Clusterización y análisis de lenguaje para caracterizar las compras públicas de la Gobernación del Valle del Cauca en beneficio de las Mujeres del departamento.

Oliveros M, Esteban.
eoliveros@valledelcauca.gov.co, estebanoliverom@gmail.com

Enlace al github
https://github.com/estebanoli8/secop_valle_mujer

Resumen - Este trabajo presenta el desarrollo de un modelo de aprendizaje no supervisado para el análisis y caracterización de los objetos contractuales de 2.493 contratos públicos ejecutados por la Gobernación del Valle desde el 2.008 hasta la actualidad, con el fin de fortalecer el control y la crítica política usando datos confiables y la identificación de riesgos de corrupción.

Palabras clave - Aprendizaje de máquina, análisis de datos, aprendizaje no supervisado, clustering, control político, contratación pública, equidad de género.

I. INTRODUCCIÓN

Enfoque del negocio Desde la Asamblea Departamental:

El sistema colombiano de vigilancia sobre los recursos públicos se fundamenta en un conjunto articulado de controles —fiscal, social y político— que operan de manera complementaria buscando la correcta gestión de los contratos estatales y de su ejecución. El **control fiscal** se concibe como una función pública autónoma, orientada a verificar si la gestión contractual se ajusta a principios de eficiencia, economía, legalidad y protección del patrimonio público [1]. Este control puede ejercerse de forma posterior y selectiva, o de manera preventiva y concomitante cuando la situación lo exige, apoyándose en tecnologías de información para realizar seguimiento en tiempo real al ciclo contractual, al uso de los recursos y al cumplimiento de los resultados esperados [2].

Paralelamente, el **control social** constituye una manifestación del derecho ciudadano a la participación democrática, habilitando a la sociedad civil para vigilar el estado, avance y destino de los recursos vinculados a los contratos públicos, mediante mecanismos electrónicos, plataformas abiertas y herramientas tecnológicas accesibles [3].

El **control político**, por su parte, es ejercido por las corporaciones públicas de elección popular (Como la Asamblea Departamental) y tiene como alcance supervisar, interpelar y solicitar información a las entidades ejecutoras cuando existan retrasos injustificados, incrementos contractuales atípicos, riesgos para la población o contratos de alto impacto social y presupuestal [4]. La jurisprudencia y el desarrollo institucional en Colombia han subrayado que estos controles no son excluyentes; por el contrario, deben operar de manera coordinada para garantizar la vigilancia integral del ciclo contractual, la detección temprana de anomalías y la protección del interés público [5].

Esta articulación, basada en la transparencia, el acceso a la información y el uso de tecnologías, refuerza la capacidad del Estado y de la ciudadanía para evaluar la calidad de la ejecución contractual, medir los resultados obtenidos y verificar el impacto real de los recursos invertidos.

Desde la labor ejercida en la Asamblea Departamental, se han promovido debates orientados a modernizar las prácticas de vigilancia mediante la sistematización, digitalización y utilización de tecnologías emergentes, con el fin de mejorar la transparencia y eficiencia del seguimiento a la ejecución contractual. La implementación de herramientas como observatorios en línea, monitoreo con drones y metodologías basadas en datos abiertos han buscado fortalecer la capacidad institucional para identificar inconsistencias, retrasos, sobrecostos o riesgos en contratos de alto impacto social.

Contexto social y público

Este enfoque tecnológico permite ampliar el alcance del control político y articularlo con el control fiscal y el control social, generando un ecosistema más robusto para supervisar sectores críticos como las obras públicas, la educación, la salud y la inversión social. Aunque la agenda de vigilancia es amplia, el presente estudio se centra particularmente en el análisis de la contratación pública con perspectiva de equidad de género, dada su relevancia social y enfoque que afectan de manera diferenciada a las mujeres del departamento.

Ley	Año	Enfoque de la ley
Ley 823	2003	Establece normas para garantizar la igualdad de oportunidades para las mujeres en los ámbitos público y privado. Define lineamientos para la incorporación del enfoque de género en políticas públicas y promueve la participación de las mujeres en la vida económica y social. [6]
Ley 1257	2008	Adopta medidas integrales para la sensibilización, prevención y sanción de la violencia y discriminación contra las mujeres. Ordena protocolos institucionales de atención a víctimas y responsabilidades específicas del Estado. [7]
Ley 1496	2011	Garantiza la igualdad salarial y la retribución laboral entre mujeres y hombres. Establece criterios técnicos para valorar puestos de trabajo y obliga a instituciones públicas y privadas a corregir brechas salariales. [8]
Ley 1761	2015	Crea el delito autónomo de feminicidio y define medidas reforzadas de prevención, investigación y sanción. Reconoce la violencia contra las mujeres como un fenómeno estructural y exige abordajes integrales. [10]
Ley 2215	2022	Crea y regula las Casas de Refugio para mujeres víctimas de violencias basadas en género. Define obligaciones territoriales, lineamientos de atención y financiación. [11]
Ley 2453	2025	Amplía las garantías de igualdad económica, participación política y acceso equitativo a servicios públicos. Extiende la exigencia de enfoque de género a las políticas de inversión y contratación estatal. [12]

Tabla 1 - Leyes que vienen fortaleciendo, orientando y obligando a las entidades del estado, a la ciudadanía y a los Gobiernos alrededor del trabajo a realizar para beneficiar a las mujeres, reducir las brechas y equilibrar la balanza social, económica y políticamente con el hombre.

Pertinencia

En este sentido es importante para el departamento dilucidar de la manera más objetiva posible los alcances y los resultados de las inversiones que responden a todas estas iniciativas provenientes del legislativo. Es por eso que se hace necesario indagar sobre la contratación pública de la Gobernación del Valle, que es la Entidad del Estado **a la cuál se debe hacer control político por parte de los Diputados del Valle**. En ese sentido, desde esta curul, es relevante poder investigar de manera masiva grupo de contratos, sus valores, sus fechas de inicio y fin, sus modos de contratación, los contratistas, y demás variables que lleven la discusión política hacia espacios de mayor objetividad y acercar la ingeniería a la política.

Relevancia

Al analizar el **presupuesto departamental para el año 2026 presentado por la Gobernadora del Valle Dilian Francisca Toro, ante la Asamblea Departamental en octubre de 2025**, se identificó inicialmente que con respecto al lo asignado en 2025, la cantidad entregada a la Secretaría de Mujer, Equidad de Género y Diversidad Sexual **bajó a $13.573.391.400 (que representa el 0,3% del total presupuestal de la Entidad)**

Y si se hace un análisis temporal más amplio, es posible evidenciar a través de la siguiente gráfica el comportamiento en decrecimiento en relación con el pico de 2023.

