Inspira Crea Transforma

UNIVERSIDAD EAFIT

ALIANZA CAOBA – UNIVERSIDAD EAFIT

Estrategia de Segmentación de Empresas

Proyecto Integrador I

Maestría en Ciencia de los Datos y Analítica

Gitlab Jupyter CAOBA

Cristian David Rojas Rincón, cdrojasr@eafit.edu.co, Economista.
Daniela Vasquez Jaramillo, dvasqu18@eafit.edu.co, Negociadora Internacional.
Yaliza Margarita Barcelo Pulgar, ymbarcelop@eafit.edu.co, Contadora - Esp. ingeniería de Software
Daniel Patiño Barraza, dpatinob@eafit.edu.co, Estudiante Finanzas y Matemáticas
David Rúa Jaramillo, druaj@eafit.edu.co, Ingeniero Administrador - MBA.

AGENDA

1. Problema a resolver:
   - ✓ Descripción del problema
   - ✓ Objetivos
   - ✓ Fuentes de datos
2. Análisis exploratorio de datos:
   - ✓ EDA - Análisis Exploratorio de Datos
3. Modelado
   - ✓ Desarrollo de modelo
   - ✓ Evaluación
   - ✓ Visualización
4. Análisis de resultados
5. Conclusión

1. PROBLEMA A RESOLVER
DESCRIPCIÓN DEL PROBLEMA

Alianza Caoba es el centro de excelencia y apropiación que apoya el uso de las tecnologías de Big Data y Data Analytics, a través de diferentes frentes que incluyen la formación del talento humano, la investigación aplicada y el desarrollo de productos a la medida.

Como estrategia de Alianza CAOBA se quiere entender el universo de empresas que tienen mayor inclinación analítica para ser contactadas y así enfocar mejor los esfuerzos de la estrategia comercial.

El líder de innovación quiere resolver la siguiente pregunta:

¿Cuáles son las empresas que se deben contactar dadas las variables asociadas con innovación en analítica internamente en las entidades?

1. PROBLEMA A RESOLVER
OBJETIVOS

Objetivo general:
Identificar y priorizar organizaciones del sector público que puedan necesitar soluciones de analítica.

Objetivos específicos:

1. Identificar cuáles empresas del sector público han realizado inversiones en proyectos de analítica en los últimos años.

2. Entregar un dashboard segmentado para identificar cuales entidades pueden estar interesadas en los servicios de analítica.

1. PROBLEMA A RESOLVER
FUENTES DE DATOS

Plan Anual Secop (Publico)

"El Plan Anual de Adquisiciones es una herramienta para: **(i)** facilitar a las Entidades Estatales identificar, registrar, programar y divulgar sus necesidades de bienes, obras y servicios; y **(ii)** diseñar estrategias de contratación basadas en agregación de la demanda que permitan incrementar la eficiencia del proceso de contratación. (…) **busca comunicar información útil y temprana a los proveedores potenciales de las Entidades Estatales, para que éstos participen de las adquisiciones que hace el Estado.** "

(colombiacompra.gov)

Actualizado
25 de marzo de 2021

Datos actualizados por última vez
25 de marzo de 2021

Última actualización de metadatos
8 de septiembre de 2020

Fecha de creación
7 de julio de 2020

Vistas
1.226

Descargas
139

Datos ofrecidos por
Colombia Compra Eficiente

Propietario de conjunto de datos
Colombia Compra Eficiente

Información de la Entidad

Área o dependencia	Subdirección de IDT
Nombre de la Entidad	Agencia Nacional de Contratación Pública Colombia Compra Eficiente
Departamento	Bogotá D.C.
Municipio	Bogotá D.C.
Orden	Nacional
Sector	Planeación

Información de Datos

Idioma	Español
Cobertura Geográfica	Nacional
Frecuencia de Actualización	Semanal
Fecha Emisión (aaaa-mm-dd)	2020-07-07

1. PROBLEMA A RESOLVER
FUENTES DE DATOS

IDENTIFICACIÓN DE VARIABLES

1. Año
2. Identificador PAA
3. Entidad
4. NIT
5. Localización
6. Descripcion/Ubicación
7. Mision/Vision
8. Perspectiva Estratégica
9. Presupuesto Menor Cuantía
10. Presupuesto Minima Cuantía
11. Presupuesto Global
12. Fecha Primera Publicación
13. Mes Proyectado
14. Identificador Item
15. Categoria Principal
16. Precio Base
17. Ultima Fecha Modificacion
18. Version
19. Referencia Contrato
20. Referencia Operación
21. Fecha Publicacion
22. Modalidad
23. Contacto
24. UNSPSC - Codigo Producto
25. UNSPSC - Nombre Producto
26. UNSPSC - Codigo Clase
27. UNSPSC - Nombre Clase
28. UNSPSC - Codigo Familia
29. UNSPSC - Nombre Familia

2. ANÁLISIS EXPLORATORIO DE DATOS

CLASIFICACIÓN DE VARIABLES

Identificación:

- ~~Identificador PAA~~
- Entidad (obj)
- NIT (int64)
- Localización(obj)
- Descripcion/Ubicación (obj)
- Mision/Vision (obj)
- Perspectiva Estrat. (obj)

Monetización:

- ~~Presupuesto Menor Cuantía~~
- ~~Presupuesto Minima Cuantía~~
- Presupuesto Global (int64)
- Precio Base (int64)

Otros:

- ~~Versión~~
- ~~Referencia Operación~~
- Referencia Contrato (obj)
- Contacto (obj)

Temporalidad:

- Año (int64)
- ~~Fecha Primera Publicación~~
- Mes Proyectado (dtime)
- Ultima Fecha Modif.(dtime)
- Fecha Publicacion (dtime)

