Septiembre 2022

Universidad de Los Andes

Jesús David Barrios 201921887

Sergio Peñuela 201922873

Jhoan Diaz 201819861

Laboratorio 2 Inteligencia de Negocios

1. Descarga y entendimiento de datos

   Se descargaron los datos entregados. Inicialmente se obtuvo información básica de los datos. A continuación, se presentan tablas e imágenes de información básica de los datos.

```
Número de filas:  1068
Número de columnas:  20
Número de columnas y sus tipos de datos:

No.                                  int64
NIT                                  int64
RAZON SOCIAL                        object
SUPERVISOR                          object
REGIÓN                              object
DEPARTAMENTO DOMICILIO              object
CIUDAD DOMICILIO                    object
CIIU                                object
MACROSECTOR                         object
INGRESOS OPERACIONALES\r\n2018*     object
GANANCIA (PERDIDA) 2018             object
TOTAL ACTIVOS 2018                  object
TOTAL PASIVOS 2018                  object
TOTAL PATRIMONIO 2018               object
INGRESOS OPERACIONALES\r\n2017*     object
GANANCIA (PERDIDA) 2017            float64
TOTAL ACTIVOS 2017                  object
TOTAL PASIVOS 2017                  object
TOTAL PATRIMONIO 2017               object
GRUPO EN NIIF                       object
dtype: object
```

Ejemplo de los datos con algunas de sus columnas:

	No.	NIT	RAZON SOCIAL	SUPERVISOR	REGIÓN	DEPARTAMENTO DOMICILIO	CIUDAD DOMICILIO	CIIU	MACROSECTOR	INGRESOS OPERACIONALES\r\n2018*	GANANCIA (PERDIDA) 2018	TOTAL ACTIVOS 2018	TOTAL PASIVOS 2018
318	319	800052534	DISTRIBUCIONES AXA SAS	SUPERSOCIEDADES	Bogotá - Cundinamarca	BOGOTA D.C.	BOGOTA-D.C.-BOGOTA D.C.	G4645 - Comercio al por mayor de productos far...	COMERCIO	2431623100	60163750	787302360	564013370
1063	1064	900292211	REFINADORA NACIONAL DE ACEITES Y GRASAS SAS	SUPERSOCIEDADES	Costa Pacífica	VALLE	CANDELARIA-VALLE	C1030 - Elaboración de aceites y grasas de ori...	MANUFACTURA	1386316790	22151990	501416690	274589220
1065	1066	860058831	avicola los cambulos sa.	SUPERSOCIEDADES	Bogotá - Cundinamarca	BOGOTA D.C.	BOGOTA-D.C.-BOGOTA D.C.	A0145 - Cría de aves de corral	AGROPECUARIO	2485960390	18906280	780480300	505029030
614	615	830131993	EFECTIVO LTDA	SUPERSOCIEDADES	Bogotá - Cundinamarca	BOGOTA D.C.	BOGOTA-D.C.-BOGOTA D.C.	H5229 - Otras actividades complementarias al t...	SERVICIOS	5200510120	483527130	3293350180	1985023190
652	653	900818642	CONSTRUCCIONES COLOMBIANAS OHL SAS	SUPERSOCIEDADES	Bogotá - Cundinamarca	BOGOTA D.C.	BOGOTA-D.C.-BOGOTA D.C.	F4210 - Construcción de carreteras y vías de f...	CONSTRUCCIÓN	1448136230	-893997020	1996355320	1986833700

Dimensiones de calidad

- Completitud: Se observa el número de valore nulos en los datos

```
Número de filas con valores nulos:  30
Número de columnas con valores nulos:  13
Lista de columnas con valores núlos y sus tipos:
 REGIÓN                              True
DEPARTAMENTO DOMICILIO              True
CIUDAD DOMICILIO                    True
CIIU                                True
MACROSECTOR                         True
INGRESOS OPERACIONALES\r\n2018*     True
GANANCIA (PERDIDA) 2018             True
TOTAL ACTIVOS 2018                  True
TOTAL PASIVOS 2018                  True
TOTAL PATRIMONIO 2018               True
INGRESOS OPERACIONALES\r\n2017*     True
GANANCIA (PERDIDA) 2017             True
TOTAL ACTIVOS 2017                  True
dtype: bool

Porcentaje de completitud de las columnas: 97.19%
```

Dado que hay datos con valores nulos, estos se manejan posteriormente para asegurar un 100% de completitud.

