Laboratorio 2

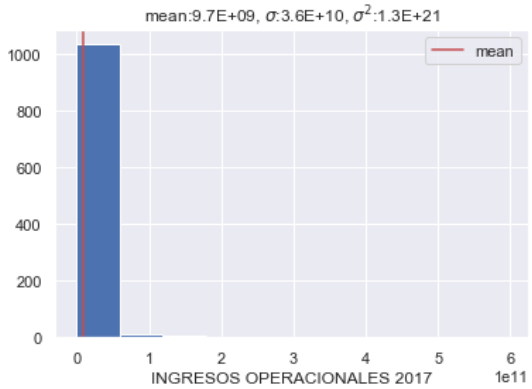
Inteligencia de negocios 2022_20

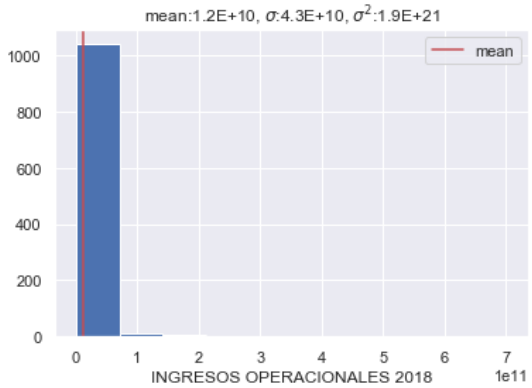
María Valentina García y
Pedro Vallejo

Entendimiento de los datos

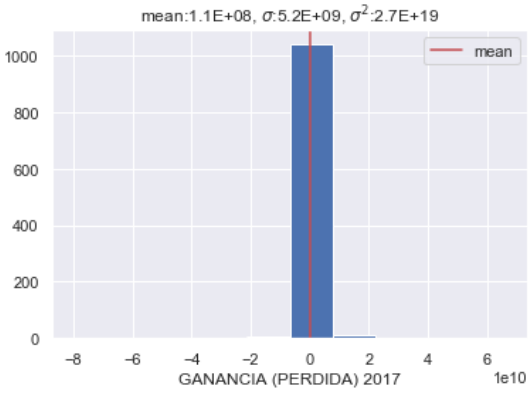
Se tienen 19 columnas y 1068 filas. Las columnas son:

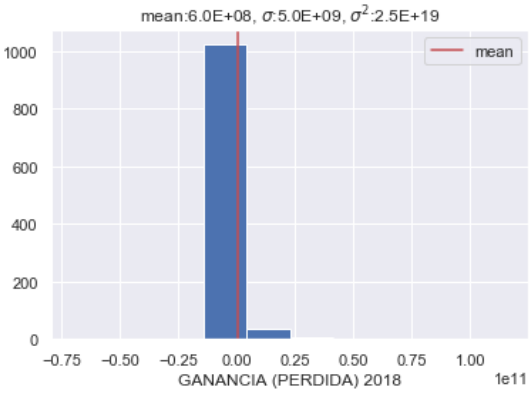
- NIT (entero, discreto): 1000 categorías
- RAZON SOCIAL (objeto): 1000 categorías
- SUPERVISOR (objeto): 6 categorías
  - SUPERFINANCIERA 17
  - SUPERSALUD 5
  - SUPERSERVICIOS 6
  - SUPERSOCIEDADES 747
  - SUPERSUCIEDADES 2
  - SUPERVIGILANCIA 7
- REGIÓN (objeto): 8 categorías
  - Antioquia 128
  - Bogotá - Cundinamarca 435
  - Centro - Oriente 25
  - Costa Atlantica 3
  - Costa Atlántica 69
  - Costa Pacífica 100
  - Eje Cafetero 19
  - Otros 5
- DEPARTAMENTO DOMICILIO (objeto): 21 categorías diferentes
- CIUDAD DOMICILIO (objeto): 73 categorías
- CIIU (objeto): 167 categorías
- MACROSECTOR (objeto): 7 CATEGORÍAS
  - AGROPECUARIO 23
  - COMERCIO 258
  - CONSTRUCCION 1
  - CONSTRUCCIÓN 51
  - MANUFACTURA 290
  - MINERO-HIDROCARBUROS 37
  - SERVICIOS 124
- INGRESOS OPERACIONALES 2017/2018 (float, continuo)