Categorización:

- ~~Identificador Item~~
- ~~Categoría Principal~~
- UNSPSC - Codigo Producto (obj)
- UNSPSC - Nombre Producto (obj)
- UNSPSC - Codigo Clase (obj)
- UNSPSC - Nombre Clase (obj)
- UNSPSC - Codigo Familia (obj)
- UNSPSC - Nombre Familia (obj)

Modalidad:

- Modalidad (obj)

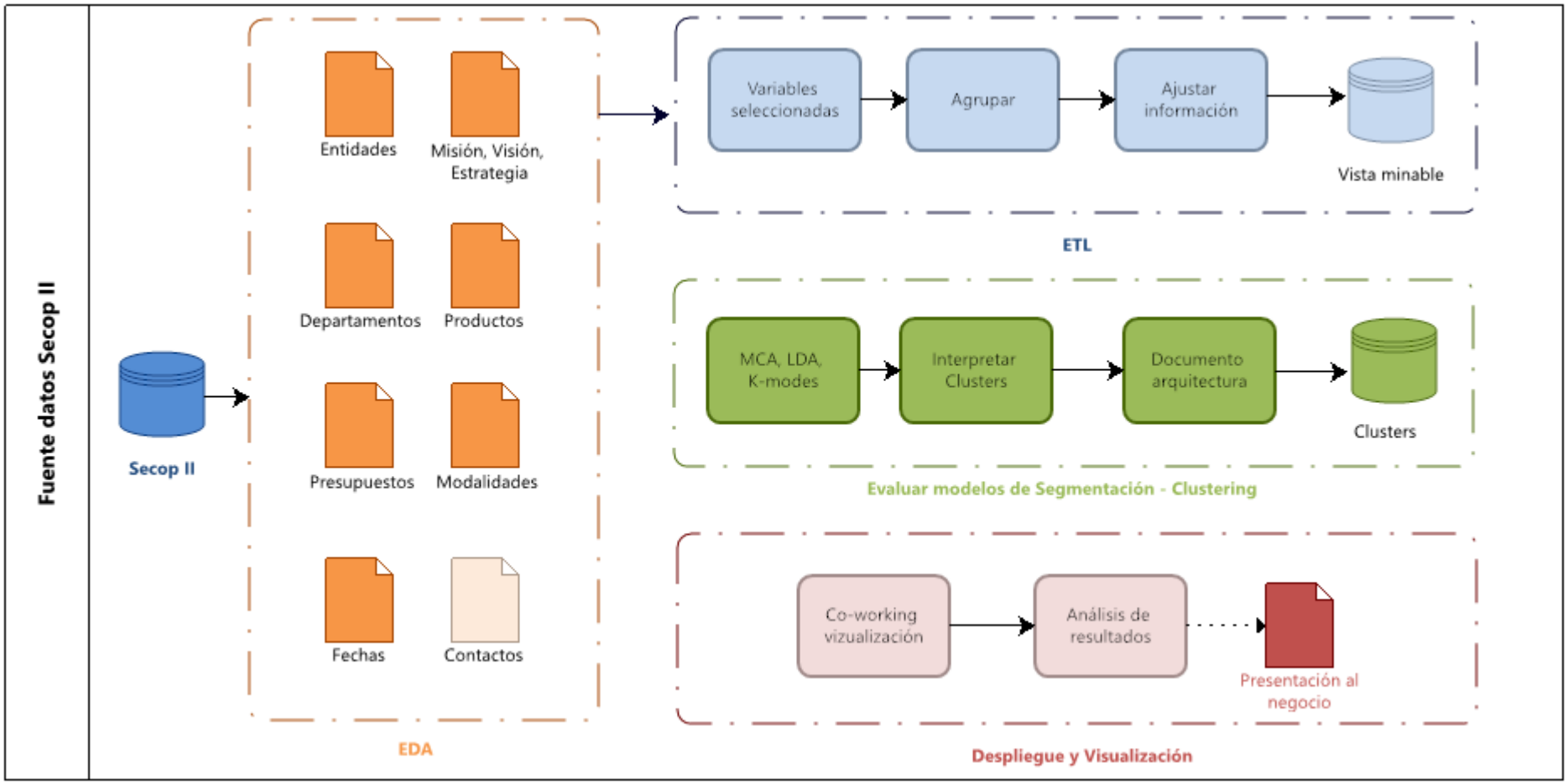
Vista minable:

Dimensión 349.671 x 24

Creación matriz_entidad
Ajuste formato de fecha
Remoción de valores nulos
Segmentación por modalidades

3. MODELADO

ARQUITECTURA DEL PROYECTO



3. MODELADO

ANÁLISIS HEURÍSTICO

1. Creación de diccionario
2. Filtrado de diccionario en nombre_producto (PivotTable)
3. Análisis de familia_producto del filtrado anterior
4. Análisis de clase_producto del filtrado anterior

DICCIONARIO:
Tecnologia, Informatica, Data, Datos, Analitica, Analizador, Dato, Prediccion