- Unicidad: Se ven el número de valores repetidos en la base

```
Número de filas con indice duplicado:  68
```

Se observa que hay 68 empresas que se repiten en la base, por lo que se eliminan posteriormente los duplicados.

- Consistencia y validez:

La consistencia hace referencia a la integridad de datos entre fuentes y observaciones. Por su parte, la validez mide si los datos hacen sentido para el contexto específico. En este sentido, se considera que en términos generales se cumple con estas métricas. Existen columnas que deberían ser numéricas, pero inicialmente no lo son, por lo que en el procesamiento de datos esto se debe manejar. También se encuentran ciertos valores que se consideran fueron por error de tabulación, por lo que se deben corregir.

2. Preparación de los datos

Todo el procedimiento se puede ver en el notebook titulado “entendimiento_limpieza.ipynb”.

Inicialmente se eliminaron las columnas “Razon social” y “No.” pues no aportan información relevante a lo que se busca dado que eran identificadores únicos de cada empresa. En la columna “Supervisor” se obtuvo el siguiente conteo de los valores dentro de la columna:

```
SUPERSOCIEDADES    892
SUPERSALUD          42
SUPERFINANCIERA     38
SUPERSERVICIOS      16
SUPERVIGILANCIA      8
SUPERSUCIEDADES      4
Name: SUPERVISOR, dtype: int64
```

Como se puede ver hay un valor con el nombre “SUPERSUCIEDADES”, se asume que es un error en la escritura y se reemplazaron esos valores por “SUPERSOCIEDADES”. Después, con respecto a la columna “Región” se obtuvieron los siguientes valores:

```
Bogotá - Cundinamarca    555
Antioquia                161
Costa Pacífica           126
Costa Atlántica           94
Centro - Oriente          30
Eje Cafetero              22
Otros                      7
Costa Atlantica            4
NaN                        1
Name: REGIÓN, dtype: int64
```

Al haber un valor nulo, para no eliminarlo se decidió reemplazar su valor con la moda de los datos. De igual manera se puede ver que hay “Costa Atlántica” y “Costa Atlantica” por lo que se utilizará el nombre sin tilde. Después los valores “Eje Cafetero” y “Centro-Oriente” al ser pocos se agruparon en una misma categoria “Otros”. Por último, se utilizó One Hot Encoder para pasar los valores categóricos de las regiones a numéricos indicando a que región pertenecen.

