Universidad EAFIT

Maestría en Ciencia de Datos y Analítica

Proyecto Integrador

Alianza Caoba: Modelo Analítico para área comercial

Autores:

Cristian David Rojas Rincon, cdrojasr@eafit.edu.co
Daniela Vasquez Jaramillo, dvasqu18@eafit.edu.co
Yaliza Margarita Barcelo Pulgar, ymbarcelop@eafit.edu.co
Daniel Patiño Barraza, dpatinob@eafit.edu.co
David Rua Jaramillo, druaj@eafit.edu.co

Junio 2021

UNIVERSIDAD EAFIT

Tabla de contenido

1. Introducción

La segmentación nos ayuda a dividir información (clientes, proveedores, empleados) en grupos, de acuerdo con una gama de características diferentes (Gončarovs, 2018). Esta es importante para entender la información desde diferentes ángulos y sus resultados se integran a aplicaciones de inteligencia de negocios para ser empleados en estrategias de mercadeo, de ventas, de promoción, de personalización, etc. El uso de distintas técnicas de agrupación en clústeres nos ayuda a segmentar las bases de datos, de acuerdo con sus particularidades (Gončarovs).

Es por medio del uso de modelos de clasificación y reconocimiento de patrones que podemos comprender la naturaleza de la relación entre las características y el resultado de la clase (Hastie, 2017). Así, los motores de predicción tipo caja negra pueden ser muy efectivos y frecuentemente se desempeñan mejor en problemas con datos reales.

En este proyecto integrador veremos procesos de Análisis Exploratorio de Datos (EDA), de Extracción, Transformación y Carga de Datos (ETL), de Análisis de Colinealidad (MCA), de Análisis Discriminante Lineal (LDA) y de K-Modes, para identificar los posibles segmentos que servirán a Caoba en su proceso de promoción y venta de servicios de Analítica.

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

2. Marco Teórico

Alianza CAOBA solicitó realizar un proyecto de analítica que tiene como objetivo principal identificar las empresas del sector privado que puedan estar interesadas en adquirir proyectos de analítica, para cumplir con este objetivo suministraron como fuente de información el Plan Anual de Adquisiciones 2020 Secop II, luego se realizó un análisis de exploración de datos que permitió evidenciar las necesidades de los modelos a implementar.

Entre las variables más relevantes se identificó la necesidad de realizar procesamiento de lenguaje natural, aplicar text mining y modelo LDA, este resultado sería una base para aplicar los otros modelos descritos (El método de clustering de k-modes y MCA).

El procesamiento de lenguaje natural (PLN) es un campo de las ciencias de la computación ampliamente estudiado en la actualidad, ya que aborda las interacciones entre las computadoras y el lenguaje humano. Para el desarrollo de este proyecto fue necesario hacer una revision de distintas técnicas de PLN para modelación de lenguaje, como Bag of Words (BOW) que modela documentos como un vector de números binarios y Latent Dirichlet Allocation (LDA) que modela tópicos en documentos.

(Blei, 2003) Es un modelo estadístico generativo que permite que conjuntos de observaciones sean explicados por grupos no observados, que justifican el hecho de que algunas partes de los datos son similares. Si las observaciones son palabras recogidas de documentos, se postula que cada documento es una mezcla de cierta cantidad de temas (tópicos) y que la presencia de cada palabra se puede atribuir a uno de estos tópicos del documento.

LDAvis permite una inspección profunda de las relaciones entre el topic-term en un LDA, al mismo tiempo que proporciona una visión global de los temas, a través de sus prevalencias y similitudes entre sí, en un espacio compacto.

En este análisis de aprendizaje no supervisado es necesario salir de las fronteras conocidas de datos continuos, pues a menudo los datos recopilados están en grupos categorizados no ordinarios, y es por esto por lo que a diferencia de la prueba de hipótesis tradicional diseñada para verificar hipótesis a priori sobre relaciones entre variables, el análisis de datos exploratorio se utiliza para identificar relaciones sistemáticas entre variables, cuando hay expectativas a priori incompletas en cuanto a la naturaleza de esas relaciones. (Soares, 2013) El análisis de correspondencia de métodos (CA), una técnica analítica de datos descriptivos (multivariante), permite simplificar datos complejos y proporciona una descripción detallada de los datos,

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

produciendo un análisis simple pero exhaustivo. Específicamente, CA múltiple (MCA) permite el análisis de variables categóricas o categorizadas que abarcan más de dos variables categóricas.