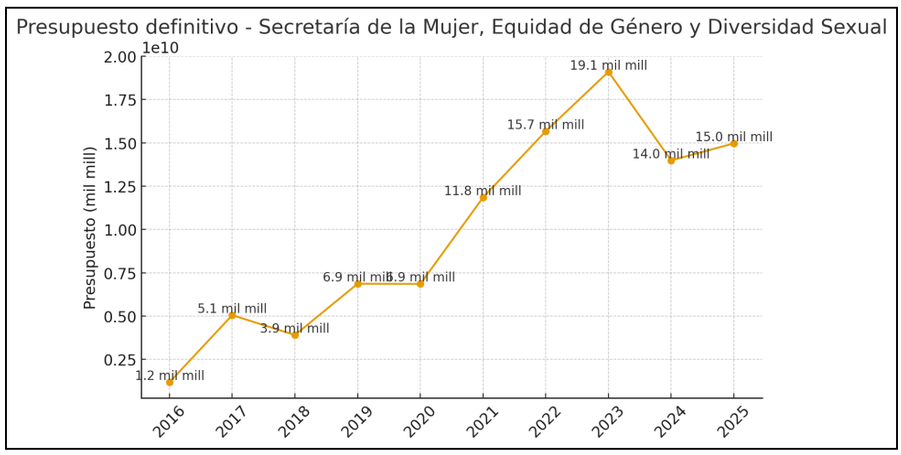

Figura 1 - Presupuesto asignada a la Secretaría de la Mujer, Equidad de Género y Diversidad Sexual

Esta información financiera, sumada a la información política de la discusión actual sobre la aplicación de la ley 2453 de 2025 establece un norte relevante para analizar la ejecución contractual de la Entidad alrededor de la Mujer vallecaucana y sus derechos.

Al validar la información disponible en SECOP **(Sistema Electrónico de Contratación Pública**, **plataforma donde se registran todos los contratos públicos en proceso, en ejecución y finalizado de todas las entidades públicas en Colombia)** encontramos disponibilidad de un servicio de consulta vía API de Sócrata para datos abiertos y públicos. Donde se pudo identificar el dataset integrado de SECOP 1 y SECOP 2 [14].

De donde donde se evidencia que las únicas variables reportadas son las presentadas en la siguiente figura.

0	nivel_entidad	object
1	codigo_entidad_en_secop	object
2	nombre_de_la_entidad	object
3	nit_de_la_entidad	object
4	departamento_entidad	object
5	municipio_entidad	object
6	estado_del_proceso	object
7	modalidad_de_contrataci_n	object
8	objeto_a_contratar	object
9	objeto_del_proceso	object
10	tipo_de_contrato	object
11	fecha_de_firma_del_contrato	datetime64[ns]
12	fecha_inicio_ejecuci_n	object
13	fecha_fin_ejecuci_n	object
14	numero_del_contrato	object
15	numero_de_proceso	object
16	valor_contrato	int64
17	nom_raz_social_contratista	object
18	url_contrato	object
19	origen	object
20	tipo_documento_proveedor	object
21	documento_proveedor	object

Figura 2 - Nombres de las variables disponibles para las consultas y su tipo. Diccionario de datos para el dataset analizado.

Información con la cual no se puede ingresar a las actividades detalladas del contrato de manera automatizada, ni se puede realizar una valoración objetiva de su cumplimento, razón por la cual se decide hacer un entrenamiento no supervisado para explorar el conjunto de datos disponible y **descubrir patrones que enriquezcan el debate público alrededor del análisis de grandes volúmenes de datos con relativo bajo esfuerzo en comparación con el análisis humano tradicional.**

Objetivo SMART

En un periodo de cuatro semanas, procesar los objetos contractuales publicados por la Gobernación del Valle del Cauca, aplicar técnicas de análisis semántico y métodos de clusterización no supervisada para identificar grupos temáticos relacionados con el enfoque de género, y generar un informe que clasifique los contratos, intente identificar patrones sospechosos de las compras públicas orientadas a beneficiar a las mujeres del departamento.

II. Materiales y métodos

Origen y naturaleza del Data Set

Se utilizó el siguiente dataset “SECOP INTEGRADO” con 20,4M Columnas con 22 columnas donde cada fila es un proceso (de contratación). Los datos se acceden a través del servicio público datos.gov.co

habilitado por el Ministerio de las TIC mediante protocolo de Sócrata.

Coding
La escritura, ejecución y debugging de todo el código fue realizado en Colab con la asistencia completa de Gemini integrado a la misma plataforma de Google.

Interacción inicial de los datos
Se establece una primera conexión con el servidor

```
# --- Nuevo Procedimiento: Interacción Inicial ---

print("--- Paso 1: Interacción inicial con el servicio ---")

import requests
import pandas as pd

# ---- Configuración base ----
DOMAIN = "https://www.datos.gov.co"
DATASET_ID = "rpmr-utcd"
RESOURCE_URL = f"{DOMAIN}/resource/{DATASET_ID}.json"     # SoQL clásico

# Usamos el token proporcionado para las cabeceras (asumiendo que es necesario
para interacciones basicas tambien)
APP_TOKEN = "tnXsP2icZ2ugAGaKdSf0QHoWU"
HEADERS = {"X-App-Token": APP_TOKEN}

# Realizar una consulta muy simple para confirmar conectividad y acceso basico
# Por ejemplo, obtener los primeros 5 registros sin filtros complejos
simple_query_params = {
    "$select": "nombre_de_la_entidad, fecha_de_firma_del_contrato,
    valor_contrato", # Seleccionar algunas columnas basicas
    "$limit": 5 # Limitar a 5 registros
}
```

Figura 3 - Establecimiento de conexión con la API de datos.gov.co. Ver todo el bloque de código en el repositorio de Github.

Se obtuvo respuesta positiva por parte del servidor como se muestra en la siguiente figura.

```
--- Paso 1: Interacción inicial con el servicio ---

  → Realizando solicitud simple para verificar conexión...
  ✅ Interacción inicial exitosa. Se recibieron datos de ejemplo.

--- Paso 1 Completado ---
```

Figura 4 - Validación de establecimiento de conexión con el servicio de datos abiertos del Estado Colombiano.

Se procede a estudiar los datos disponibles en la página del repositorio. [15]

Planeación y construcción de la Consulta:
Se preparan los dos primeros filtros de la consulta. El primero es el número de NIT de la Gobernación escrito de diversas maneras. Este campo es esencial para identificar la entidad pues es identitario de la misma. A continuación se configura el segundo filtro para la variable Nombre de la Entidad.

```
# Variaciones de NIT para la Gobernación del Valle del Cauca
NIT_VARIANTS = [
    "890399029", "890399029-5", "890.399.029-5", "NIT 890399029-5", "NIT 890.
    399.029-5", "8903990295" # Added back "8903990295"
]

# Variaciones del Nombre de la Entidad para la Gobernación del Valle del Cauca
ENTITY_NAME_VARIANTS = [
    "gobernación del valle",
    "gobernacion del valle", # Sin tilde
    "gobernación del valle del cauca",
    "gobernacion del valle del cauca", # Sin tilde
    "Gobernación del Valle", # Capitalización inicio
    "Gobernacion del Valle", # Capitalización inicio sin tilde
    "Gobernación del Valle del Cauca", # Capitalización inicio
    "Gobernacion del Valle del Cauca", # Capitalización inicio sin tilde
    "GOBERNACIÓN DEL VALLE", # Todo mayúsculas
    "GOBERNACION DEL VALLE", # Todo mayúsculas sin tilde
    "GOBERNACIÓN DEL VALLE DEL CAUCA", # Todo mayúsculas
    "GOBERNACION DEL VALLE DEL CAUCA" # Todo mayúsculas sin tilde
]
```

Figura 5 - Establecimiento de los dos primeros filtros por NIT y por Nombre de la entidad.

