a que el modelo identifique y trabaje con las unidades básicas de significado dentro del texto. Luego se aplica un proceso de DataLoader que produce lotes de datos que facilitan el entrenamiento de modelos al volverlos

más eficientes (PyTorch, 2023) para finalmente aplicar un modelo BETO pre-entrenado para clasificación de secuencias mencionado anteriormente

Al realizar el ajuste del modelo a tres épocas encontramos que el modelo Transformers se ajusta bastante bien desde la primera época tanto para la población de entrenamiento como testeo sin que tengamos síntomas de overfitting, como se puede ver en la ilustración 4.

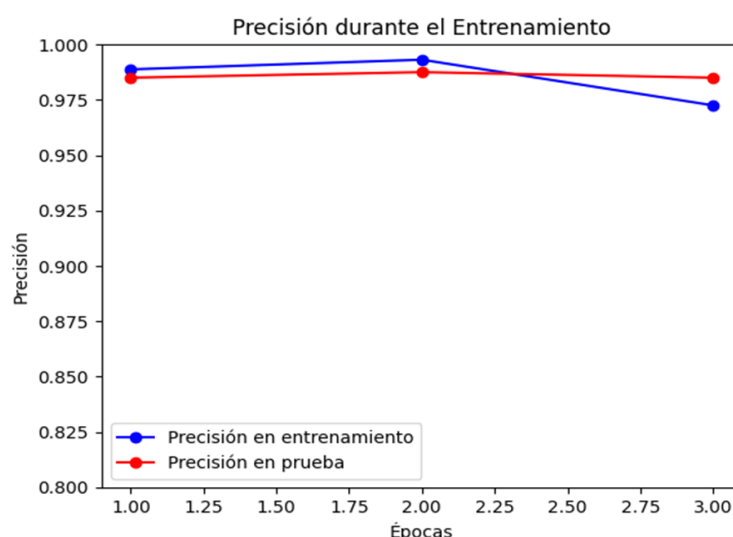


Ilustración 4. Precisión del modelo Transformers

Al aplicar el modelo entrenado a las 352.373 descripciones de adición sin identificar, se logra la clasificación como adición de 10.422 contratos y 341.951 que no lo son. Por último, podemos ver los resultados mencionados anteriormente en la ilustración 5 con la identificación de 450.522 contratos con adición en valor y prorrogación en tiempo.

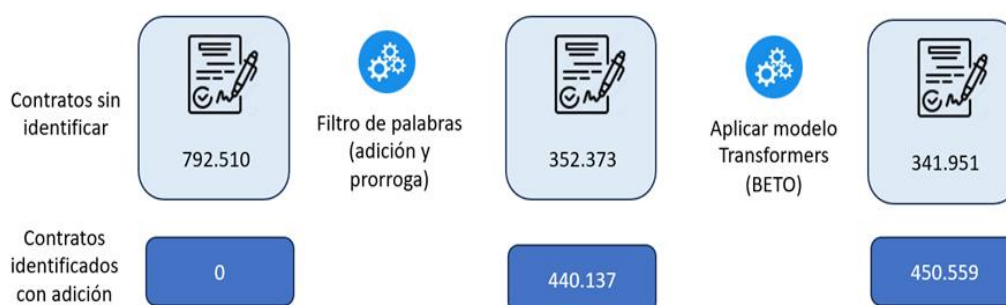


Ilustración 5. Proceso identificación de adiciones

En conclusión, el modelo de clasificación de texto BETO identificó en la mayoría de los casos las descripciones que efectivamente son adición, incluso en situaciones donde no se mencionaba explícitamente una adición en valor o monto. Sin embargo, también hubo instancias en las que el modelo clasificó erróneamente contratos que no correspondían a adiciones, como se detalla en la Tabla 3. Estos resultados subrayan tanto la eficacia como las limitaciones del modelo en la identificación precisa de diferentes tipos de contratos.

Descripción contratos identificados como adición	Descripción contratos identificados como adición, pero no son adición
“Se continúa presentando la necesidad”	“Se realiza cambio de la minuta del contrato 012 de 2020 correccion nit del establecimiento”
“Que persiste la necesidad de cumplir las obligaciones pactadas dentro de los contratos interadministrativos suscritos con el ejercito y las fuerzas militares que habiendo realizado reservas presupuestales a fin de cumplir las obligaciones contraídas en los contratos interadminiostativos antes mencionados”	“Se requiere generar códigos de autorización para plan de pagos”
“Se tiene la necesidad de garantizar la continuidad de la prestación de servicio de las piscinas con calidad y en excelentes condiciones de funcionamiento a nuestros usuarios y beneficiarios del centro vacacional de tolú y centro recreativo de sincelejo”	“Corrección de fechas”

Tabla 3. Descripción de los contratos marcados como adición.

Además, el modelo mostró una notable eficiencia en la identificación de contratos que no correspondían a adiciones, manteniendo así la precisión general observada en análisis anteriores. No obstante, se observaron algunas excepciones a esta tendencia. En un número reducido de instancias, el modelo no pudo distinguir adecuadamente aquellos contratos que en realidad eran adiciones, clasificándolos incorrectamente como no adiciones. Este patrón sugiere que, en estos casos específicos, el modelo enfrentó dificultades para generalizar de manera efectiva, posiblemente debido a descripciones mal redactadas o excesivamente breves en los prompts como se muestra en la Tabla 4. Esta limitación resalta la importancia de contar con descripciones claras y detalladas para mejorar la precisión en la clasificación de contratos.

Descripción contratos identificados cómo no adición	Descripción contratos identificados cómo no adición, pero son adición
“Se modifican las cláusulas 4 y 6 del contrato de acuerdo a documento adjunto”	“Ps 6025 2021 adi no 1 pro no 1”
“Se efectua la modificación del contrato respecto a unos valores erréneos en la cláusula décimoquinta de la imputación presupuestal”	“Ps 3940 2023 adi no 1 pro no 1 modif no 1”
“Se realiza la aclaratoria del valor total del contrato”	“Se realiza mod del valor del contrato según los documentos anecos”

Tabla 4. Descripción contratos mal redactados.

En conclusión, la variable objetivo se definirá como dicotómica. Se asignará el valor 1 para identificar casos con adición en valor o prorroga en tiempo, y el valor 0 se utilizará para señalar los casos donde no hay adición.

3.4 Análisis Descriptivo Y Exploratorio.

En este apartado se exploran las diferentes características de la base de datos. Para las variables cuantitativas se obtuvieron algunas medidas de resumen y gráficos de cajas y bigotes como acercamiento general a sus distribuciones. Por otro lado, para las variables de naturaleza categórica, se obtuvieron las frecuencias del rango de valores posibles de cada variable.

Con esto, se busca obtener un entendimiento del comportamiento a nivel discriminado entre los valores de la variable objetivo, es decir, los contratos con adiciones (1) y los que no tuvieron adiciones (0).

3.4.1 No Prestación De Servicios.

- **Variable objetivo (adiciones):** Para el caso de la base “otros” (no prestación), la base final cuenta con 32.094 registros, de los cuales aproximadamente el 65% (20.746) corresponden a contratos sin adiciones y los restantes a contratos que han tenido adiciones. Estas frecuencias, permiten concluir que hay un número suficiente de registros para cada caso y, en consecuencia, descartamos la realización de algún proceso de balanceo.
- **Variables continuas:** Se presenta en este apartado, la distribución sobre las variables continuas consideradas más relevantes:

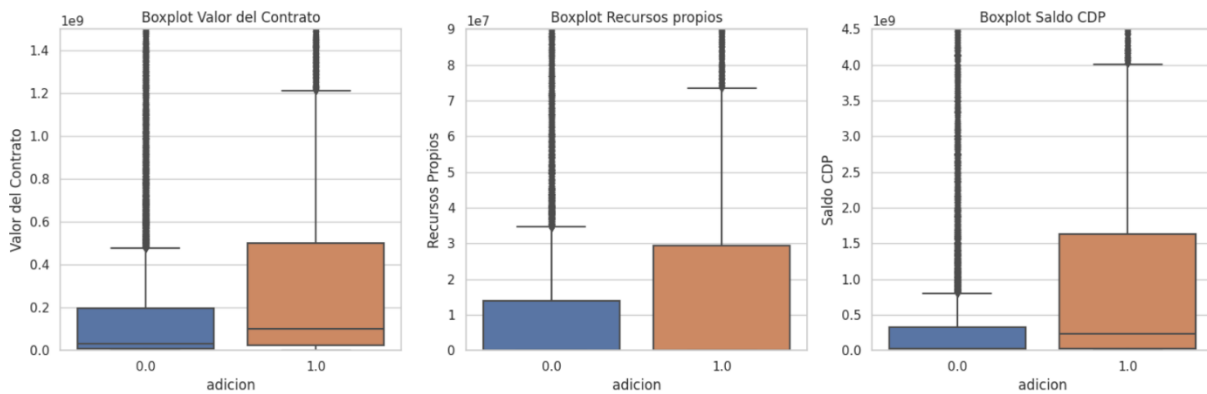


Ilustración 6. Variables continuas relevantes No Prestación(Valor contrato, Recursos propios, Saldo CDP)

Como es natural en variables medidas en dinero, presentan una asimetría significativa a la derecha. En los gráficos se observan valores en magnitud $1e9$ y $1e7$, existen valores mucho mayores a los observados, sin embargo, se han cortado para poder obtener una mejor representación de los cuartiles.

Cabe mencionar que existen diferencias en el comportamiento de estas variables para los contratos con adiciones y los de no adiciones. Adicionalmente, se han calculado las medidas de resumen de la variable sin discriminar, en ellas se observa un comportamiento también asimétrico y con alta concentración en valores cero, se incluyen en el [Anexo 1](#).

- **Variables categóricas:** En cuanto a las variables categóricas consideradas más relevantes se incluyen tablas de contingencia de las variables frente a la variable objetivo (adiciones) en el [Anexo 1](#), para ver las diferencias en las distribuciones de sus frecuencias.

En términos generales, se puede observar algunas diferencias entre los contratos con adiciones y los de no adiciones en algunas variables, mientras que para otras la distribución de las frecuencias es aproximadamente igual.

Por último, sobre todas las tablas de contingencia se calcula el estadístico chi-cuadrado, del cual se obtuvieron los siguientes resultados:

Variable	Chi_Cuadrado No Prestación
Sector	1965.60
Rama	361.03
Entidad Centralizada	1109.74
Tipo de Contrato	1809.38
Modalidad de Contratación	1555.37
Justificación Modalidad de Contra.	1708.75
Habilita Pago Adelantado	925.72
Obligación Ambiental	26.12
Obligaciones Postconsumo	6.80
Reversion	1.38
Año BPIN	3149.31
Es Post Conflicto	3.88
Destino Gasto	188.77
Origen de los Recursos	6.72
Puntos del Acuerdo	7.49
Pilares del Acuerdo	7.49

Tabla 5. Tabla de contingencia No Prestación

Preliminarmente se presume que algunas de estas variables están relacionadas con la variable adición y presentan un valor del estadístico chi-cuadrado grande.