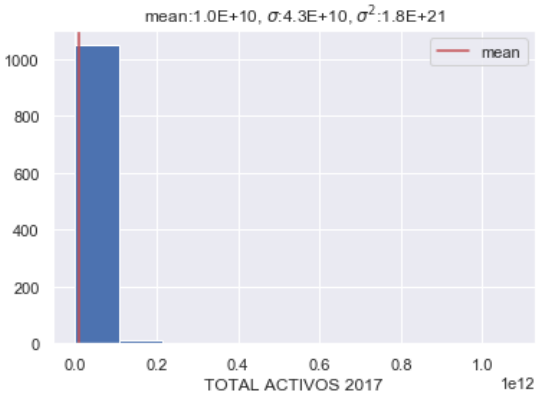


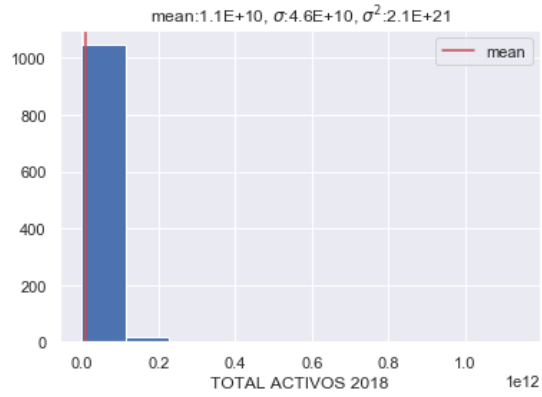
- GANANCIA (PERDIDA) 2017/2018 (float, continuo)



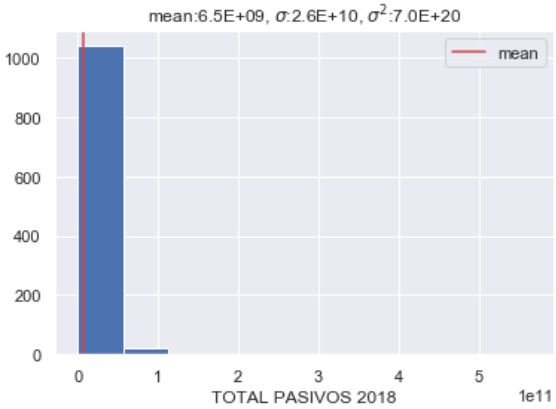


- TOTAL ACTIVOS 2017/2018 (float, continuo)

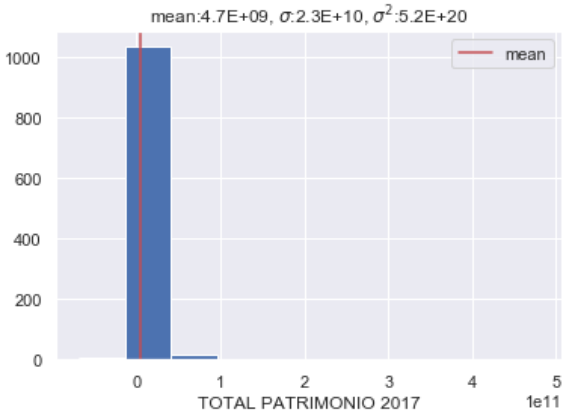


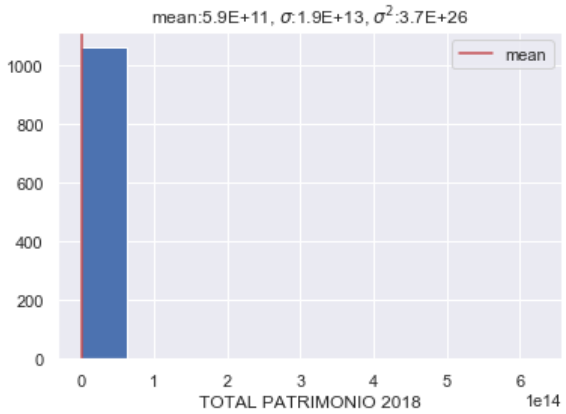


- TOTAL PASIVOS 2017/2018 (objeto/float, continuo)



