Dossier. Analítica de datos

Detección de contratistas multiobjeto mediante minería de textos para focalizar el ejercicio del control y vigilancia fiscal

Detection of Multi-object Contractors through Text Mining to Targeting the Exercise of Fiscal Control and Surveillance

Dulce Vanegas, Manuel Francisco; Beltrán Gómez, Adam

Manuel Francisco Dulce Vanegas
manuelf.dulcev@utadeo.edu.co
Universidad de Bogotá Jorge Tadeo Lozano , Colombia
Adam Beltrán Gómez
revista.mutis@utadeo.edu.co
Universidad Católica de Colombia, Colombia

Revista Mutis
Universidad de Bogotá Jorge Tadeo Lozano, Colombia
ISSN: 2256-1498
Periodicidad: Semestral
vol. 11, núm. 1, 2021
revista.mutis@utadeo.edu.co

Recepción: 18 Noviembre 2020
Aprobación: 15 Marzo 2021

URL: http://portal.amelica.org/ameli/journal/193/1933515005/

DOI: https://doi.org/10.21789/22561498.1732

Resumen: Las entidades fiscalizadoras superiores, y en específico su ente rector, la Organización Internacional de las Entidades Fiscalizadoras Superiores (INTOSAI), han impulsado en los últimos cuatro años iniciativas encaminadas al uso de tecnologías y métodos para sus procesos de vigilancia y fiscalización que sean replicables y que generen resultados tangibles en el contexto fiscal. En este sentido, la Contraloría General de la República de Colombia viene fortaleciendo su infraestructura tecnológica y capacidades técnicas con mirar a mejorar y optimizar sus esfuerzos en cuanto a la vigilancia de los recursos de los colombianos. Aunque dicha tarea no es sencilla, esta entidad ha logrado detectar patrones de aquellos contratistas que acaparan la contratación estatal, logrando estar en diferentes sectores económicos sin tener probablemente la competencia técnica para cumplir el objeto contractual estipulado. A estos se les conoce en el ámbito de la Contraloría General como contratistas "multiobjeto". En el presente artículo se muestra la construcción de un conjunto de datos de 1.998 registros etiquetado por expertos, que corresponden a contratos del sector educativo en Colombia. Con este instrumento se llevó a cabo el entrenamiento y las pruebas sobre un clasificador automático construido para los objetos contractuales a fin de detectar presuntos contratistas "multiobjeto". Adicionalmente, se encontró que el mejor algoritmo de clasificación fue "Máquina de Soporte Vectorial Lineal", con una exactitud de 84 %, el cual permitió finalmente listar por agrupamiento los presuntos contratistas de este tipo.

Palabras clave: minería de texto, aprendizaje de má-.

Abstract: Supreme audit institutions, and specifically its governing body, the International Organization of Supreme Audit Institutions (intosai), have promoted during the last four years a series of initiatives in the fiscal context aimed at the use of technologies and methods that are replicable and generate tangible results, thus reinforcing the surveillance and auditing processes carried out by supreme audit institutions. In this sense, the Comptroller General of the Republic of Colombia has been strengthening its technological infrastructure and technical capacities in order to improve and optimize its efforts in the monitoring of the resources of Colombian citizens. Although this task is not an easy one, this entity has managed to detect patterns of contractors who monopolize state contracting and

are inserted into different economic sectors, without probably having the technical competence to fulfill stipulated contractual deeds. These subjects are known in the field of the General Comptroller's office as "multi-object" contractors. This article explains the construction of a data set of 1,998 records labeled by experts that correspond to education sector contracts. Training and tests were carried out with this tool on an automatic classifier built for the contractual objects in order to detect suspected "multi-object" contractors. It was found that the best classification algorithm was the "Linear Vector Support Machine," with an accuracy of 84%, which will eventually find presumed multi-object contractors by grouping.

Keywords: Text mining, machine learning, state procurement, fiscal control and surveillance.

INTRODUCCIÓN

Entre los diferentes procedimientos para la detección del fraude, en el marco de la Séptima Conferencia de Criminología Financiera (Othman et al., 2015) se manifestó que las operaciones de auditoría siguen siendo las técnicas más relevantes para su detección, lo cual revela la importancia del rol de los auditores en la detección de este fenómeno. Aunque las operaciones de auditoría se encuentren supeditadas a la legislación o las características sociodemográficas de cada país, existen métodos aplicables a diferentes contextos donde el proceso auditor requiera realizar la búsqueda de patrones relacionados con la incidencia de actividades fraudulentas. Entre estos se encuentra la minería de datos, que permite realizar clasificaciones, clusterización y segmentación de datos de manera semiautomatizada, como ocurrió en Rumania para detectar patrones de fraude en el sistema de salud de este país (Bologa et al., 2010).

Mientras tanto, la Organización Internacional de las Entidades Fiscalizadoras Superiores (intosai), ente rector de las entidades fiscalizadores superiores (efs), cuyo objetivo principal es "promover el intercambio de ideas, experiencias y conocimientos entre las efs de países alrededor del mundo" (intosai, s. f.), realiza importantes esfuerzos para facilitar el acceso y aprovechamiento de las iniciativas de desarrollo de capacidades para que las efs cumplan con su labor en cada nación. Según Harib Saeed Al-Amimi, presidente de intosai, "la automatización robótica de procesos tiene el potencial de consumir trabajo de auditoría repetitivo y hacerlo de manera más precisa, confiable e incansable, en una fracción del tiempo" (2020, p. 5).

Debido a lo anterior, se han suscitado iniciativas en esta materia dentro de las efs, tales como la identificación de riesgos en la ejecución de los recursos públicos asociados bien sea al contrato, la entidad pública contratante o al contratista. Por ejemplo, la Contraloría de Bogotá (2018), entidad fiscalizadora territorial, expuso un ejercicio de uso de técnicas de minería de datos en las labores de auditoría fiscal durante el Concurso Regional sobre Buena Gobernanza, edición 2018 (Giraldo-Polanía et al., 2018). En dicho ejercicio, la Contraloría de Bogotá logró establecer un proceso de responsabilidad fiscal por USD 72.000 en dos hallazgos, tras identificar fraude en la asignación de bonos de alimentos para las vigencias 2016 y 2017 mediante el cruce de información y la extracción de textos.

De igual forma, la Contraloría General de la República de Colombia (CGR), ente superior de control fiscal a nivel nacional, dio a conocer en 2019 la iniciativa Océano (ahora Dirección de Análisis, Información y Reacción Inmediata), que surge con el fin de brindar una solución a la brecha entre la labor de vigilancia en la correcta ejecución de los recursos públicos y la capacidad tecnológica para poder cumplir con su labor constitucional. Dentro de los logros obtenidos, se pudo establecer con certeza las cifras de contratación, los riesgos del proceso (mediante mallas empresariales) y los cruces con otras fuentes de información, con lo cual

ha sido posible establecer inhabilidades de contratación y analizar a los contratistas multiobjeto, que como indica Carlos Felipe Córdoba Larrarte (2019), contralor general de la república, son aquellos contratistas que reflejan patrones en la forma de asociarse temporalmente con personas o empresas con experiencia para conseguir contratos en diferentes sectores, por lo cual generan riesgos de posible cartelización, atrasos en la ejecución y, en últimas, redundan en incumplimientos contractuales.