3.4.2 Prestación De Servicios.

- **Variable objetivo (adiciones):** Para el caso de la base de prestación de servicios, la base final cuenta con 437.180 registros, de los cuales aproximadamente el 66,4% (290.093) corresponden a contratos sin adiciones y los restantes a contratos que han tenido adiciones. Estas frecuencias, permiten concluir que se mantiene aproximadamente la misma proporción que en la base de no prestación de servicios y que también hay un número suficiente de registros para cada caso.
- **Variables continuas:** Se presenta en este apartado, la distribución sobre las variables continuas consideradas más relevantes:

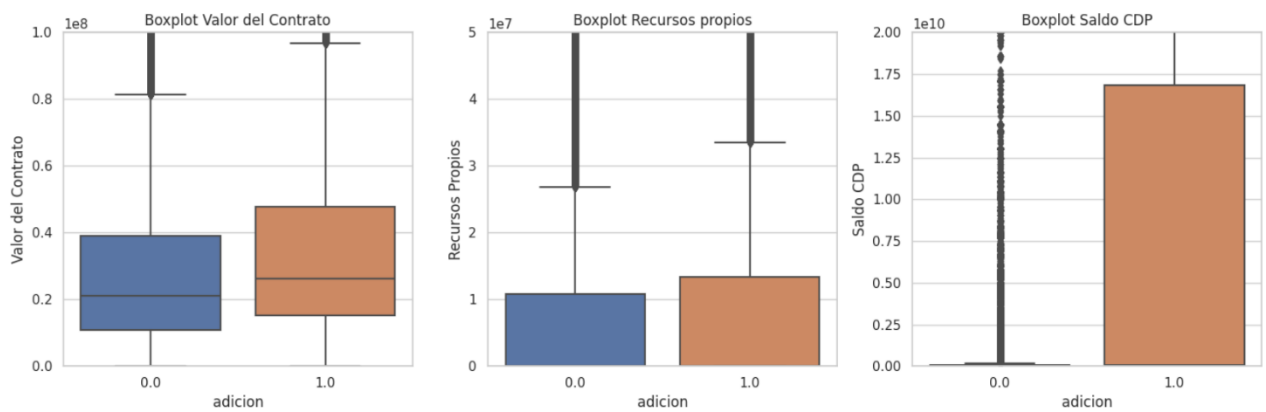


Ilustración 7. Variables continuas relevantes Prestación (Valor contrato, Recursos propios, Saldo CDP)

Nuevamente, se observa una asimetría significativa a la derecha, sin embargo, es mucho mayor a la presentada en los contratos de no prestación de servicios. En los gráficos se observan valores en magnitud entre $1e10$ y $1e7$, existen valores mayores a los observados, sin embargo, se han cortado para poder obtener una mejor representación de los cuartiles.

Cabe mencionar que para las variables Valor del Contrato y Recursos propios la distribución es similar tanto para los contratos de adiciones (1) como para los no adiciones (0), sin embargo, para la variable Saldo CDP, existen diferencias significativas en el comportamiento de estas variables.

- **Variables categóricas:** En este caso, igual que en el caso de no prestación, se incluyen las tablas de contingencia de las variables frente a la variable objetivo (adiciones) en el Anexo I. para ver las diferencias en las distribuciones de sus frecuencias.

En términos generales, se puede observar algunas diferencias entre los contratos con adiciones y los de no adiciones en algunas variables, mientras que para otras la distribución de las frecuencias es aproximadamente igual.

Por último, sobre todas las tablas de contingencia se calcula el estadístico chi-cuadrado, del cual se obtuvieron los siguientes resultados:

Variable	Chi_Cuadrado Prestación
Sector	30570.86
Rama	10045.26
Entidad Centralizada	16804.07
Tipo de Contrato	9707.73
Modalidad de Contratación	17202.44
Justificación Modalidad de Contra.	37992.57
Habilita Pago Adelantado	3863.50
Obligación Ambiental	190.10
Obligaciones Postconsumo	0.53
Reversion	0.12
Año BPIN	11362.94
Es Post Conflicto	11.05
Destino Gasto	14202.88
Origen de los Recursos	552.33
Puntos del Acuerdo	6.90
Pilares del Acuerdo	9.78

Tabla 6. Tabla de contingencia Prestación

3.5 Ingeniería De Características.

La ingeniería de características es una técnica de aprendizaje automático que se utiliza para crear nuevas variables con la información disponible y/o transformándolos en una forma que sea más fácil de interpretar por los modelos de aprendizaje automático. Por esta razón, se realizó clasificación de texto con la herramienta de NLP (natural language processing), que permite extraer información lingüística, semántica, identificación de temas claves; consiguiendo aumentar el número de variables descriptoras y mejorando el aprendizaje del modelo.

Una variable no estructurada con información relevante es el “*Objeto del contrato*”, donde se puede identificar el contenido esencial del mismo. Por tal razón, Se trabajó con [NLTK](#) y [SpaCy](#) que contienen bibliotecas de procesamiento de texto para extraer información que encapsulara acciones y características significativas. Posteriormente, se aplicó la técnica de “Term Frequency - Inverse Document Frequency” (TF-IDF) en conjunto con NLTK para transformar y calcular que tan relevante es una palabra dentro de dicho texto. Este paso fue fundamental para preparar los datos para el análisis de clustering, donde se desarrolló un modelo no supervisado K-means con el objetivo de descubrir relaciones intrínsecas entre diversos elementos del objeto de contrato procesado, lo cual permite identificar patrones subyacentes y agrupaciones naturales sin la necesidad de etiquetas predefinidas. A fin de determinar el número óptimo de clusters, se aplicaron tres técnicas distintas: el Método del Codo para evaluar la inercia, el Score de Silueta y el Índice de Davies-Bouldin, proporcionando así un enfoque multifacético para la optimización del modelo (Ver [anexo 2](#)).

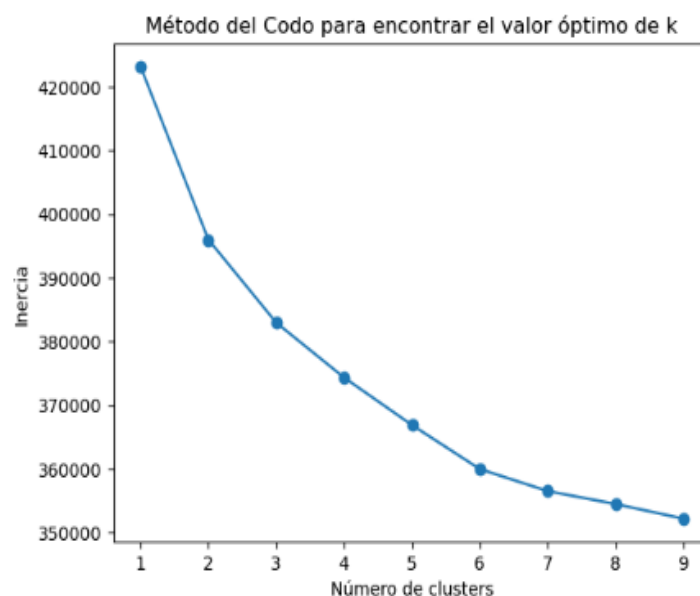


Ilustración 8. Método del Codo

Se puede observar en la ilustración 8 que el cambio en la inercia disminuye más lentamente después del 3 clúster, por lo que se asignó este valor para el número de agrupaciones a usar en los modelos siguientes.

Continuando con el análisis de la información del objeto del contrato con el fin de identificar temas y crear agrupaciones en los datos, se ejecutaron varios modelos. En primera instancia se desarrolló un modelo de K-Means aplicando como pasos previos la Vectorización de Texto TF-IDF y luego la Descomposición en Valores Singulares Truncada (TruncatedSVD) que se utiliza para disminuir la cantidad de características de un conjunto de datos manteniendo la mayor cantidad posible de información original, en este caso se usaron 1000 componentes que explicaban el 76.6% de la variabilidad en los datos. Por otro lado, se realizó una red neuronal (autoencoder) para reducir la dimensionalidad de datos de texto y, posteriormente, realizar clustering con el algoritmo K-Means. Al comparar estas dos aproximaciones se observó que la segunda opción proporcionaba agrupaciones que tenían un mayor sentido desde el punto de vista del negocio. Esto dado que al aplicar clustering en las representaciones de baja dimensión, se espera que los clusters sean más significativos y representativos de las estructuras latentes en los datos de texto.

Después de obtener los resultados del modelo de clustering se procedió a caracterizar cada uno de cluster resultante por los temas que se mostraban relevantes. Con el objetivo de que la visualización de estos temas fuera clara y comprensible, se recurrió al uso de nubes de palabras. Esta técnica se aplicó sobre la columna previamente procesada donde se extrajeron los tokens del objeto del contrato, permitiendo así una representación gráfica que destacó de forma efectiva las palabras clave y conceptos predominantes en cada cluster como se muestra en la ilustración 20 (ver [anexo 2](#)), un ejemplo de esta representación se muestra a continuación en la ilustración 9:



Agrupación	Nombre
Cluster 1	Alimentación
Cluster 2	Local
Cluster 3	Salud

Otra variable no estructurada es el Nombre de la entidad, que se vuelve relevante para los contratos de prestación de servicios. Hay entidades distritales que engloban la mayor cantidad de los contratos, como las Subredes de Salud. Sin embargo, la variable contiene 115 entidades distintas, que si bien, es un número bajo si se tiene en cuenta la cantidad de contratos que tiene esta población, igual es problemático a la hora de usar la variable. Además, entidades como colegios e instituciones educativas tienen pocos contratos, pero una diversidad alta de nombres.

Se realiza la misma operación con la variable de Nacionalidad del representante legal, terminando en las siguientes etiquetas. Aquellas nacionalidades diferentes a Latam en conjunto, están por debajo del 0.2% de la población. De hecho, las vacías pesan un poco más, aunque el 96% son de nacionalidad colombiana.

Nacionalidad	Presenta adición	
	No	Si
colombiana	63.62%	33.28%
latam	0.16%	0.10%
otro	0.10%	0.04%
vacio	2.48%	0.23%

Tabla 8. Nacionalidad Representante legal

Se resalta que al realizar una prueba de independencia o de tabla de contingencia Chi², al 5% de significancia se concluye que hay una dependencia entre la nacionalidad como se establece en la tabla anterior y en tener adiciones en el contrato de servicios profesionales personales. Las correcciones con expresiones regulares se pueden ver en el [Anexo 1](#)

3.5.1 Categorizaciones.

En el análisis exploratorio concluimos que las variables de naturaleza continua en general presentan casuísticas como la concentración en cero y la presencia de bastantes valores “atípicos”. Con el fin de mitigar el efecto de estas condiciones, se agrupan las variables en dos o más categorías, siguiendo como criterio que cada una de ellas tenga una volumetría de registros razonablemente suficiente.

