Indicador de Justa Asignación: Riesgos y Contrataciones Irregulares en Infraestructura

Grupo de trabajo: Alan Yovany Gasca, Christian David González, Jolman Salinas

Proyecto aplicado en analítica de datos

MIAD

Universidad de los Andes

En Colombia, la contratación pública es un proceso que se realiza a través de la Agencia Nacional de Contratación Pública (ANCP), la cual es la entidad encargada de la gestión de la contratación pública en el país. La ANCP es una entidad descentralizada del orden nacional, con personería jurídica, autonomía administrativa y financiera, y patrimonio propio, que tiene como función principal la gestión de la contratación pública en el país, en los términos de la Ley 1150 de 2007 y demás normas que la modifiquen, adicionen o complementen. La ANCP es la única entidad del orden nacional que tiene la facultad de expedir normas y reglamentos para la contratación pública, así como de establecer los procedimientos y mecanismos para la gestión de la contratación pública en el país. La ANCP es la única entidad del orden nacional que tiene la facultad de expedir normas y reglamentos para la contratación pública, así como de establecer los procedimientos y mecanismos para la gestión de la contratación pública en el país.

La vigilancia de la contratación pública es el instrumento para promover la confianza en el Estado y en la administración pública, y para garantizar que los recursos públicos se utilicen de manera eficiente y transparente, prevenir que las insituciones públicas sean utilizadas para fines de corrupción y garantizar que los recursos públicos se destinen a los fines para los cuales fueron asignados es crucial para el desarrollo del país.

La oportunidad de contar con datos abiertos sobre contratación pública en Colombia, y la posibilidad de analizarlos para responder preguntas de interés, es un reto que se debe abordar con la participación de actores públicos y privados, pero también es una oportunidad para mejorar la transparencia y la eficiencia de la contratación pública en el país mediante la aplicación de técnicas de análisis de datos y la generación de conocimiento.

> El presente proyecto busca asistir a las entidades gubernamentales de Colombia en la detección y prevención de prácticas indebidas en la asignación de contratos públicos de infraestructura. A través del análisis de datos provenientes de tres fuentes principales - SECOP Integrado, PACO y datos públicos de transparencia, lucha contra la corrupción, sanciones y multas - el objetivo es minimizar la exposición de estas entidades a contratos poco transparentes, incumplidos o corruptos.

Proceso ETL

Datos de contratación

Para llevar a cabo este análisis, se implementa un proceso ETL (Extracción, Transformación y Carga) que consta de varias etapas:

- **Extracción de datos**: La función read_csv_with_pyspark se utiliza para leer archivos CSV almacenados en la carpeta 'data', utilizando el separador '|'. Esta función devuelve un DataFrame de PySpark con la información del archivo CSV.
- **Análisis de calidad de datos**: La función analyze_data_quality se encarga de realizar un análisis básico de la calidad de los datos en el DataFrame proporcionado. Esto incluye el conteo de registros, registros nulos y duplicados, estadísticas descriptivas para columnas numéricas y la identificación de valores atípicos basados en el rango intercuartil.
- **Limpieza de nulos y duplicados**: La función limpiar_nulos_y_duplicados recibe un DataFrame y una lista de columnas. Esta función elimina los registros que contienen valores nulos en las columnas especificadas y también elimina los registros duplicados.
- **Filtrado de categorías de infraestructura**: Se realiza un filtrado de datos en base a las categorías de interés relacionadas con la infraestructura, como "VIVIENDA CIUDAD Y TERRITORIO", "TRANSPORTE", "MINAS Y ENERGIA" y "AMBIENTE Y DESARROLLO SOSTENIBLE".
- **Análisis de calidad de datos en el DataFrame filtrado**: Se vuelve a realizar un análisis de calidad de datos en el DataFrame filtrado para identificar posibles problemas en los datos después de la transformación.
- **Limpieza de nulos y duplicados en el DataFrame filtrado**: Se utiliza la función limpiar_nulos_y_duplicados con las columnas de referencia 'REFERENCIA_CONTRATO' y 'NIT_ENTIDAD' para eliminar registros nulos y duplicados en el DataFrame paco_secop_df. Se guarda el resultado en infraestructura_df_limpio.
- **Filtrado de los últimos años**: Se aplica la función filtrar_ultimos_anos al DataFrame infraestructura_df_limpio para conservar únicamente los registros correspondientes a los últimos 2 años. Se guarda el resultado en infraestructura_df_limpio_2_anos.
- **Selección de columnas relevantes**: Se utiliza la función seleccionar_columnas con una lista de columnas de interés (columnas_a_conservar) para reducir el número de columnas en el DataFrame infraestructura_df_limpio_2_anos. Se guarda el resultado en infraestructura_df_seleccionado.
- **Escritura de resultados en CSV**: Finalmente, el DataFrame resultante, infraestructura_df_seleccionado, se guarda en un archivo CSV llamado "contratos_infraestructura_df.csv" en la carpeta "etl_data".