La segmentación proporciona una vista basada en datos para examinar los grupos significativos que permitan tomar acciones específicas y mejorar los resultados empresariales (Dutta, 2020). Muchas empresas corren el riesgo de tomar decisiones basadas en generalizaciones, ya que utilizan un enfoque único para evaluar su entorno empresarial. La segmentación mejora la toma de decisiones al proporcionar múltiples puntos de vista para separar los datos y tomar medidas.

El método de clustering de k-modes ha llamado mucho la atención, ya que trabaja bien con datos categóricos. Sin embargo, su desempeño es especialmente sensible a la selección inicial de centroides (Jiang, 2015). En general, el algoritmo de k-modes es más rápido que el de k-means, ya que necesita menos iteraciones para converger. La base de datos empleada tiene muchas variables categóricas, por lo cual, utilizaremos este algoritmo.

En el contexto del aprendizaje no supervisado, no existe una medida directa del éxito. Es difícil determinar la validez de las inferencias extraídas de la mayoría de los algoritmos de aprendizaje no supervisado. Se debe recurrir a argumentos heurísticos no sólo para motivar los algoritmos, sino también para juicios en cuanto a la calidad de los resultados (Hastie).

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

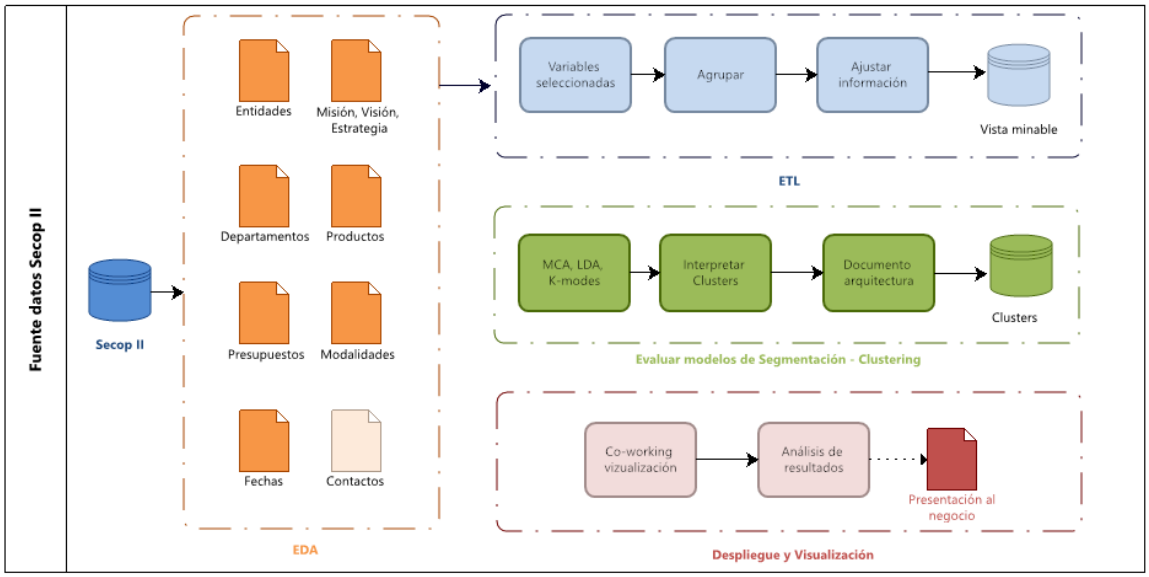
EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

3. Desarrollo Metodológico

3.1 Entendimiento del problema, pregunta de negocio o hipótesis

El problema planteado implica generar estrategias desde la ciencia de datos que nos permitan identificar las empresas del sector público que puedan estar interesadas en servicios de analítica, los cuales Caoba ofrece.

El siguiente flujograma muestra el proceso a desarrollar, desde la ingesta de datos hasta la presentación al negocio; este abarca aspectos como el EDA, ETL, evaluación de modelos y despliegue y visualización.



3.2 Análisis Exploratorio de Datos

La base de datos utilizada corresponde al Plan Anual de Adquisiciones 2020 Secop II, que se encuentra en la página web de Datos Abiertos (www.datos.gov.co), es una base de acceso público y gratis. Las Entidades Estatales crean, evalúan y adjudican Procesos de Contratación. Los Proveedores pueden hacer comentarios a los documentos del proceso, presentar ofertas y seguir el proceso de selección en línea.

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 (www.colombiacompra.gov.co).