Posteriormente se trabajó el tercer filtro que busca que las siguientes palabras se encuentren en el objeto a contratar.

ACOSO SEXUAL	CÁNCER DE MAMA	Equidad de genero	LACTANCIAS
Acoso sexual	CÁNCER DE SENO	Equidad de género	Lactancia
BRECHA SALARIAL	CÁNCER UTERINO	FEMENINA	Lactancias
Brecha salarial	Cáncer de mama	FEMENINO	MADRES
CANCER DE MAMA	Cáncer de seno	FEMINICIDIO	MADRES CABEZA
CANCER DE SENO	Cáncer uterino	Femenina	MAMOGRAFIA
CANCER UTERINO	EMBARAZO	Femenino	MAMOGRAFÍA
CITOLOGIA	EMPODERAMIENTO FEMENINO	Feminicidio	MATENA
CITOLOGÍA	ENFOQUE DE GENERO	GENERO	MATERNO
CUELLO UTERINO	ENFOQUE DE GÉNERO	GESTACION	MENSTRUACION
Cancer de mama	EQUIDAD DE GENERO	GESTACIÓN	MENSTRUACIÓN
Cancer de seno	EQUIDAD DE GÉNERO	Genero	MENSTRUAL
Cancer uterino	Embarazo	Gestacion	MUJER
Citologia	Empoderamiento femenino	Gestación	MUJERES
Citología	Enfoque de genero	GÉNERO	
Cuello uterino	Enfoque de género	Género	

Figura 6 - Conjunto 1 de palabras con diferentes formas de escribirlas para buscarse en el objeto contractual

Madres	Perspectiva de género	embarazo	mamografia
Madres cabeza	VIOLENCIA DE GENERO	empoderamiento femenino	mamografía
Mamografia	VIOLENCIA DE GÉNERO	enfoque de genero	matena
Mamografía	Violencia de genero	enfoque de género	materno
Matena	Violencia de género	equidad de genero	menstruacion
Materno	acoso sexual	equidad de género	menstruación
Menstruacion	brecha salarial	femenina	menstrual
Menstruación	cancer de mama	femenino	mujer
Menstrual	cancer de seno	feminicidio	mujeres
Mujer	cancer uterino	genero	paridad
Mujeres	citologia	gestacion	perspectiva de genero
PARIDAD	citología	gestación	perspectiva de género
PERSPECTIVA DE GENERO	cuello uterino	género	violencia de genero
PERSPECTIVA DE GÉNERO	cáncer de mama	lactancias	violencia de género
Paridad	cáncer de seno	madres	
Perspectiva de genero	cáncer uterino	madres cabeza	

Figura 7 - Conjunto 2 de palabras con diferentes formas de escribirlas para buscarse en el objeto contractual

Y también se agregó un filtro de exclusión con las siguiente palabras que pueden generar confusión en la consulta dado que se encontraron coincidencias en una primera exploración que llevaban a errores de enfoque en la búsqueda. Por ejemplo, “chalecos de protección balística talla femenina” de la secretaría de Seguridad y Convivencia.

```
# Frases a excluir de la búsqueda
EXCLUDED_PHRASES = [
    "protección balística", "genero musical", "genero
    cinematográfico", "genero literario",
    # Variaciones de "balística" añadidas
    "balistica", "Balistica", "BALISTICA", "balística",
    "Balística", "BALÍSTICA"
]
```

Figura 8 - Palabras a excluir de la búsqueda en el objeto contractual.

Posterior a la configuración de la consulta se procede a ejecutar donde resultaron 3.837 contratos que coincidieron con el NIT, el Nombre de la Entidad Gobernación del Valle y sus variaciones y alguna de las palabras en el objeto contractual.

Descarga de datos:

Posterior a la respuesta del servidor se procede a descargar la información seleccionada obedeciendo a los requerimientos de paginación y temporización del servicio de la API de donde se obtiene el dataframe `df_contracts` de dimensiones de 3837 x 22.

0	Nacional	709416515	GOBERNACION DEL VALLE- SECRETARIA DE MUJER EQU...	890399029
1	Nacional	709416515	GOBERNACION DEL VALLE- SECRETARIA DE MUJER EQU...	890399029
2	Nacional	709416515	GOBERNACION DEL VALLE- SECRETARIA DE MUJER EQU...	890399029
3	Nacional	709416515	GOBERNACION DEL VALLE- SECRETARIA DE MUJER EQU...	890399029
4	Nacional	709416515	GOBERNACION DEL VALLE- SECRETARIA DE MUJER EQU...	890399029

5 rows × 22 columns

```
Dimensiones del DataFrame: (3837, 22)
```

Figura 9 - Algunos elementos de la consulta ya descargados y almacenados en el Dataframe.

Análisis Exploratorio de los Datos:

Eliminación de duplicados: Por conocimiento del sector público y del funcionamiento del SECOP I y II se procedió inmediatamente a identificar y eliminar los duplicados de los registros, pasando de 3837 a 2481. Se visualiza en la siguiente gráfica la distribución de esos registros en las secretarías encargadas. Se realizó también una abreviación y una nueva variable en el dataset para el nombre de cada secretaría para poder graficar y visualizar con facilidad. Como era de esperarse se concentraron principalmente en la secretaría de Mujer, Equidad y Diversidad Sexual.

nombre_entidad_corto	count
Sec. Mujer, Equidad y Diversidad Sexual	1831
Sec. Paz Territorial y Reconciliación	326
Gobernación Valle del Cauca	297
Sec. Convivencia y Seguridad Ciudadana	14
Sec. Educación	5
Sec. Cultura	5
DADI	4
Sec. Salud	3
Depto. Administrativo de Planeación	3
GOBERNACIÓN DEL VALLE DEL CAUCA - SECRETARIA DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES	3
Sec. Asuntos Étnicos	1
Sec. Desarrollo Rural, Agricultura y Pesca	1

Figura 10 - Distribución de los contratos por secretaría.

Evidenciándose también en la siguiente gráfica la distribución del valor de contrato que es una variable muy relevante pues permite enfocar las acciones de control en sumas cuantiosas donde tanto el alcance como el riesgo es mayor. Se percibe una concentración

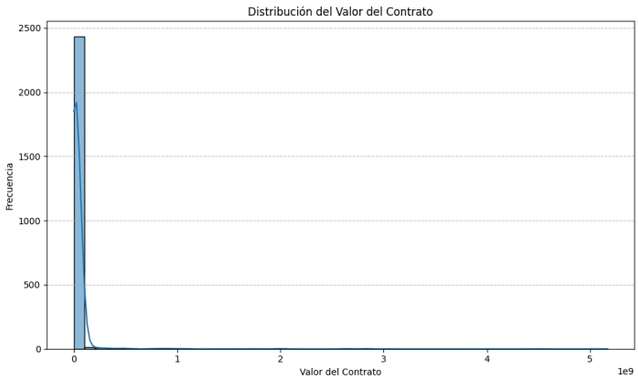

Figura 11 - Histograma del valor del contrato

Para intentar hacer una separación inicial de los tipos de contratos se establecieron con criterio de experto 3 rangos de valores de contratos para visualizar su distribución. Los rangos son: valor de contrato menor a $200millones de pesos, entre $200 y $1.000 millones de pesos y mayor a $1.000 millones de pesos. Se obtuvieron en los últimos dos rangos 28 y 18 contratos respectivamente. Lo cual permite inferir que gran cantidad de los contratos podrían estar relacionados con la prestación de servicios profesionales de

personas naturales.