REFERENCIA_CONTRATO	MUNICIPIO	DEPARTAMENTO	ESTADO_DEL_PROCESO	FECHA_INICIO_CONTRATO	FECHA_FIN_CONTRATO	CLASE_PROCESO	TIPO_PROCESO	TIPO_CONTRATO	
CDSECGRAL1692021	CARTAGENA	BOLIVAR	MODIFICADO	2021-01-20 00:00:00	2021-07-19 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	B
70092021	BOGOTA	BOGOTA	EN EJECUCION	2021-09-01 00:00:00	2021-11-30 00:00:00	REGIMEN ESPECIAL	PRESTACION DE SER...	CONTRATACION REGI...	B
CO1PCCNTR2239141	NO DEFINIDO	CAQUETA	EN EJECUCION	2021-02-15 00:00:00	2021-12-18 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	S
CO1PCCNTR2628519	DUITAMA	BOYACA	EN EJECUCION	2021-06-30 00:00:00	2021-06-30 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	B
CCS10202021	BOGOTA	BOGOTA	ACTIVO	1900-01-01 00:00:00	2021-12-31 00:00:00	CONCURSO DE MERITOS	CONSULTORIA	CONCURSO DE MERIT...	M
202100000063	NO DEFINIDO	CAQUETA	EN EJECUCION	2021-02-23 00:00:00	2021-10-31 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	C
0602 DE 2021	VILLAVICENCIO	META	TERMINADO	2021-03-10 00:00:00	2021-08-09 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	D
BOG3372021	BOGOTA	BOGOTA	EN EJECUCION	2021-03-02 00:00:00	2021-11-02 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	E
41542021	BOGOTA	BOGOTA	MODIFICADO	2021-03-06 00:00:00	2021-10-31 00:00:00	REGIMEN ESPECIAL	PRESTACION DE SER...	CONTRATACION REGI...	B
22382021	POPAYAN	CAUCA	EN EJECUCION	2021-09-28 00:00:00	2021-12-31 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	
CMACD25989083436 2021	MEDELLIN	ANTIOQUIA	EN EJECUCION	2021-08-06 00:00:00	2021-12-31 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	A
CPS3642021	CHIA	CUNDINAMARCA	EN EJECUCION	2021-04-22 00:00:00	2021-11-21 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	C
CO1PCCNTR2670534	CALARCA	QUINDIO	EN EJECUCION	2021-07-15 00:00:00	2021-10-14 00:00:00	CONTRATACION DIRECTA	OTRO	CONTRATACION DIRECTA	Q
4112060261024	CALI	VALLE DEL CAUCA	EN EJECUCION	2021-01-26 00:00:00	2021-04-30 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	S
20112021	POPAYAN	CAUCA	EN EJECUCION	2021-08-10 00:00:00	2021-12-24 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	
PLA1314PSP2021	ARMENIA	QUINDIO	EN EJECUCION	2021-06-24 00:00:00	2021-09-21 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	Q
IND210053	CALI	VALLE DEL CAUCA	EN EJECUCION	2021-04-29 00:00:00	2021-08-31 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	I
CO1PCCNTR2207421	BOGOTA	BOGOTA	EN EJECUCION	2021-02-03 00:00:00	2021-07-31 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	D
CO1PCCNTR2608311	SAN ANDRES	SAN ANDRES PROVID...	EN EJECUCION	2021-06-28 00:00:00	2021-10-27 00:00:00	CONTRATACION DIRECTA	PRESTACION DE SER...	CONTRATACION DIRECTA	G
62192021	BOGOTA	BOGOTA	MODIFICADO	2021-06-21 00:00:00	2021-07-31 00:00:00	REGIMEN ESPECIAL	PRESTACION DE SER...	CONTRATACION REGI...	B

Figura 1. Vista del DataFrame de contratos transformado

A través de este proceso ETL, se garantiza la calidad y relevancia de los datos analizados, permitiendo a las entidades gubernamentales colombianas identificar y prevenir situaciones de riesgo en la asignación de contratos públicos de infraestructura. El resultado de este análisis es un archivo CSV que contiene información detallada y depurada sobre los contratos de infraestructura en los últimos dos años, facilitando la toma de decisiones y el monitoreo de posibles irregularidades en el proceso de contratación.

Datos de entidades

En el proceso de ETL segenera tambien un dataframe de entidades que permita perfilarlas para los análisis posteriores, para ello se implementan las siguientes funciones:

- **agregar_por_nit_entidad**: Esta función recibe un DataFrame y realiza agregaciones a nivel de entidad (NIT_ENTIDAD y NOMBRE_ENTIDAD) utilizando diversas métricas, como el número de contratos, la suma y el promedio del valor total de los contratos, el último contrato firmado, la cantidad de departamentos, estados de proceso, clases de proceso, tipos de proceso, familias y clases involucradas. La función también calcula la cantidad de meses transcurridos desde el último contrato.
- **pivotar_por_columna**: Esta función recibe un DataFrame y una columna, y realiza una operación de pivoteo en función de los valores distintos presentes en la columna especificada. El resultado es un DataFrame con una columna por cada valor distinto encontrado, y el conteo de registros por entidad para cada valor.
- **unir_dataframes**: Esta función recibe dos DataFrames y realiza una unión 'inner' entre ellos utilizando las columnas 'NIT_ENTIDAD' y 'NOMBRE_ENTIDAD' como claves de unión.
- **aggregate_multas_data**: Esta función recibe un DataFrame de multas y realiza agregaciones a nivel de entidad (nit_entidad), calculando el número de multas, la suma y el promedio del valor de las sanciones, y los meses transcurridos desde la última multa.
- **left_join_dataframes**: Esta función recibe dos DataFrames y las columnas clave para realizar una unión 'left' entre ellos.

Posteriormente, se aplican estas funciones al DataFrame infraestructura_df_seleccionado para generar un perfil detallado de las entidades involucradas en los contratos de infraestructura. Se realiza el pivoteo y la unión de los DataFrames en función de las columnas 'DEPARTAMENTO', 'ESTADO_DEL_PROCESO', 'CLASE_PROCESO', 'TIPO_PROCESO' y 'NOMBRE_CLASE', obteniendo así un DataFrame agregado y pivotado (infraestructura_df_agregado_y_pivotado) que contiene información clave sobre las entidades y su relación con los contratos de infraestructura.

NIT_ENTIDAD	NOMBRE_ENTIDAD	num_contratos	suma_valor_total_contrato	promedio_valor_total_contrato	num_departamentos	num_estados_proceso	num_clases_proceso	num_tipos_proceso	n
899999239	ICBF REGIONAL SUCRE	169	77094589633	4.561810037455621 4E8	1	4	3	3	
899999239	INSTITUTO COLOMBI...	9004	3862769591782	4.29005951971124E8	27	6	7	13	
899999027	DEPARTAMENTO ADMI...	2218	22864702285	1.0308702563119927E7	1	5	2	5	
899999027	DEPARTAMENTO ADMI...	5635	87548446400	1.5536547719609583E7	2	7	5	7	
899999027	DEPARTAMENTO ADMI...	1541	16428195762	1.0660737029201817E7	1	6	2	3	

Figura 2. Vista del DataFrame de entidades transformado

Finalmente, se realiza una unión 'left' entre el DataFrame de entidades agregado y pivotado y el DataFrame de multas agregado, utilizando la columna 'nit_entidad' como clave de unión. El resultado es un DataFrame que contiene información detallada y depurada sobre las entidades involucradas en contratos de

infraestructura, incluyendo información sobre multas y sanciones, lo que facilita la toma de decisiones y el monitoreo de posibles irregularidades en el proceso de contratación.