En algunos casos, se crean las categorías simplemente haciendo una marca de la presencia de valores cero (0) y valores diferentes de cero (1). En otros casos, se crea la categoría (0) para los valores en cero, y posteriormente se obtiene la mediana de los valores diferentes de cero. A partir de ella se crean dos categorías adicionales: una en la que están los valores diferentes de cero y “bajos” (1), y otra en la que quedan los valores diferentes de cero y “altos” (2). Por último, en otros casos se parte de la división en deciles o cuartiles.

Este tratamiento se aplica para las variables: valor del contrato, saldo CDP, saldo vigencia, presupuesto general de la nación, sistema general de participación, sistema general de regalías, recursos propios de entidades (alcaldías, gobernaciones y resguardos indígenas), recursos de crédito y recursos propios, las tablas y gráficas resultantes se incluyen en el [Anexo 3](#) respectivamente para las variables consideradas más relevantes.

Una vez creada la categorización en grupos para las variables, se calculó el valor del estadístico chi-cuadrado para la tabla de contingencia de la variable contra la variable objetivo (marca de adición). Se obtuvieron los siguientes resultados:

Variable	Chi_Cuadrado Prestación	Chi_Cuadrado No Prestación
Grupos Valor Contrato	8249.65	1874.34
Grupos Saldo CDP	47687.96	2222.81
Grupos Saldo Vigencia	1332.57	29.89
Grupos PGN	2091.01	14.41
Grupos Sis. General Participación	5464.72	185.19
Grupos Sis. General Regalias	10.17	0.12
Grupos Recursos Propios Entidades	518.92	245.48
Grupos Recursos de Credito	60.85	3.33
Grupos Recursos Propios Entidades	3285.42	526.04

Tabla 9. Tabla de contingencia de la variable contra la variable objetivo

De la tabla anterior, se observa que preliminarmente hay algunas de estas agrupaciones que puedes resultar significativas en la etapa del modelado, pues tienen un valor de chi-cuadrado significativo. Por otro lado, también hay algunas variables que parecen no ser explicativas de la marca de adición. En términos generales, para algunas variables se observa una distribución de frecuencias diferenciada para los contratos con adiciones de los contratos sin adiciones, para otras el comportamiento de las frecuencias es similar.

3.6 Aplicación Modelos

Recordando el objetivo: identificar el riesgo de un contrato para tener adiciones futuras, basándose en sus condiciones precontractuales, se ha decidido crear un Pipeline que permita el pre-procesamiento y posterior entrenamiento. Además, se pretende comparar Regresiones Logísticas, Support Vector Machine, Random Forest y Regresión Logística, acompañados de selección de variables (Lasso) y optimización de hiperparámetros por medio de la búsqueda de grilla y reforzando esta última con validación cruzada de 5 dobleces (folds), para lograr el reto, se utilizan las dos bases de prestación y no prestación de servicios que se dividieron en la etapa de preprocesamiento.

El pre-procesamiento aplica: en variables numéricas la estandarización mediante un StandarScaler de scikit-learn y a categóricas OneHotEncoding, de la misma librería. Para Logística se elimina una variable para todas las categóricas, pero para otros modelos sólo quitan una en las binarias. Para no prestación el tratamiento resultó en 155 variables, aplicando una selección de variables dejando apenas 97 y finalmente estableciendo un umbral de selección de los parámetros (threshold) de 0.04, con el objetivo de afinar más para esa población. Este ajuste culminó en la identificación de 41 variables. Tabla 10 del [Anexo 1](#), titulado " Variables Significativas y sus Coeficientes", presenta un desglose detallado de estas variables seleccionadas, exhibiendo la magnitud de sus coeficientes y, por ende, su relevancia dentro del modelo.

En el [Anexo 1](#) en las Tablas 31 y 32 se presenta el detalle de los modelos implementados y los hiperparámetros probados. Cabe resaltar el conjunto de datos de prestación 13 veces más grande que no-prestación. A pesar de tener una máquina con mayor potencia en procesamiento, no convergía el modelo, determinando realizar un muestreo aleatorio, por la mitad de la muestra original. Luego, al aplicar selección de variables no lograba elegir un

conjunto de features, por lo que este método fue desistido. Así las cosas, se utilizó el modelo con todas las características que presentaron alguna incidencia estadística sobre la variable objetivo, con las pruebas de independencia/contingencias realizadas previamente.

Tras seleccionar los hiperparámetros, se termina de estructurar un pipeline, para incorporar la búsqueda de grilla. Durante esta fase se consideraron múltiples métricas de desempeño para determinar el modelo más eficaz, con las bases de entrenamiento y prueba. Los resultados obtenidos revelaron los modelos más destacados, tanto en el conjunto completo de variables, como en de selección de características. Los mejores modelos se encuentran en el Anexo 1 en Tabla 33 y 34 , con y sin prestación, respectivamente.

Los resultados indican que el modelo XGBoost sobresale en términos de desempeño en, con y sin prestación. En no-prestación se implementaron modelos con todas las características, y otros aplicando selección de características. Por el principio de parsimonia, que favorece modelos más simples frente a complejos con resultados similares, prima aquel menos complejo. Al adoptar este enfoque, buscamos mantener un equilibrio entre la simplicidad y su capacidad predictiva, con una generalización robusta y sin complejidad innecesaria.

Al comparar los modelos evaluados, con y sin prestación, en gráficos de caja (Ilustración 10 y 11) se destaca el desempeño sobresaliente del XGBoost, tanto en entrenamiento como en prueba superando significativamente a los demás.

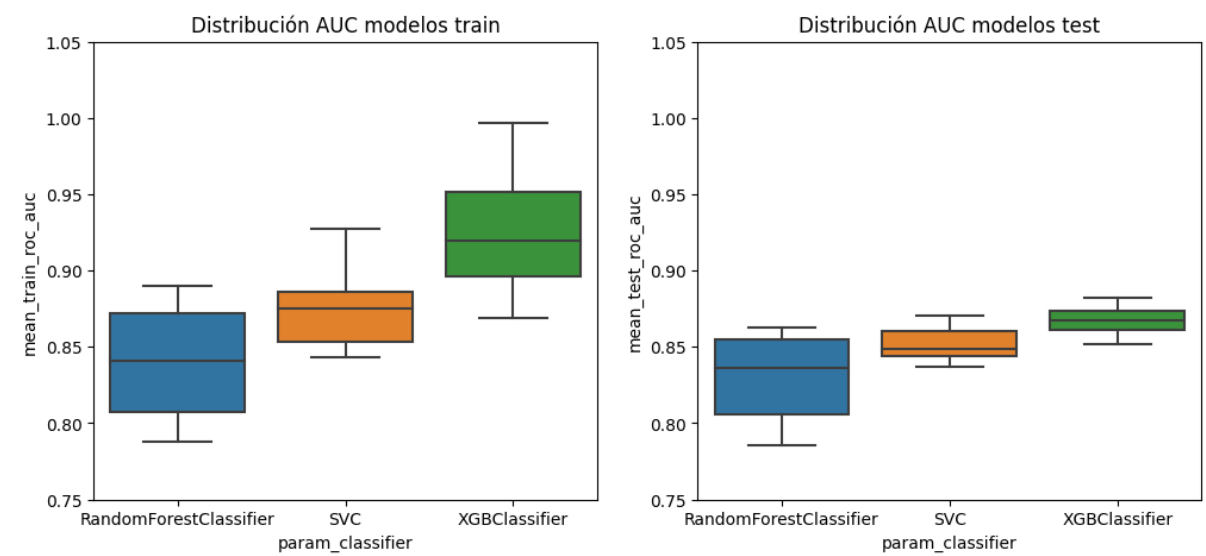


Ilustración 10. Desempeño de los modelos No Prestación

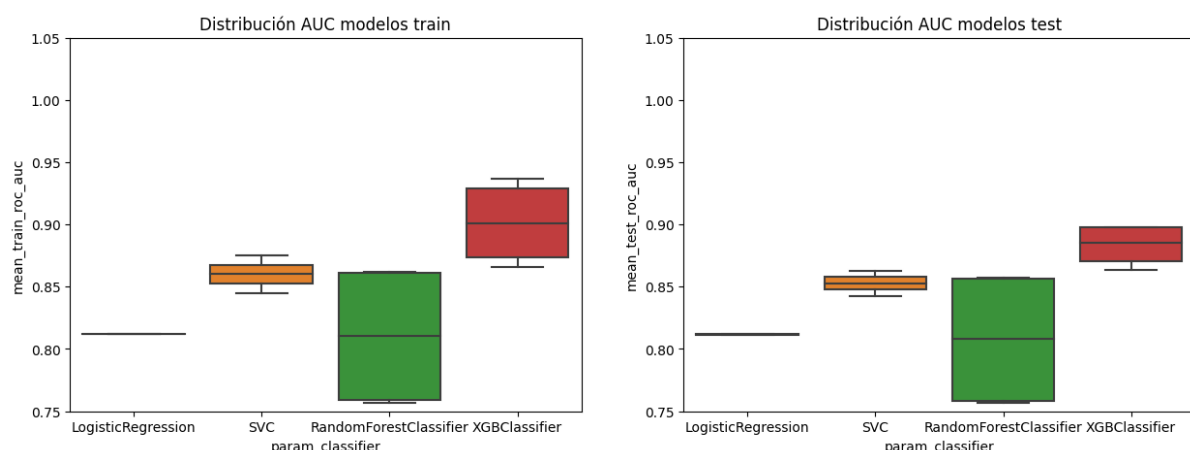


Ilustración 11. Desempeño de los modelos Prestación

Elegido el mejor modelo de clasificación, procede evaluar su poder de predicción, tanto para casos verdaderos (positivos) como falsos (negativos). La matriz de confusión es una herramienta esencial en este análisis, acompañada de algunas de sus métricas clave: sensibilidad y especificidad. La sensibilidad fue de 0.64 para no prestación y de 0.71 para el segundo. La especificidad fue de 0.86 para no prestación y 0.93 para el segundo. Como el objetivo es identificar eficazmente contratos con adición futura (verdaderos positivos), la alta sensibilidad es de gran importancia. Para mejorar esta métrica, podemos ajustar el umbral de clasificación de los valores positivos, que por defecto se establece en 0.5, pero puede ser modificado para alinearse mejor con las necesidades específicas del negocio. Las matrices, con y sin prestación se encuentran en el [Anexo 3](#) En las Ilustraciones 26 y 27.