3. MODELADO

ANÁLISIS HEURÍSTICO

Familia

nit	397
entidad_matriz	396
year	5
entidad	474
localizacion	107
localizacion_desc	30
mision_vision	730
pers_estrategica	682
ppto_global	857
mes_proyectado	12
precio_base	11275
date_last_publication	327
ref_contrato	37815
date_publised	191
modalidad	4
contacto	687
cod_producto	701
nombre_producto	701
cod_clase	185
nombre_clase	185
cod_familia	32
nombre_familia	32
diff_dates	289
year_publised	5

dtype: int64

Clase

nit	323
entidad_matriz	322
year	5
entidad	383
localizacion	88
localizacion_desc	28
mision_vision	573
pers_estrategica	532
ppto_global	670
mes_proyectado	12
precio_base	7406
date_last_publication	272
ref_contrato	28745
date_publised	154
modalidad	4
contacto	540
cod_producto	261
nombre_producto	261
cod_clase	44
nombre_clase	44
cod_familia	32
nombre_familia	32
diff_dates	237
year_publised	5

dtype: int64

Producto

nit	145
entidad_matriz	145
year	4
entidad	165
localizacion	35
localizacion_desc	15
mision_vision	219
pers_estrategica	198
ppto_global	243
mes_proyectado	12
precio_base	772
date_last_publication	124
ref_contrato	1178
date_publised	71
modalidad	3
contacto	205
cod_producto	67
nombre_producto	67
cod_clase	44
nombre_clase	44
cod_familia	32
nombre_familia	32
diff_dates	117
year_publised	4

dtype: int64

3. MODELADO

ANÁLISIS DE COLINEALIDAD (MCA)

1. Vista minable
2. Análisis univariado
3. Análisis bivariante
4. Análisis de anova de Precio Base con:
   - 4.1 Familia
   - 4.2 Localización
   - 4.3 Mes proyectado
   - 4.4 Modalidad
5. Análisis de independencia – Tabla de contingencia entre variables categoricas
   - 5.1 código producto – modalidad
   - 5.2 código producto – localización
   - 5.3 código producto – mes proyectado
   - 5.4 precio base por categorías

6. **MCA :**
   - 6.1 Localización
   - 6.2 Familia
   - 6.3 Categoria Precio Base
   - 6.4 Modalidad

3. MODELADO

ANÁLISIS DE COLINEALIDAD (MCA)

- Análisis Univariado

Info

```
Int64Index: 1285 entries, 233 to 858930
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   nit                   1285 non-null   int64
 1   entidad_matriz        1285 non-null   object
 2   year                  1285 non-null   int64
 3   entidad               1285 non-null   object
 4   localizacion          1285 non-null   object
 5   localizacion_desc     1285 non-null   object
 6   mision_vision         1285 non-null   object
 7   pers_estrategica      1285 non-null   object
 8   ppto_global           1285 non-null   int64
 9   mes_proyectado        1285 non-null   object
 10  precio_base           1285 non-null   float64
 11  date_last_publication 1285 non-null   datetime64[ns]
 12  ref_contrato          1285 non-null   object
 13  date_publised         1285 non-null   datetime64[ns]
 14  modalidad             1285 non-null   object
 15  contacto              1285 non-null   object
 16  cod_producto          1285 non-null   object
 17  nombre_producto       1285 non-null   object
 18  cod_clase             1285 non-null   object
 19  nombre_clase          1285 non-null   object
 20  cod_familia           1285 non-null   object
 21  nombre_familia        1285 non-null   object
 22  diff_dates            1285 non-null   int64
 23  year_publised         1285 non-null   int64
dtypes: datetime64[ns](2), float64(1), int64(5), object(16)
memory usage: 251.0+ KB
```

Describe

	entidad_matriz	entidad	localizacion	localizacion_desc	mision_vision	pers_estrategica	mes_proyectado	ref_contrato
count	1285	1285	1285	1285	1285	1285	1285	1285
unique	145	165	35	15	219	198	12	1178
top	BOGOTA DISTRITO CAPITAL	INSTITUTO GEOGRÁFICO AGUSTÍN CODAZZI	CO-DC	Distrito Capital de Bogotá	Producir, proveer y divulgar información y con...	De acuerdo con las (6) estrategias establecida...	Enero	ENV-07-09-026-18
freq	101	90	510	991	90	83	735	16

Precio Base:
Variable numérica objetivo

Media: $397’020.789
Mediana: $43’827.960
Moda: $43’827.960

DS: $1.847’282.413.21

3. MODELADO

ANÁLISIS DE COLINEALIDAD (MCA)

- **Análisis Bivariante**

✓ Boxplot Precio Base y Modalidad
✓ Histograma Precio Base
✓ Distribución de participación por modalidad
✓ Diagrama de frecuencias por Localización

- **Análisis Independencia (Precio Base)**
  (Tablas de contingencia)

✓ Modalidad
✓ Localización
✓ Mes proyectado

modalidad	CCE-02-Licitacion_Publica	CCE-06-Seleccion_Abreviada_Menor_Cuantia	CCE-16-Servicios_profesionales_gestion	All
nombre_producto				
Administradores permanentes de bases de datos o de sistemas de tecnologías de la información	3	0	14	17
Administradores temporales de bases de datos o de sistemas de tecnologías de la información	0	0	47	47
Analizador de datos	0	0	2	2
Centros de información	0	0	20	20
Copias de seguridad y almacenamiento de datos	0	0	1	1
...	...	...	...	...
Software de recuperación o búsqueda de información	0	0	1	1
Software de reportes de bases de datos	0	2	4	6
Software de sistemas de manejo de base datos	0	1	1	2
servicios de almacenamiento de datos	0	0	19	19
All	51	107	1127	1285

3. MODELADO

ANÁLISIS DE COLINEALIDAD (MCA)

- **Análisis de la Varianza (ANOVA)**

Familia

	sum_sq	df	F	PR(>F)
nombre_familia	5.020026e+20	31.0	5.230099	2.755968e-18
Residual	3.879586e+21	1253.0	NaN	NaN

Localización

	sum_sq	df	F	PR(>F)
localizacion	4.001801e+19	34.0	0.338875	0.99988
Residual	4.341571e+21	1250.0	NaN	NaN