Reducción de dimensionalidad

En el presente estudio, nos enfrentamos al desafío de analizar un conjunto de datos voluminoso y de alta dimensión que abarca información detallada sobre las entidades gubernamentales de Colombia y los contratos públicos de infraestructura que otorgan. Dado el volumen y la complejidad de los datos, decidimos aplicar técnicas de reducción de dimensionalidad para facilitar el análisis y mejorar la eficiencia computacional sin comprometer la calidad de la información obtenida.

La reducción de dimensionalidad es un enfoque ampliamente utilizado en la ciencia de datos para simplificar conjuntos de datos de alta dimensión, conservando la mayor parte de la información relevante y manteniendo las relaciones subyacentes entre las variables. Al reducir la cantidad de características que representan a las entidades y sus contratos, nuestro objetivo es optimizar el tiempo de cálculo y mejorar la capacidad para perfilar las entidades según las características de los contratos asignados.

Para lograr esto, empleamos diversas técnicas de reducción de dimensionalidad, incluidas PCA (Análisis de Componentes Principales), LDA (Análisis Discriminante Lineal), t-SNE (Incorporación de Vecinos Estocásticos Distribuidos en T) y UMAP (Proyección Uniforme de Aproximación de Manifold), con el fin de evaluar y comparar su eficacia en función de la varianza explicada y la preservación de las relaciones de proximidad entre los puntos. A través de una cuidadosa selección de la técnica más apropiada, nos esforzamos por conservar la mayor proporción de información en el conjunto de datos reducido, permitiendo una interpretación significativa y una toma de decisiones informada en el contexto de la contratación pública en Colombia.

Al aplicar estos métodos de reducción de dimensionalidad, esperamos obtener información valiosa sobre las entidades y sus prácticas de contratación, lo que permitirá a las autoridades detectar y prevenir la exposición a prácticas poco transparentes, incumplidas o corruptas en la asignación de contratos públicos de infraestructura. Al equilibrar la necesidad de mantener la información relevante y reducir el tiempo de cómputo, este enfoque nos permite abordar de manera efectiva el problema de la transparencia y la integridad en la contratación pública, contribuyendo al desarrollo sostenible y al bienestar de la sociedad colombiana.

> El proceso de reducción de dimensionalidad es un componente clave en el proyecto de detección y prevención de prácticas indebidas en la asignación de contratos públicos de infraestructura en Colombia. Este enfoque se emplea para disminuir la complejidad del conjunto de datos, facilitando el análisis y la visualización de los patrones subyacentes en los datos. En el proyecto, se aplican diversas técnicas de reducción de dimensionalidad, como UMAP, PCA, t-SNE y LDA, para lograr este propósito.

Importación de datos y preprocesamiento

Inicialmente, se lee el archivo CSV resultante del proceso ETL y se crea un DataFrame de pandas con la información. Luego, se aplica la función apply_standard_scaler para escalar las características numéricas y estandarizarlas, lo que facilita la comparación entre ellas.

UMAP

Uniform Manifold Approximation and Projection (UMAP) es una técnica de reducción de dimensionalidad no lineal que permite visualizar estructuras de alta dimensión en un espacio bidimensional o tridimensional. En este caso, se emplea la función apply_umap para aplicar el algoritmo UMAP al DataFrame escalado, ajustando el modelo y transformando los datos en función de sus componentes principales. La varianza explicada por cada componente se calcula y se representa gráficamente.

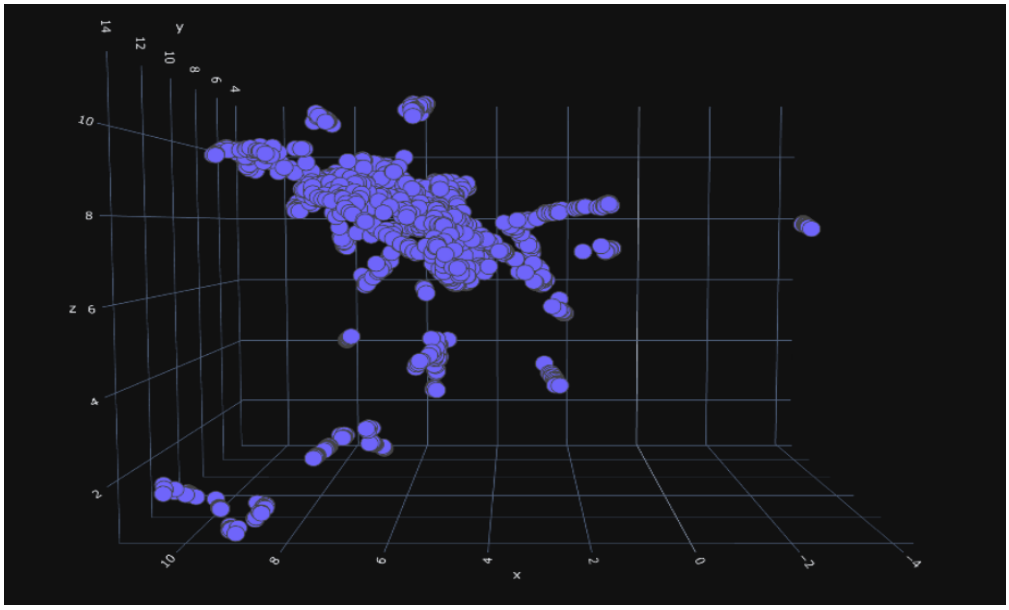

Figura 3. Vista de los 3 primeros componentes UMAP

PCA

Principal Component Analysis (PCA) es una técnica lineal de reducción de dimensionalidad que busca encontrar las direcciones de mayor varianza en los datos de alta dimensión. Se aplica la función apply_pca al DataFrame escalado, ajustando el modelo PCA y transformando los datos en función de sus componentes principales. La varianza explicada acumulada se calcula y se representa gráficamente.