Un enfoque efectivo para encontrar el umbral óptimo es utilizar el índice de Youden, también conocido como Youden's J statistic (Brownlee, 2021), que busca maximizar la diferencia entre la tasa de verdaderos positivos y la tasa de falsos positivos, equilibrándolas. Al aplicarlo el umbral más adecuado es de 0.39 y 0.33, sin y con prestación, respectivamente; mejorando la capacidad de detectar contratos propensos a adiciones.

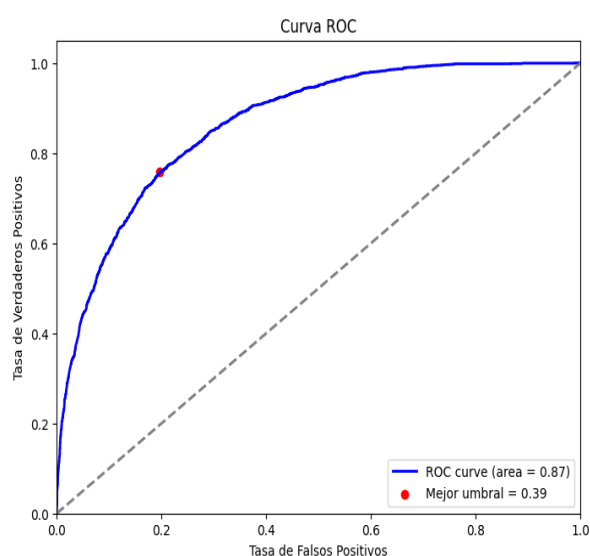


Ilustración 12. Curva ROC No Prestación

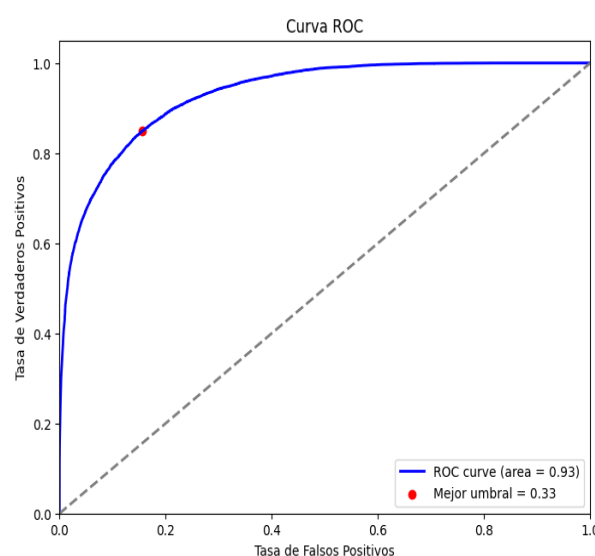


Ilustración 13. Curva ROC Prestación

En la ilustración 12 y 13, con y si prestación, se muestra e la gráfica del ROC un punto rojo que representa el índice del umbral que maximiza la diferencia entre la tasa de verdaderos

positivos y la tasa de falsos positivos. Luego de realizar este cambio, volvemos a calcular la matriz de confusión, presentada en el [Anexo 3](#) En las Ilustraciones 28 y 29.

Tras ajustarlo, se mejora la sensibilidad, alcanzando ahora valores de 0.75 (+0.11) y 0.85 (+0.14) (sin y con prestación, respectivamente), con un trade-off en especificidad, terminando en niveles aceptables de 0.80 (-0.06) y 0.84 (-0.09) (sin y con prestación, respectivamente). Aunque hay una ligera disminución, no representa una reducción considerable para discernir correctamente los contratos sin adiciones; pero es más adecuada para el objetivo del modelo.

4 Resultados Y Análisis.

Se utiliza la técnica shap value que nos indica la importancia de las características en el modelo y su impacto tanto positivo o negativo, como se muestra en el Ilustración 14 y 15, para los 2 tipos de modelos.

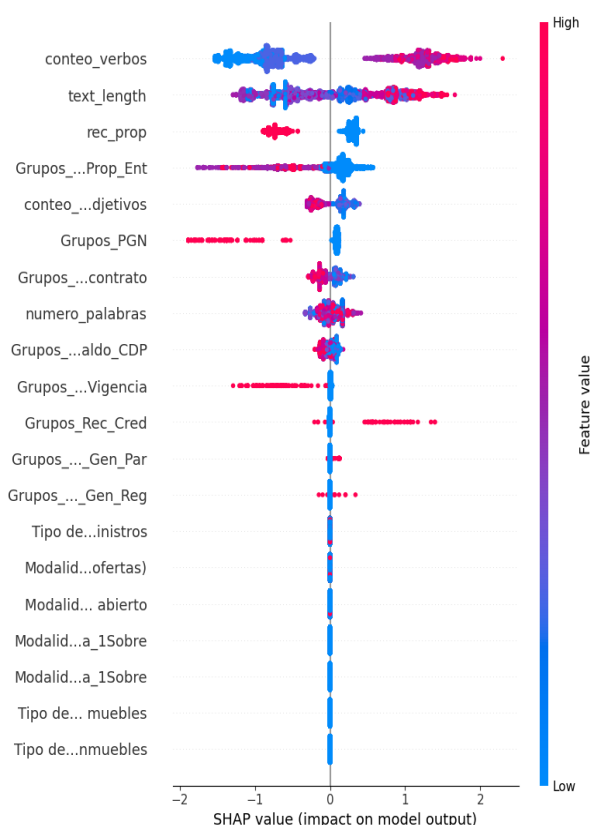


Ilustración 14. Shap Values No Prestación

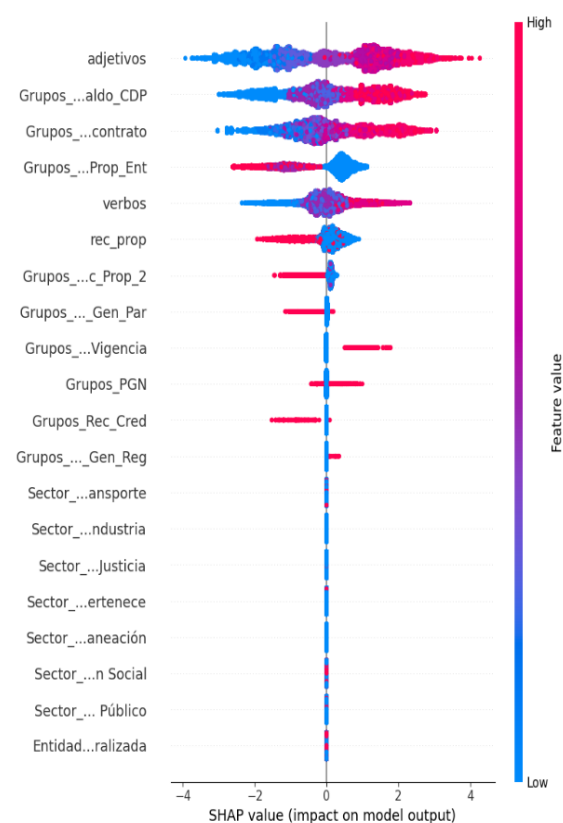


Ilustración 15. Shap Values Prestación

El análisis de las variables revela hallazgos importantes. El eje Y proporciona una visión clara de su importancia para el modelo. Aquí, “conteo verbos” destaca como la más influyente. Además, describe su influencia positiva o negativa en la probabilidad de adición de un contrato, resaltando que “text_length” sugiere que un texto más extenso en la Objeto del contrato aumenta el riesgo de adición visceversa

A partir de los gráficos, con y sin prestación, se concluyen insights muy parecidos:

- Descripciones detalladas y extensas parecen estar asociadas a mayor complejidad e incertidumbre, que aumenta el riesgo de adición. Igual que para el modelo anterior.

- Cuando una entidad presenta saldo de CDP (Certificado de Disponibilidad Presupuestaria), que no es más que una provisión intocable y exclusiva para la ejecución de una actividad, se manifiesta su compromiso con los recursos destinados a acabar ese contrato y además reduce su riesgo de fondeo y con este el riesgo de adición.
- Cuando el valor del contrato es alto, se tiene un impacto negativo en el riesgo de adición, para contratos que no son de prestación; con sentido al tener un mayor control de las entidades y otros entes. Contratos con montos pequeños pueden pasar inadvertidos y tener una mala gestión en la asignación de estos recursos. Sin embargo, para prestación los contratos más costosos presentan mayor riesgo, en contravía de anterior. Aquellos roles con cuantías importantes se prorrogan sucesivamente.
- La participación de entidades nacionales, alcaldías y gobernaciones en la asignación de parte del presupuesto para proyectos reduce el riesgo de adición. Esta práctica se relaciona con la responsabilidad administrativa y la importancia de cumplir eficientemente con la ejecución y entrega de contratos, evitando así repercusiones negativas tanto en el ámbito administrativo como económico.
- Cuando los recursos de la Entidad están en riesgo, suelen tener menos adiciones, ya que son más rigurosas en la exigencia del correcto cumplimiento de sus contratos.
- Contratos con recursos a crédito generan un mayor número de adiciones. Parece enajenarse o "relajarse" la responsabilidad de la Entidad por el capital mayor que está en juego.

4.1 Herramienta Dashboard

Se desarrolló una herramienta en Plotly-Dash-HTML (Ver Anexo IV.) que muestra la probabilidad, las características esenciales y permite el ajuste interactivo de los valores, facilitando la observación del impacto estimado en la probabilidad. Como los modelos con mejores métricas fueron de "caja negra", no se tiene una medida del impacto que genera cada variable. Sin embargo, su función es el de brindar a los usuarios una comprensión transparente, accesible y de guía para ver cómo mejorar su riesgo de tener adiciones.

Como se indicó previamente, el diseño se efectúa utilizando Plotly, una herramienta que facilita la creación de gráficos interactivos. A través de una plantilla específica, se pueden cargar las variables precontractuales esenciales de un contrato. En el backend, se disponen de modelos predictivos previamente entrenados y almacenados en formato pickle, los cuales se encargan de calcular la estimación de la probabilidad de que se produzca una adición en el contrato, en función de las variables proporcionadas por el usuario.

Dicha probabilidad se visualiza mediante un gráfico en forma de velocímetro que se divide en tres colores representativos de diferentes niveles de probabilidad (Ver Anexo IV). Si la probabilidad supera el 0.4, se considera que existe un riesgo significativo de adición, lo que sugiere la necesidad de implementar mecanismos de control para asegurar el cumplimiento del contrato. La frecuencia y la intensidad de estos controles pueden variar según el indicador de riesgo que refleja el velocímetro: un color amarillo justificaría realizar revisiones trimestrales, mientras que una probabilidad mayor a 0.7 podría activar una alerta de alto riesgo que requiere atención inmediata.