Dimensión: 863.969 x 29

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

3.3 Selección de modelos, Ingeniería de Características, Entrenamiento, Evaluación

Modelos, Entrenamiento y Evaluación

Partiendo del objetivo principal del análisis realizado, en el cual se debe entregar un listado de empresas que estén dispuestas a adquirir proyectos de analítica, y seguido al análisis exploratorio de datos, se pudo evidenciar que la información requiere de modelos no supervisados.

De acuerdo con las variables seleccionadas se pretende encontrar patrones que nos permitan hacer agrupaciones de aquellas empresas que de acuerdo con la información de la fuente se puedan segmentar en las que estarán dispuestas a adquirir proyectos de tecnología.

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

Por tanto, se seleccionaron los siguientes modelos:

Modelo Heurístico:

Lo heurístico se describe como el arte del descubrimiento y la invención, de aquí nace el nombre de este modelo. Inicialmente y de forma mas intuitiva cogimos nuestra fuente de datos en su vista minable, la revisamos en la versión .CSV y para un análisis de esta generamos una tabla dinámica.

Desde la concepción misma del proyecto sabíamos que había unas temáticas de nuestro interés que nos importaba revisar y es por medio de estas que creamos un diccionario con las palabras: tecnología, informática, data, datos, analítica, analizador, dato y predicción. Eran entonces los servicios que ofreciesen alguna de estas palabras claves los que nos entregarían información directa y acertada de lo que se buscaba, por tal motivo se filtró en nombre_producto todos los registros que éstas incluyeran.

Al reconocer los cod_nombres correspondientes a este filtro, identificamos paralelamente los cod_familia y cod_clase para analizar las familias y clases a los cuales estos productos hacían referencia.

Así corrimos un modelo donde analizábamos la influencia de los mismos en la cantidad de entidades que había, y encontramos las siguientes diferencias según su respectiva segmentación:

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	

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

Como se puede observar, de las familias relacionadas con el filtro hay 397 entidades que se ven asociadas a la búsqueda, correspondiente a 701 productos diferentes. Sin embargo, solo el 9.6% de esos productos son relevantes para el modelo a estudiar y por tanto el filtrado por productos permite una segmentación de entidades del 36.5% de las entidades totales presentadas en el filtrado por familias, generando así una segmentación final de 145 entidades.

Modelo MCA (Análisis de Correspondencia Múltiple):

El análisis de correspondencia múltiple es un método utilizado para el análisis de correspondencia entre variables cualitativas, homólogo al análisis que solemos hacer con PCA (Análisis de Componentes Principales) cuando tenemos variables numéricas.

Ahora que tenemos una segmentación por nombre_producto, identificamos que era relevante analizar la correlación entre las variables que componían la fuente de datos que incluían las entidades relevantes a revisar, y por eso hicimos una elección de variables prominentes según el análisis exploratorio previo.

Dividimos este análisis en: análisis univariante, bivariante, ANOVAS, dependencias y finalmente según los registros previos identificaríamos las variables a proceder para el CMA.

En los análisis univariantes y bivariantes tenemos un entendimiento mejor de los datos bajo esta vista minable que tenemos, aquí hacemos un análisis inicial de cuáles son los registros finales, sus valores únicos, el tipo de variables que manejamos y una descripción de estos:

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

Información:

```
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
```

Descripción:

	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

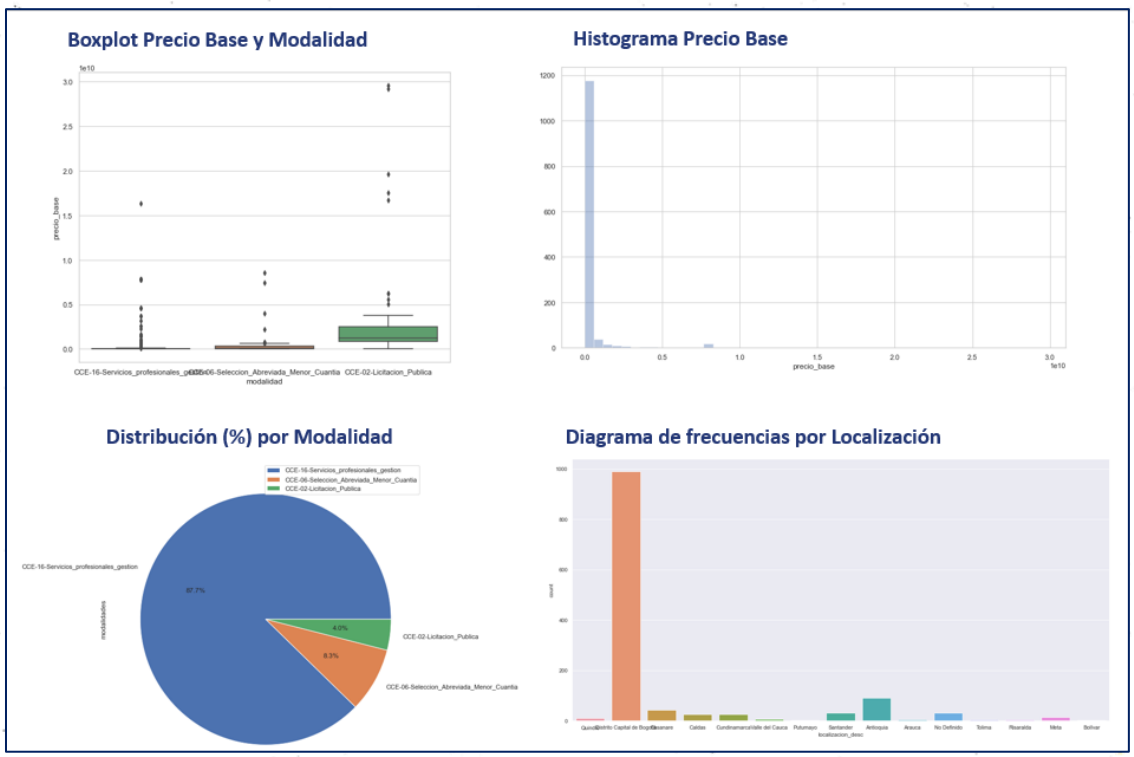
Adicionalmente, desde el análisis de vista minable previamente realizado habíamos identificado la variable “Precio Base” como aquella que bajo este modelo no-supervisado nos podría dar una dirección a cuánto tienen permitido estas entidades invertir en proyectos relacionados con analítica. Por lo tanto, es a partir de esta donde hacemos el estudio de sus medidas de tendencia central:

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

Según esto podemos tener un indicio inicial que los contratos promedios tienen un valor de 397 millones de inversión, sin embargo, con el análisis de la mediana y la moda confirmamos que el 50% de los contratos se centran en 43.8 millones que coincide con el valor más frecuente. La diferencia entre media y mediana se ve claramente explicada

cuando al hacer un análisis de la variación estándar identificamos un valor de 1.800 millones, dato considerable para la variable a analizar.

Igualmente, el análisis multivariante lo hacemos entre múltiples variables por medio de análisis graficas que nos dan información del comportamiento de las variables entre sí:



En el tercer procedimiento de este análisis de correlación entre variables aplicamos Análisis de Varianza (ANOVA), el cuál es un método estadístico utilizado para evaluar si existen diferencias significativas entre las medias de dos o más grupos. ANOVA calcula la desviación de las medias entre los grupos y genera una puntuación F, posteriormente calcula el valor P para identificar que tan estadísticamente significativo es el valor.

Por medio de Statsmodel e importando OLS podemos hacer un modelo de análisis que se ajuste a ANOVA y calcularemos en este caso el Precio Base como variable de interés con familia, localización, proyectado y modalidad:

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

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

Según este análisis el significativo F es mayor que 1 en Familia, Mes proyectado y Modalidad con un valor P cercano al 0, por lo tanto, se rechaza la hipótesis de igualdad de medias entre las variables y se concluye que hay diferencias entre las medias de estos valores.

Finalmente se hace un análisis de tablas de contingencia que permiten evaluar la relación entre dos variables. Esta se hace con relación a la variable producto con modalidad, localización y mes proyectado, donde arroja que ninguna de las variables tiene asociación alguna y por tanto son independientes.