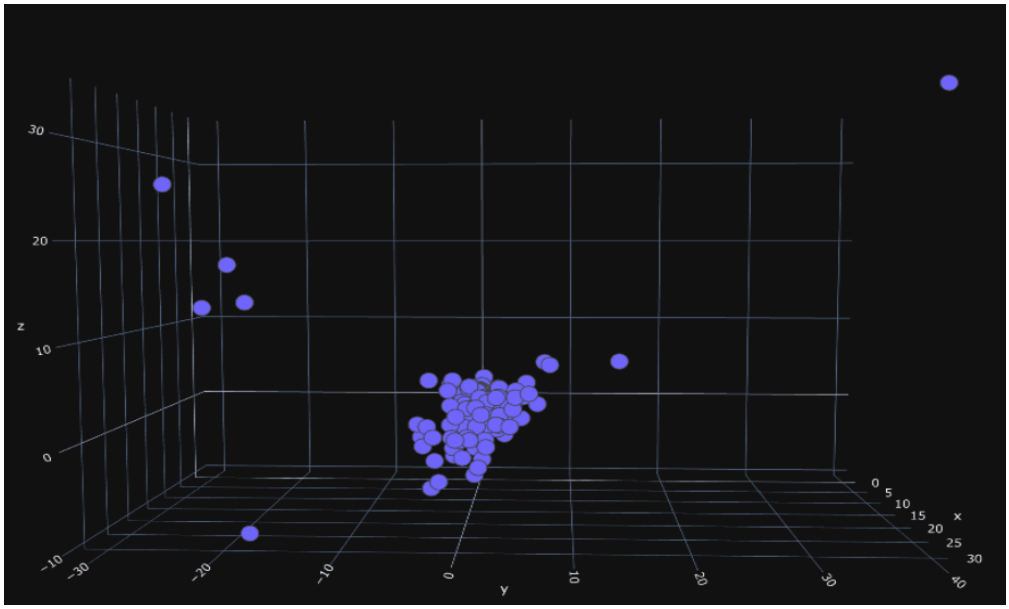

Figura 4. Vista de los 3 primeros componentes PCA

t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) es una técnica no lineal de reducción de dimensionalidad que busca preservar las relaciones de proximidad entre los puntos en un espacio de alta dimensión. Se aplica la función apply_tsne al DataFrame escalado para ajustar el modelo t-SNE y transformar los datos en función de sus componentes principales.

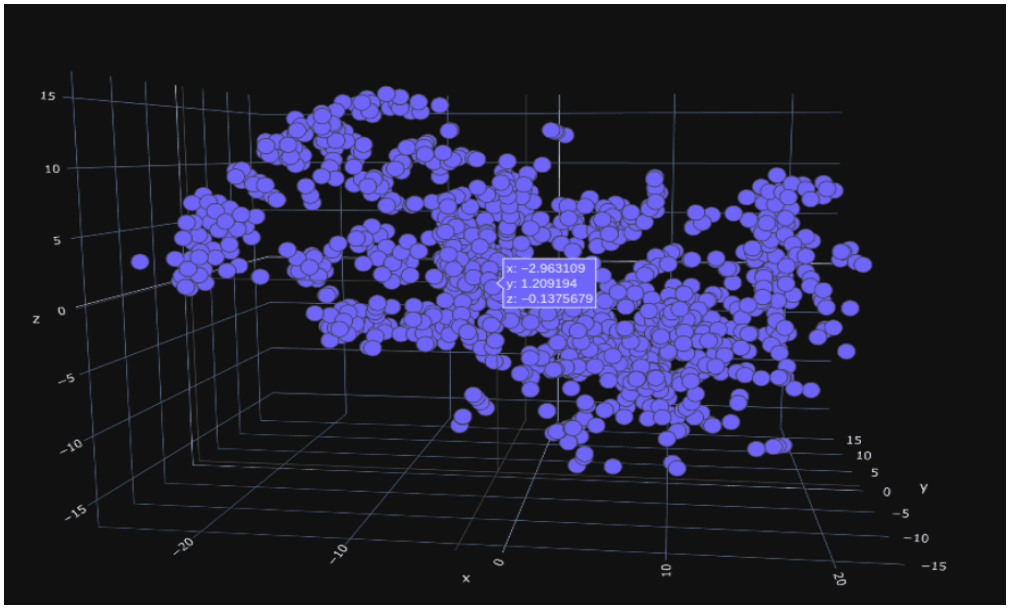

Figura 5. Vista de los 3 primeros componentes t-SNE

LDA

Linear Discriminant Analysis (LDA) es una técnica lineal de reducción de dimensionalidad que busca maximizar la separabilidad entre clases en un espacio de alta dimensión. Se aplica la función apply_lda al DataFrame escalado, proporcionando las etiquetas de clase y ajustando el modelo LDA. La varianza explicada acumulada se calcula y se representa gráficamente.

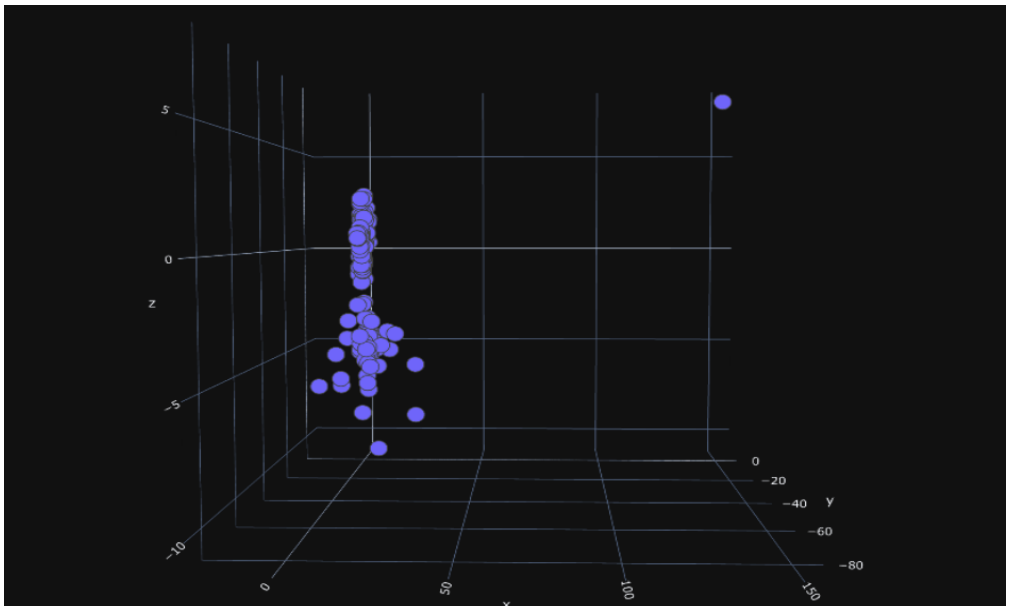

Figura 6. Vista de los 3 primeros componentes LDA

Después de un exhaustivo proceso de evaluación y comparación, hemos seleccionado t-SNE como el método más adecuado para nuestro propósito, ya que maximiza la varianza acumulada y permite detectar patrones de agrupación de manera más eficiente que las otras técnicas evaluadas. La capacidad de t-SNE

para preservar las relaciones de proximidad entre puntos en un espacio de menor dimensión facilita la identificación de grupos y la comprensión de las relaciones subyacentes entre las entidades y sus contratos.