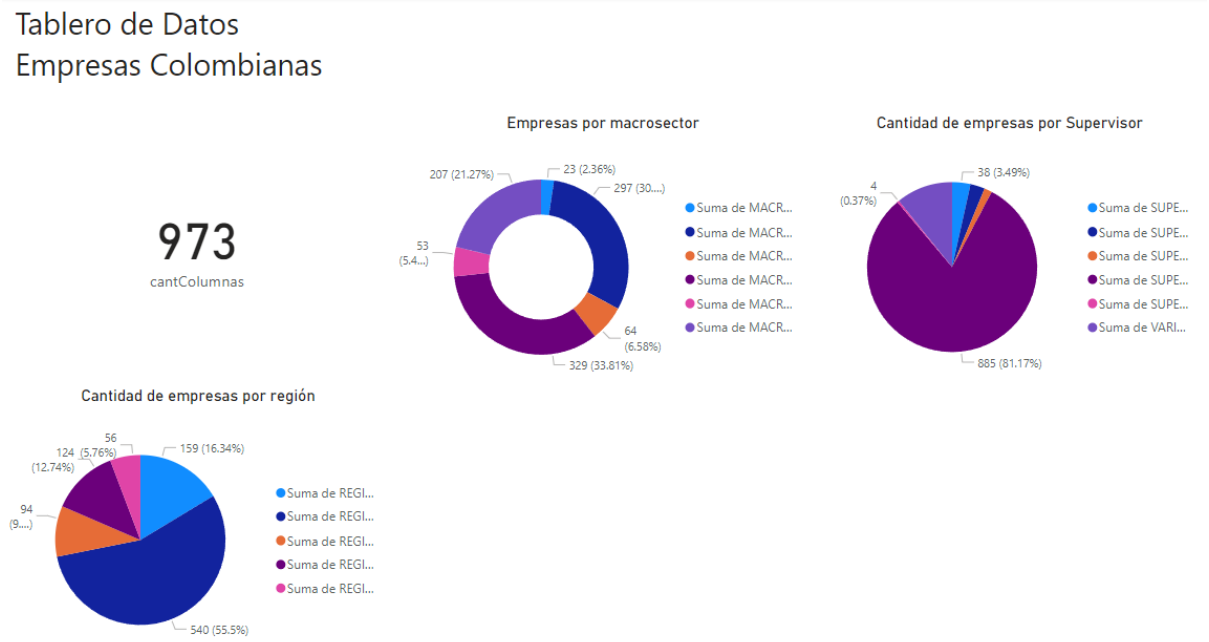
Las columnas “Departamento Domicilio” y “Ciudad Domicilio” se decidieron eliminar al tener la columna “Región” que nos indica a donde pertenece la empresa. Con esto, se busca evitar la redundancia entre los datos. La columna “CIIU” se eliminó debido a la redundancia entre esta columna y la columna “Macrosector”. Para esta última se obtuvieron los siguientes valores:

```
MANUFACTURA             328
COMERCIO                301
SERVICIOS               225
CONSTRUCCIÓN             63
MINERO-HIDROCARBUROS     54
AGROPECUARIO             23
CONSTRUCCION              3
NaN                       3
Name: MACROSECTOR, dtype: int64
```

Para los datos nulos se decidió cambiarlos por la moda y se utilizó el valor “Construccion” en lugar de “Construcción”. En el caso de Total Activos, Total Pasivos, y Total Patrimonio, debido

a que Total Activos es igual a la suma de pasivos y patrimonio se decidió eliminar los último dos.

Finalmente, en la columna “Grupo en NIIF” se decidió eliminar los datos de “Régimen R 414 de 2014 – GCN” por solo tener 8 datos y ser considerados outliers.

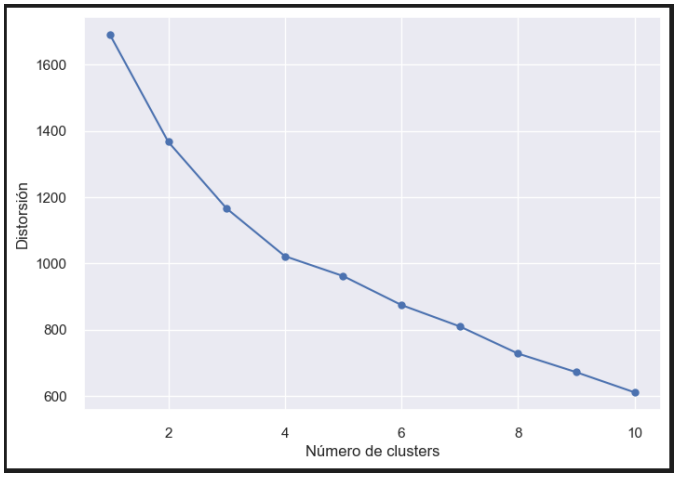


3. Modelamiento, validación y visualización

Para este caso se utilizaron 3 algoritmos de clustering diferentes. Se utilizó K-means, DBscan y GaussianMixture Cada uno utilizó el mismo conjunto de datos.

K-Means

El modelo de K-means fue implementado por Sergio Peñuela. Este método funciona ubicando una k cantidad de centroides de manera aleatoria, donde cada centroide equivale a un clúster diferente, y a partir de esos clústeres se clasifican los datos dependiendo del clúster al que estén más cerca del centroide. Un dato va a hacer parte del clúster en el cual esté más cerca del centroide, este proceso se repitió 10 veces o hasta obtener un modelado óptimo. El único hiperparámetro que tiene este modelo es la cantidad de clústeres k que se utilizarán. Para obtener el número de clústeres se utilizó el método del codo donde se fue aumentando la cantidad de clúster desde 1 hasta 11 y se comparó con el nivel de distorsión de los datos y se obtuvo la siguiente gráfica:



Como se puede ver no hay un punto claro en el cual la reducción de la distorsión se ralentice de manera evidente, pero a partir de 4 clústeres su reducción es menor entre mayor cantidad de clústeres por lo que se utilizarán 4 clústeres para el modelo. Después de esto, debido a la diferencia en la magnitud de algunos valores de los datos se procedió a estandarizarlos para poder tener una mejor comparación entre los diferentes valores y que la función de distancia no se viera afectada.