A pesar de los esfuerzos en cuanto a la detección de fraudes antes mencionados, aún no existen trabajos que automaticen la identificación de contratistas multiobjeto. Por este motivo, el presente trabajo aporta al proceso de detección de este tipo de contratistas, con el fin de facilitar la labor del proceso auditor mediante un modelo multiclase. Lo anterior se encuentra soportado en estudios que sugieren que "la detección del fraude mediante el empleo de un proceso automático permite clasificar de forma masiva operaciones o sujetos e identificar casos de alto riesgo" (Álvarez Jareño et al., 2018, p. 4), basados principalmente en cuatro pasos: la detección del fraude, la investigación del fraude, la confirmación de este y su prevención, lo cual ha sido fundamental para frenar procesos de fraude y corrupción en entidades financieras o gubernamentales, incluso antes de que ocurran.

Para lograr el propósito mencionado, en esta investigación se utilizaron los algoritmos de mayor uso en la clasificación multiclase, que según Mohamed (2005) son: árboles de decisión, k-vecinos cercanos, Bayes ingenuo, máquinas de soporte vectorial y redes neuronales perceptrón multicapa. Los árboles de decisión (Song & Lu, 2015) permiten clasificar los ejemplos, tomándolos principalmente desde la raíz hasta algún nodo hoja y particionando los datos en función del valor del atributo. K-vecinos cercanos (García et al., 2018) es un algoritmo de clasificación que usa los datos directamente para la clasificación, por lo que no se hace necesario considerar detalles de la construcción del modelo para estimar la pertenencia o no en una clase puntual, siendo la cantidad de vecinos cercanos k su único parámetro ajustable. Por su parte, Bayes ingenuo multinomial (García et al., 2018) es un clasificador probabilístico basado en la predicción de la probabilidad de posibles resultados en una categoría, para los que se requieren ejemplos previamente definidos. De otro lado, máquinas de soporte vectorial (Rennie & Rifkin) son algoritmos de aprendizaje supervisado para problemas de regresión y reconocimiento de patrones que construyen su solución basándose en un subconjunto de los datos de entrenamiento; estos se clasifican en svm lineal y no lineal, donde el modelo lineal parte de un conjunto de puntos etiquetados para entrenamiento linealmente separable y el no lineal resuelve el problema de los datos no separables hasta cierto grado, guardando un margen de error en la clasificación. Por último, las redes neuronales de tipo perceptrón multicapa (Hsu, 2020) son un tipo especial de red neuronal en el que se apilan varias capas de perceptrones, que generan una predicción para una entrada de acuerdo con una función de activación.

En cuanto a la estructura del presente documento, la sección después de esta introducción muestra el uso del marco metodológico crisp-dm bajo el cual se realizó el entendimiento de los datos, así como el preprocesamiento y la construcción del modelo. Luego, se muestran los resultados del ejercicio realizado, donde se compara el desempeño de los diferentes algoritmos utilizados. En la sección posterior se presenta la discusión respecto a las dificultades en los datos, las categorías y la interpretación que el modelo tiene respecto a los mismos. Finalmente, se formulan las conclusiones del ejercicio, las lecciones aprendidas y los aspectos a tener en cuenta para el mantenimiento y la mejora del presente modelo.

MATERIALES Y MÉTODOS

La metodología utilizada para el desarrollo del presente proyecto fue CRISP-DM. Debido al uso generalizado de esta metodología en gran parte de los proyectos de minería de datos, "el modelo de proceso CRISP-DM tiene como objetivo hacer grandes proyectos de minería de datos, menos costosos, más confiables, repetibles, manejables y de rápida ejecución" (Wirth & Hipp, 2000, p. 2). En cada fase explicaremos las actividades desarrolladas para este proyecto, las cuales se encuentran resumidas en la figura 1.

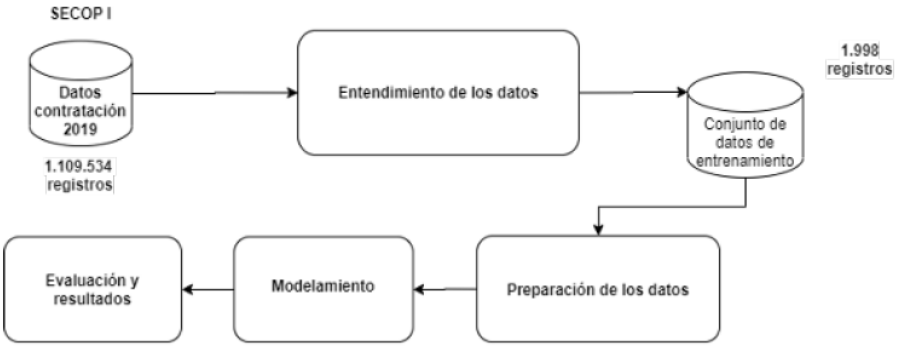

FIGURA 1.
Metodología del proyecto
Fuente: elaboración propia.

Gran parte del desarrollo se fundamentó en el ejercicio de "clasificación de las quejas de financiación del consumidor" (Li, 2018), el cual permite realizar una clasificación multiclase a partir del texto de las quejas de clientes y la categorización del producto financiero referido. Este ejercicio puede ser replicable, salvo en algunos aspectos del negocio propios del contexto del control fiscal.

En cuanto a las herramientas, se utilizó Python 3 en Google Colab, con el uso de bibliotecas como scikit-learn (Pedregosa et al., 2011), la cual tiene un ecosistema de algoritmos de clasificación propicio para el desarrollo del proyecto, y nltk para el entendimiento de lenguaje natural en el procesamiento y extracción de los textos de los objetos contractuales. En la fase de exploración de los datos también se utilizó la herramienta de minería de datos llamada knime, que permite la integración y realización de pruebas en la construcción del conjunto de datos preentrenado.

Entendimiento del negocio

En esta primera fase se brinda un contexto más amplio del estado en el que se encuentra la cgr en materia de implementación de tecnologías para el seguimiento de la contratación, así como de la manera en que este proyecto responde a la necesidad de la entidad dentro de su proceso auditor (control fiscal micro).

La CGR (2018) plantea en el objetivo 5 de su plan estratégico la búsqueda del fortalecimiento de las capacidades y los servicios tecnológicos con miras a impulsar la transformación digital. A partir de ello, la entidad ha implementado el uso de tecnologías disruptivas con el fin de analizar información masivamente y detectar patrones de comportamiento atípico que generen riesgos en la contratación y ejecución presupuestal del recurso público. En la información recopilada para sus procesos de auditoría fiscal, la cgr ha logrado recopilar varias fuentes de información contractual para contar con una base de datos más cercana a la realidad, para lo cual es necesario involucrar técnicas avanzadas de minería de datos que permitan generar valor y brindar agilidad a los procesos auditores.

En la figura 2 se observa de manera resumida el proceso auditor que realiza la CGR como ente de fiscalización, el cual consiste en la verificación de la gestión financiera, contable, administrativa y contractual de las entidades sujetas a vigilancia por parte de este órgano de control. Como en todo proceso de auditoría, es necesario que el equipo auditor conozca y planee la auditoría teniendo previo conocimiento de la información con la que cuenta la entidad a inspeccionar, la evaluación de riesgos y el alcance de la auditoría; fase que culmina con la construcción del plan y el programa de auditoría. Como siguiente etapa, se realiza la auditoría *in situ* de acuerdo con la planeación establecida y siguiendo las guías de auditoría de la CGR, culminando con la identificación de hallazgos y la elaboración del informe final. En este contexto, el aporte de este proyecto

se fundamenta en brindar las herramientas necesarias para la fase de planeación, específicamente en cuanto al conocimiento en detalle y el esquema de auditoría requerido para los contratistas multiobjeto.