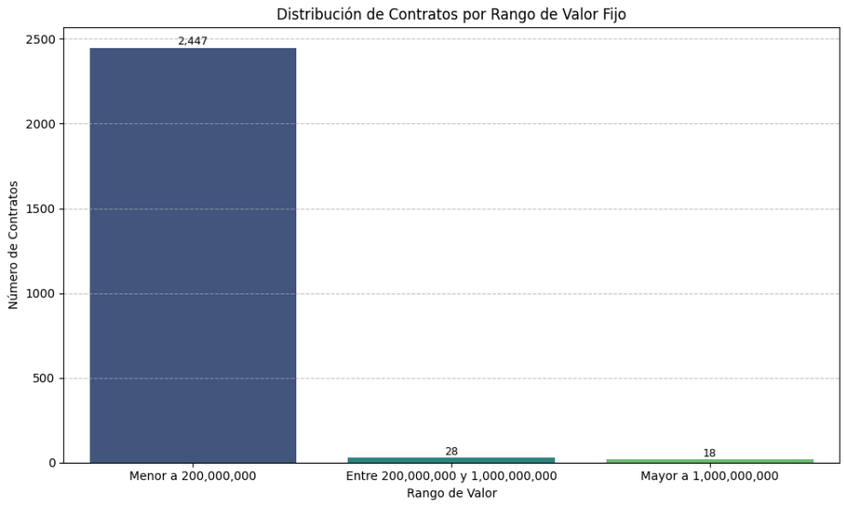

Figura 12 - División por rangos de valor de contrato de los registros.

Se procedió a visualizar la inversión total por año de los objetos contractuales, resultando esta gráfica que coincide morfológicamente entre 2016 y 2025 con la asignación de presupuesto a SOLAMENTE la secretaría de la Mujer, equidad de género y diversidad sexula (Figura 1) es decir, se percibe en la contratación global de la Gobernación una baja desde 2023.

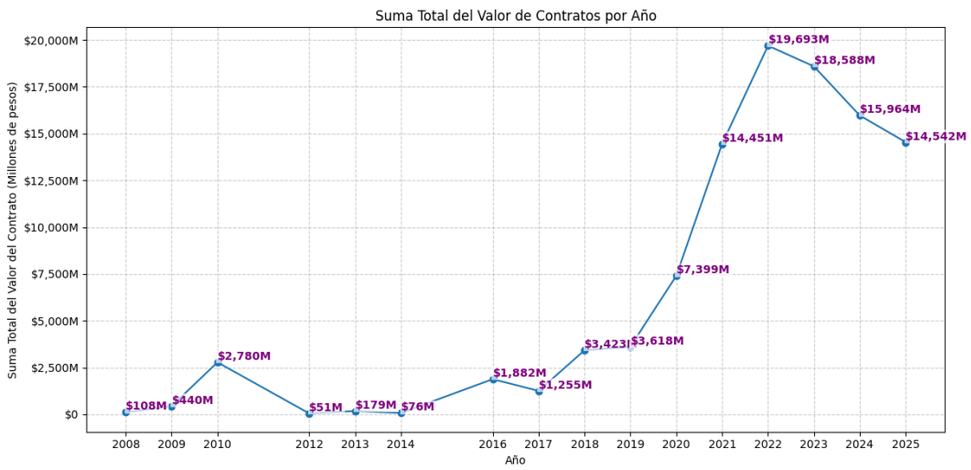

Figura 13 - Acumulado de valores de contratos por año.

Evaluando la variable Modalidad de selección, presentada en la siguiente figura, se percibe que la contratación directa es el principal mecanismo de selección de los proveedores de bienes, productos, servicios u obras en el Departamento del Valle en materia de beneficios para la Mujer.

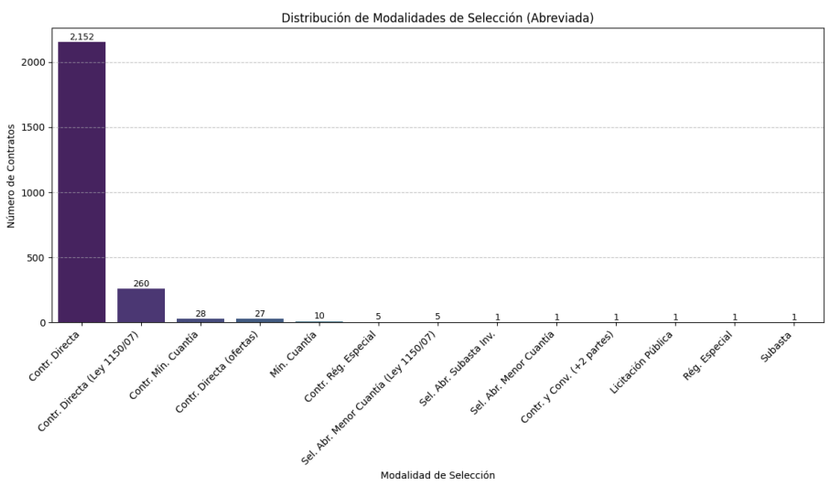

Figura 14 - Histograma de modalidad de selección de los contratistas.

A continuación se muestra el crecimiento de cantidad de contratos por año que a la fecha a la fecha en 2025 se han publicado 507 contratos y que el más antiguo identificado es del 2008.

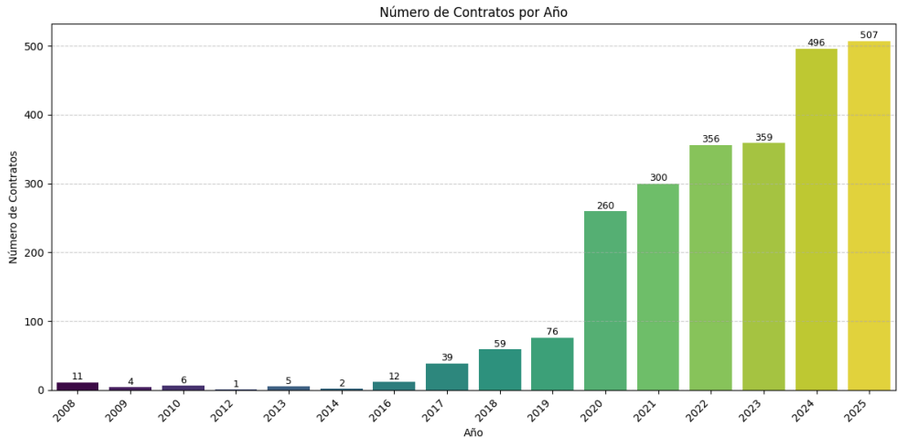

Figura 15 - Histograma de cantidad de contratos por año

Al realizar una exploración de los nombres de los contratistas y su número de identificación se determina el siguiente gráfico con los que más contratos han tenido en la entidad. Se observan que la mayor recurrencia son de personas naturales lo cual coincide con el comportamiento de la secretaría de la Mujer, ya que es bien sabido que hay personas que se convierten con el tiempo en parte de la institucionalidad sin ser de carrera administrativa.