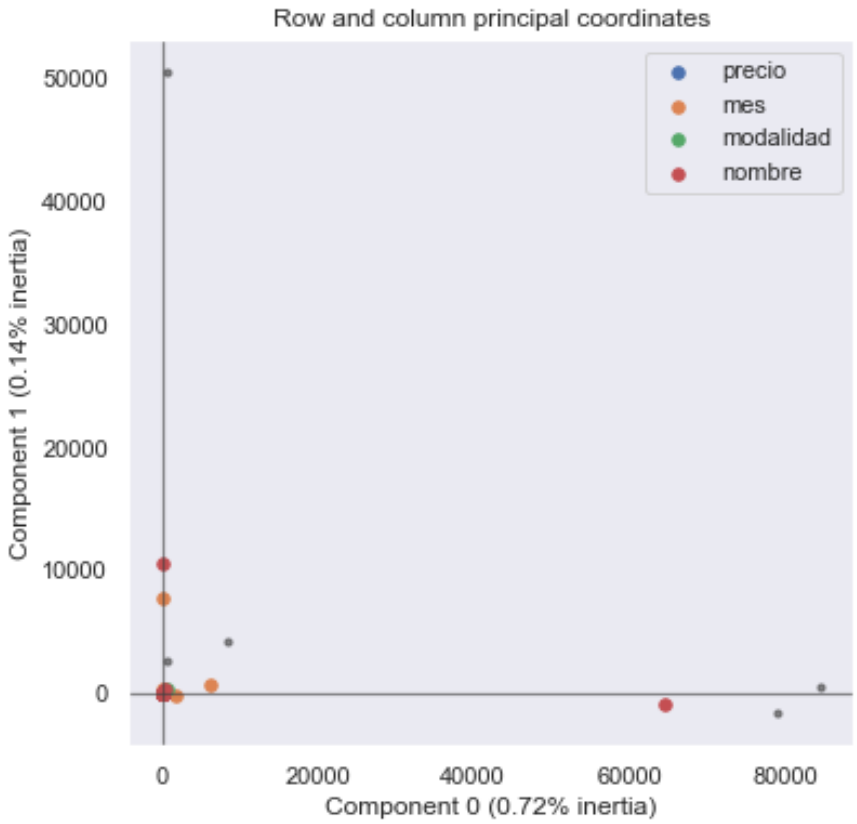
modalidad nombre_producto	CCE-02-Licitacion_Publica	CCE-06-Seleccion_Abreviada_Menor_Cuantia	CCE-16-Servicios_profesionales_gestion	All
Administradores permanentes de bases de datos o de sistemas de tecnologías de la información	3	0	14	17
Administradores temporales de bases de datos o de sistemas de 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

Finalmente, por medio de la librería Prince aplicamos MCA en el modelo, y según el análisis previo de todas las variables por lo diferentes métodos utilizamos esta vez en

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

conjunto las variables categóricas mes proyectado, precio base, modalidad y familia. El resultado es:

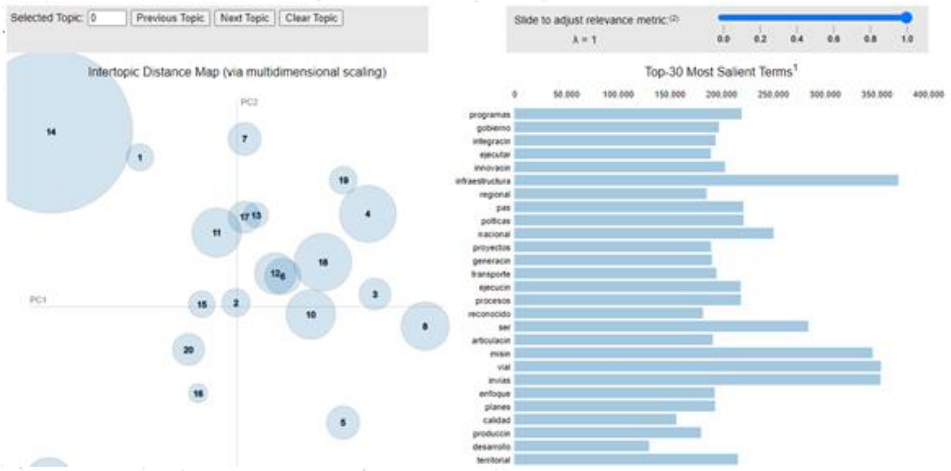


Analizando los valores de este podemos identificar que las asociaciones de estas 4 variables no permiten describir la variabilidad total por una de ellas, ni ayudan en la predicción complementaria entre sí.

Modelo LDA (Latent Dirichlet Allocation) :