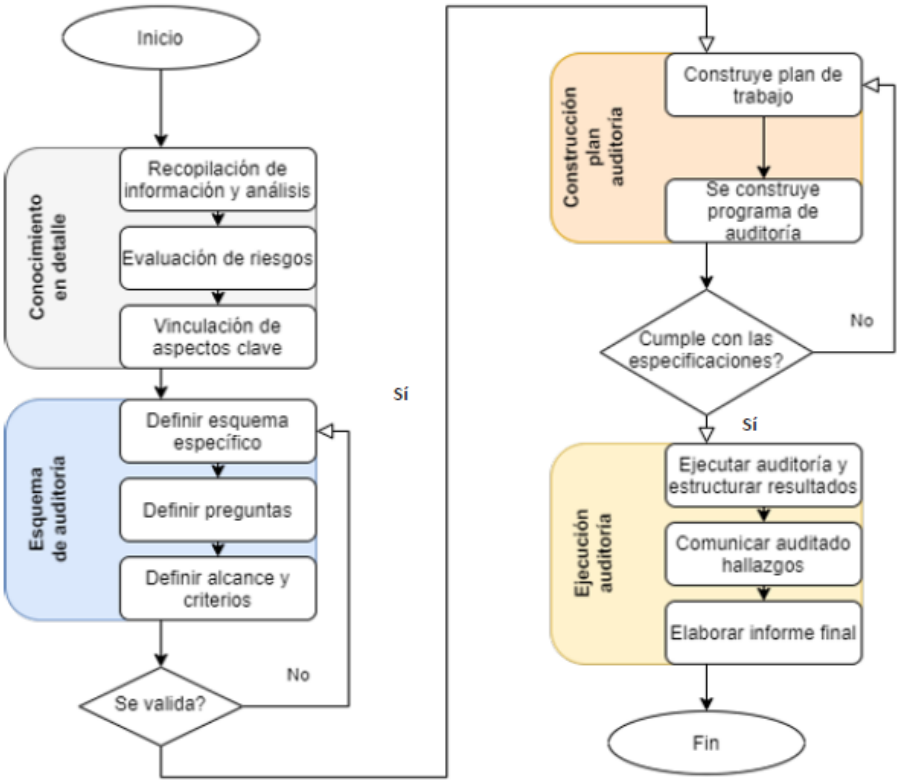

FIGURA 2.
Diagrama resumen del proceso auditor en la CGR
Fuente: elaboración propia.

Para realizar lo anterior es necesario clasificar los objetos contractuales de los registros de contratación del Estado colombiano reportados en la fuente secop i (Sistema Electrónico de Contratación Pública) para el sector de Educación, disponible en la página web de Colombia Compra Eficiente.[1] Paso seguido, mediante la puesta en práctica de técnicas de minería y clasificación de textos, se busca identificar los contratistas que tienen la característica de multiobjeto, es decir, aquellos que en la clasificación resultante tengan dos o más categorías asignadas.

Así, para el desarrollo de esta investigación se propu- sieron las siguientes actividades principales:

- Identificar las categorías óptimas de clasificación que se asignen a los objetos contractuales.
- Realizar el comparativo de validación de modelos de clasificación de texto para determinar el que más generalice la realidad.
- Ejecutar el modelo seleccionado y clasificar los objetos contractuales.
- Establecer los contratistas con multiplicidad de objetos.

Entendimiento de los datos

En esta segunda fase se describe el proceso de obtención y análisis preliminar de los datos que se utilizaron en el ejercicio. De igual forma, se realiza la identificación de las categorías utilizadas en el conjunto de datos de entrenamiento.

Se realizó la descarga de 1.109.534 registros de contratos suscritos en 2019 desde la plataforma de datos abiertos Colombia Compra Eficiente, en archivo plano (csv). Dicho archivo tiene un total de 71 atributos (tabla 1), entre los que se incluyen las variables de interés de nuestro ejercicio:

- *uid* (tipo texto): valor compuesto para identificar de manera individual cada registro.
- *Nombre de la Entidad* (tipo texto): nombre de la entidad del estado a la que corresponde el proceso, de la cual seleccionaremos los registros pertenecientes a entidades del sector educación.
- *Detalle del Objeto a Contratar* (tipo texto): detalle de la definición del bien o servicio a adquirir dentro del proceso.

TABLA 1.
Atributos del conjunto de datos de entrada y atributos de interés

Nombre campo (71)	Tipo
UID	Cadena
Anno Cargue SECOP	Número (entero)
Anno Firma del Contrato	Número (entero)
Nivel Entidad	Cadena
Orden Entidad	Cadena
Nombre de la Entidad	Cadena
NIT de la Entidad	Cadena
Código de la Entidad	Número (entero)
ID Tipo de Proceso	Número (entero)
Tipo de Proceso	Cadena
Estado del Proceso	Cadena
Causal de Otras Formas de Contratación Directa	Cadena
ID de Regimen de Contratación	Número (entero)
Regimen de Contratación	Cadena
ID Objeto a Contratar	Número (entero)
Objeto a Contratar	Cadena
Detalle del Objeto a Contratar	Cadena
Tipo de Contrato	Cadena
Municipio Obtención	Cadena
Municipio Entrega	Cadena
Municipios Ejecución	Cadena
Fecha de Cargue en el SECOP	Cadena
Número de Constancia	Cadena

Nombre campo (71)	Tipo
Número de Proceso	Cadena
Número del Contrato	Cadena
Cuantía Proceso	Cadena
ID Grupo	Cadena
Nombre Grupo	Cadena
ID Familia	Número (entero)
Nombre Familia	Cadena
ID Clase	Número (entero)
Nombre Clase	Cadena
ID Ajudicación	Número (entero)
Tipo Identificación del Contratista	Cadena
Identificación del Contratista	Cadena
Nombre Razón Social Contratista	Cadena
Departamento y Municipio Contratista	Cadena
Tipo Documento Representante Legal	Cadena
Identificación del Representante Legal	Cadena
Nombre del Representante Legal	Cadena
Fecha de Firma del Contrato	Cadena
Fecha Ini Ejecución Contrato	Cadena
Plazo de Ejecución del Contrato	Número (entero)
Rango de Ejecución del Contrato	Cadena
Tiempo Adiciones en Dias	Número (entero)
Tiempo Adiciones en Meses	Número (entero)
Fecha Fin Ejecución Contrato	Cadena
Compromiso Presupuestal	Número (entero)
Cuantía Contrato	Cadena
Valor Total de Adiciones	Cadena
Valor Contrato con Adiciones	Cadena
Objeto del Contrato a la Firma	Cadena
ID Origen de los Recursos	Número (entero)
Origen de los Recursos	Cadena
Código BPIN	Número (entero)
Proponentes Seleccionados	Cadena
Calificación Definitiva	Cadena
ID Sub Unidad Ejecutora	Cadena
Nombre Sub Unidad Ejecutora	Cadena
Ruta Proceso en SECOP I	Cadena
Moneda	Cadena
EsPostConflicto	Cadena
Marcación Adiciones	Cadena
Posición Rubro	Cadena
Nombre Rubro	Cadena
Valor Rubro	Cadena

Nombre campo (71)	Tipo
Sexo Rep Legal Entidad	Cadena
Pilar Acuerdo Paz	Cadena
Punto Acuerdo Paz	Cadena
Municipio Entidad	Cadena
Departamento Entidad	Cadena

Fuente: elaboración propia.