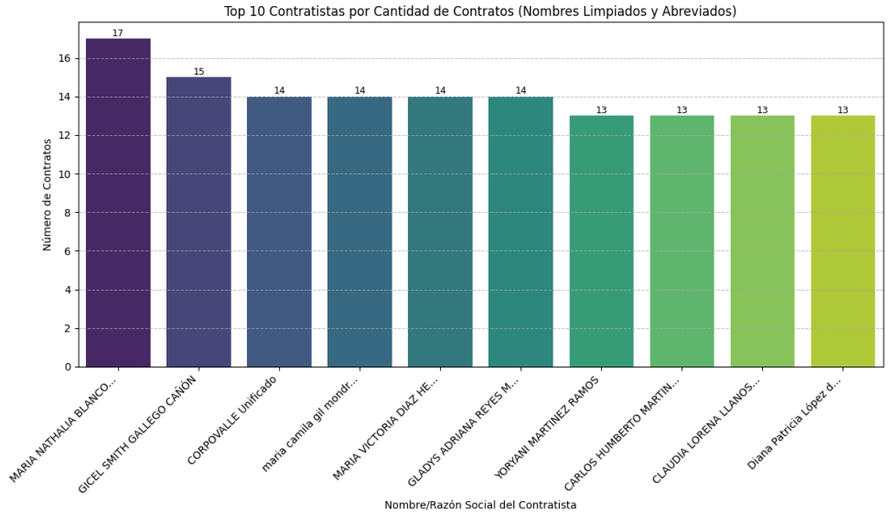

Figura 16 - Top 10 de contratistas por cantidad de contratos en todo el dataset.

Y para cerrar la exploración de los datos se realizó el análisis de los contratistas a lso que se les ha asignado mayor cantidad acumulada de presupuesto desde el 2008 hasta la fecha. Se realizó una unificación de los nombres de los contratistas pues se detectó previamente que se pueden escribir de muchas maneras diferentes, dependiendo de la persona que esté operando el sistema y subiendo la información.

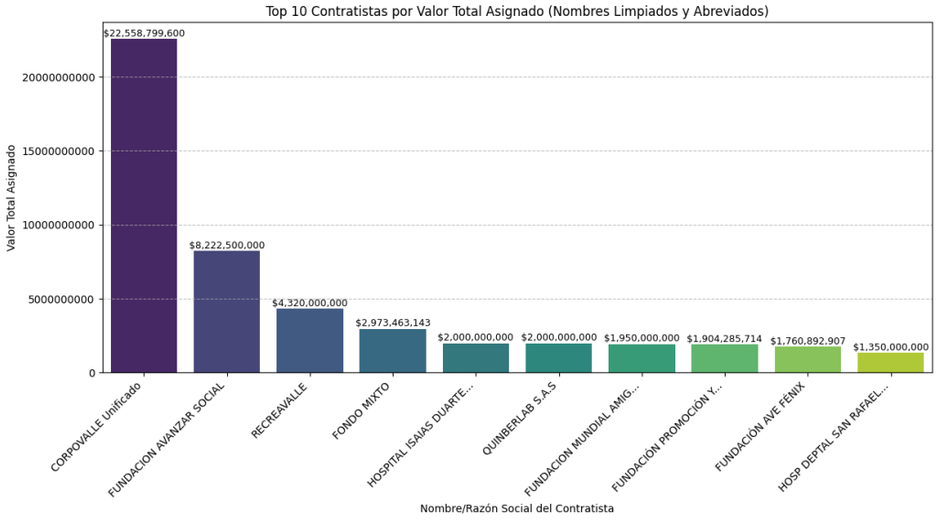

Figura 17 - Top 10 de contratistas por cantidad de recursos asignados.

En el preprocesamiento y análisis exploratorio se transformaron algunas de las variables originales para dar origen a estas 6 nuevas variables. El tamaño del **dataset previo al entrenamiento es de 545.5 KB**

```
22  nombre_entidad_corto              2493 non-null   object
23  valor_final                       2493 non-null   int64
24  Rango_Valor_Fijo                  2493 non-null   object
25  Año_Contrato                      2493 non-null   int64
26  modalidad_corta                   2493 non-null   object
27  nom_raz_social_contratista_corto  2493 non-null   object
```

Figura 18 - Nuevas variables procedentes de la transformación de las originales.

No se realizó una matriz de correlación de las variables porque la **única variable numérica contínua es el valor de contrato.** El resto de variables son categóricas nominales y de sin embargo se realizaron análisis bivariados que permiten dar cuenta del relacionamiento entre ellas.

III. RESULTADOS

Entrenamiento No Supervisado

Se realizaron una serie de experimentos de aprendizaje no supervisado primero con un Pipeline básico con técnicas que no satisficieron las necesidades del trabajo por lo cual se procedió a emplear otro Pipeline Avanzado con mayor capacidad de clusterización.

Pipeline Básico TF-IDF + K-Means + PCA

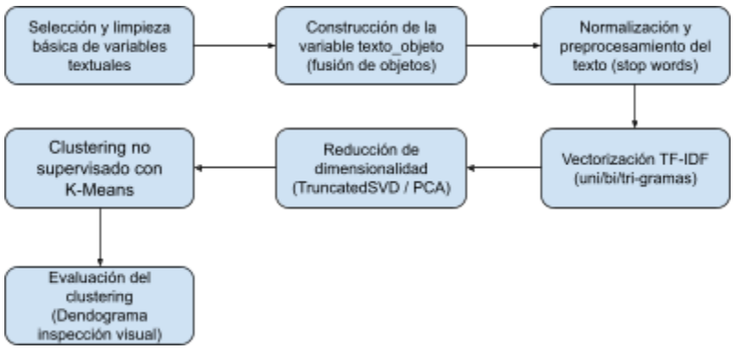

Figura 19 - Pipeline básico

TF-IDF (Term Frequency – Inverse Document Frequency)

Con TF-IDF se creó un representación vectorial que cuantifica la importancia de términos dentro de un conjunto de documentos, penalizando palabras comunes y destacando aquellas que discriminan textos.
En el contexto contractual, permite identificar vocabulario característico de cada tipo de objeto contractual, permitiendo separar el lenguaje administrativo del misional. Se hizo un análisis exploratorio inicial de textos relativamente homogéneos como los objetos SECOP.

K-Means

Se utilizó para agrupar documentos en *k* conglomerados según la similitud geométrica de sus vectores creados del texto. En el análisis de contratación pública permite identificar categorías temáticas globales aunque fragmenta o mezcla temas cuando los textos son no son de fácil discriminación.

TruncatedSVD

Esta técnica se usó para reducir la dimensionalidad de matrices dispersas generadas por TF-IDF preservando relaciones entre términos y documentos. Suaviza “ruido” léxico y revelar componentes semánticos principales de los objetos contractuales. Incrementa la estabilidad de K-Means y ayuda a capturar estructuras temáticas subyacentes.

Los resultados se ven en la siguiente figura donde se observa que la distribución de los puntos presenta una forma alargada y segmentada en “brazos”, lo que indica que la reducción lineal mediante SVD

comprimió excesivamente la estructura semántica del texto.