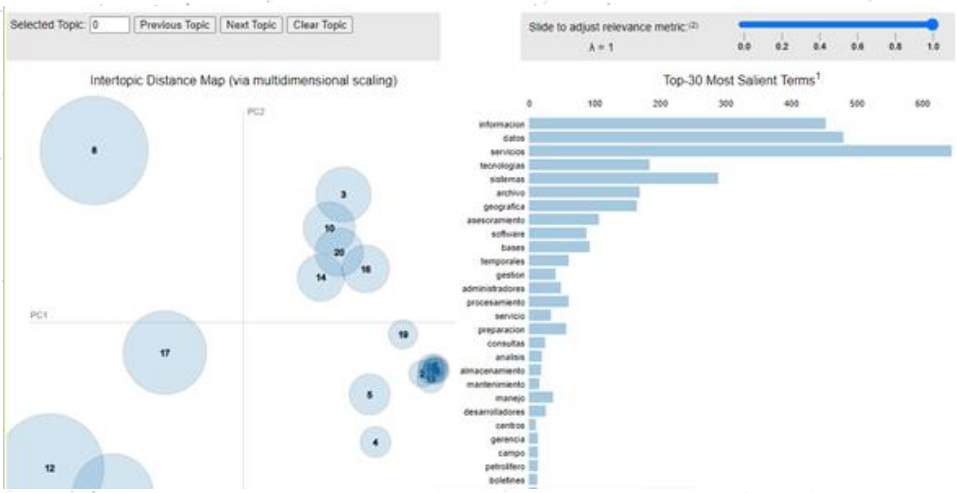
Se identificaron variables con información relevante para el objetivo del proyecto, las cuales requieren de un proceso de text mining, con la finalidad de simplificar la manipulación e interpretación de los datos, esto reduce la cantidad de esfuerzo a emplear para que nuestro análisis de datos sea eficiente, consistente y mejore la capacidad de predicción.

Seguido se aplicó el modelo de LDA, el cual, mediante la identificación de ciertos términos de interés en la búsqueda generó un grupo de clústeres, analizamos el resultado de cada uno de los clústeres e identificamos el clúster que contiene las palabras que presentan mayor relevancia de acuerdo con lo buscado, para de esa forma poder identificar las empresas relacionadas con este, y con base en el clúster elegido se hizo el filtrado de 10 palabras claves.



En este resultado se puede evidenciar que el resultado general presentaba fallas en algunas palabras como errores al eliminar tildes,

Seguido a este resultado se generó un nuevo dataframe que se utiliza como base para el modelo de clústeres.



Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

Modelo de Clusters (K-Modes):

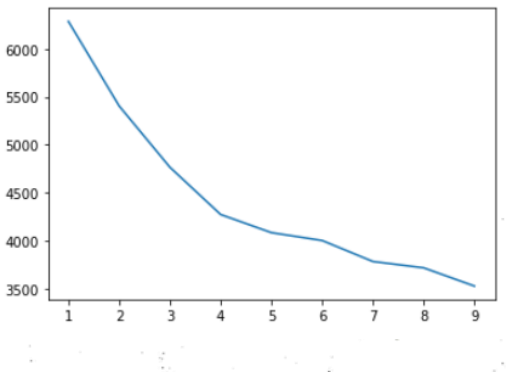
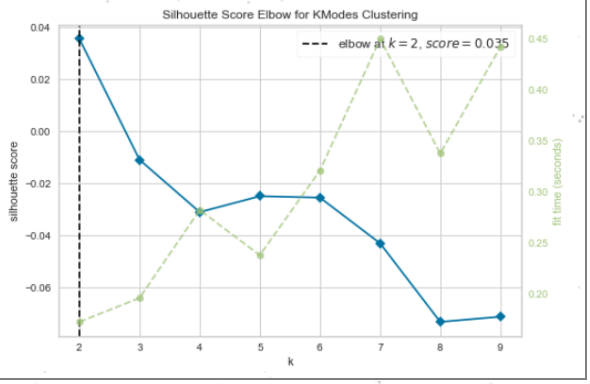
Luego de generar el nuevo dataframe se puede realizar algunos clústeres para diferentes posibilidades de análisis de los datos. Estas diferentes opciones de agrupamiento, ayuda al cliente a identificar las empresas según ciertas características, por ejemplo: agrupar las empresas por las palabras claves en temas de analítica, y luego agruparlas por ciudad, y luego analizar los montos presupuestales que tienen asignado, esto le dará a la empresa una visión más clara de los clientes a los que desea llegar.

En aprendizaje no supervisado existen varios métodos para seleccionar el número de clusters óptimo, entre los que se encuentran: método de Codo, coeficiente de Silueta, índice de Calinski-Harabasz, índice de Davies-Bouldin, Dendrograma y criterio de información bayesiano (Dutta). Empleamos el método de Codo y el coeficiente de Silueta para evaluar el *k* óptimo.

El *método de Codo* se basa en calcular la suma de errores cuadrados dentro del grupo (WSS) para diferentes números de grupos (k) y seleccionar el k para el cual el cambio en WSS primero comienza a disminuir.