5 Conclusiones.

En definitiva, los modelos de aprendizaje de máquina, que ahora se observan embebidos en toda clase de herramientas denominadas inteligencias artificiales, ya entrega valor en todo tipo de organizaciones, tanto privadas como públicas. En esta oportunidad y de manera práctica, se logra con éxito ejecutar una combinación de modelos y metodologías tradicionales, como una regresión logística y la estandarización de valores, como algunos de vanguardia, como redes neuronales, BERT y Transformers, en una aplicación pública y con datos de libre acceso. Además, si bien se requirió de una capacidad computacional sobre la media de un computador personal, esto no imposibilita la ejecución de modelos con métricas decentes con los recursos y el conocimiento a la mano.

El modelo XGBoost tiene una gran capacidad de obtener resultados robustos para modelos de clasificación, a pesar de parecer del sobre ajuste a los datos de entrenamiento, sobre todo al incluir muchos parámetros. Para el presente este tipo de modelo nos entregó unos resultados exitosos, para identificar cómo las características de uno (o varios) modelo(s) se ven representados en una alta o baja probabilidad de tener adiciones en el futuro. Por ello es posible identificar las variables cuyas modificaciones pueden tener implicaciones directas en la reducción de costos adicionales futuros en contratos. Al mejorar la sensibilidad del modelo para identificar con mayor precisión los contratos propensos a adiciones, las entidades pueden anticiparse a posibles sobrecostos y gestionarlos de manera proactiva. Esta capacidad predictiva fortalece la planificación financiera y la asignación de recursos, permitiendo a las empresas optimizar sus presupuestos y evitar gastos imprevistos en proyectos contractuales.

Los análisis de los datos han proporcionado información valiosa sobre los contratos de prestación y no prestación de servicios personales. Estos análisis nos ayudan a comprender los patrones de comportamiento de las entidades en diversas situaciones, destacando la importancia de la planificación y gestión de recursos para asegurar una ejecución fluida de los proyectos. La falta de liquidez, por ejemplo, puede ser un factor crucial que detenga un proyecto, incluso si este se desarrolla según lo previsto.

Por otro lado, la inversión y gestión eficiente de los recursos propios de una entidad contribuyen al éxito de los contratos. En el caso de los contratos de prestación de servicios, al ser de menor valor suelen tener un menor riesgo individual, lo cual tienden a tener adiciones más frecuentes, posiblemente debido a una menor supervisión. En contraste, los contratos de no prestación de servicios que implican mayores cantidades o requieren habilidades técnicas especializadas suelen tener una mayor tendencia a ser extendidos o modificados, reflejando la necesidad de adaptabilidad en la gestión de proyectos de mayor envergadura.

Lo anterior se puede lograr con el análisis de las variables de mayor impacto en su etapa precontractual, esto resalta la importancia de una evaluación exhaustiva en las etapas tempranas de la formación de contratos. Variables como la naturaleza de los recursos financieros se revelan como indicadores clave en determinar la adición de un contrato. Este conocimiento empodera a los gerentes y a los equipos de contratación para tomar decisiones

más informadas y estratégicas desde el comienzo del proceso contractual, lo que puede resultar en contratos más sólidos y menos susceptibles a modificaciones costosas.

6 Trabajo Futuro.

Para futuras investigaciones, se propone el diseño de mecanismos de control ajustados a la probabilidad de adición del contrato. Estos mecanismos deberán incluir planes de acción detallados para cada escenario, que podrían basarse en el color indicado por el velocímetro de riesgo. Por ejemplo, cada tonalidad del indicador estaría asociada con una serie de medidas preventivas o correctivas específicas.

Otra línea de trabajo a desarrollar, más enfocada en el análisis, consiste en incorporar estimaciones de variables de monto y duración para predecir posibles extensiones en tiempo y coste de un contrato. La integración de estas variables en el tablero interactivo ya mencionado podría proporcionar estimaciones más precisas sobre el impacto económico de las adiciones para la entidad. Además, al ajustar ciertas variables en el dashboard, se podrían identificar y mitigar proactivamente los factores que contribuyen a la necesidad de dichas adiciones.

7 Referencias.

1. Transparencia por Colombia. (2020). Contratación pública y plataformas para acceder a la información. Recuperado de https://transparenciacolombia.org.co/Documentos/Publicaciones/cuidado-paz/caqueta/Contratacion_p%C3%BAblica_plataformas_acceder_informacion.pdf
2. Colombia Compra Eficiente. (2021, noviembre 17). SECOP- Sistema Electrónico de Contratación Pública. Recuperado de <https://www.colombiacompra.gov.co/secop/secop>
3. Colombia Compra Eficiente. (2022, agosto 2). Relatoria: [Título del documento]. Recuperado de <https://relatoria.colombiacompra.gov.co/relatoria/5/2566/4201912000007298.docx>
4. Legis. (2021, mayo 27). Etapas de contratación estatal. Recuperado de <https://blog.legis.com.co/juridico/etapas-contratacion-estatal>
5. OECD. (s.f.). Contratación pública. Recuperado de <https://www.oecd.org/gov/contratacion-publica/>
6. Oficina de las Naciones Unidas contra la Droga y el Delito (ONUDD). (2014, noviembre). Guía sobre medidas contra la corrupción en la contratación pública y en la gestión de la hacienda pública. Recuperado de https://www.unodc.org/documents/corruption/Publications/2014/14-04933_ebook.pdf

7. Servicio Civil Distrital. (2022, marzo 21). ¿Qué es la adición y prórroga de contrato? Recuperado de <https://serviciocivil.gov.co/content/%C2%BFqu%C3%A9-la-adici%C3%B3n-y-prorroga-de-contrato>
8. Cañete, J., Chaperon, G., Fuentes, R., Ho, J., Kang, H., & Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. Recuperado de <https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased>
9. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
10. Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.). Pearson.
11. Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
12. Indeed.com. Job titles starting with A. Recuperado de <https://www.indeed.com/browsejobs/Title/A>
13. Murilo [pseudónimo en Stack Overflow], Ingeniero de Software en Mint Creative Lab. (2019, diciembre 17). Recuperado de <https://stackoverflow.com/questions/59185739/getting-jobs-list-from-linkedin-api>
14. Rae. Definición de Objeto del Contrato - Diccionario panhispánico del español jurídico - Real Academia Española. Recuperado de <https://dpej.rae.es/lema/objeto-del-contrato>
15. Bird, S., Loper, E., & Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc. Recuperado de <https://www.nltk.org/>
16. Spacy. Industrial-strength Natural Language Processing in Python. Recuperado de <https://spacy.io/>
17. Brownlee, J. (2021, enero 5). A Gentle Introduction to Threshold-Moving for Imbalanced Classification - MachineLearningMastery.com. Recuperado el 23 de noviembre de 2023, de <https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/>
18. Dalgaard, P. (2008). Introductory Statistics with R. Springer New York.
19. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (2nd ed.). Springer.
20. Molnar, C. (2023, agosto 21). Interpretable Machine Learning. Recuperado el 24 de noviembre de 2023, de <https://christophm.github.io/interpretable-ml-book/>
21. Nielsen, M. (2015). Neural Networks and Deep Learning. Recuperado el 15 de noviembre de 2023, de <http://neuralnetworksanddeeplearning.com/>

22. Ramírez, L. (2023, enero 5). Algoritmo k-means: ¿Qué es y cómo funciona? Recuperado el 14 de noviembre de 2023, de <https://www.iebschool.com/blog/algoritmo-k-means-que-es-y-como-funciona-big-data/>
23. Rodríguez, F. (2023, febrero 17). ¿Qué es el TF-IDF Vectorizer? Recuperado el 14 de noviembre de 2023, de <https://keepcoding.io/blog/que-es-el-algoritmo-tf-idf-vectorizer/>
24. SECRETARÍA DISTRITAL DE GOBIERNO. (2014). Contratacion - Bogotá. Recuperado el 15 de noviembre de 2023, de <https://www.gobiernobogota.gov.co/transparencia/contratacion>
25. Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). Introduction to Data Mining. Pearson.
26. Chen, T., & Guestrin, C. (2016, enero 29). XGBoost: A Scalable Tree Boosting System. Recuperado de <https://ar5iv.labs.arxiv.org/html/1603.02754>
<https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf>

8 Anexos.

8.1 Anexo 1. Tablas.

Variables	coef
Grupo_Entidad_colegio	0.333999
Cluster2_salud	0.285937
Justificacion Modalidad de Contratacion_No Esp...	0.250914
Sector_Ley de Justicia	0.177887
Origen de los Recursos_Distribuido	0.166951
Sector_agricultura	0.166723
Sector_Vivienda, Ciudad y Territorio	0.160120
Anno BPIN_2023	0.145013
Sector_Ambiente y Desarrollo Sostenible	0.134543
Justificacion Modalidad de Contratacion_Acquis...	0.130811
Anno BPIN_2022	0.111857
Justificacion Modalidad de Contratacion_Artícu...	0.104856
Grupos_valor_contrato	0.092952

Variables	coef
Habilita Pago Adelantado_No	0.089133
Tipo de Identificación Representante Legal_Sin...	0.072008
Tipo de Contrato_Arrendamiento de muebles	0.069266
Grupo_Entidad_subred_salud	0.065661
Modalidad de Contratacion_Selección Abreviada ...	0.064320
Sector_Servicio Público	0.061613
Grupos_Saldo_CDP	0.056693
Tipo de Contrato_No Especificado	0.053272
Grupo_Entidad_entretenimiento_distrital	0.049473
Entidad Centralizada_Descentralizada	0.049227
Sector_No aplica/No pertenece	0.047077
Justificacion Modalidad de Contratacion_Arrend...	0.046580
Anno BPIN_2020	0.041511
Justificacion Modalidad de Contratacion_No exi...	0.040184

Tabla 10. Variables Significativas y sus Coeficientes

NO PRESTACIÓN.

○ Valor del contrato.

VALOR DEL CONTRATO								
Media	Desviacion	Varianza	Minimo	1Q	Mediana	3Q	Máximo	Asimetría
2.39E+11	4.07E+13	1.65E+27	0	1.19E+07	4.69E+07	3.00E+08	7.28E+15	178.6

Tabla 11. Valor del contrato No Prestación

○ Recursos Propios.

RECURSOS PROPIOS								
Media	Desviacion	Varianza	Minimo	1Q	Mediana	3Q	Máximo	Asimetría
4.61E+08	6.58E+09	4.33E+19	0	0	0	1.77E+07	5.00E+11	45.39

Tabla 12. Recursos Propios No Prestación

○ Saldo CDP.