Posteriormente, se seleccionaron los registros del sector Educación (10.120 registros), como se muestra en la figura 3.

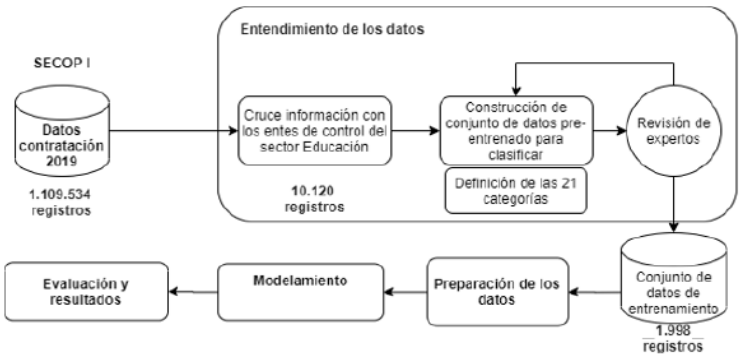

FIGURA 3.
Esquema de entendimiento de los datos
Fuente: elaboración propia.

Mediante juicio de expertos, se realizaron 3 iteraciones para discutir y establecer cuáles serían las categorías óptimas para clasificar los objetos en el sector educación. En la primera iteración se establecieron 6 categorías principales. En las siguientes iteraciones se determinó que el modelo tendría resultados más acertados al utilizar un segundo nivel, para así reducir la probabilidad de que el clasificador tuviese ambigüedades debido a la existencia de clases demasiado genéricas (por ejemplo, para el caso de los contratos de prestación de servicios, como se explica en este apartado). La tabla 2 muestra las 21 categorías definitivas para el desarrollo del modelo.

TABLA 2.
Categorías iniciales y subcategorías definitivas

Categorías Iniciales	Subcategorías Derivadas
Arrendamiento	Arriendo inmueble
	Otros arriendos
Construcción y adecuaciones	Interventoría
	Obra civil
Convenios interadministrativos	Concurso público
	Convenio gratuidad educativa
	Convenios en actividades deportivas
	Otros convenios

Categorías Iniciales	Subcategorías Derivadas
Logística, publicidad y eventos	Formación y capacitación
	Logística y transporte
Prestación de servicios	Apoyo jurídico
	Prestación de servicios
	Servicios auxiliares, técnicos y de apoyo a la gestión
	Servicios de alimentación
	Servicios de docencia
	Servicios de mantenimiento
	Servicios profesionales
Suministros	Elementos oficina y aseo
	Otros suministros
	Software y equipos de cómputo
	Compraventa

Fuente: elaboración propia.

Se seleccionó una muestra de 1.000 registros de los 10.120 disponibles para el sector analizado. Esta muestra fue validada en las primeras 6 categorías generales por medio de expresiones regulares y palabras clave en cada una de ellas. Sin embargo, durante el análisis y juicio de expertos se logró determinar una coincidencia de 78 % en la clasificación humana respecto a las reglas de expresiones regulares, identificada por los 3 expertos a cargo. Además, se detectó que las ambigüedades entre los anotadores se encuentran sesgadas hacia prestación de servicios o suministros, por lo cual se hizo necesario subcategorizar esas dos clases a fin de reducir el sesgo y balancear la muestra (figura 4).

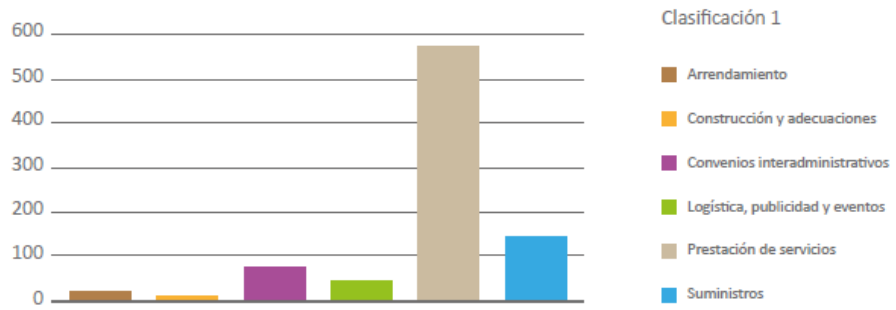

FIGURA 4.
Distribución de los datos para las 6 categorías en primer nivel
Fuente: elaboración propia en Python.

Debido a lo anterior, se optó por realizar subcategorías junto con el juicio de expertos. A partir de las categorías principales se definieron 21 subcategorías que se tomarán para reducir ambigüedades en la

clasificación. Para esto se tomó una muestra de 100 registros por subcategoría, cuando la cantidad de registros así lo permitió, los cuales fueron clasificados manualmente e incluidos en el modelo, como se muestra en la figura 5.

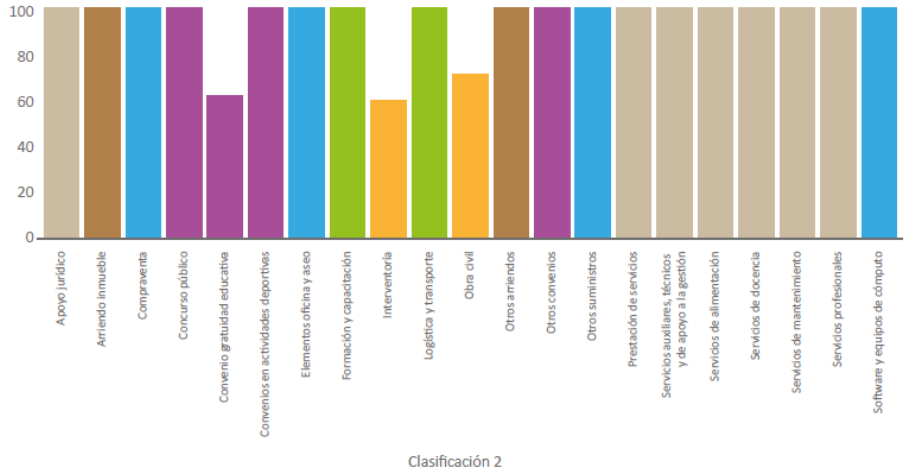

FIGURA 5.
Distribución de los datos para las 21 categorías en segundo nivel
Fuente: elaboración propia en Python.

Finalmente, el conjunto de datos construido para el entrenamiento de los modelos se compone de 1.998 registros con 16 campos, de los cuales 3 serán seleccionados para la construcción del modelo: el objeto contractual (variable de entrada), clasificación2 (variable objetivo) y el uid (identificador único para integrar al conjunto de datos nuevamente y lograr identificar los contratistas multiobjeto).

Preparación de los datos

En esta fase se realiza la limpieza y mejora de la calidad de los datos incluidos en el corpus. El procesamiento de los datos, aspecto fundamental en el desarrollo de este proyecto, se muestra de manera resumida en la figura 6. Este incluyó la construcción de funciones iterables a través de expresiones regulares y el uso de otras expresiones, contenidas en la librería nltk y Scikitlearn, para la limpieza y el preprocesamiento de los datos en nuestra variable de entrada, con lo que se retiraron principalmente objetos contractuales vacíos, números, tildes, diéresis y espacios en blanco, además de reducir las letras minúsculas. La construcción de la lista de las palabras vacías en español se realizó tomando como base la lista contenida por defecto en la librería nltk, que fue complementada con palabras que generan ruido a nivel de negocio y de las geopalabras (nombres de municipios y departamentos).