Finalmente, a partir de los 4 clústeres armados se utilizó el coeficiente de la silueta para ver que tanto los datos pertenecían al clúster en el cual fueron clasificados. Este coeficiente varía entre -1 y 1. Siendo -1 un valor que indica que los datos no están bien emparejados o no tienen relación alguna con el clúster al cual fueron asignados, y 1 es un valor que indica que el dato sí tiene relación con el clúster en el que fue asignado. En este caso el valor del coeficiente obtenido fue de 0.408.

DBScan

El modelo de DBScan fue implementado por Jesús David Barrios. El agrupamiento espacial basado en densidad de aplicaciones con ruido, conocido como DBScan, es un método de aprendizaje no supervisado para clustering basado en la distancia entre los puntos más cercanos. En algoritmos como K-means dado que los grupos dependen del valor medio de los datos, cada dato desempeña un papel en la formación de los clusters. Un ligero cambio en los datos de datos podría afectar el resultado de la agrupación (Singh Chauhan, 2022). Por ello, se decidió utilizar un algoritmo basado en densidad como DBScan, el cual reduce este riesgo basándose principalmente en la densidad de puntos. Además, con este método no se tiene la necesidad de especificar el número de clusters, por lo que se reduce el sesgo en ese sentido.

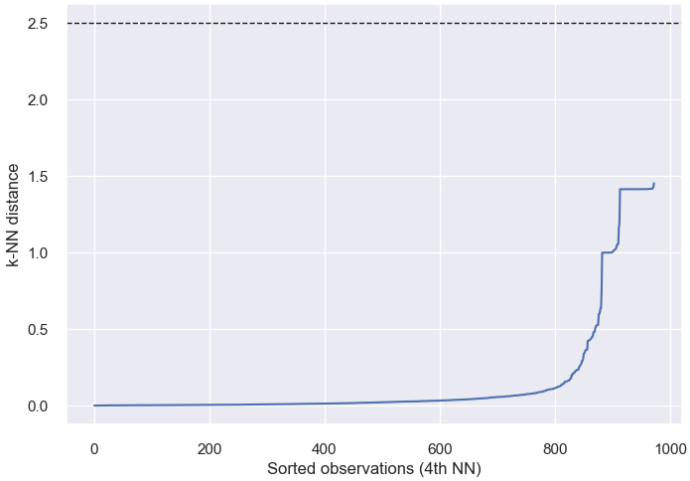
El DBScan tiene dos hiperparametros principales: (1) el mínimo número de puntos para que un grupo sea considerado una región densa (cluster); y (2) el parámetro épsilon, el cual hace referencia a una medida de distancia que se utiliza para ubicar los puntos en la vecindad de cualquier punto (Singh Chauhan, 2022). Esta metodología, en términos generales está compuesta por los siguientes pasos:

1) Escoger puntos centrales aleatorios para buscar clusters
2) Identificar puntos cercanos teniendo en cuenta los parámetros establecidos
3) Si se cumple con las características crear el cluster, de lo contrario se sigue con el siguiente punto.
4) Se repite el proceso hasta que todos los puntos son visitados

(Bedre, 2022)

En primer lugar, para implementarlo, se procedió a estandarizar los datos. Esto, dadas las altas diferencias en magnitud entre las columnas. Luego, se procedió a determinar los parámetros para llevar a cabo el proceso de clustering. Para el mínimo número de puntos, se suele tener un valor mayor o igual al número de dimensiones que se tienen en los datos (Bedre, 2022). En este caso, en los datos procesados se tienen 19 dimensiones. Luego de iterar, se define un valor de 19 para minPts dado que es el que da mejores resultados.