Método	Varianza Acumulada (%)
PCA	67%
LDA	65%
t-SNE	N/A
UMAP	N/A

Al elegir t-SNE como nuestro enfoque principal de reducción de dimensionalidad, buscamos aprovechar sus ventajas en términos de maximización de la varianza acumulada y detección de patrones de agrupación. Esto nos permitirá analizar de manera efectiva el conjunto de datos y descubrir cualquier patrón que pueda indicar prácticas inadecuadas en la contratación pública. Al identificar estos patrones, podemos proporcionar información valiosa a las autoridades para que tomen medidas preventivas y correctivas, lo que contribuirá a mejorar la transparencia y la integridad en la contratación pública en Colombia.

Visualización en 3D

La función plot_3d se utiliza para visualizar las tres primeras componentes principales de los datos transformados en un gráfico tridimensional. Esto facilita la identificación de agrupaciones y patrones en los datos, lo que a su vez puede ser útil para detectar y prevenir prácticas indebidas en la asignación de contratos públicos de infraestructura en Colombia.

A medida que se aplican estas técnicas de reducción de dimensionalidad, es importante evaluar y comparar su eficacia en función de la varianza explicada y la capacidad para preservar las relaciones de proximidad entre los puntos. Algunas técnicas, como PCA y LDA, son lineales y pueden no ser adecuadas para capturar relaciones no lineales en los datos. Por otro lado, UMAP y t-SNE son técnicas no lineales que pueden ser más efectivas en estos casos. Sin embargo, es fundamental realizar pruebas y comparaciones exhaustivas para determinar cuál de estas técnicas es la más adecuada en función de los objetivos y las características del conjunto de datos específico.

Además, la visualización en 3D de las componentes principales resultantes puede ser útil para identificar agrupaciones y patrones en los datos. Estos patrones pueden proporcionar información valiosa para las entidades gubernamentales de Colombia que otorgan contratos públicos de infraestructura, ayudándoles a detectar y prevenir la exposición a prácticas indebidas en la asignación de contratos. En última instancia, este enfoque de reducción de dimensionalidad y análisis puede contribuir a garantizar una mayor transparencia, eficiencia y rendición de cuentas en la contratación pública en Colombia.

Metodologías de clustering

Se aplicaron tres técnicas de clustering: K-means, Agglomerative Clustering y OPTICS. A continuación, se presenta una descripción detallada de cada técnica y su aplicación en el proyecto:

K-means:

K-means es un algoritmo de clustering basado en la partición que agrupa los datos en K clusters, minimizando la suma de las distancias al cuadrado entre los puntos y los centroides de los clusters a los que

pertenecen. La función find_optimal_clusters se utiliza para determinar el número óptimo de clusters a utilizar en el algoritmo K-means. Se calcula la suma de las distancias al cuadrado dentro del cluster (WCSS) para un rango de números de clusters, y se grafica la relación entre el número de clusters y la WCSS en un gráfico de codo. El número óptimo de clusters se elige en el punto en el que se observa un cambio significativo en la tasa de disminución de la WCSS.

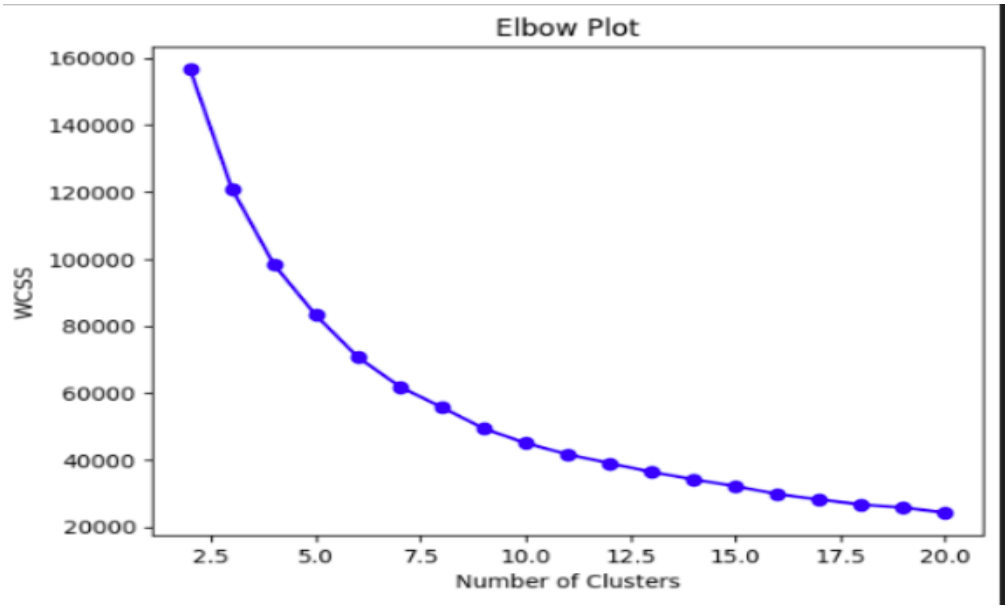

Figura 7. Elbow Kmeans

se utilizó el método del codo (Elbow Method), que consiste en calcular la suma de las distancias al cuadrado (WCSS) entre los puntos de cada cluster y su centroide para diferentes valores de K. A medida que aumenta el número de clusters, Al aplicar el método del codo en este análisis, se graficó la WCSS en función del número de clusters, y se observó que el punto de inflexión, es decir, el punto en el que el cambio en la WCSS se vuelve menos pronunciado, ocurrió en K=7. Por lo tanto, se decidió utilizar 7 clusters para el algoritmo K-means en este estudio.

Con base en esta selección, se aplicó el algoritmo K-means con 7 clusters y se obtuvieron agrupaciones de entidades gubernamentales que permiten identificar patrones y perfiles de riesgo en la contratación pública en Colombia.