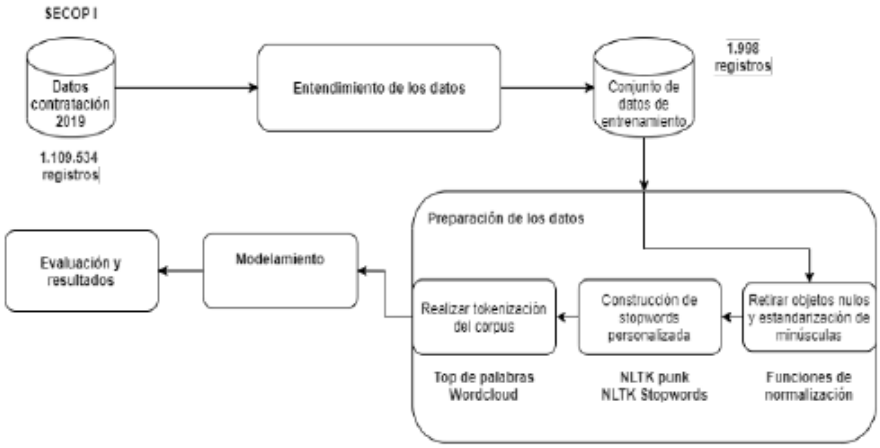

FIGURA 6.
Proceso de preparación de los datos
Fuente: elaboración propia.

Posteriormente, se realizó el proceso de división de cadenas de texto para determinar, de acuerdo con la frecuencia de aparición en el conjunto de datos, su relevancia dentro del corpus. Así mismo, se construyó una nube de palabras por cada categoría para determinar los términos más representativos dentro de dichas categorías, como se aprecia en las figuras 7 y 8.

FIGURA 7.
Nube de palabras de algunas categorías
Fuente: elaboración propia en Python

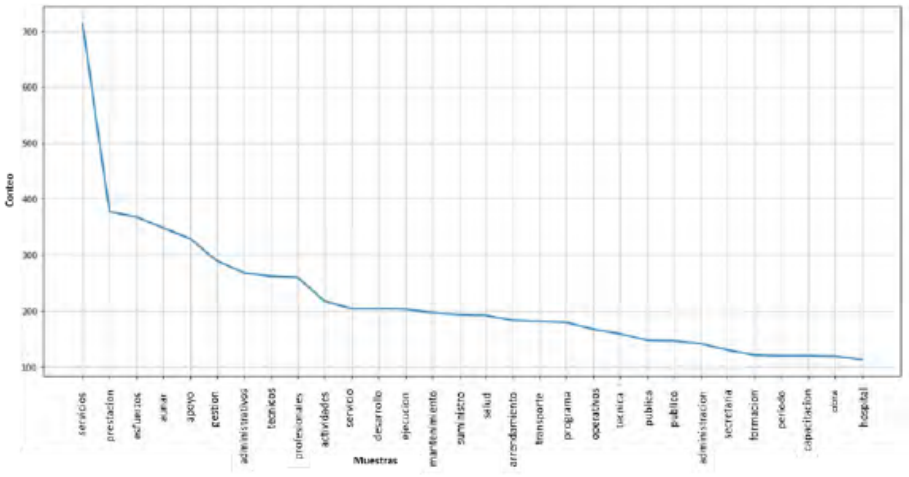

FIGURA 8.
Distribución de principales palabras en el conjunto de datos
Fuente: elaboración propia en Python.

Modelamiento

En esta fase se construye el modelo a partir de la matriz de frecuencia de términos y se realiza una exploración por medio de los ngramas seleccionados, para luego realizar una evaluación del comportamiento de los algoritmos y así identificar aquel con la mejor métrica de generalización (figura 9).

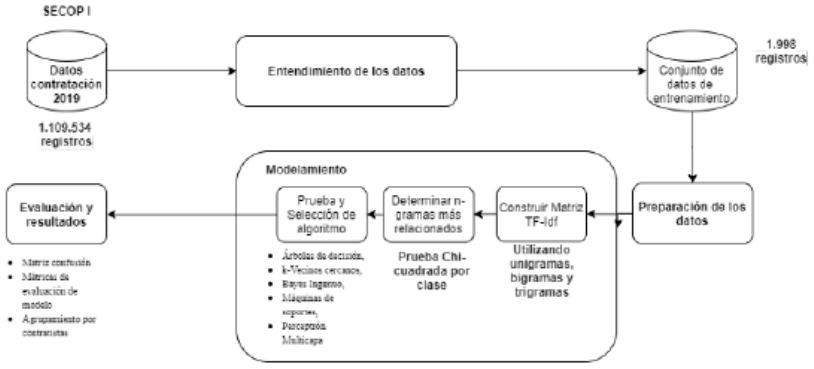

FIGURA 9.
Proceso de modelamiento
Fuente: elaboración propia.

Debido a que los algoritmos de aprendizaje y los clasificadores no pueden procesar los textos directamente, se construye una matriz de frecuencia de términosfrecuencia inversa de documentos (tfidf) con el vector de cada uno de los objetos contractuales. Para ello se establecieron los siguientes parámetros:

- sublinear_df = True, para utilizarlo de forma logarítmica.
- min_df = 6, número mínimo de documentos que debe contener una palabra para conservarla.
- norm = l2, para utilizarlo con forma euclidiana 1.
- ngram_range = (1,3), puesto que vamos a utili zar, unigramas, bigramas y trigramas.

stop_words = quitar palabras que contienen, como se indicó en la fase anterior, las palabras en español de la librería nltk y las geopalabras en Colombia.

El resultado fue una matriz [1.998, 1.750], es decir, 1.998 registros de objetos contractuales con 1.750 características que representan la puntuación para los unigramas, bigramas y trigramas.

Para revisar cuáles son los términos más relacionados en cada categoría por unigramas, bigramas y trigramas, se realizó una prueba chicuadrada. En la figura 10 se muestran algunos resultados de esta verificación.

```
*'APOYO JURIDICO':
  . unigramas más correlacionados:
       . abogado
       . juridica
       . juridico
  . bigramas más correlacionados:
       . asesoria juridica
       . profesionales abogado
       . apoyo juridico
  . trigramas más correlacionados:
       . apoyo juridico procesos
       . prestacion servicios profesionales
       . servicios profesionales abogado
*'ARRIENDO INMUEBLE':
  . unigramas más correlacionados:
       . arrendamiento
       . ubicado
       . inmueble
  . bigramas más correlacionados:
       . ubicado carrera
       . arrendamiento bien
       . arrendamiento inmueble
  . trigramas más correlacionados:
       . inmueble local comercial
       . arrendamiento casa verdad
       . arrendamiento bien inmueble
```

FIGURA 10:
Resultados prueba Chicuadrado para los términos más relevantes
Fuente: elaboración propia en Python.

Posteriormente, se realizó la partición del conjunto de entrenamiento y pruebas, el cual se definió en un 30 % a través del método estratificado para balancear las 3 clases con menos registros: "Convenio Gratuidad Educativa", "Interventoría" y "Obra civil".