SALDO CDP								
Media	Desviacion	Varianza	Minimo	1Q	Mediana	3Q	Máximo	Asimetría
1.07E+10	1.76E+11	4.33E+19	0	3.45E+06	5.80E+07	7.00E+08	1.32E+13	33.43

Tabla 13. Saldo CDP No Prestación

○ **Sector.**

Categoría	SECTOR	
	Sin Adición	Con Adición
Ambiente y Desarrollo Sostenible	6.07	5.87
Cultura	5.47	4.17
Educación Nacional	9.73	3.12
Hacienda y Crédito Público	2.14	1.39
Inclusión Social y Reconciliación	9.43	5.52
Industria	5.77	0.93
Ley de Justicia	1.15	1.92
No aplica/No pertenece	2.42	3.34
Planeación	1.01	0.68
Salud y Protección Social	23.83	41.67
Servicio Público	23.45	23.10
Transporte	5.15	3.08
Vivienda, Ciudad y Territorio	1.82	2.17
Agricultura	0.30	0.29
Defensa	1.88	2.24
Interior	0.39	0.52
	100	100

Tabla 14. Distribución Adición / No adición por sector No Prestación

○ **Rama.**

Categoría	RAMA	
	Sin Adición	Con Adición
Corporación Autónoma	17.42	9.57
Ejecutivo	82.58	90.43
	100	100

Tabla 15. Rama No Prestación

○ **Tipo de Contrato.**

Categoría	TIPO DE CONTRATO	
	Sin Adición	Con Adición
Acuerdo de Cooperación	0.00	0.01
Arrendamiento de Inmuebles	6.07	3.94
Arrendamiento de Muebles	0.40	0.87
Asociación Público-Privada	0.01	0.05
Comisión	0.09	0.21

Comodato	4.43	0.45
Compraventa	17.03	8.23
Concesión	0.20	0.14
Consultoría	2.09	2.93
DecreeLaw092/2017	16.54	13.56
Emprestito	0.15	0.04
Interventoría	3.22	5.73
Negocio Fiduciario	0.05	0.07
No Especificado	8.31	5.36
Obra	4.28	6.45
Otro	20.47	22.87
Seguros	1.49	2.80
Servicios Financieros	0.13	0.13
Suministros	15.00	26.15
Venta Inmuebles	0.00	0.00
Venta Muebles	0.01	0.00
	100	100

Tabla 16. Tipo de Contrato No Prestación

○ **Modalidad de contratación.**

Categoría	MODALIDAD DE CONTRATACION	
	Sin Adición	Con Adición
CCE-19-Concurso_Meritos_Con_Lista_Corta_1Sobre	0.01	0.01
CCE-20-Concurso_Meritos_Sin_Lista_Corta_1Sobre	1.81	1.66
Concurso de Méritos Abierto	3.00	6.48
Contratación Directa (con ofertas)	6.40	9.00
Contratación Directa	17.07	9.26
Contratación Régimen Especial	24.86	23.42
Contratación Régimen Especial (con ofertas)	11.68	18.36
Enajenación de Bienes con Sobre Cerrado	0.01	0.00
Enajenación de Bienes con Subasta	0.01	0.00
Licitación Pública	1.02	2.34
Licitación Pública Obra Pública	1.75	4.27
Minima Cuantía	14.76	7.96
No Definido	8.03	4.97
Selección Abreviada Menor Cuantia Sin Manifestación Interes	0.24	0.29
Selección Abreviada de Menor Cuantía	1.89	3.50
Selección Abreviada Subasta Inversa	7.46	8.49
	100	100

Tabla 17. Modalidad de Contratación No Prestación

- Entidad centralizada.

ENTIDAD CENTRALIZADA	Categoría	Centralizada	Descentralizada	No Definido
	Sin Adición	43.85	50.19	5.95
	Con Adición	29.78	68.53	1.68

Tabla 18. Entidad Centralizada No Prestación

- Destino gasto.

DESTINO GASTO	Categoría	Funcionamiento	Inversión	No Definido
	Sin Adición	47.99	49.86	2.15
	Con Adición	53.57	45.88	0.55

Tabla 19. Destino del gasto No Prestación

PRESTACION.

- Valor del contrato.

VALOR DEL CONTRATO								
Media	Desviacion	Varianza	Minimo	1Q	Mediana	3Q	Máximo	Asimetría
9.18E+10	1.92E+13	3.69E+26	0	1.23E+07	2.28E+07	4.19E+07	6.90E+15	292.95

Tabla 20. Valor del contrato Prestación

- Recursos propios.

VALOR DEL CONTRATO								
Media	Desviacion	Varianza	Minimo	1Q	Mediana	3Q	Máximo	Asimetría
1.33E+10	8.69E+12	7.56E+25	0	0	0	1.17E+07	5.76E+15	662.64

Tabla 21. Recursos Propios Prestación

- Saldo CDP.

VALOR DEL CONTRATO								
Media	Desviacion	Varianza	Minimo	1Q	Mediana	3Q	Máximo	Asimetría
1.23E+11	6.86E+13	4.70E+27	0	1.42E+07	3.87E+07	3.59E+08	4.53E+16	657.68

Tabla 22. Saldo CDP Prestación

○ **Sector.**

Categoría	SECTOR	
	Sin Adición	Con Adición
agricultura	0.24	0.05
defensa	1.59	0.32
interior	0.22	0.41
Planeación	1.12	0.49
Ley de Justicia	0.78	0.64
Industria	2.38	1.02
Hacienda y Crédito Público	2.34	1.19
Educación Nacional	5.29	2.0
Vivienda, Ciudad y Territorio	2.73	2.52
Transporte	6.27	2.97
No aplica/No pertenece	2.21	3.25
Cultura	6.65	4.95
Ambiente y Desarrollo Sostenible	4.80	5.67
Inclusión Social y Reconciliación	16.90	5.84
Servicio Público	13.20	13.02
Salud y Protección Social	33.25	55.66
	100	100

Tabla 23. Distribución Adición / No adición por sector Prestación

○ **Rama.**

Categoría	RAMA	
	Sin Adición	Con Adición
Corporación Autónoma	18.75	7.35
Ejecutivo	81.24	92.65
	100	100

Tabla 24. Rama Prestación

○ **Tipo de Contrato.**

Categoría	TIPO DE CONTRATO	
	Sin Adición	Con Adición
Arrendamiento de muebles	0.0	0.0
Comisión	0.0	0.0
Compraventa	0.02	0.01
DecreeLaw092/2017	7.15	0.4
Obra	0.0	0.0
Otro	0.03	0.03
Prestación de servicios	92.74	99.44
Seguros	0.0	0.03
Servicios financieros	0.0	0.01
Suministros	0.05	0.08
	100	100

Tabla 25. Tipo de Contrato Prestación

- **Modalidad de contratación.**

Categoría	MODALIDAD DE CONTRATACIÓN	
	Sin Adición	Con Adición
Contratación Directa (con ofertas)	0.16	0.05
Contratación directa	70.34	50.66
Contratación régimen especial	28.06	46.88
Contratación régimen especial (con ofertas)	0.12	0.37
Licitación pública	0.14	0.47
Mínima cuantía	0.74	0.70
Selección Abreviada Menor Cuantía Sin Manifestación Interés	0.01	0.02
Selección Abreviada de Menor Cuantía	0.31	0.61
Selección abreviada subasta inversa	0.13	0.24
	100	100

Tabla 26. Modalidad de Contratación Prestación.

- **Entidad centralizada.**

ENTIDAD CENTRALIZADA	Categoría	Centralizada	Descentralizada	No Definido
	Sin Adición	40.46	57.26	2.28
	Con Adición	21.78	77.13	1.08

Tabla 27. Entidad Centralizada Prestación

- **Destino gasto.**

DESTINO GASTO	Categoría	Funcionamiento	Inversión	No Definido
	Sin Adición	36.93	62.51	0.56
	Con Adición	55.76	43.71	0.54

Tabla 28. Destino del Gasto Prestación.

Agrupación etiquetas grupo entidad por expresiones regulares

Etiquetas	Entidades	Regular Expression
alcaldia_local	Alcaldías locales	(?i).*alcald.a.*local.*
colegio	Colegios	(?i).*colegio.*
	Instituciones Educativas	(?i).*instituci.n.*educa.*
entretenimiento_distrital	Jardín Botánico	(?i).*bot.nico.*
	Orquesta Filarmónica de Bogotá	(?i).*filarm.nica.*
	Canal Capital	(?i)canal.capital.*
fondo_desarrollo_local	Fondo de Desarrollo Local	(?i).*fondo.*local.*
instituto_distrital	Institutos Distritales	(?i).*instituto.*distr.*

Etiquetas	Entidades	Regular Expression
	IDU - Instituto de Desarrollo Urbano	(?i)instituto.*desar.*urb.*
secretaria_distrital	Secretarías Distritales	(?i).*secretar.a.*distr.*
	Secretaría General de Bogotá	(?i).*secretar.a.*gene.*
	Fondo Distrital de Salud - Dependiente	(?i).*fondo.*distr.*salud.*
	Secretaría Distrital de Salud	(?i).*fondo.*distr.*salud.*
servicios_publicos	Empresas que terminan en ESP - Empresa Servicios Públicos	(?i).*esp(?: \$)
	Empresas que dicen ser de servicios públicos	(?i).*serv.*p.bli.*
subred_salud	Subredes de salud	(?i).*subred.*salud.*
transporte_publico	Transmilenio	(?i).*trans?milenio.*
	Metro de Bogotá	(?i).*metro.*bogot+.*
unidad_admin_especial	Unidades Administrativas Especiales	(?i).*unidad.*admin.*
	Unidad Admin. Es. de Catastro Distrital	UAECD
otros	Las demás entidades	

Tabla 29. Agrupación etiquetas grupo entidad por expresiones regulares.