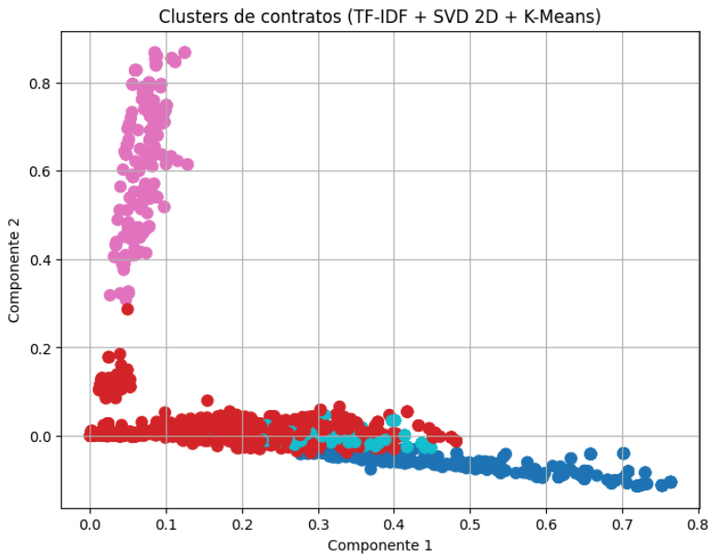

Figura 20 - Clusterización inicial 2493 contratos con pipeline básico

De igual manera se observa que el dendrograma muestra una estructura altamente comprimida en la base, con ramas muy cortas y casi sin separación clara entre la mayoría de los documentos, lo que indica una **baja variabilidad** capturada por la representación SVD 2D. Solo dos grandes divisiones aparecen a mayor distancia, reflejando una separación muy gruesa

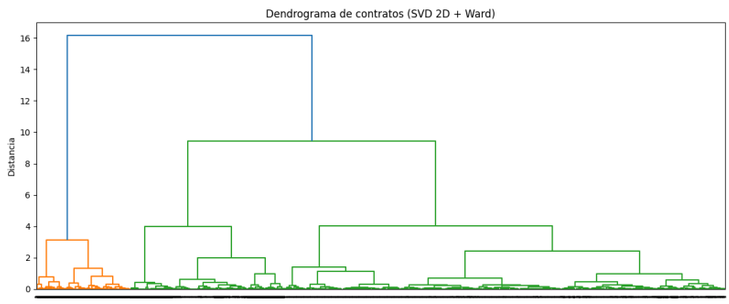

Figura 21 - Dendrograma de clusterización inicial 2493 contratos con pipeline básico

Se evidencia que el enfoque basado en TF-IDF, SVD y clustering lineal no captura adecuadamente la complejidad semántica de los 2.493 objetos contractuales, generando separaciones artificiales y jerarquías poco informativas. La estructura alargada de la proyección y el dendrograma comprimido muestran que la variabilidad real del texto no puede representarse eficazmente en espacios lineales de baja dimensión. Esto establece las limitaciones del modelo clásico para este tipo de lenguaje administrativo altamente repetitivo. Por ello, se justifica avanzar hacia el pipeline semántico avanzado con embeddings, UMAP y HDBSCAN e inetentar obtener clusters más coherentes y explicables.

Pipeline (Embeddings + UMAP + HDBSCAN)

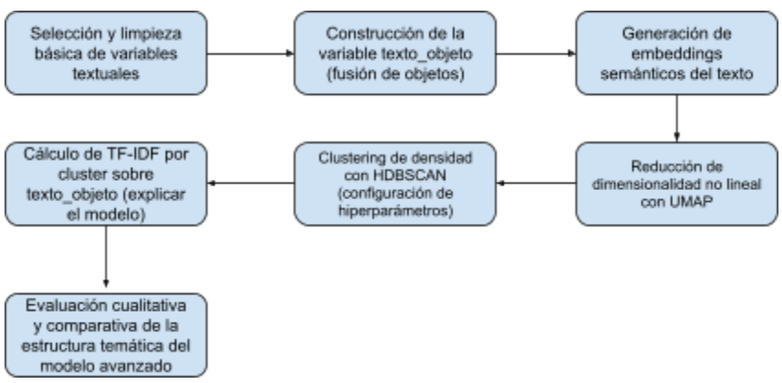

Figura 22 - Pipeline avanzado

Embeddings semánticos (Mini Modelos de Lenguaje Preentrenado)

Con esta técnica se generaron representaciones más densas y continuas del texto, donde palabras y frases con significados similares quedaron más cercanas entre sí en el espacio vectorial. Se capturó la semántica contextual más allá de la frecuencia de términos, permitiendo diferenciar contratos con lenguaje similar pero finalidades distintas.

El modelo usado fue el **all-MiniLM-L6-v2**, perteneciente a la familia **SentenceTransformers**, basado en la arquitectura **MiniLM**, una variante compacta de los modelos Transformer.

Configuración: **vectores de 384 dimensiones**, generados a partir de un encoder con **6 capas Transformer**, aproximadamente **22 millones de parámetros**. Fue **preentrenado utilizando técnicas de aprendizaje contrastivo** sobre más de **1.000 millones de pares de sentencias**, provenientes de corpus multilingües y tareas como inferencia textual (NLI), detección de parafraseo y recuperación semántica.

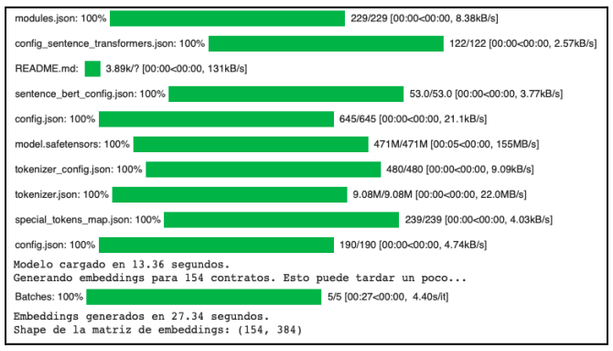

Figura 23 - Ejecución y uso en Colab del all-MiniLM-L6-v2

UMAP (Uniform Manifold Approximation and Projection)

Se empleó UMAP como método de reducción de dimensionalidad no lineal que preserva tanto la estructura local como global del espacio de embeddings. Esto con el fin de visualizar y organizar grandes volúmenes de textos en regiones densas donde emergen temas naturales sin imponer formas esféricas.
Para el caso de contratación pública, UMAP ayudó a separar gradualmente los ejes temáticos incluso cuando el lenguaje es repetitivo

HDBSCAN (Hierarchical Density-Based Clustering)

Esta técnica se eligió porque puede agrupar puntos según densidades de datos, detectando clusters de diferentes tamaños y formas sin requerir un número de clusters predefinido.
También puede hacerlo con grupos temáticos robustos y marca como “ruido” contratos con objetos atípicos, ambiguos o extremadamente genéricos, algo frecuente en SECOP.

Pipeline Avanzado para Todos los 2493 contratos

Con los parámetros por defecto se obtuvieron estos resultados, donde se aprecia la incapacidad del modelo de clusterizar y expresar en dos dimensiones una separación de los datos requeridos.