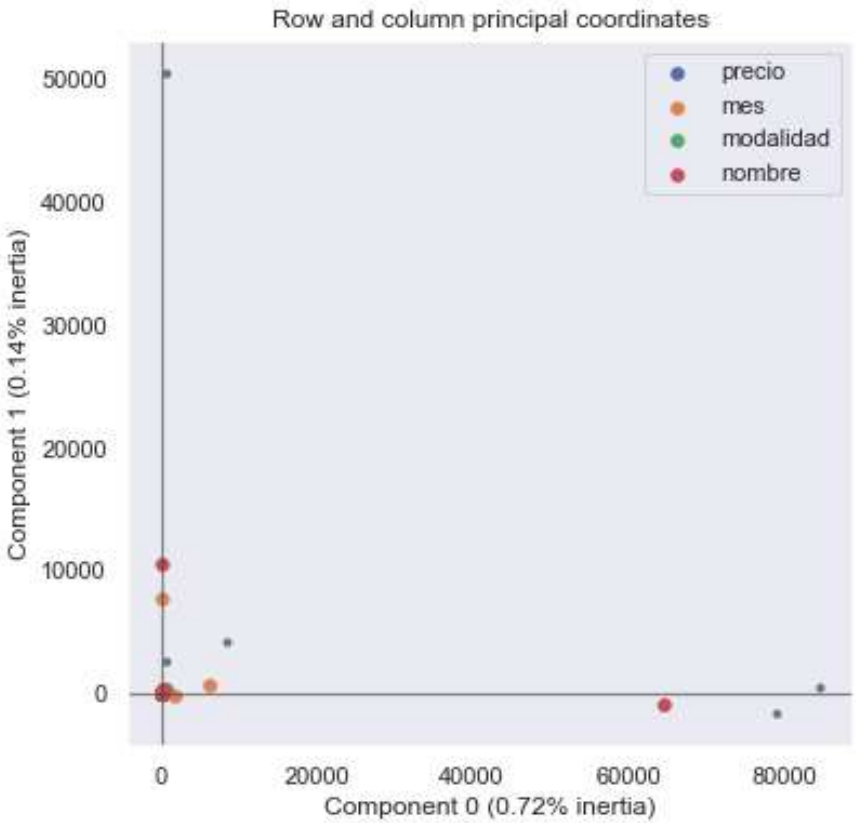
Mes proyectado

	sum_sq	df	F	PR(>F)
mes_proyectado	1.259183e+20	11.0	3.424181	0.000105
Residual	4.255670e+21	1273.0	NaN	NaN

Modalidad

	sum_sq	df	F	PR(>F)
modalidad	6.012724e+20	2.0	101.95326	8.122319e-42
Residual	3.780316e+21	1282.0	NaN	NaN

MCA:



3. MODELADO

LDA Y HDP

1. Vista minable
2. Remover caracteres especiales
3. Tokenizar
4. Remover Stop Words, Common Words
5. Reducir dimensionalidad, eliminando tokens de alta y baja frecuencia
6. Lematización
7. Cálculo de bi-grams y tri-grams
8. Aplicamos LDA con LDAvis y HDP
9. Graficar por topics
10. Análisis del experto
    - 10.1 Identificación de palabras por tópicos
    - 10.2 Identificación de palabras en dif tópicos
    - 10.3 Mejoramiento de análisis LDA desde 4 y 5

3. MODELADO

LDA Y HDP

Frecuencia de palabras

Bigrams

```
finder.nbest(bigram_measures.likelihood_ratio, 10)
```

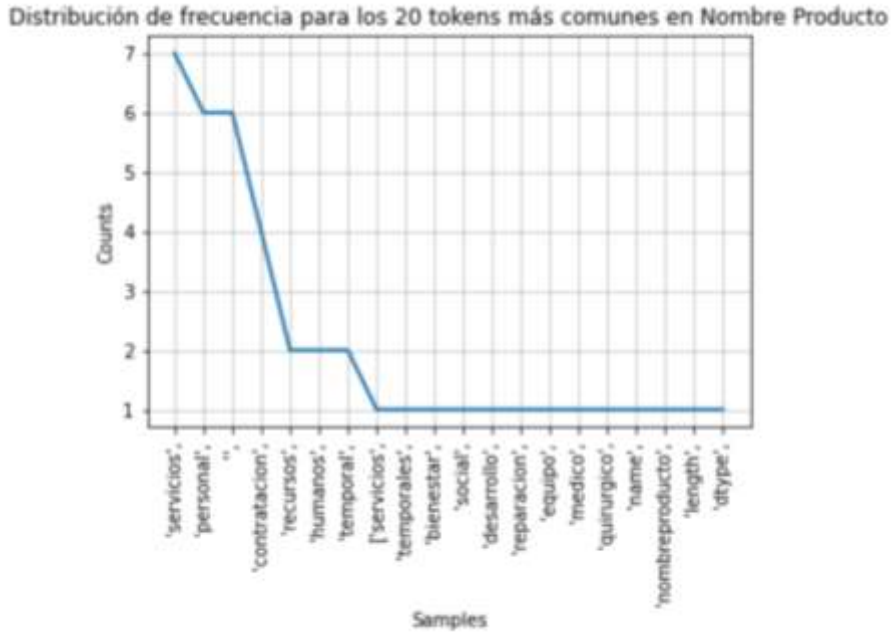
```
[("'',", "'servicios',"),
 ("'servicios',", "'contratacion',"),
 ("'',", "'personal',"),
 ("'personal',", "'',"),
 ("'humanos',", "'servicios',"),
 ("'recursos',", "'servicios',"),
 ("'servicios',", "'',"),
 ("'servicios',", "'personal',"),
 ("'personal',", "'contratacion',"),
 ("'personal',", "'servicios',")]
```