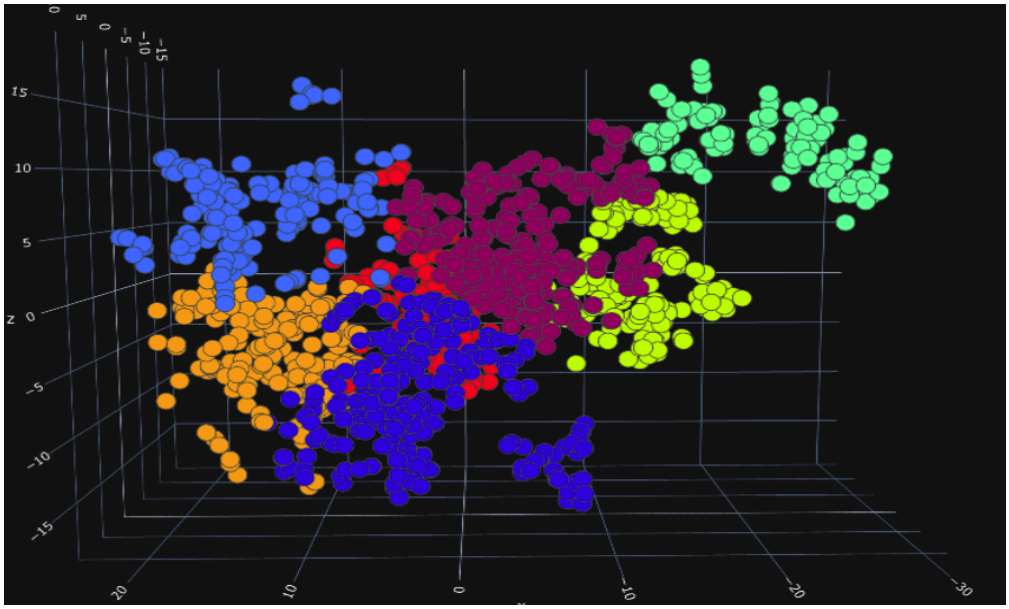

Figura 8. clusters_kmeans

Agglomerative Clustering:

El clustering aglomerativo es un enfoque jerárquico que construye un árbol de clusters (dendrograma) fusionando los clusters más cercanos en cada etapa. La función perform_agglomerative_clustering_and_plot_dendrogram se utiliza para aplicar el algoritmo de clustering aglomerativo y trazar el dendrograma resultante. El número óptimo de clusters se elige en función de la estructura del dendrograma, y se utiliza la técnica de 'ward' para calcular las distancias entre los clusters.

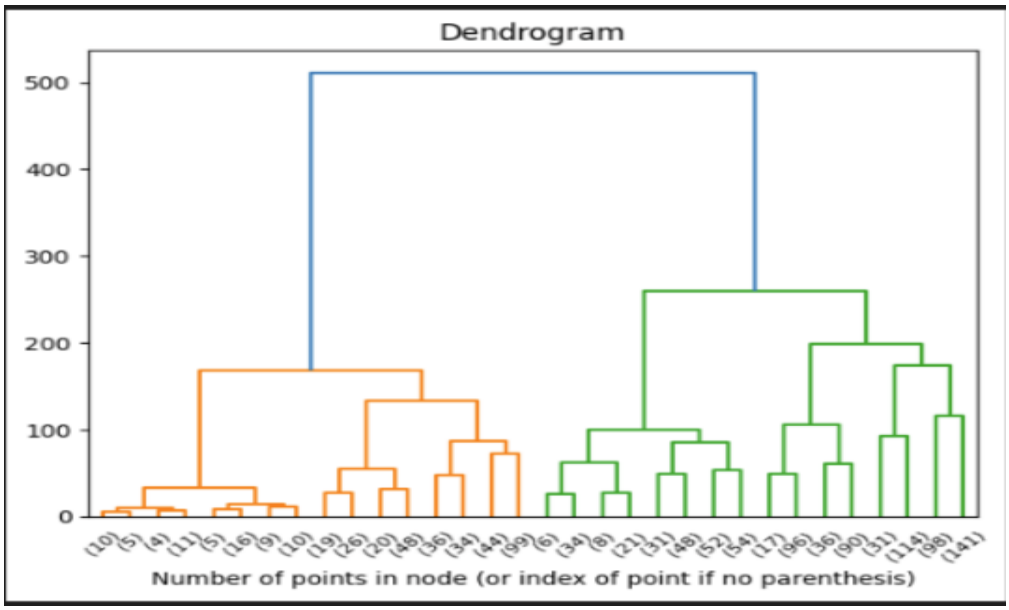

Figura 9. dendrograma

Para determinar el número óptimo de clusters en el enfoque de clustering jerárquico, se utilizó el dendrograma, que es una representación gráfica en forma de árbol que ilustra la disposición de los clusters y las relaciones jerárquicas entre ellos. Al observar el dendrograma, se pueden identificar los clusters mediante la selección de un nivel de corte en el eje vertical, que indica la distancia o disimilitud en la que se fusionan los clusters.

En este análisis, se empleó el enfoque de clustering jerárquico aglomerativo con la medida de distancia euclidiana y el método de enlace de Ward, que minimiza la varianza dentro de los clusters. Al graficar el dendrograma y analizar las fusiones de los clusters, se determinó que el número óptimo de clusters para este caso era de 7. Esta selección se basó en la observación de la altura en la que se producían fusiones significativas y la coherencia de los clusters resultantes.