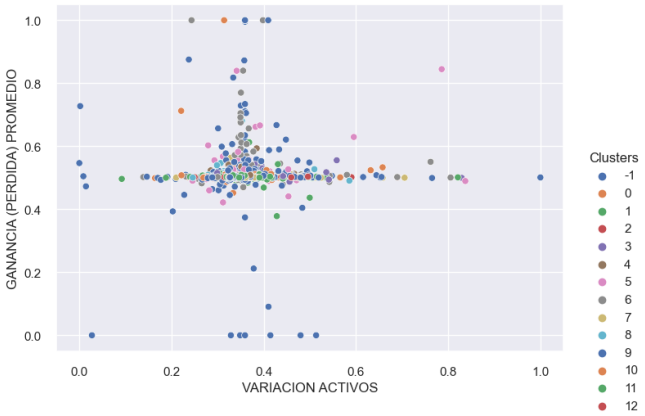
Para el parámetro épsilon, por su parte, se calcularon las distancias con los k (12) vecinos más cercanos mediante knn. Luego, se obtuvo la columna k-ésima ordenada (distancias con k vecinos) y se graficaron las distancias. Acá, se buscó el punto de inflexión de la curva para encontrar el valor optimo del parámetro. Los datos por debajo de este punto pertenecen a un clúster y el resto son ruido, lejanos al resto de puntos. Así, se estimó una épsilon de 0.4. De acá, se puede ver que más de 100 observaciones pueden ser consideradas ruido. A continuación, se muestra este gráfico.

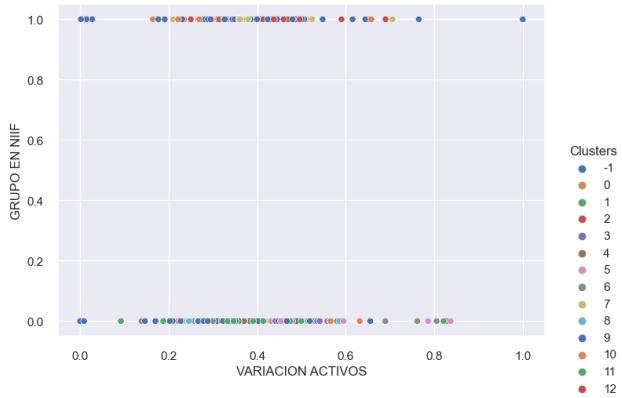


Con los parámetros ya definidos, se procedió a crear el modelo y los clusters. Se obtuvieron 12 clusters, con la siguiente distribución:

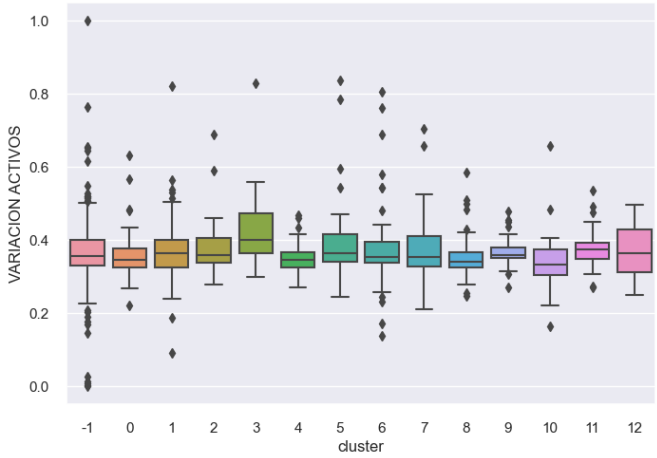
```
Counter({-1: 304, 1: 138, 0: 112, 6: 90, 4: 53, 8: 52, 7: 38, 5: 36, 9: 34, 3: 28, 11: 24, 10: 22, 12: 22, 2: 20})
```