El *coeficiente de silueta* nos dice si los puntos individuales están asignados correctamente a sus grupos. Podemos usar las siguientes reglas de pulgar mientras usamos el coeficiente de silueta: cerca de 0 significa que el punto está entre dos grupos; cerca de -1, sería mejor asignarlo a los otros grupos; cerca de 1, entonces el punto pertenece al cluster "correcto".

Al ejecutar el código con las variables categóricas de la base de datos depurada, obtuvimos los siguientes resultados:

Método de Codo	Coeficiente de silueta
	

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

Al buscar un cambio de pendiente, de empinada a poca profundidad (elbow), para determinar el número óptimo de clústeres, encontramos el primero en K=2, el segundo en k=3 y el tercero en k=4 .	Observamos que el valor positivo más alejado de cero se da en K=2, aunque los valores son cercanos a cero y se presenta ambigüedad en el k óptimo.

Características e Ingeniería de Características:

Con el objetivo de incrementar la eficacia de los modelos a aplicar, durante el proceso de modelado identificamos algunas variables que aplicando una característica mejora considerablemente el resultado esperado.

Cabe destacar que es muy importante el conocimiento previo de la data que se está analizando para aplicar ingeniería de características, si se desea tener un resultado óptimo, ya que la función de aplicar característica es reducir la dimensionalidad y mejorar el aprendizaje del modelo.

Por tanto, al identificar las variables importantes en el modelado se aplicaron características que mejoraron la extracción y clasificación de los datos relevantes.

3.4 Análisis y Conclusiones

El método de MCA con sus análisis previos de contingencias y varianza podemos concluir que las variables manejadas no presentan dependencia o correlaciones directas entre ellas, son variables independientes que no se explican entre sí ni permiten la reducción de la dimensionalidad por este medio.

Al ejecutar el método del codo en el algoritmo de k-modes, pudimos ver que el cambio importante en la pendiente se da en K=2, 3, 4, por lo cual acudimos al segundo método (coeficiente de silueta), presentando el mejor resultado en k=2. Tomaremos este valor como referente para segmentar los datos, aunque con las depuraciones de los métodos anteriores nos brindan una vista minable que se puede entregar a Caoba como resultado final.

4. Tecnología: Ingeniería de Datos y uso de Tecnología

4.1 Desarrollo del proyecto

Como herramienta principal para satisfacer la necesidad de analítica y presentación de datos, escogimos Jupyter Notebook, debido a que facilita la visualización de los datos y se integra perfectamente al lenguaje de programación utilizado en nuestro proyecto (Python); de igual manera permite utilizar otros tipos de lenguajes dada su versatilidad.

Una vez se obtuvieron los datos a analizar desde el api de Socrata Open Data API (SODA) a través del siguiente enlace, https://www.datos.gov.co/resource/edfp-tdwk.csv obteniendo una fuente de datos semi-estructurada y realizando una ingesta de datos por lotes (batch), los cuales fueron almacenados en un archivo tipo CSV para su posterior análisis.

Luego realizamos el análisis exploratorio de datos e identificamos una serie de librerías requeridas para realizar el procesamiento de los datos, según lo requerido en este proyecto.

4.2 Despliegue del proyecto

El despliegue del proyecto se realizó en Microsoft Power BI, actualmente se cuenta con una visualización, y al momento de realizar la entrega al cliente se realizará en una publicación en un link de acceso web, o en un sitio de escogencia de CAOBA.

4.3 Fuentes de datos, Ingesta de datos y almacenamiento

Requerimientos Software

- Microsoft Windows
- Anaconda Individual Edition
- Python
- Jupyter NoteBook

Librerías Empleadas

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Datetime

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

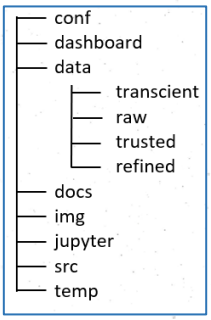
- Pickle
- Sklearn
- Scipy
- Smart-open
- Gensim
- Nltk
- PyLDAvis
- Prince

Requerimientos Hardware

Para un óptimo procesamiento de datos y para una buena ejecución del análisis de modelos de clúster se recomienda un equipo con las siguientes características:

- Versión de Windows tipo datacenter.
- Memoria RAM superior a 16Gb
- Procesador tipo Opteron, Xeon o Athlon con frecuencia mínima de 2GHz.
- Disco duro estado sólido con capacidad para albergar la fuente de datos y las aplicaciones requeridas.