- TOTAL PATRIMONIO 2017/2018 (float, continuo)





- GRUPO EN NIIF (objeto): 3 categorías
  - NIIF PLENAS-GRUPO 1 812

- NIIF PYMES-GRUPO 2 246
- REGIMEN R 414 de 2014 - CGN 10

Se puede notar que hay duplicados en los datos categóricos NIT y RAZON SOCIAL pues al ser únicos, deberían ser la misma cantidad de filas. También hay problemas de consistencia de los datos pues hay datos (que son la misma palabra) con tilde y sin tilde, por ejemplo, Costa Atlantica y Costa Atlántica. Además, hay 30 columnas con datos nulos. Hay algunas inconsistencias en la columna PATRIMONIO pues esta no corresponde con el cálculo Activos – Pasivos y en la columna SUPERVISOR pues la entidad SUPERSUCIEDADES no existe.

Preparación de datos

Se eliminaron todas las filas con datos nulos y los duplicados. Se eliminaron los outliers (datos menores al cuantil 1 y mayores al cuantil 99). Se arreglaron los datos inconsistentes. Se calculó la diferencia de dinero ('CAMBIO EN INGRESOS OPERACIONALES 2017-2018', 'CAMBIO EN GANANCIA (PERDIDA) 2017-2018', 'CAMBIO TOTAL PATRIMONIO 2017-2018', 'CAMBIO TOTAL ACTIVOS 2017-2018', 'CAMBIO TOTAL PASIVOS 2017-2018'). Se normalizaron estos datos y finalmente, nuestro dataset de tiene 5 columnas y 790 filas.

Modelamiento

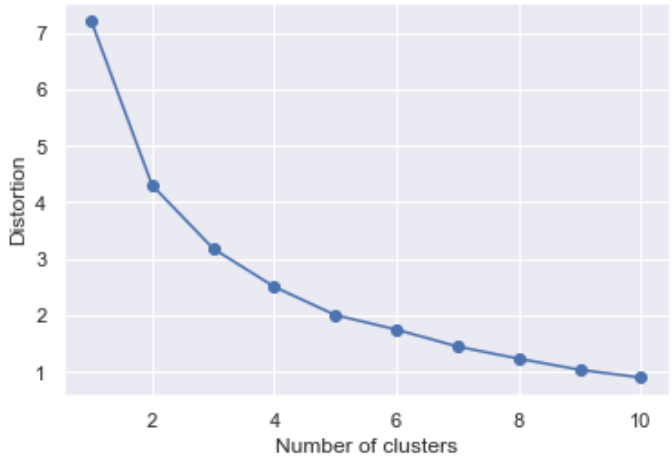
Se implementaron los algoritmos K-Means, DBSCAN y Clustering Jerárquico.

K-MEANS (Pedro)

K-Means es un algoritmo no supervisado que permite agrupar al detectar las distintas categorías de grupos en los conjuntos de datos sin etiquetas por sí mismo. Este es un algoritmo basado en centroides, de modo que cada grupo está conectado a un centroide mientras sigue el objetivo de minimizar la suma de distancias entre los puntos de datos y sus grupos correspondientes. Como entrada, el algoritmo consume un conjunto de datos sin etiquetar, divide el conjunto de datos completo en un número k de grupos e itera el proceso para cumplir con los grupos correctos

Elección de hiperparámetros

El único hiperparametro es k que es el número de clusters, una manera de obtener el mejor K, es a través del método del codo. En este método, se varía el número de K y para cada valor, se esta la suma de la distancia al cuadrado entre cada punto y el centroide en un grupo. Cuando se traza esta distancia contra el valor K, el gráfico parece un codo. A medida que aumenta el número de clústeres, el valor de del cálculo comenzará a disminuir. Desde el punto del codo, el gráfico comienza a moverse casi paralelo al eje X, y ese punto del codo es el k recomendado por el método.