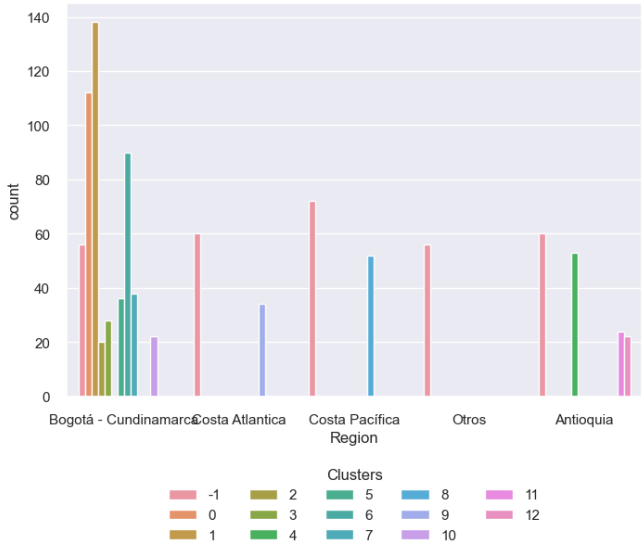
El cluster -1 hace referencia a los datos considerados ruido, por lo que se confirma que el nivel de ruido o datos atípicos es alto. Con el modelo implementado se procede a obtener el Silhouette Score y se obtiene un valor de 0.537. En este sentido, teniendo un valor superior a 0.5, se considera que los datos son cercanos dentro del clúster al que pertenece y están lejos de los otros clústeres. Sin embargo, en la visualización no se observa esto y se es difícil de interpretar.





En las gráficas de dispersión no se alcanzan a visualizar diferencias destacables entre los clusters, sin embargo, al graficar de manera individual las dimensiones enfrentadas a los clusters se pueden identificar ciertos patrones. En la sección de conclusiones se profundiza.

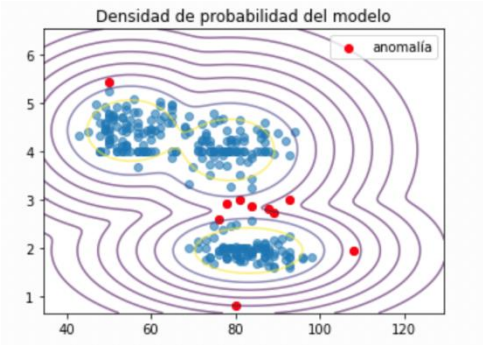




GaussianMixture

El modelo de GaussianMixture fue implementado Jhoan Diaz, es un modelo el cual puede entenderse como una generalización de K-means con la que, en ligar de asignar cada observación a un único clúster, se obtiene una distribución de probabilidad de pertenencia a

cada uno. Al momento de ajustar un GMM se necesitará estimar los parámetros que definan la función de distribución de cada componente: la media y la matriz de covarianza



4. Conclusión

En conclusión, el mejor modelo de segmentación obtenido para los datos dados por ConsultAndes fue el de DBScan, obteniendo un valor de 0.537 en el coeficiente de la silueta lo que demuestra que las agrupaciones hechas tienen datos muy parecidos entre ellos y de igual manera diferentes a los otros clústeres.

A partir de la segmentación, se pueden ver ciertas características diferenciales de las empresas. En cuanto a variación de activos, por ejemplo, se observa que el cluster 3 tiene en general un mayor crecimiento de activos y una baja cantidad de datos atípicos por lo que podría considerarse una inversión segura. Por su parte, las empresas en el cluster 6 tienden a tener más varianza por lo que hay empresas que presentaron un crecimiento significativo, pero a su vez unas que evidenciaron un decrecimiento. En cuanto a ganancias, también se ve un comportamiento parecido en cuanto a varianza. Se destaca, además, que aquellas con mayor crecimiento (visto en ganancias y varianza de los activos), son empresas grandes (grupo 1 en NIIF) y de Bogotá. En la presentación se incluyen gráficos que permite observar esta diferencia entre los clusters. Con esto, se evidencia que la segmentación realizada puede ser de utilidad para ConsultAlpes para clasificar empresas para inversión. Cabe resaltar que de tenerse más dimensiones esta segmentación podría ser más valiosa.

5. Referencias

Bedre, R. (1 de Mayo de 2022). *DBSCAN in Python (with example dataset)*. Obtenido de Data Science Blog: https://www.reneshbedre.com/blog/dbscan-python.html

Singh Chauhan, N. (4 de Abril de 2022). *DBSCAN Clustering Algorithm in Machine Learning*. Obtenido de KD Nuggets: https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html