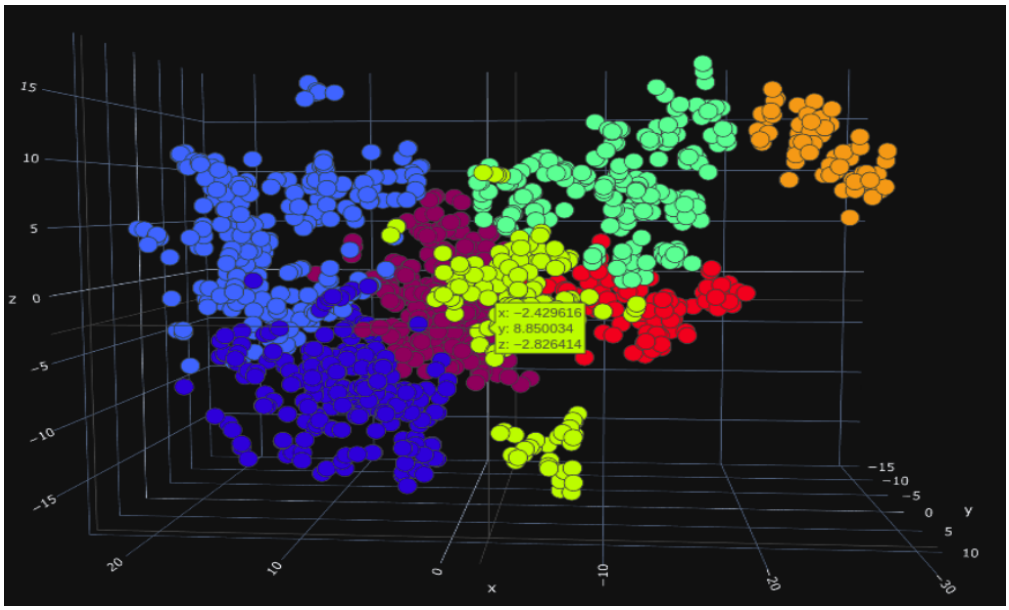

Figura 10. clusters hiherarchic

OPTICS (Ordering Points To Identify the Clustering Structure):

El algoritmo OPTICS es un enfoque basado en la densidad que identifica automáticamente los clusters en función de la densidad de los puntos en el espacio de datos. La función perform_optics_clustering se utiliza para aplicar el algoritmo OPTICS. Para determinar el valor óptimo de eps (para el alcance de búsqueda local) y min_samples (para el número mínimo de puntos en un cluster), se emplea un enfoque basado en el cálculo de las distancias k-vecinas. Se calcula la matriz de distancias k-vecinas y se grafica la relación entre las distancias k-vecinas y los puntos en el espacio de datos. El valor óptimo de eps se selecciona en función de la gráfica de distancias k-vecinas, mientras que min_samples se establece en función del tamaño del conjunto de datos (un 1% del total de puntos).

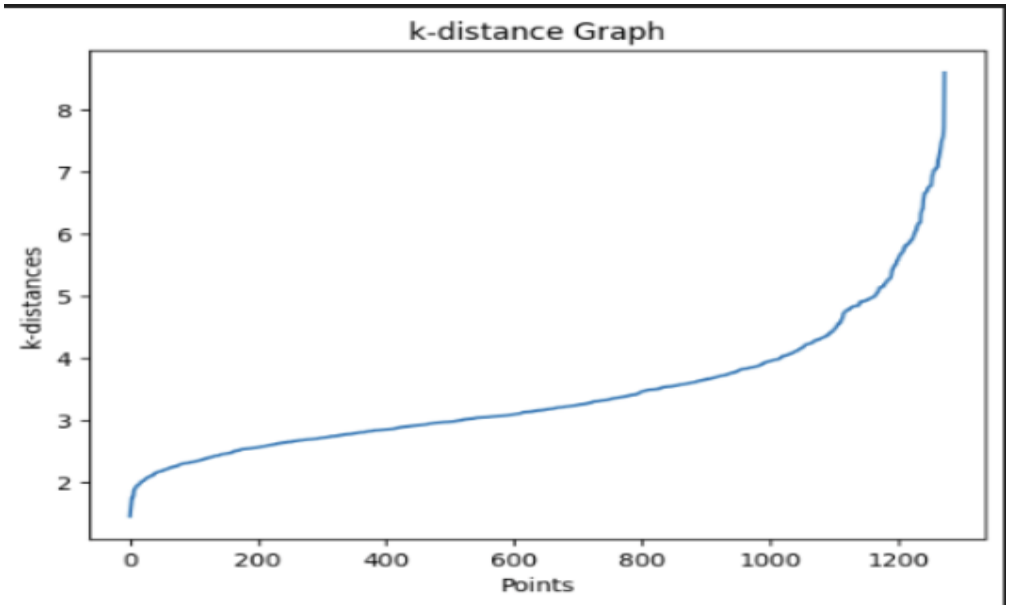

Figura 11. k_distance_optics

Un aspecto clave en la aplicación del algoritmo OPTICS es la determinación del valor óptimo de epsilon (ε), que es el radio máximo alrededor de un punto para considerar a otros puntos como parte de su vecindario. Para encontrar este valor, se utilizó un enfoque basado en k vecinos. Concretamente, se empleó el método

de los k vecinos más cercanos (k-nearest neighbors, k-NN) para calcular las distancias entre las observaciones del conjunto de datos.

Se ajustó un modelo de k vecinos más cercanos utilizando un valor predefinido de k, y se calcularon las distancias entre cada punto y su k-ésimo vecino más cercano. Estas distancias se ordenaron de manera ascendente y se graficaron en un gráfico de k-distancias. Al observar este gráfico, se identificó un "codo" o punto de inflexión, que indicaba el valor de epsilon óptimo para ser utilizado en el algoritmo OPTICS.