Como se puede ver en la imagen, para el caso del laboratorio se obtiene un valor de k alrededor de 3. Sin embargo, se hicieron pruebas con los valores 2, 3, 4 ya que el método del

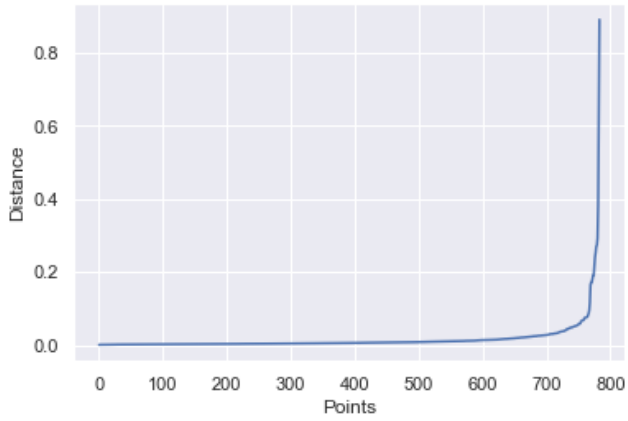
codo no es exacto y es recomendable probar con los valores que lo rodean, y se obtuvo un mejor resultado con un valor de 4 clusters.

DBSCAN (Valentina)

DBSCAN requiere dos parámetros: qué tan cerca deben estar los puntos entre sí para ser considerados parte de un grupo (eps) y el número mínimo de puntos para formar un grupo (minPts). El algoritmo comienza por un punto arbitrario que no haya sido visitado. Se encuentran los puntos que son alcanzables según la distancia eps y si contiene suficientes puntos, se inicia un clúster sobre el mismo. Luego, para cada uno de estos puntos, se realiza esto mismo y así sucesivamente hasta no encontrar más puntos alcanzables. Cuando esto sucede, un nuevo punto no visitado se visita y procesa con el objetivo de descubrir otro clúster. Al final, los grupos de puntos que no son lo suficientemente densos, se marcan como ruido.

Elección de hiperparámetros

- minPts: Se toma igual o mayor a la dimensionalidad de los datos. En nuestro caso minPts = 5, que es igua al número de columnas del dataset de entrenamiento.
- eps: Se utilizan las distancias al vecino más cercano a cada punto que calcula el algoritmo KNN. Se grafican las distancias contra el número de puntos y se identifica el codo de esta curva.



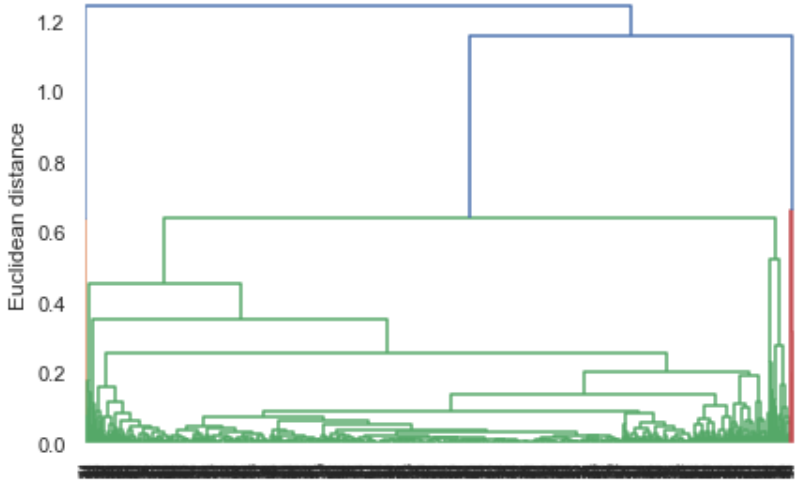
En nuestro caso, vemos que el codo correspondiente al parámetro eps, está alrededor de 0.1.

JERARQUICO (Pedro)

Para el algoritmo jerárquico se escogió el acercamiento de aglomeración donde cada dato empieza desde su propio grupo y a medida que avanza el algoritmo, estos datos se van juntando según jerarquías a un mismo grupo. Este algoritmo también se basa en distancias euclideanas.

Elección de hiperparámetros

El algoritmo necesita, al igual que k-means, el número de clústeres como entrada. Para esto primero se construye una matriz de distancias con la cuál después se construye un dendograma basado en jerarquías el cual muestra el número de clusters indicado.