Una vez se tienen las representaciones de los obje tos en vectores, estos son utilizados en los modelos seleccionados para este proyecto. La estrategia utilizada para garantizar la independencia de los datos de entrenamiento y testeo fue utilizar validación cruzada a 5 pliegues, que además permitió aumentar la robustez del estimador y mantener la rigurosidad metodológica en la construcción del modelo. La construcción del proceso descrito del modelamiento se resume en la figura 11.

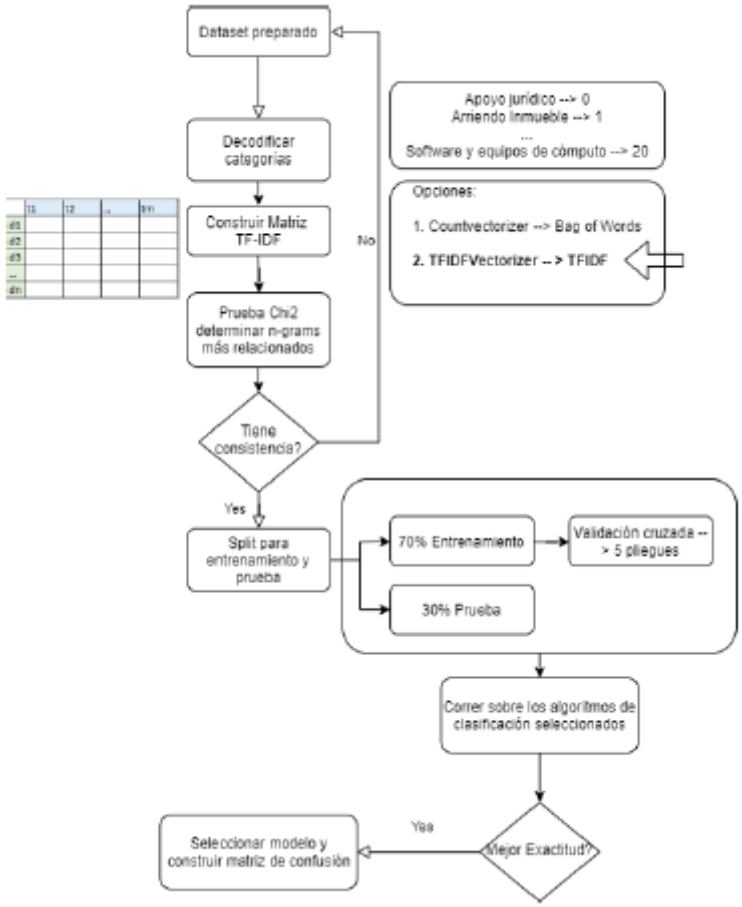

FIGURA 11.
Construcción del modelamiento del proyecto
Fuente: elaboración propia.

Con este modelo se obtendrá un avance en la identificación contextual de los objetos contractuales para las métricas en el sector educación, principalmente, que además puede ser aplicable a otros sectores de interés al extender el conjunto de categorías.

RESULTADOS

Los hiperparámetros seleccionados para los modelos comparados fueron los siguientes:

- MLPClassifier (hidden_layer_sizes=(80)) → Se establecieron 80 neuronas en 1 capa oculta.
- LinearSVC (random_state=0) → Se conservaron los parámetros establecidos por defecto.[2]
- MultinomialNB () → Se conservaron los parámetros por defecto.
- DecisionTreeClassifier (random_state=0) → Se conservaron los parámetros por defecto.
- KNeighborsClassifier (n_neighbors=100) → Se conservaron 100 vecinos cercanos.

Una vez seleccionados los modelos, se realizó una comparación para determinar cuál de ellos tiene mejor comportamiento por la métrica de Exactitud (figura 12), permitiendo establecer el siguiente orden: máquina de soporte lineal (81,68 %), árbol de decisión (81,08 %) y perceptrón multicapa (73,28 %), como se muestra en la tabla 3.

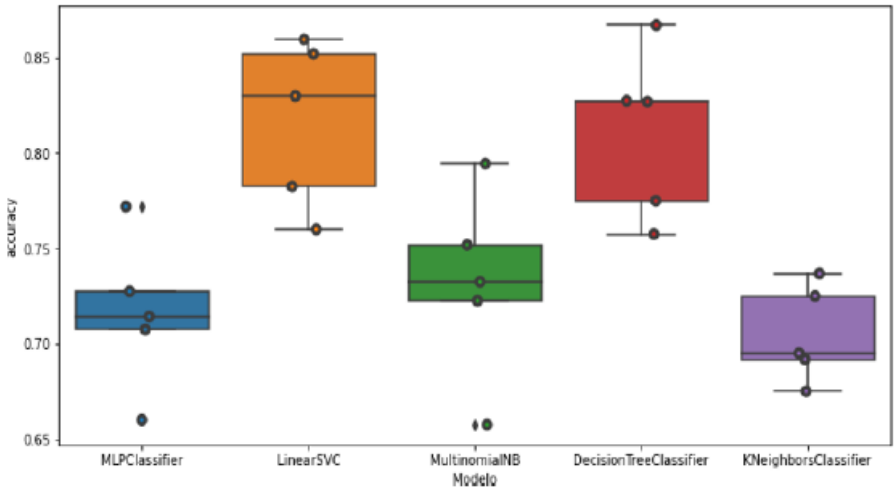

FIGURA 12.
Exactitud de los modelos utilizados en el proyecto
Fuente: elaboración propia en Python.

TABLA 3.
Porcentajes de exactitud por modelo

Clasificador	Exactitud (%)
Máquinas de soporte vectorial tipo lineal	81,68
Árboles de decisión	81,08
Clasificador perceptrón multicapa	73,28
Bayes ingenuo multinomial	73,18
K-vecinos cercanos	70,47

Fuente: elaboración propia.

Se seleccionó la máquina de soporte vectorial tipo lineal debido a que este tiene los mejores resultados. Al realizar la prueba del modelo respecto al yprueba/ypredicción, se obtuvo un comportamiento adecuado en la matriz de confusión (figura 13), señalando que su nivel de generalización es adecuado, sin presentar sobreajuste. Sin embargo, se presentan algunos casos de falsos positivos en la clasificación para las categorías "Servicios auxiliares de apoyo a la gestión", "elementos de oficina y aseo" y "otros convenios", principalmente.

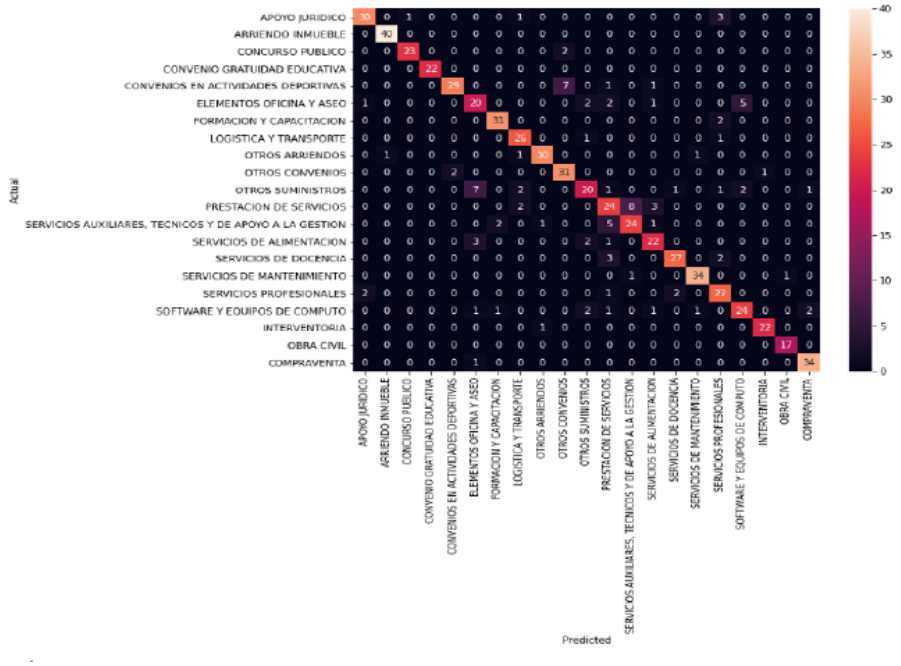

FIGURA 13.
Matriz de confusión de valores predichos vs. valores de prueba
Fuente: elaboración propia en Python.