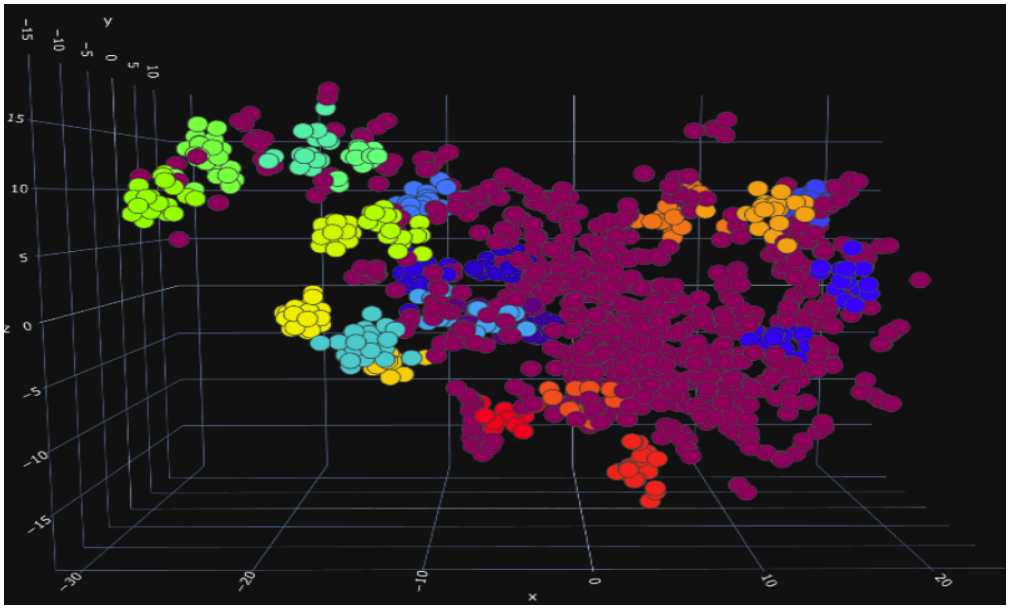

Figura 12. clusters optics

Una vez que se han aplicado las técnicas de clustering, se evalúa la calidad y la interpretabilidad de los resultados obtenidos. Para ello, se emplean métricas como el coeficiente de silueta y se analizan las características de los clusters identificados.

Algoritmo de Clustering	Coeficiente de Silueta
K-means	73%
Clustering Jerárquico	72%
OPTICS	45%

Tras la aplicación de los algoritmos K-means y clustering jerárquico, se observó que los resultados de los coeficientes de silueta y la estructura de los clusters eran muy similares entre ambos métodos. Esto indica que ambos algoritmos lograron identificar patrones y agrupaciones consistentes en el conjunto de datos.

Por otro lado, el algoritmo OPTICS también fue capaz de detectar grupos pequeños y realmente diferentes en el conjunto de datos. Sin embargo, este algoritmo dejó una masa muy grande de observaciones sin diferenciar de manera adecuada, lo que sugiere que OPTICS no fue tan efectivo para segmentar y extraer información útil de este conjunto de datos en particular.

Teniendo en cuenta estos resultados, se decidió seleccionar el algoritmo K-means como el método de agrupamiento final para este proyecto. La elección de K-means se basa en su capacidad para generar clusters cohesivos y bien separados, como lo demuestran los coeficientes de silueta similares a los del clustering jerárquico, y en su eficiencia computacional en comparación con los otros dos algoritmos. Además,

K-means permite una mayor diferenciación de la masa de observaciones que no pudo ser segmentada adecuadamente por el algoritmo OPTICS.

Descrpción de los clusters

Cluster 1

> Pequeños Contratos con Bajas Multas: Este cluster agrupa a entidades que otorgan en promedio 35 contratos con un valor total aproximado de 4,73 mil millones de pesos. Han sido multados en promedio 2 veces con un valor total de multas de alrededor de 260 mil pesos. El tiempo desde la última multa es de aproximadamente 20 meses. Operan principalmente en Santander, Cauca y Antioquia.

Cluster 2

> Contratos Medianos con Multas Moderadas: En este cluster, las entidades otorgan alrededor de 148 contratos con un valor total de 56 mil millones de pesos. Han sido multadas en promedio 7 veces con un valor total de multas de alrededor de 9,6 millones de pesos. El tiempo desde la última multa es de aproximadamente 19 meses. Operan principalmente en Antioquia, Bogotá y Valle del Cauca.

Cluster 3

> Grandes Contratos con Altas Multas: Este cluster incluye entidades que otorgan en promedio 458 contratos con un valor total de 71,7 mil millones de pesos. Han sido multadas en promedio 5 veces con un valor total de multas de alrededor de 511 millones de pesos. El tiempo desde la última multa es de aproximadamente 19 meses. Operan principalmente en Bogotá, Antioquia y Valle del Cauca.

Cluster 4

> Contratos Mínimos con Mínimas Multas: Las entidades en este cluster otorgan en promedio 4 contratos con un valor total de aproximadamente 453 millones de pesos. Han sido multadas en promedio 1 vez con un valor total de multas de alrededor de 116 mil pesos. El tiempo desde la última multa es de aproximadamente 24 meses. Operan principalmente en Santander y Nariño.

Cluster 5

> Contratos Bajos con Multas Bajas: Este cluster agrupa a entidades que otorgan en promedio 9 contratos con un valor total de aproximadamente 576 millones de pesos. Han sido multadas en promedio 2 veces con un valor total de multas de alrededor de 1,16 millones de pesos. El tiempo desde la última multa es de aproximadamente 20 meses. Operan principalmente en Bogotá, Antioquia y Tolima.

Cluster 6

> Contratos Múltiples con Multas Múltiples: Las entidades en este cluster otorgan en promedio 300 contratos con un valor total de aproximadamente 29,3 mil millones de pesos. Han sido multadas en promedio 15 veces con un valor total de multas de alrededor de 70,6 millones de pesos. El tiempo

> desde la última multa es de aproximadamente 14 meses. Operan principalmente en Bogotá, Antioquia y Valle del Cauca.

Cluster 7

> Contratos Moderados con Multas Significativas: En este cluster, las entidades otorgan alrededor de 75 contratos con un valor total de 11,7 mil millones de pesos. Han sido multadas en promedio 10 veces con un valor total de multas de alrededor de 152 millones de pesos. El tiempo desde la última multa es de aproximadamente 16 meses. Operan principalmente en Antioquia, Bogotá y Atlántico.

Clusters y sus características principales

Cluster	Descripción	Contratos	Valor Total (COP)	Multas	Valor Total Multas (COP)	Tiempo desde última multa (meses)	Principales regiones
1	Pequeños Contratos con Bajas Multas	35	4,73 mil millones	2	260 mil	20	Santander, Cauca, Antioquia
2	Contratos Medianos con Multas Moderadas	148	56 mil millones	7	9,6 millones	19	Antioquia, Bogotá, Valle del Cauca
3	Grandes Contratos con Altas Multas	458	71,7 mil millones	5	511 millones	19	Bogotá, Antioquia, Valle del Cauca
4	Contratos Mínimos con Mínimas Multas	4	453 millones	1	116 mil	24	Santander, Nariño
5	Contratos Bajos con Multas Bajas	9	576 millones	2	1,16 millones	20	Bogotá, Antioquia, Tolima
6	Contratos Múltiples con Multas Múltiples	300	29,3 mil millones	15	70,6 millones	14	Bogotá, Antioquia, Valle del Cauca
7	Contratos Moderados con Multas Significativas	75	11,7 mil millones	10	152 millones	16	Antioquia, Bogotá, Atlántico