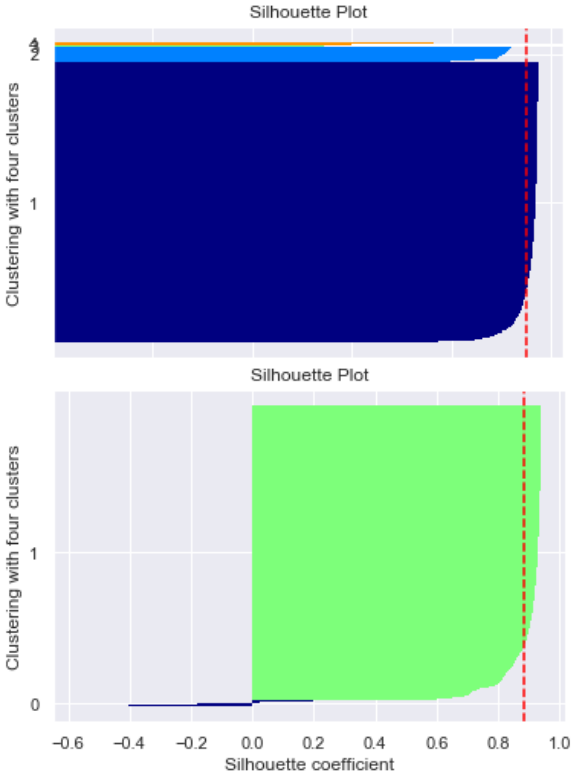


El dendograma obtenido de los datos muestra un número de 3 clústeres. Sin embargo, nuevamente, como este no es un acercamiento exacto, se probó con un valor de 2 y 4 clústeres, y el mejor resultado se obtuvo con 4 clústeres.

Validación

Validación cuantitativa

K-MEANS (Pedro)

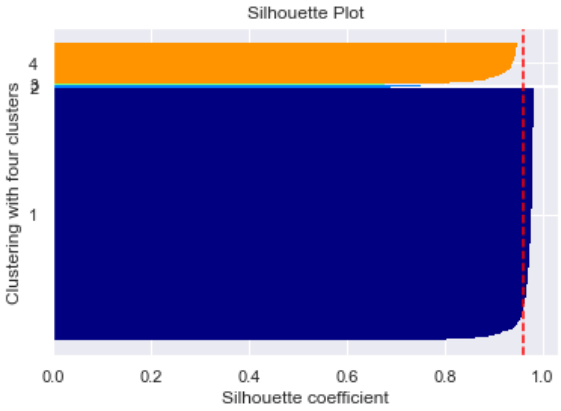


El método de la silueta muestra en general un buen resultado de clustering donde el coeficiente de silueta promedio es 0.95. Esto significa que en general hay una alta cohesión entre los datos dentro de cada cluster y un alto distanciamiento entre ellos.

DBSCAN (Valentina)

El coeficiente de silueta tiene un valor de 0.88. A pesar de que el coeficiente de Silueta general es cercano a 1, la clasificación de los outliers no es tan buena dado que tiene valores del coeficiente negativos.