En adición se realizó un control de versiones con GitHub para todos los archivos utilizados durante la realización del proyecto, conservando esta estructura:



Utilizando las mejores prácticas, implementamos las zonas para realizar una separación lógica de los datos para tener un entorno seguro, ágil y organizado.

Nombre	Último cambio	Última actualización
..		
01 - Trascient	chore: Agregando avances de Fuente1 hasta ...	4 days ago
02 - Raw	chore: actualización experimento 1	1 day ago
03 - Trusted	Chore: Actualización vista minable 3 cdrojas	1 day ago
04 - Refined	chore: ajuste Visualizacion a Trusted y texto p...	1 week ago

4.4 Ambiente de procesamiento

El ambiente de procesamiento se realizó con Notebook de Jupyter bajo el ambiente de Python y Microsoft Power BI el cual trabaja con Power Query

4.5 Aplicaciones

Para el despliegue del proyecto y la visualización del resultado de la analítica de datos aplicada a la fuente suministrada, se podría plantear un escenario en el que dichos resultados se almacenen en una base de datos transaccional para mejorar los tiempos de respuesta en el acceso y presentación de los datos. Como herramienta para visualización y presentación de datos, se puede integrar perfectamente Microsoft Power BI a esta fuente transaccional, para mostrar los resultados de manera dinámica, rápida y gráfica.

En el contexto del proyecto se utilizó como fuente de datos el archivo ".pickle" (archivo serializado), para poder realizar la analítica y visualización de este.

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

5. Conclusiones

El proceso de EDA es muy importante en un proyecto de analítica, ya que nos permite entender cada variable, los tipos de datos, los rangos, algunos valores atípicos y realizar análisis heurístico, entre otros. Uno de los principales hallazgos del grupo es que el tiempo invertido en el EDA fue mayor al esperado y con esto logramos entender mejor la información y disminuir reprocesos en etapas posteriores.

Para el análisis de texto escoger LDA es una alternativa ágil que nos permite hacer agrupaciones dentro del lenguaje natural, segmentando tópicos sin necesidad de filtrados manuales, que se entrena progresivamente en torno a las palabras claves y clústeres naturales de texto. Sin embargo, es necesario hacer una limpieza progresiva e iteraciones del proceso para llegar a un resultado óptimo como modelo base para otros análisis.

Adicionalmente, es importante una vez realizado el EDA entender el comportamiento y relacionamiento real de los datos. Los análisis de correlación, univariantes y bivariantes aumentan el espectro de entendimiento de los datos y permiten priorizar variables. Sin embargo, es evidente que el aprendizaje no-supervisado con la mayoría de las variables categóricas comprometen un esfuerzo mayor, la interpretación no numérica de los mismos obligan a métricas y modelos que han sido menos tratados y que aún deben perfeccionarse.

El análisis de clúster con información categórica nos brindó perspectivas de posibles segmentos en la base de datos, aunque como fue previamente mencionado, al tratarse de aprendizaje no supervisado nos genera dudas de la precisión o resultado óptimo para cada grupo.

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115

6. Referencias

Cheat sheet for implementing 7 methods for selecting the optimal number of clusters in Python; Dutta, Indraneel; 2020; https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad.

Initialization of K-modes clustering using outlier detection techniques; Jiang, Feng; 2015; Qingdao University of Science and Technology.

The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables in Healthy Ageing; Patrício Soares Costa, Nadine Correia Santos, Pedro Cunha, Jorge Cotter, Nuno Sousa; Journal of Aging Research, vol. 2013, Article ID 302163, 12 pages, 2013. https://doi.org/10.1155/2013/302163

The Journal of Machine Learning Research; Blei David, Ng Andrew, Jordan Michael; 2003; p. 993-1022.

Using Data Analytics for Customers Segmentation: Experimental Study at a Financial Institution; Gončarovs, Pāvels; 2018; Riga Technical University.

Unsupervised Learning. The Elements of Statistical Learning; Hastie Trevor, Tibshirani Robert, Friedman Jerome; Springer; 2017; P 485-585.

Universidad EAFIT-Campus principal
Carrera 49 7 Sur 50, avenida Las Vegas
Medellín-Colombia
Teléfonos: (57) (4) 2619500-4489500
Apartado Aéreo: 3300 I Fax: 3120649
Nit: 890.901.389-5

EAFIT Llanogrande
Teléfono: (57) (4) 2619500 exts. 9562-9188
EAFIT Bogotá
Teléfonos: (57) (1) 6114523-6114618
EAFIT Pereira
Teléfono: (57) (6) 3214115