Tipo Entidad	Regular Expresion
colombiana	(?i)col.*
	(?i).*bian.*
	(?i).*mbia.*
	(?i)bogot.*
	(?i)cundin.*
latam	(?i)venez.*
	(?i)bra.*
	(?i)ecua.*
	(?i)salvad.*
	(?i)cuba.*
	(?i)arge.*
	(?i)boliv.*
	(?i)mexi.*
	(?i)meji.*
	(?i)chil.*
	(?i)peru.*

Tipo Entidad	Regular Expresion
	(?i)haiti.*
	(?i)costarr.*
	(?i)domini.*
	(?i)hondur.*
vacio	(?i).*no def.*
	(?i).*nada.\$
	(?i).*no apl.*
	(?i).*sin des.*
Otros	Los demás

Tabla 30. Clasificación Tipo entidad Expresiones regulares

Modelo	Hiper-parámetro	Descripción	Grilla
Random Forest	max_depth	Profundidad máxima del árbol.	[3, 6, 10]
	min_samples_leaf	Número mínimo de muestras en una hoja	[1, 2]
	n_estimators	Cantidad de árboles	[10, 100, 1000]
XGBoost	max_depth	Profundidad máxima del árbol.	[3, 6, 10]
	subsample	Fracción de muestras a utilizar para entrenar cada árbol de decisión	[0.5, 1]
	n_estimators	Cantidad de árboles	[100, 200]
Support Vector Machine Classifier (SVC)	C	Parámetro de regularización	[0.1, 1, 10]

Tabla 31. Hiperparámetros modelos no prestación servicios

Modelo	Hiper-parámetro	Descripción	Grilla
--------	-----------------	-------------	--------

Random Forest	max_depth	Profundidad máxima del árbol.	[3,10]
	n_estimators	Cantidad de árboles	[100,200]
XGBoost	max_depth	Profundidad máxima del árbol.	[3, 10]
	n_estimators	Cantidad de árboles	[100, 200]
Support Vector Machine Classifier (SVC)	C	Parámetro de regularización	[0.1, 10]
Regresión Logística	C	Parámetro de regularización	[0.1, 1, 10]

Tabla 32. Hiperparámetros modelos prestación servicios.

Mejores Resultados No Prestación

Modelo	Parámetros	Precisión		Recall		F1		AUC	
		test	train	test	train	test	train	test	train
XGBoost	max_depth: 6; n_estimators: 100; subsample:1	0,760	0,849	0,690	0,786	0,723	0,816	0,882	0,946
SVC	C:10	0,749	0,827	0,690	0,767	0,718	0,796	0,870	0,927
Random Forest	max_depth: 10; min_samples_leaf: 1; n_estimators: 1000	0,764	0,789	0,587	0,611	0,664	0,689	0,858	0,878

Tabla 33. Mejores resultados No prestación

Mejores Resultados Prestación

Modelo	Parámetros	Precisión	Recall	F1	AUC
--------	------------	-----------	--------	----	-----

		test	train	test	train	test	train	test	train
XGBoost	max_depth: 10; n_estimators: 100;	0.840	0.797	0.701	0.661	0.723	0.764	0.898	0.926
SVC	C:10	0.802	0.791	0.624	0.614	0.691	0.702	0.863	0.875
Random Forest	max_depth: 10; n_estimators: 100	0.828	0.824	0.466	0.463	0.592	0.597	0.857	0.861
Regresión Logística	C: 1 (sin regularización)	0.710	0.709	0.530	0.530	0.607	0.607	0.812	0.812

Tabla 34. Mejores resultados Prestación.

8.2 Anexo 2. Figuras.

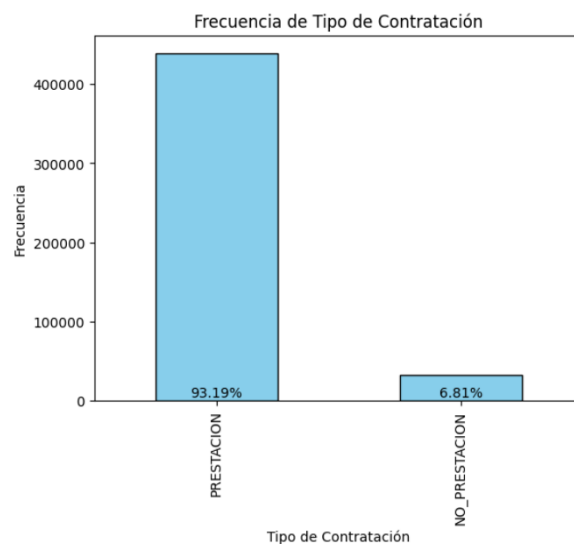


Ilustración 16. Tipos de contratos

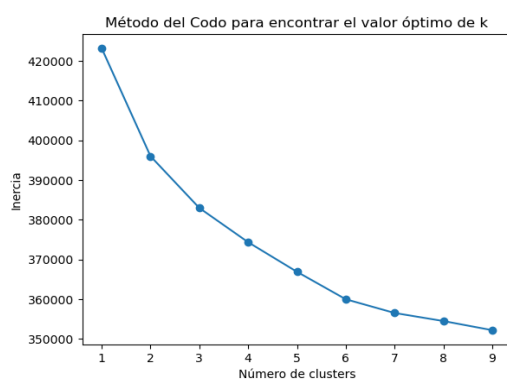


Ilustración 17. Método del codo selección de número de Clusters

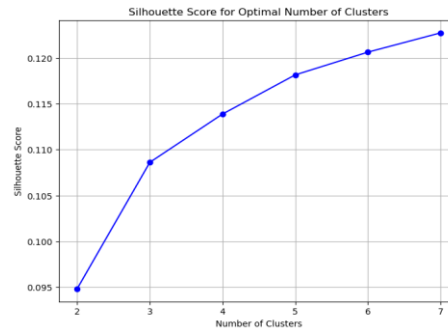


Ilustración 18. Score Silhouette

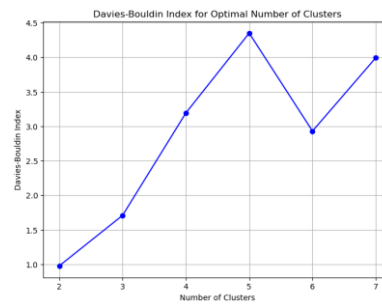


Ilustración 19. Davis-Bouldin Index



Ilustración 20. Nube de palabras presentes en el objeto del contrato Cluster 1,2,3 respectivamente

8.3 Anexo 3. Categorizaciones.

NO PRESTACIÓN.

A continuación, se presentan los resultados de algunas variables que se consideran relevantes una vez hecha la categorización.

- **Valor del contrato:** Se parte de la división de la variable en deciles, de esta división obtenemos la siguiente tabla de contingencia:

Grupos	0	1	2	3	4	5	6	7	8	9
Sin Adición	14.0	11.8	10.5	10.5	10.4	9.3	8.4	8.4	8.2	8.1
Con Adición	2.6	6.7	8.9	8.9	9.2	11.1	12.9	12.8	13.2	13.3

Tabla 35. Categorización Valor del contrato.

Se observan diferencias en la distribución de los grupos entre los registros identificados con adiciones (1) y los que no fueron identificados con adiciones (0). Adicionalmente, se observa que la proporción de contratos que no tuvieron adición disminuye a medida que el valor del contrato es más alto, a diferencia del comportamiento en los contratos con adición.

- **Recursos propios:** Para esta variable, la categoría (0) contiene todos los registros sin recursos propios. Para los demás registros se calcula la media que tuvo un valor aproximado de 54 millones para recursos propios y 95 millones para recursos propios de Alcaldías, Gobernaciones y Resguardos Indígenas. Así, con los contratos con recursos propios por debajo de la media se conforma la categoría (1) y con los mayores a ese valor la categoría (2).

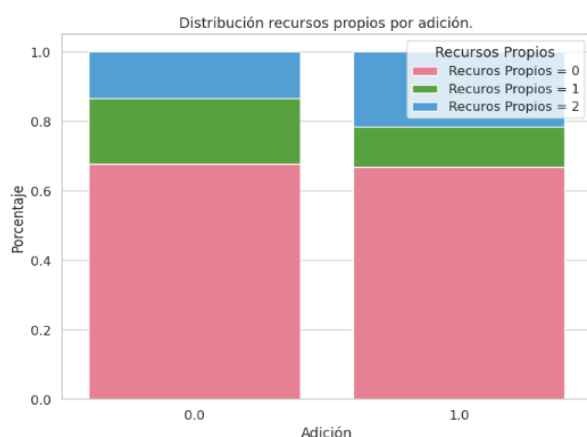


Ilustración 21. Distribución recursos Propios por Adición

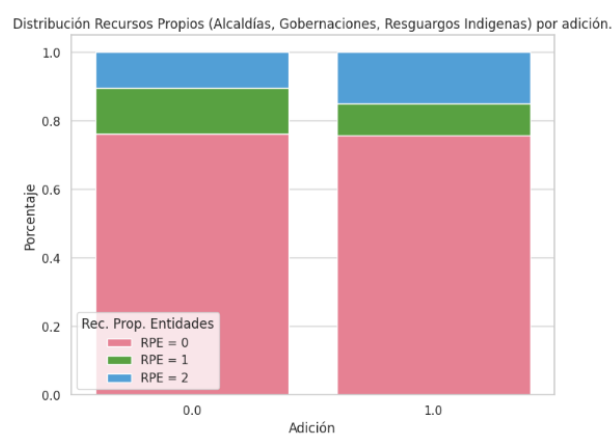


Ilustración 22. Distribución recursos Propios (Alcaldías, Gobernaciones) por Adición

Se observa que los contratos con adiciones y los que no tuvieron adiciones tienen un porcentaje similar de contratos sin recursos propios, sin embargo, para los contratos con adiciones hay un porcentaje más alto de altos recursos.

- **Presupuesto General de la Nación PGN:** Para esta variable, se han marcado como (1) los contratos que tienen algún valor diferente de cero.

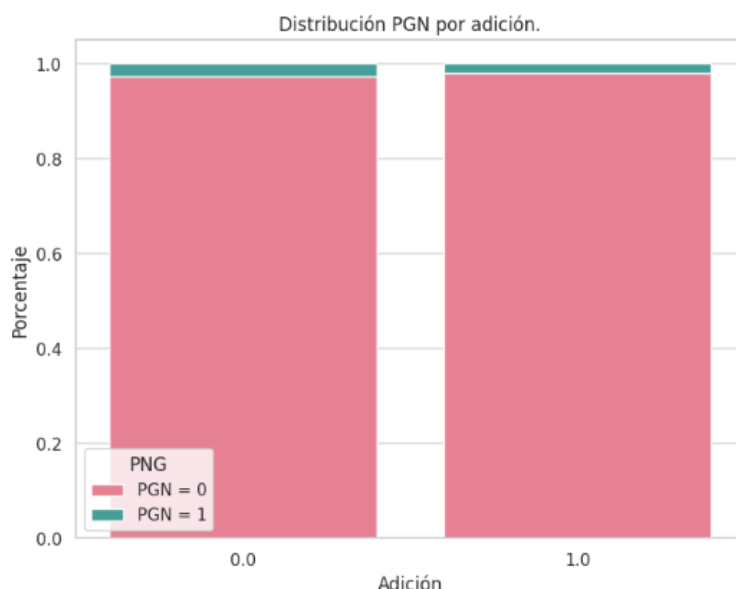


Ilustración 23. Presupuesto General de la Nación por Adición

Aunque es ligera, se observa diferencia en la proporción de contratos con PGN entre los registros de adiciones y no adiciones.

PRESTACIÓN.