En el análisis de los casos que en la categoría “convenios de actividades deportivas” fueron erróneamente clasificados como “otros convenios”, se observa que el contenido gramatical sí pertenece al conjunto de convenios, por lo que este nivel de error puede ser aceptable en el contexto del negocio, así como los contratos de prestación de servicios que fueron pronosticados como “servicios auxiliares, técnicos y de apoyo a la gestión”, lo cual se debe a que esta es una categoría que sigue estando en la misma línea categórica. Para el caso de los registros que siendo de “otros suministros” fueron pronosticados como “Elementos de Oficina y Aseo”, se podría considerar aumentar el tamaño de muestra para que el corpus de la matriz “documentos términos” contemple vocablos que permitan clasificar mejor esta categoría en una siguiente iteración. Finalmente, se realizó una exploración de los términos más relacionados mediante una prueba chicuadrada, encontrando que los términos se clasificaron acordes con las categorías predichas.

La tabla 4 muestra las categorías con el resultado del modelo en las métricas de Puntuación F1, Sensibilidad y Precisión.

TABLA 4.
Métricas de comportamiento del modelo

Categorías	Precisión	Sensibilidad	Puntuación F1	Casos
Apoyo jurídico	0,91	0,86	0,88	35
Arriendo inmueble	0,98	1,00	0,99	40
Concurso público	0,96	0,92	0,94	25
Convenio gratuidad educativa	1,00	1,00	1,00	22
Convenios en actividades deportivas	0,94	0,76	0,84	38
Elementos oficina y aseo	0,62	0,65	0,63	31
Formación y capacitación	0,91	0,94	0,93	33
Logística y transporte	0,81	0,93	0,87	28
Otros arriendos	0,94	0,91	0,92	33
Otros convenios	0,78	0,91	0,84	34
Otros suministros	0,74	0,57	0,65	35
Prestación de servicios	0,62	0,65	0,63	37
Servicios auxiliares, técnicos y de apoyo a la gestión	0,73	0,73	0,73	33
Servicios de alimentación	0,76	0,79	0,77	28
Servicios de docencia	0,90	0,84	0,87	32
Servicios de mantenimiento	0,94	0,94	0,94	36
Servicios profesionales	0,75	0,84	0,79	32
Software y equipos de cómputo	0,77	0,73	0,75	33
Interventoría	0,96	0,96	0,96	23
Obra civil	0,94	1,00	0,97	17
Compraventa	0,92	0,97	0,94	35
Exactitud			0,84	660
Promedio macro	0,85	0,85	0,85	660
Promedio ponderado	0,85	0,84	0,84	660

Fuente: elaboración propia en Python.

Las clases con mejor PuntuaciónF1 son “CONVENIO Y GRATUIDAD EDUCATIVA” (100#%), “ARRIENDO INMUEBLE” (99#%), “OBRA CIVIL” (97#%) e “INTERVENTORÍA” (96 %). Es muy probable que esto se deba a que dichas categorías son bastante específicas dentro del contexto del contrato y, por ende, se generalizan adecuadamente. Por el contrario, las clases con menor PuntuaciónF1 son “ELEMENTOS DE OFICINA Y ASEO” (63#%), “PRESTACIÓN DE SERVICIOS” (63#%) y “OTROS SUMINISTROS” (65#%), puesto que, como se expuso previamente, estas pueden crear ambigüedades con otras categorías que se encuentren en un contexto similar. La exactitud general para el modelo lineal con los datos de predicción es de 84#%, lo cual es bastante bueno para el propósito del proyecto.

Como el objetivo final es identificar los contratistas multiobjeto, o aquellos que están presentes en más de dos categorías, se realiza un agrupamiento por identificador de contratista y el número de categorías que se

les asignó de acuerdo al clasificador construido. Se realizó el conteo bajo estos parámetros para el conjunto de datos de los 10.120 registros del sector educación, como se aprecia en la tabla 5.

TABLA 5.
Conteo de cantidad de categorías por identificador del contratista

	Identificacion del Contratista	Clasificación2
0	899999063	12
1	900761702	11
2	900568326	9
3	30284542	8
4	800157163	8
5	860070374	8
6	900600059	8
7	901104562	7
8	900176990	7
9	901274854	7
10	810004774	7

Fuente: elaboración propia en Python.

Por último, se realiza un histograma (figura 14) para comprobar la distribución de los contratistas que tienen dos o más categorías, con el fin de que el proceso auditor pueda focalizar sus esfuerzos. Es evidente que este agrupamiento permitirá revisar los contratos sobre aquellos contratistas que tienen 3 o más categorías y así revisar su correcta ejecución.

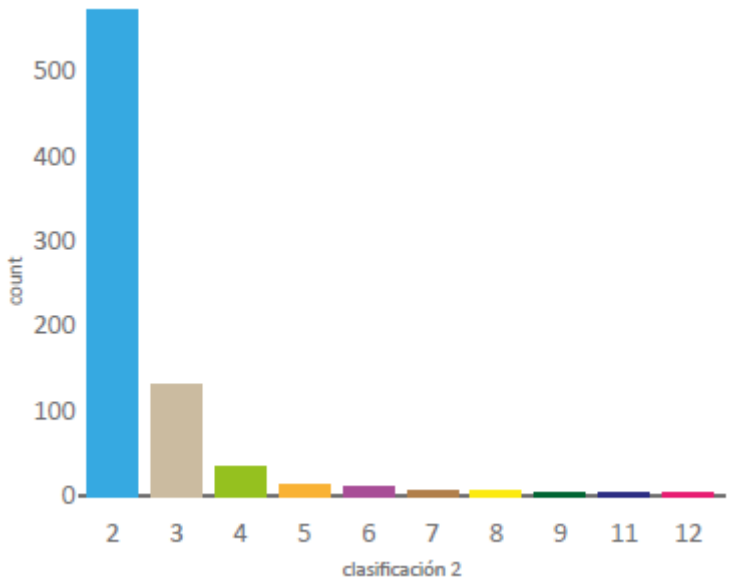

FIGURA 14.
Métricas de comportamiento del modelo
Fuente: elaboración propia en Python.

DISCUSIÓN

En el ámbito del aprendizaje automático, es de amplio conocimiento que el éxito de un modelo depende en gran medida de la calidad de los datos de entrada, más aún en el caso de aquellos enfocados a minería de texto para clasificación. Para el presente caso, el conjunto de datos de contratos en la variable de entrada, que es el objeto contractual, presenta inconvenientes de ambigüedad respecto al contexto del fin contractual; es decir, la finalidad del objeto a contratar puede confundir al algoritmo debido a la redacción del mismo, llevando a clasificarlo equivocadamente en otro (por ejemplo, "SERVICIOS PRESTAR LOS PROFESIONALES PARA BRINDAR APOYO LOGÍSTICO EN LAS OLIMPIADAS DEPORTIVAS DEL MUNICIPIO XYZ", en el cual se perciben 3 categorías: prestación de servicios, logística y transporte y convenios en actividades deportivas).