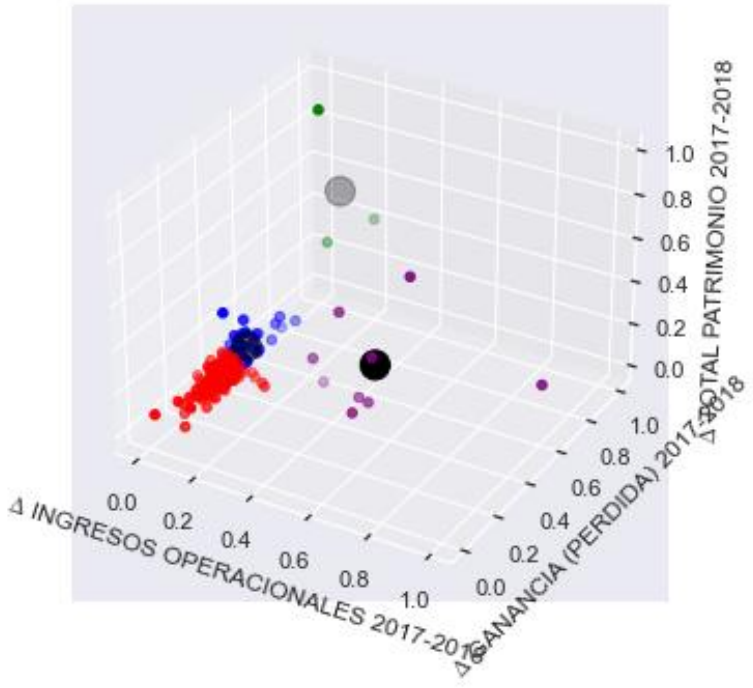
JERARQUICO



En este caso se obtiene el mejor coeficiente de silueta siendo igual a 0.96, muy cercano a 1. Lo que indica que se tiene una muy buena división entre los grupos identificados. Este modelo es el que logra dividir de mejor manera los datos.

Validación cualitativa

K-MEANS



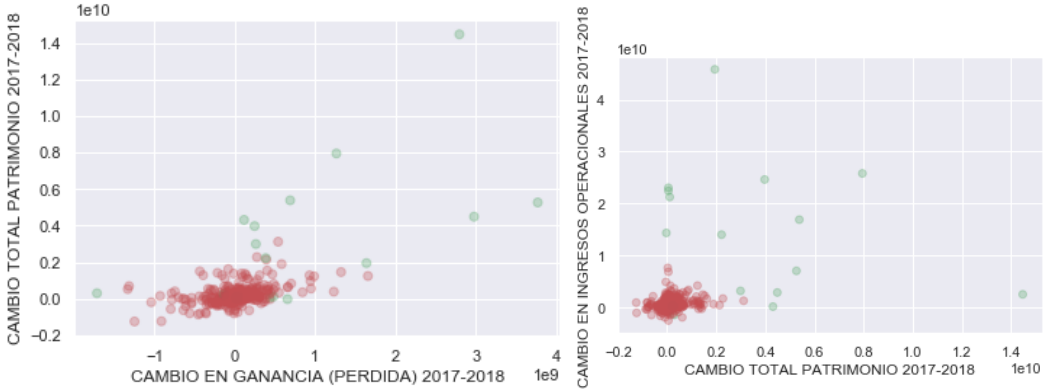
Cluster	Num. elementos
1	737
2	42
3	3
4	9

Se puede ver como el algoritmo creo 4 clusters distintivos. En el cual la mayoría de los datos se encuentran en el primer cluster.

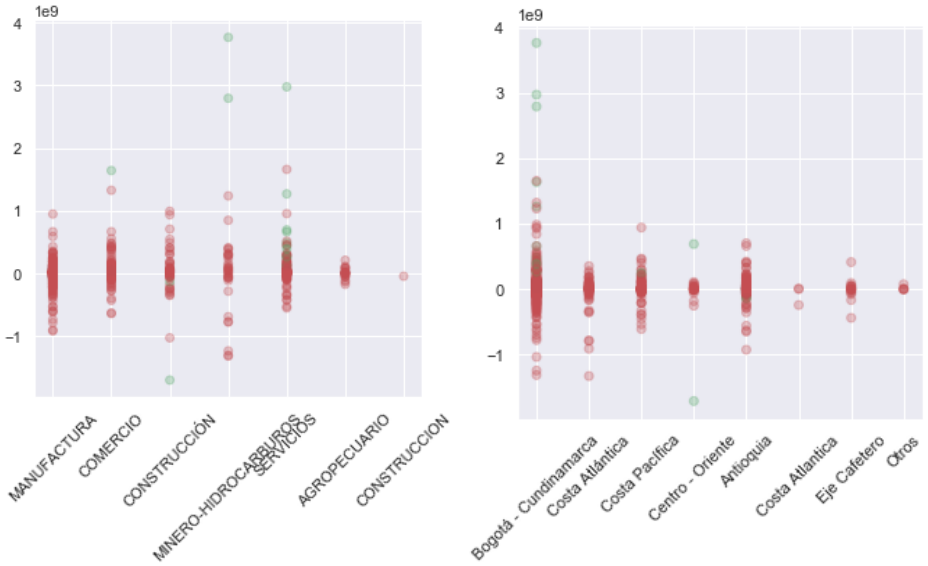
Se puede notar que los clusters 1 y 2 no se encuentran muy separados el uno del otro, lo cual no es deseable para esta tarea de aprendizaje. Sin embargo, estos clusters si muestran un alto grado de cohesion dentro de sus datos ya que es notable que en su mayoría los datos que pertenecen a estos clusters se encuntran muy cercanos unos de los otros