Trigrams

```
finder = nltk.collocations.TrigramCollocationFinder.from_words(tokens)
```

```
finder.ngram_fd.most_common(5)
```

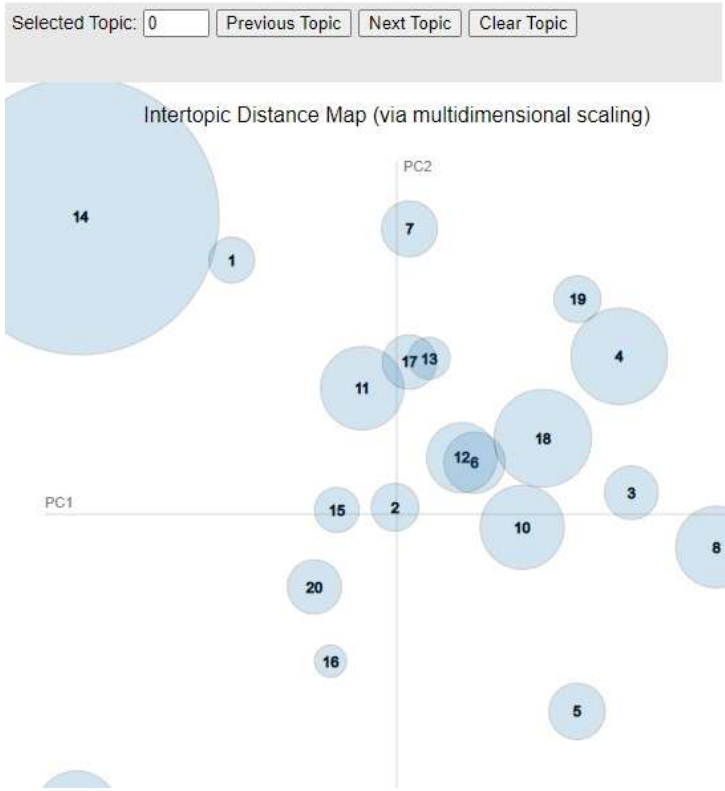
```
[(("'',", "'servicios',", "'contratacion',"), 4),
 (("'servicios',", "'contratacion',", "'personal',"), 4),
 (("'contratacion',", "'personal',", "'',"), 4),
 (("'personal',", "'',", "'servicios',"), 3),
 (("'recursos',", "'humanos',", "'servicios',"), 2)]
```