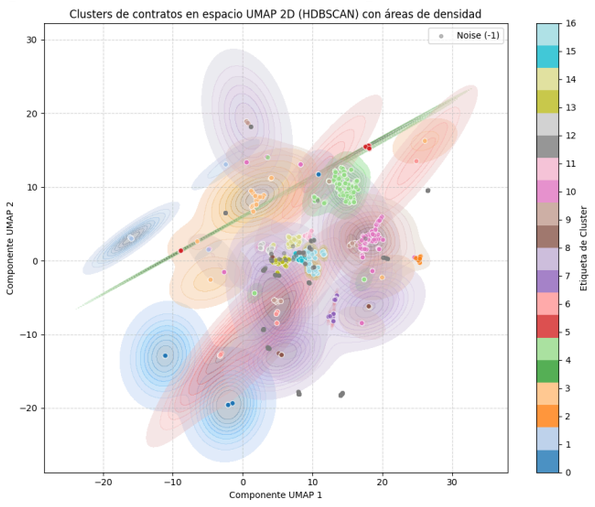

Figura 24 - Clusterización de 16 clases para 2493 objetos contractuales usando pipeline avanzando

Al realizar variaciones heurísticas no se percibió mejora en la clusterización por lo que se procedió a realizar una partición de los datos de entrenamientos separando los 154 contratos contratados con empresas y los 2339 contratados con personas naturales, esto debido a que por conocimiento del sector, los objetos contractuales de personas naturales son muy repetitivos semánticamente y tiene características especiales que al intentar modelar junto con los de empresas privadas se tendrá mayor complejidad. Es también apropiado volver a mencionar que, en general, los valores de los contratos por prestación de servicios son mucho menores que los contratados a las empresas privadas, razón por la cual desde la óptica del negocio es una separación que puede tener buenos resultados.

Solo los de prestación de Servicio

Al correr el mismo algoritmo con los siguientes parámetros (variados heurísticamente)

```
umap_n_neighbors: 90
  umap_n_components: 3
  umap_min_dist: 0.3
  umap_metric: euclidean
  hdbscan_min_cluster_size: 200
  hdbscan_min_samples: 60
  hdbscan_metric: euclidean
  hdbscan_cluster_selection_method: leaf
  hdbscan_prediction_data: True
```

Se observa el descubrimiento de 5 clases (incluyendo el “ruido” que también es una clase válida para este estudio)

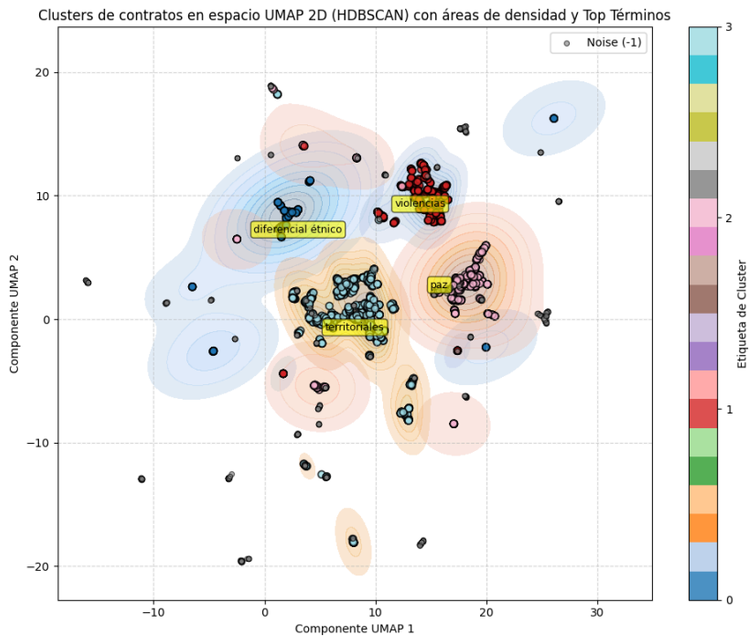

Figura 25 - Clusterización de 5 clases (incluyendo ruido) clases para 2339 objetos contractuales usando pipeline avanzando

Solo los de Empresas

Al correr el mismo algoritmo con los siguientes parámetros (variados heurísticamente)

```
umap_n_neighbors: 90
  umap_n_components: 3
  umap_min_dist: 0.3
  umap_metric: euclidean
  hdbscan_min_cluster_size: 200
  hdbscan_min_samples: 60
  hdbscan_metric: euclidean
  hdbscan_cluster_selection_method: leaf
  hdbscan_prediction_data: True
```

Se observa el descubrimiento de 20 clases (incluyendo el “ruido” que también es una clase válida para este estudio). La siguiente gráfica es el mejor resultado obtenido y genera gran valor para la curúl de oposición, porque esc alra, abre muchas preguntas y sobretodo valida el funcionamiento de estas herramientas para llegar a nuevos concoimeintos entre grandes volúmenes de datos.

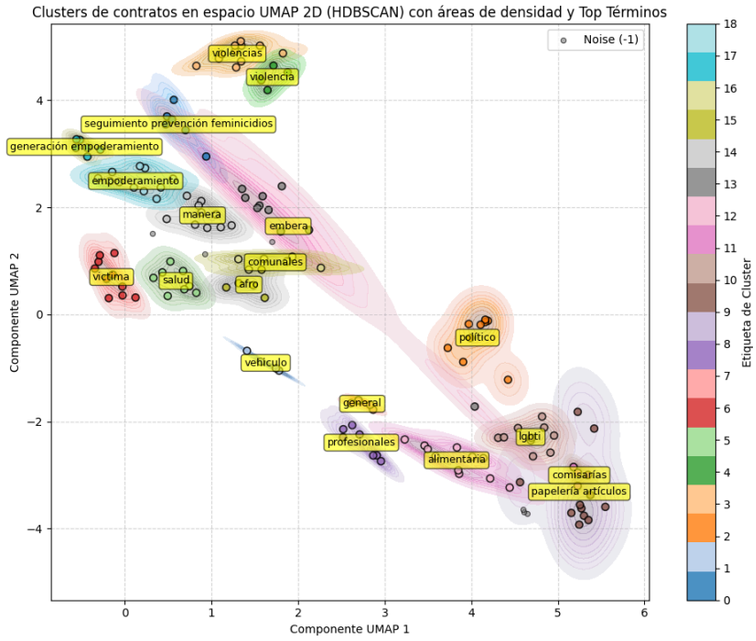

Figura 26 - Clusterización de 20 clases (incluyendo ruido) clases para 154 objetos contractuales usando pipeline avanzando

Este resultado es el más satisfactorio desde el punto de vista del sector público pues devela con claridad la separación de clases que son reconocidas en la administración pública. Se observa la atención, prevención de la violencia y prevención de feminicidios de manera cercana. Otra observación importante es el de salud y víctimas pues uno de los mandatos es a dar una atención diferencial en urgencias a mujeres víctimas de violencia intrafamiliar o de género. Se observan las políticas públicas cerca de las comisarías que son las institucionalidad, y lo administrativos como papelería e insumos. Además es muy llamativo la capacidad uqe tuvo el modelo de discriminar grupos poblacionales como comunales, afro, emera y lgtbi.

IV. Discusión

- La limpieza del texto para hacer stop_words y luego la interpretabilidad del modelo para etiquetar las clases depende mucho del expertise en el sector. Es una desventaja para una eventual socialización y apropiación masiva de este tipo de técnicas.