- **Valor del contrato:** Se parte de la división de la variable en deciles, de esta división obtenemos la siguiente tabla de contingencia:

Grupos	0	1	2	3	4	5	6	7	8	9
Sin Adición	12.0	10.9	10.4	10.4	9.8	9.6	10.1	9.3	9.3	8.2
Con Adición	5.9	8.3	9.2	9.3	10.5	10.7	9.9	11.4	11.4	13.5

Tabla 36. Deciles valor del contrato prestación

Se observan diferencias en la distribución de los grupos entre los registros identificados con adiciones (1) y los que no fueron identificados con adiciones (0). Adicionalmente, se observa que la proporción de contratos que no tuvieron adición disminuye a medida que el valor del contrato es más alto, a diferencia del comportamiento en los contratos con adición.

- **Recursos propios (Alcaldías, Gobernaciones, Resguardos Indígenas):** Para esta variable, la categoría (0) contiene todos los registros sin recursos propios. Para los demás registros se calcula la media que tuvo un valor aproximado de 26 millones. Así, con los contratos con recursos propios por debajo de la media se conforma la categoría (1) y con los mayores a ese valor la categoría (2).

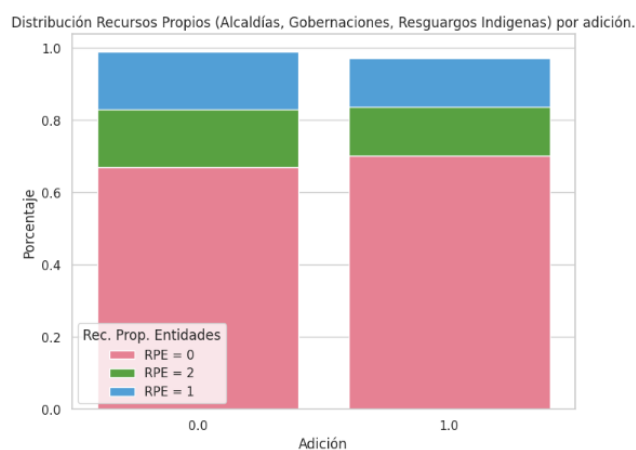


Ilustración 24. Distribución de recursos propios por Adición Prestación.

- **Presupuesto General de la Nación PGN:** Para esta variable, se han marcado como (1) los contratos que tienen algún valor diferente de cero.

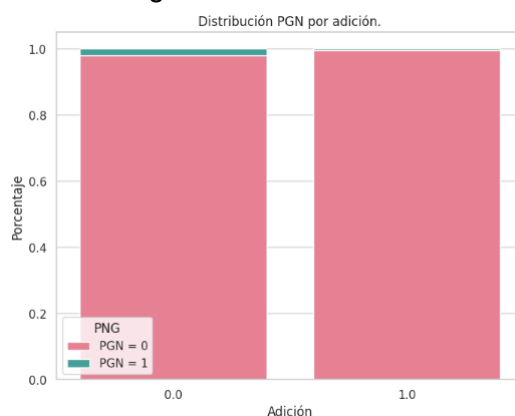


Ilustración 25. Distribución PGN por adición Prestación

Se observa diferencia en la proporción de contratos con PGN entre los registros de adiciones y no adiciones, para estos últimos el PGN es casi nulo.

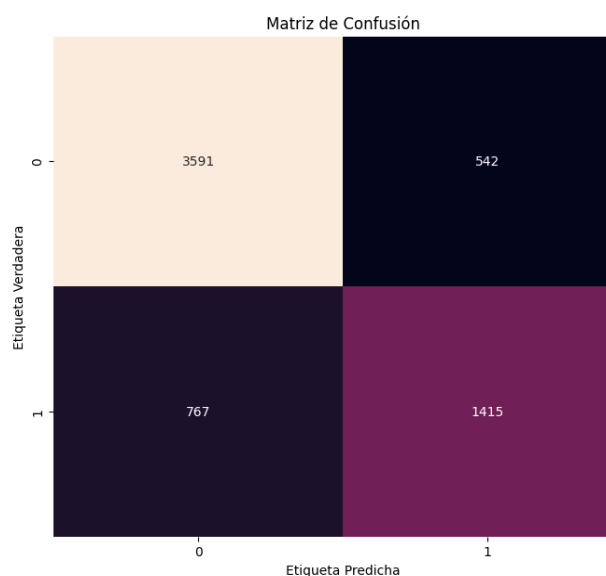


Ilustración 26. Matriz de confusión No prestación

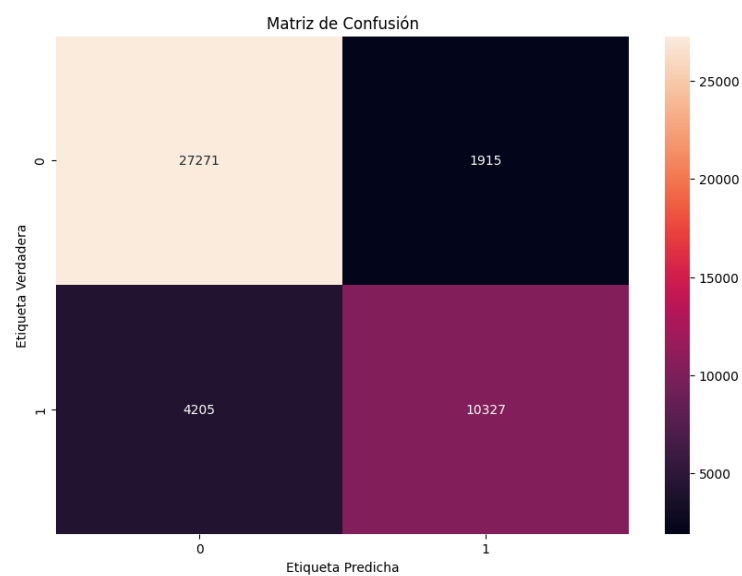


Ilustración 27. Matriz de confusión prestación

Matrices de confusion ajustadas

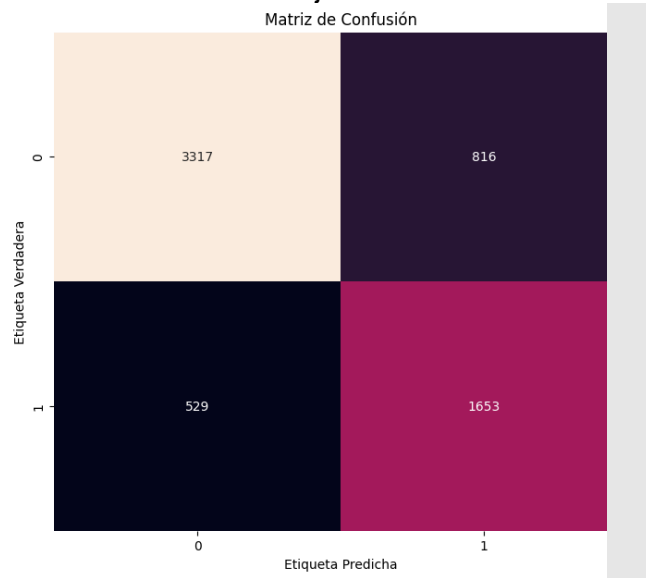


Ilustración 28. Matriz de confusión ajustada No prestación.

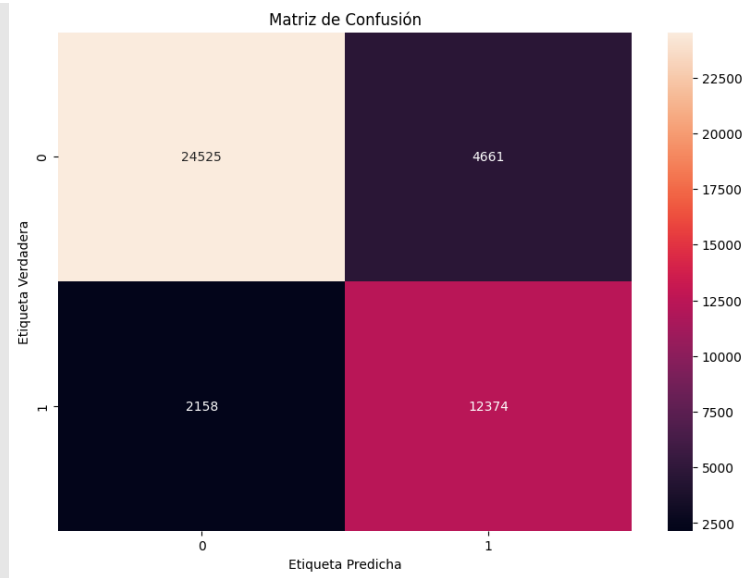




Ilustración 29. Matriz de confusión ajustada prestación.

8.4 Anexo 4. Ilustración de la Herramienta de cálculo de probabilidad en Dash.




Colombia
COMPRA EFICIENTE
Agencia Nacional de Contratación Pública



Detección de contratos públicos susceptibles a adiciones

Proyecto de Grado Maestría

Oscar Otálora | Camilo Céspedes | Julián Gómez | Juan Caballero



MAESTRÍA EN INTELIGENCIA ANALÍTICA
PARA LA TOMA DE DECISIONES
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
Universidad de los Andes
Facultad de Ingeniería
Analytics FORUM

Descargar Formato CSV

Subir Archivo CSV

Sector	Rama	Entidad Centralizada
Ambiente y Desarrollo Sostenible	Ejecutivo	Descentralizada
Tipo de Contrato	Modalidad de Contratación	Justificación Modalidad de Contratación
Acuerdo de cooperación	CCE-20-Concurso_Meritos_Sin_Lista_Corta_1 Sobre	Arrendamiento de inmuebles
Habilita Pago Adelantado	Obligación Ambiental	Obligaciones Postconsumo
No Definido	Si	Si
Reversion	Año BPIN	ExPostConflicto
No	2017	No
Destino Gasto	Origen de los Recursos	Puntos del Acuerdo
Funcionamiento	Recursos Propios	ReformaRuralIntegral
Pilares del Acuerdo	Tipo de Identificación Representante Legal	Género Representante Legal
OSPRUDS	Cédula de Ciudadanía	Mujer
Recursos Propios	Saldo CDP	Saldo Vigencia
100000000	500000000	75000000
Presupuesto General de la Nación – PGN	Sistema General de Participaciones	Sistema General de Regalías
350000000	250000000	40000000
Recursos Propios (Alcaldías, Gobernaciones y Resguardos Indígenas)	Recursos de Crédito	
50000000	100000000	

Nombre Entidad

Ingrese el nombre de la entidad. Cuide la ortografía

Descripción del Proceso

Puede dejar este espacio vacío. Tenga en cuenta que esto impacta el resultado.

Probabilidad de Adición

0.2

▲0.3