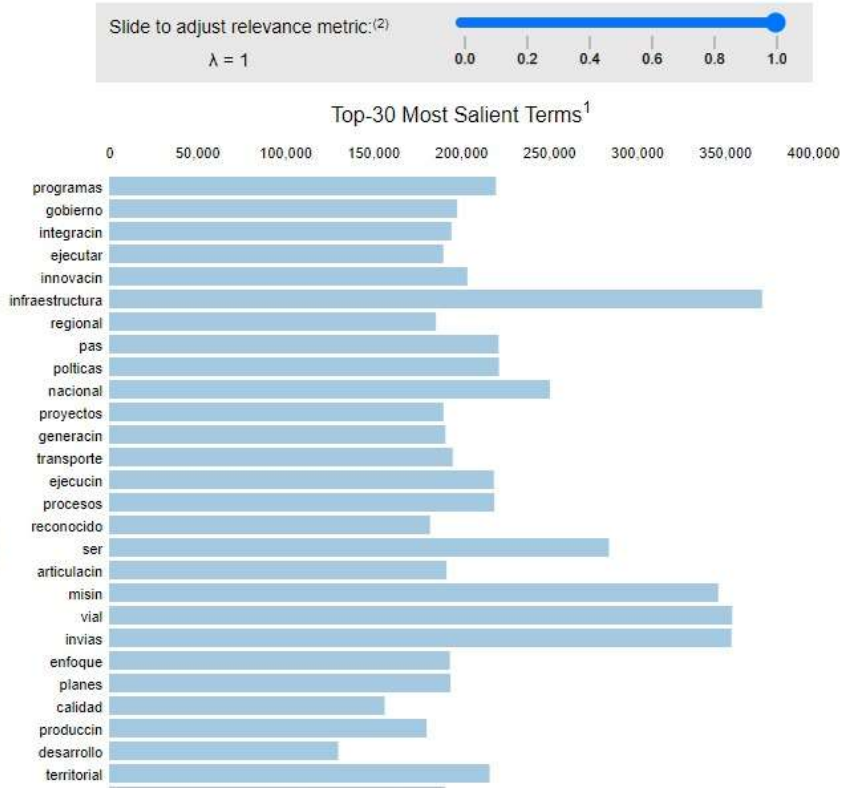


3. MODELADO

LDA Y HDP

Visualización LDA Nombre_Producto

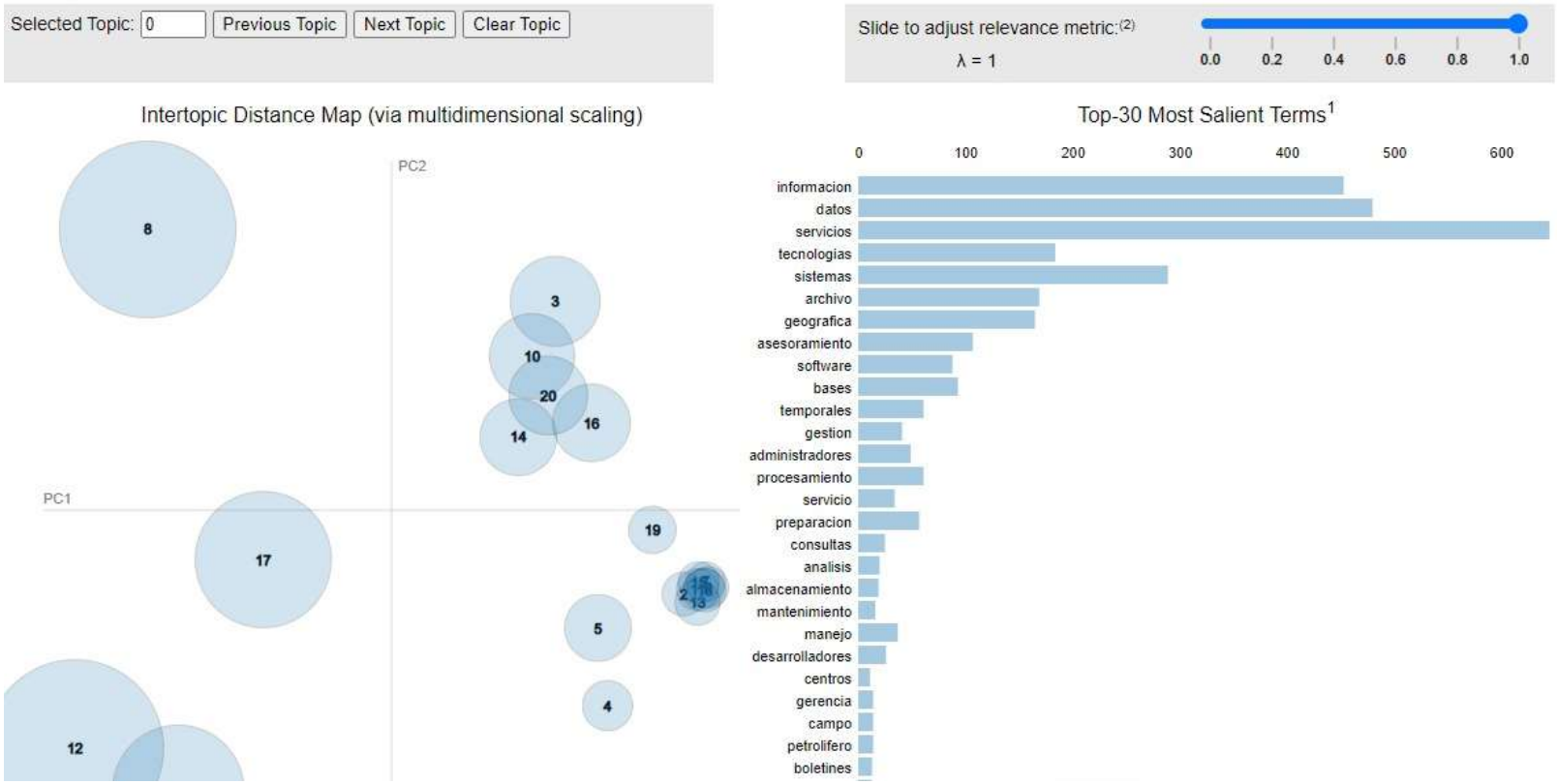




3. MODELADO

LDA Y HDP

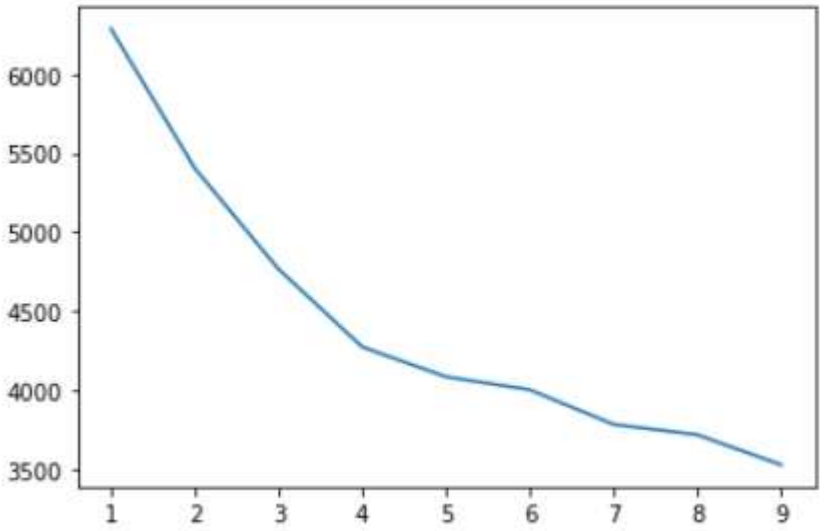
Visualización LDA Nombre_Producto (CON AJUSTE)