- 2493 contratos es una cifra relativamente baja en materia de contratación pública para ese rango de tiempo. Se deben hacer pruebas con mayor cantidad de contratos, ajustar y validar el desempeño de este mismo modelo.

- Como trabajo futuro queda obtener más info de SECOP para lograr hacer lecturas de los PDF que contienen los contratos y poder explorar mayor detalle de la ejecución, porque en este punto solo se están analizando los objetos de los, faltaría analizar las actividades ejecutadas y los informes de supervisión para poder caracterizar su impacto.

- Los resultados indican que el pipeline avanzado es superior al básico porque incluye un modelo de lenguaje con capacidades semánticas. Pese a que esto implica que se depende de un tercero en la actualidad es totalmente factible este tipo de ejecuciones de manera costo/eficientes y replicable.

- Queda pendiente como trabajo futuro avanzar en la asociación de los contratistas vinculados por termas/clases y la cantidad de dinero invertida en cada uno.

V. Conclusiones

- Este ejercicio habilitó una nueva capacidad para el análisis, discusión y control político en el Valle del Cauca y en Colombia pues se evidencia que a través de este tipo de clusterización se puede hacer una identificación de las acciones de los políticos que toman decisión alrededor.

- Con el pipeline avanzado se evidenciaron visualmente la clusterización de 5 clases bien separadas y bien explicadas para los 2339 contratos de prestación de servicio.

- Con el pipeline avanzado se evidenciaron visualmente la clusterización de 20 clases bien separadas y bien explicadas para los 154 contratos de empresas

- El sistema implementado puede considerarse como un detector de palabras clave en grandes volúmenes de datos alrededor de la lucha de las mujeres, como equidad, feminicidio, violencia sexual y que después relaciona esas palabras con los enfoques de la Gobernación para resolver dichas problemáticas.

- El análisis semántico de los objetos contractuales para el conjunto de datos de 2493 contratos usando solamente TD-IDF no fue suficiente debido a la heterogeneidad de los temas.

- Con el pipeline básico no se puede realizar una discriminación confiable de los temas en los objetos contractuales para esta cantidad de contratos.

- El uso de técnicas más avanzadas como Embeddings abrió la posibilidad de clusterizar e interpretar de manera automática los temas principales en la contratación.

- El costo computacional de esta implementación es bajo, lo cual permite proyectar una escalabilidad del sistema manteniendo costos razonables

- Como trabajo futuro queda pendiente pendiente medir y hacer una comparación de cuánto tiempo se ahorra con el uso de este tipo de algoritmos en comparación con la realización de la misma tarea por un equipo humano, pero que por conocimiento del sector se puede inferir que este tipo de tecnologías acelera sustancialmente el análisis y permite descubrir información que de otra manera sería muy compleja de determinar.

En conclusión, el modelo propuesto facilita, caracteriza y sistematiza el análisis de un conjunto de 2493 contratos que, de manera manual, hubiera tomado mucho más tiempo en ejecutarse. **Se abre una gran campo de estudio y práctica para la exploración de la contratación pública en Colombia.** El norte debería ser poder contrastar los contratos, la inversión y los enfoques contra las promesas, los anuncios y la propaganda que son comúnmente usados en el sector público. **El afinamiento del control fiscal, social y político debe ser uno de los caminos para reducir los riesgos de corrupción** y la ineficiencia administrativa. Apropiar e implementar este tipo de tecnologías por parte de la ciudadanía, los entes de control y los mismos políticos puede ayudar a la implementación de políticas públicas más eficientes y transparentes.

Esta curul de **oposición en la Asamblea departamental del Partido Alianza Verde** seguirá aplicando este tipo de tecnologías para afinar el ejercicio de control para seguir **acercando la Ingeniería a la Política.**

Reconocimiento

Agradecimientos a los miembros de la Unidad de Apoyo Normativo de la Asamblea Departamental del Valle del diputado Esteban Oliveros Montoya y el equipo legislativo del Congresista Duvalier Sánchez Arango.

Agradecimiento artificial a ChatGPT 5.1, Gemini 3 PRO que ayudaron en el desarrollo del código para identificar, descargar, procesar y analizar los datos de contratación pública del departamento. Sin estas herramientas no hubiera sido posible avanzar tan rápido en estos 2 meses.

REFERENCIAS

[1] Asamblea Departamental del Valle del Cauca, Ordenanza Anti-Elefantes Blancos. Disponible en: https://www.scribd.com/document/777893661/Ordenanza-Anti-Elefantes-Blancos, pp. 1–70. Sección: Control Fiscal, definición y alcance.

[2] Asamblea Departamental del Valle del Cauca, Ordenanza Anti-Elefantes Blancos. Disponible en: https://www.scribd.com/document/777893661/Ordenanza-Anti-Elefantes-Blancos, pp. 1–70. Fragmento sobre control preventivo y concomitante, uso de tecnologías de información.

[3] Asamblea Departamental del Valle del Cauca, Ordenanza Anti-Elefantes Blancos. Disponible en: https://www.scribd.com/document/777893661/Ordenanza-Anti-Elefantes-Blancos, pp. 1–70. Fragmento sobre participación ciudadana y control social.

[4] Asamblea Departamental del Valle del Cauca, Ordenanza Anti-Elefantes Blancos. Disponible en: https://www.scribd.com/document/777893661/Ordenanza-Anti-Elefantes-Blancos, pp. 1–70. Sección sobre competencias del control político y facultades de las corporaciones públicas.

[5] Asamblea Departamental del Valle del Cauca, Ordenanza Anti-Elefantes Blancos. Disponible en: https://www.scribd.com/document/777893661/Ordenanza-Anti-Elefantes-Blancos, pp. 1–70. Fragmento sobre complementariedad entre control fiscal, social, político e interno.

[6] Congreso de la República de Colombia. Ley 823 de 2003: Por la cual se dictan normas sobre igualdad de oportunidades para las mujeres. disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_0823_2003.html

[7] Congreso de la República de Colombia. Ley 1257 de 2008: Por la cual se dictan normas de sensibilización, prevención y sanción de formas de violencia y discriminación contra las mujeres. disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_1257_2008.html

[8] Congreso de la República de Colombia. Ley 1496 de 2011: Por medio de la cual se garantiza la igualdad salarial y de retribución laboral entre mujeres y hombres. disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_1496_2011.html

[10] Congreso de la República de Colombia. Ley 1761 de 2015: Por la cual se crea el tipo penal de feminicidio como delito autónomo y se dictan otras disposiciones (Ley Rosa Elvira Cely). disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_1761_2015.html

[11] Congreso de la República de Colombia. Ley 2215 de 2022: Por medio de la cual se establecen las Casas de Refugio en el marco de la Ley 1257 de 2008. disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_2215_2022.html

[12] Congreso de la República de Colombia. Ley 2453 de 2025: Por medio de la cual se establecen medidas para prevenir, atender, rechazar y sancionar la violencia contra las mujeres en política. disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=259640

[13] Colombia Compra Eficiente. Data set SECOP INTEGRADO disponible en: https://www.datos.gov.co/Estad-sticas-Nacionales/SECOP-Integrado/rpmr-utcd/about_data

[14] Diccionario de datos Secop Integrado https://www.datos.gov.co/Estad-sticas-Nacionales/SECOP-Integrado/rpmr-utcd/about_data