Los clusters 3 y 4 por el otro lado tienen un comportamiento completamente opuesto. Estos clusters tienen una separación notable de los demas clusters, lo cual es deseado. Pero por el otro lado, los datos dentro de estos clusters muestran una baja cohesión, lo cuál no es deseado para esta tarea de aprendizaje.

DBSCAN (Valentina)

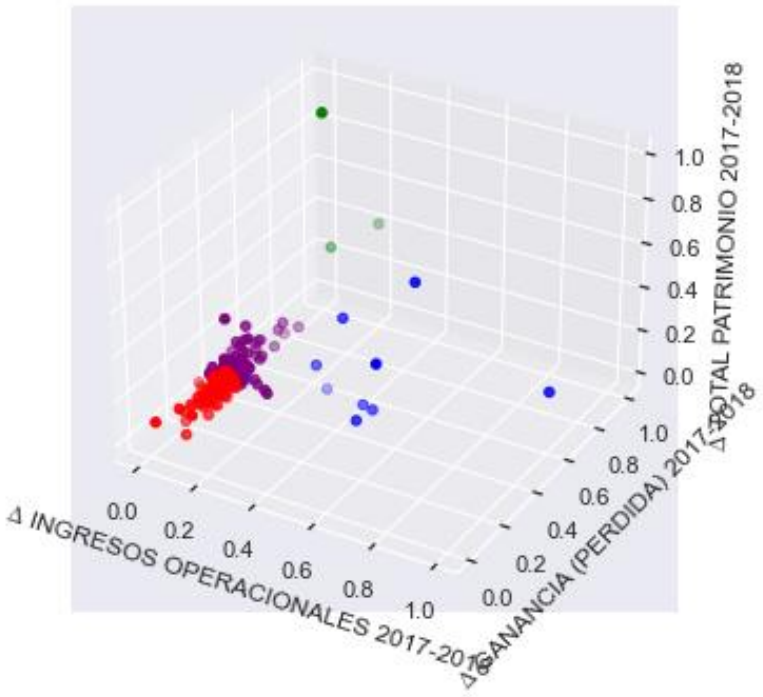


Como se ve en las gráficas, DBSCAN distingue dos grupos. Los que no tuvieron cambio o tuvieron un cambio pequeño, y los outliers que son los que tuvieron un cambio muy grande o muy pequeño.



Se puede identificar que los outilers con ganacias pertenecen a la industria de comercio minero y servicios, y con pérdidas al sector de construcción. Todos estos se encuentran en Bogotá y el Centro-Oriente del país.

JERARQUICO (Pedro)



Cluster	Num. elementos
1	668
2	9
3	3
4	111

Se puede ver como el algoritmo creo 4 clústeres distintivos. En el cual la mayoría de los datos se encuentran en el primer clúster al igual que en k-means. Sin embargo, es notable que este algoritmo hizo una división un poco mas equitativa entre los dos clústeres más parecidos.

Se puede notar que los clústeres 1 y 4 no se encuentran muy separados el uno del otro, lo cual no es deseable para esta tarea de aprendizaje. Sin embargo, estos clústeres si muestran un alto grado de cohesión dentro de sus datos ya que es notable que en su mayoría los datos que pertenecen a estos clústeres se encuentran muy cercanos unos de los otros

Los clústeres 2 y 3 por el otro lado tienen un comportamiento completamente opuesto. Estos clústeres tienen una separación notable de los demás clústeres, lo cual es deseado. Pero por el otro lado, los datos dentro de estos clústeres muestran una baja cohesión, lo cuál no es deseado para esta tarea de aprendizaje.

Visualización

Se presentan los resultados del mejor modelo, el Clustering Jerárquico en el tablero de control.