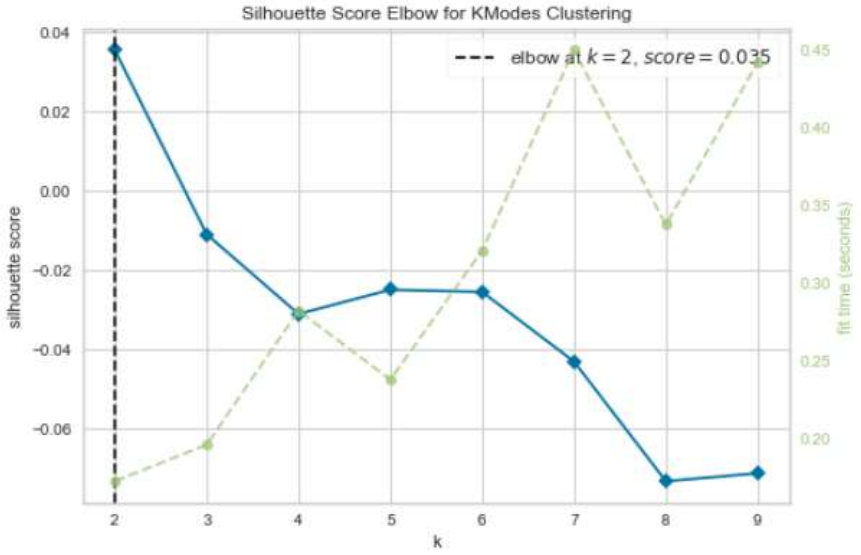
3. MODELADO

K-MODES

Análisis de Codo



Análisis de Silueta

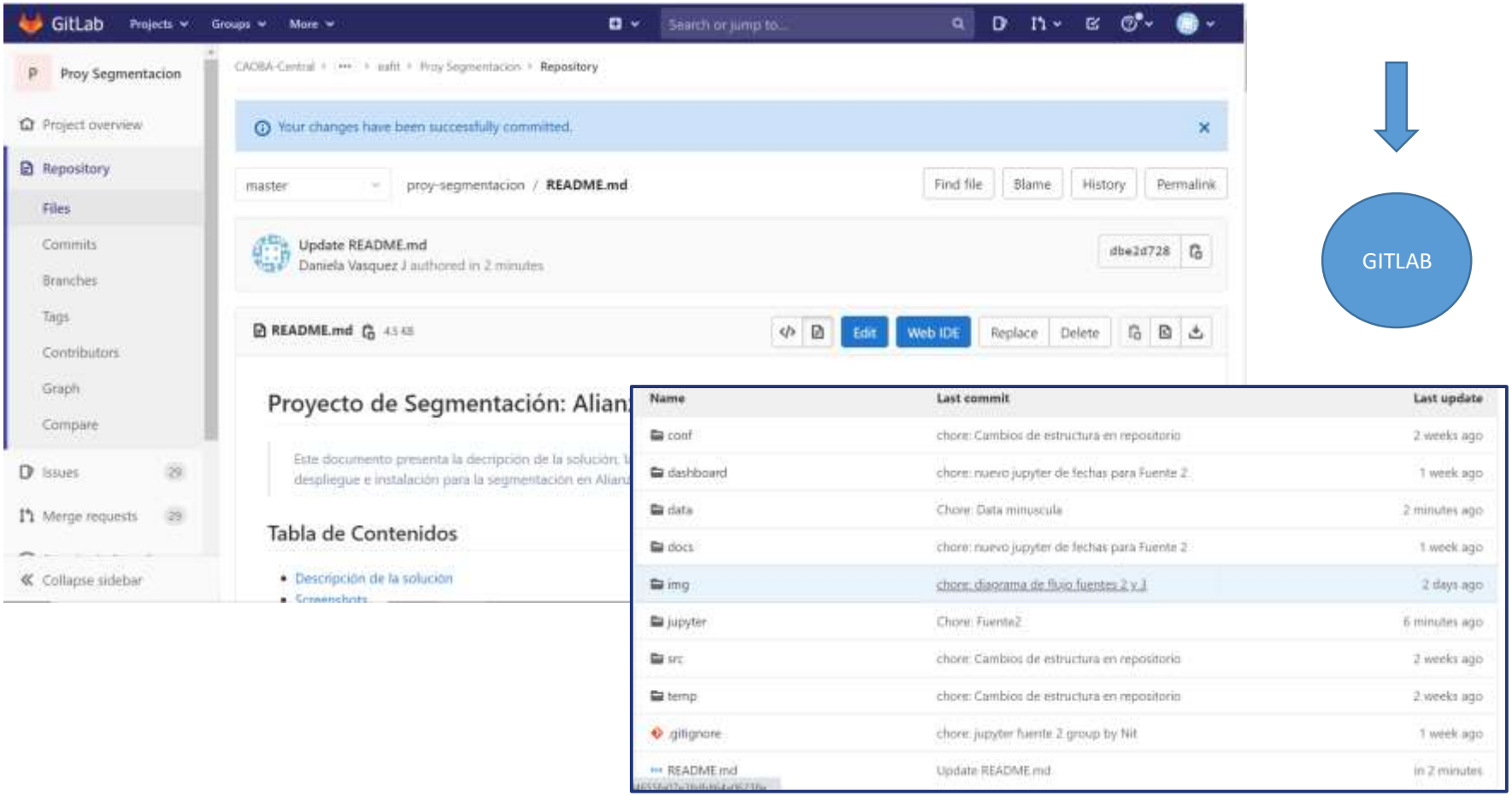


3. MODELADO

K-MODES

1. Vista minable.
2. Seleccionar variables para técnica K-modes.
3. Experimentación de k =2.
4. Experimentación de K 1 a 10.
   - 4.1 Análisis de codo.
   - 4.2 Análisis de silueta.
5. Análisis gráfico de resultados de cluster (Bar-plot).
6. Descripción y análisis de centroides.