Otro aspecto que se detectó en la construcción de este modelo, y que puede mejorarse en futuras iteraciones, es la cantidad de información respecto a la descripción del objeto, que en algunos casos puede ser insuficiente para que el modelo tome una decisión acertada al momento de clasificar (por ejemplo, objeto = "MÉDICO GENERAL", "ENFERMERA").

Por otra parte, el volumen de datos por categoría hace parte esencial de este tipo de modelos, por lo cual se recomienda tener la mayor cantidad de registros preclasificados por categoría, y que estos se encuentren balanceados, pues a pesar de utilizar la estrategia de estratificación dentro de los parámetros de particionamiento en entrenamiento y prueba, es preferible que los datos se encuentren balanceados desde el inicio para reducir el nivel de sesgo en la clasificación.

En general, es necesario crear categorías específicas para cada sector, salvo aquellas que sean transversales (como apoyo jurídico, entre otros). Para ello, se recomienda que el ajuste se lleve a cabo tras un análisis preliminar de los objetos y los contextos con un grupo de expertos del negocio, para así determinar de manera focalizada las categorías óptimas del sector estudiado.

Es posible mejorar el comportamiento y la capacidad de generalización de los modelos mediante la exploración de parámetros optimizados, lo cual probablemente no solo permitirá aumentar una métrica como la eficacia, sino incluso marcar la diferencia de comportamientos entre los dos modelos que mayor exactitud obtuvieron (máquina de soporte vectorial y árbol de decisión).

Respecto al comportamiento de ngramas, los resultados muestran que son los unigramas y bigramas los que aportan más a la generalización del modelo seleccionado.

Finalmente, en el agrupamiento de los casos de presuntos multiobjeto se evidencian también clasificaciones de índole similar en la clase, situación que puede aumentar la presencia de falsos positivos. En una próxima iteración, estos últimos pueden mejorar la asertividad de la selección de la muestra mediante reglas de asignación de scores de incompatibilidad entre clases para que el auditor revise aquellos contratistas que además de tener más de 2 clases en su conteo tengan presencia de clases incompatibles en contratos suscritos.

CONCLUSIONES

Ampliar el número de las 6 categorías inicialmente seleccionadas mejoró el desempeño del algoritmo debido a la ambigüedad en el etiquetado manual realizado por los anotadores expertos del negocio para los conceptos de prestación de servicios y suministros, lo que ocasionaba que el conjunto de datos para entrenamiento no fuera lo suficientemente adecuado. Lo anterior se subsanó mediante la creación de 21 subcategorías específicas a partir de las 6 iniciales, que los anotadores etiquetaron de una manera más precisa, pues ya no se creaba la ambigüedad en las dos categorías que lo generaban.

Los algoritmos de máquinas de soporte vectorial tipo lineal fueron los de mayor desempeño al momento de clasificar las categorías de los textos de los objetos contractuales. Esto es consistente con el estado del arte sobre el tema, donde se considera que es uno de los mejores algoritmos para este tipo de ejercicios.

El hecho de que un contratista se encuentre en dos o más categorías permitiría la revisión del asunto por parte del proceso auditor de la Contraloría General de la República, permitiendo identificar un posible contratista multiobjeto y además analizar comportamientos de posible cartelización. Sin embargo, vale la pena mencionar que se realizó un ejercicio exploratorio con el primer contratista que obtuvo presencia en 12 de las 21 categorías, el cual resultó ser una institución universitaria que dentro de su actividad misional desarrolla más de una actividad en diferentes ramas (estudios, consultoría, entre otros). Al respecto, resulta pertinente aclarar que en otro tipo de organización esto no sería un hecho común, por lo que resulta conveniente revisar este tipo de casos con mayor detalle.

REFERENCIAS

Al-Amini, H. S. (2020). The future of public sector auditing: Living in times of change. International Journal of Government Auditing, 47(1), 4-5. http://intosaijournal.org/wp-content/ uploads/2020/02/INTOSAI-Journal_Winter-2020. pdf

Álvarez-Jareño, J. A., Badal-Valero, E., & Pavía, J. M. (2018). Aplicación de métodos estadísticos,económicos y de aprendizaje automático para la detección de la corrupción. Revista Internacional de Transparencia e Integridad, 9, 1-11. https://dialnet.unirioja.es/servlet/ articulo?codigo=6977094

Bologa, A. R., Bologa, R., & Flores, A. (2010). Big data and specific analysis methods for insurance fraud detection. *Database Systems Journal, 1*(1), 30-39.

Contraloría General de la República de Colombia [CGR]. (2018). Plan Estratégico cgr 2018-2022. CGR.

Córdoba-Larrarte, C. F. (2019). Océano: monitoreo eficiente en la contratación pública. *Economía Colombiana, 356*, 4-5. https://www. economiacolombiana.co/revista/oceano-393

García, J., Molina, J. M., Berlanga, A., Patricio, M. A., Bustamante, A. L., & Padilla, W. R. (2018).Ciencia de datos. Técnicas analíticas y aprendizaje estadístico. Alfaomega.

Giraldo-Polanía, L. A, Parra-Ortiz, J. W., CotrinoGarcía, Y., Dulce-Vanegas, M. F., & Tafur-Díaz, J. (2018). Big data. Análisis de caso en la Contraloría de Bogotá con la entrega de bonos. Contraloría de Bogotá. https:// www.olacefs.com/ wp-content/uploads/2018/10/1%c2%b0-PremioBogot%c3%a1-Colombia.pdf

Hsu, B. M. (2020). Comparison of supervised classification models on textual data. *Mathematics, 8*(5). https://doi.org/10.3390/MATH8050851

Li, S. (2018). Multi-class text classification with scikit-learn. https://towardsdatascience.com/ multi-class-text-classification-with-scikit-learn-12f1e60e0a9f

Mohamed, A. (2005). Survey on multiclass classification methods. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.175.107&rep=rep1&type=pdf

Organización Internacional de Entidades Fiscalizadoras Superiores [intosai] (s.f.) About us. https://www.intosai.org/about-us

Othman, R., Aris, N. A., Mardziyah, A., Zainan, N., & Amin, N. M. (2015). Fraud detection and prevention methods in the Malaysian public sector: Accountants' and internal auditors' perceptions. Procedia Economics and Finance, 28(April), 59-67. https://doi.org/10.1016/ s2212-5671(15)01082-5

Pedregosa, F., Varoquaux, G., Gramfor, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Rennie, J. D. M., & Rifkin, R. (2001). Improving multiclass text classification with the support vector machine. Massachusetts Institute of Technology.

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. Shanghai Archives of Psychiatry, 27(2), 130-135. https://doi.org/10.11919/j.issn.1002-0829.2

Wirth, R., & Hipp, J. (2000). crisp-dm: Towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining. http://citeseerx.ist.psu.edu/viewdoc/summary? https:// doi.org/10.1.1.198.5133

Notas

1 https://www.colombiacompra.gov.co/

2 Los parámetros por defecto, son aquellos con los cuales vienen predeterminados cada algoritmo en Python.