ACCESO DE GITLAB



GITLAB

ACCESO DE DASHBOARD BI

Alianza CAOBA

Precio Base

Precio Base

0 — 29,550,000,000

Modalidad

- [ ] Seleccionar todo
- [ ] CCE-02-Licitacion_Publica
- [ ] CCE-06-Seleccion_Abreviada_Menor_Cuantia
- [ ] CCE-16-Servicios_profesionales_gestion

Año Publicado

2017 | 2018 | 2019 | 2020

Entidad	Ubicación	Contacto	Referencia Contrato
CONTRALORIA MUNICIPAL ITAGUI	Antioquia	ADRIANA PATRICIA GRISALES RENDÓN	CD004-2019
DEPARTAMENTO DE ANTIOQUIA	Antioquia	ALVARO URIBE MORENO	8718
DEPARTAMENTO DE ANTIOQUIA	Antioquia	ALVARO URIBE MORENO	8819
DEPARTAMENTO DE ANTIOQUIA	Antioquia	ALVARO URIBE MORENO	9207
DEPARTAMENTO DE ANTIOQUIA	Antioquia	ALVARO URIBE MORENO	9233
DEPARTAMENTO DE ANTIOQUIA	Antioquia	ALVARO URIBE MORENO	9813
MUNICIPIO DE MEDELLIN	Antioquia	Gustavo Alonso Deossa Lastra	0009013509
MUNICIPIO DE MEDELLIN	Antioquia	Gustavo Alonso Deossa Lastra	25756
MUNICIPIO DE MEDELLIN	Antioquia	Gustavo Alonso Deossa Lastra	25764
MUNICIPIO DE MEDELLIN	Antioquia	Gustavo Alonso Deossa Lastra	25981
MUNICIPIO DE MEDELLIN	Antioquia	Gustavo Alonso Deossa Lastra	26175
MUNICIPIO DE MEDELLIN	Antioquia	Gustavo Alonso Deossa Lastra	26271
MUNICIPIO DE MEDELLIN	Antioquia	Gustavo Alonso Deossa Lastra	26317
MUNICIPIO DE MEDELLIN	Antioquia	Gustavo Alonso Deossa Lastra	26318
MUNICIPIO DE MEDELLIN	Antioquia	Gustavo Alonso Deossa Lastra	26398

ACCESO DE DASHBOARD BI

Alianza CAOBA

Clusters

145
Entidades

Modalidad de Contratación
- [x] Seleccionar todo
- [x] CCE-02-Licitacion_Publica
- [x] CCE-06-Seleccion_Abreviada_Menor_Cuantia
- [x] CCE-16-Servicios_profesionales_gestion

Horizonte Temporal (años)
2017 — 2020

Departamento	# Entidades
Distrito Capital de Bogotá	98
Antioquia	8
Cundinamarca	7
Santander	7
Caldas	4
Casanare	4
No Definido	4
Quindío	3
Tolima	3
Valle del Cauca	3
Meta	2
Arauca	1
Bolívar	1
Putumayo	1
Risaralda	1
Total	**145**

Precio base de

397.02 mill.
Precio Base (prom)

0
Precio Base (min)

30 mil M
Precio Base (max)

Jerarquía productos

32
Familias

44
Clases

67
Productos

Clusters
0 | 1

Rango Precio
- 0-21.000.000
- 21.000.001-43.827.960
- 43.827.961-81804454
- 81804454-2.9550.000.000

ACCESO DE DASHBOARD BI

Alianza CAOBA

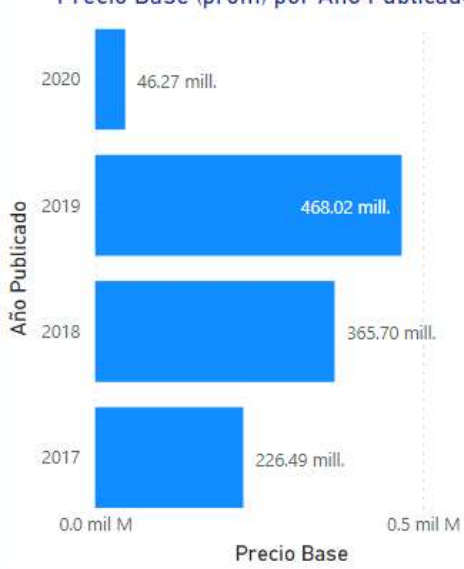
Productos

Productos a contratar

Nombre Entidad	Departamento
ALCALDÍA MUNICIPAL DE COCORNÁ	Antioquia
CONTRALORIA GENERAL DE ANTIOQUIA	Antioquia
CONTRALORIA MUNICIPAL ITAGUI	Antioquia
CORPORACION AUTONOMA REGIONAL DE LAS CUENCAS DE LOS RIOS NEGRO - NARE	Antioquia
DEPARTAMENTO DE ANTIOQUIA	Antioquia
MUNICIPIO DE ENVIGADO 2020	Antioquia
MUNICIPIO DE MEDELLIN	Antioquia
PARQUES NACIONALES NATURALES DE COLOMBIA - DIRECCION TERRITORIAL ANDES OCCIDENTALES	Antioquia
BATALLÓN DE INGENIEROS No 18 'GENERAL RAFAEL NAVAS PARDO'	Arauca
DGSM-DISAN-ARC-HONAC	Bolívar
ASOCIACION AEROPUERTO DEL CAFE	Caldas
CORPORACION AUTONOMA REGIONAL DE CALDAS CORPOCALDAS	Caldas

145
Entidades

Precio Base (prom) por Año Publicado



CONCLUSIONES

- Entre todas las etapas que incluye la analítica, el EDA es fundamental para el entendimiento de los datos y el buen desarrollo de los modelos.
- Escoger LDA es una estrategia ágil que permite hacer agrupaciones dentro del lenguaje natural, es necesario hacer una limpieza progresiva e iteraciones del proceso para lograr un resultado óptimo.
- Los modelos de aprendizaje no supervisado con variables categóricas requieren algoritmos y métricas más especializados para este tipo de datos.
- El análisis de clúster nos brindó posibles segmentos de la base de datos, aunque genera dudas sobre la precisión o resultado óptimo de grupos.