DBSCAN

El algoritmo DBSCAN es un método de clústering basado en densidad de aplicaciones con ruido. Este método permite agrupar los datos cuando estos presentan formas complejas, así como es un método robusto frente a la presencia de outliers. Para realizar el algoritmo DBSCAN se emplean sólo las variables numéricas normalizadas.

DBSCAN parte de dos parámetros que son: - Épsilon: distancia máxima a la que debe haber otra observación para ser considerar que cumple con el criterio de "estar cerca" - Mínimo de puntos: parámetro que controla la densidad mínima requerida para que un punto sea considerado un núcleo y se incluya en un grupo/clúster.

Cálculo de mínimo de puntos

Para calcularlo de manera empírica, diremos que el mínimo de puntos sea igual al 0.2 % - 0.25 % del total de los datos teniendo en cuenta que:

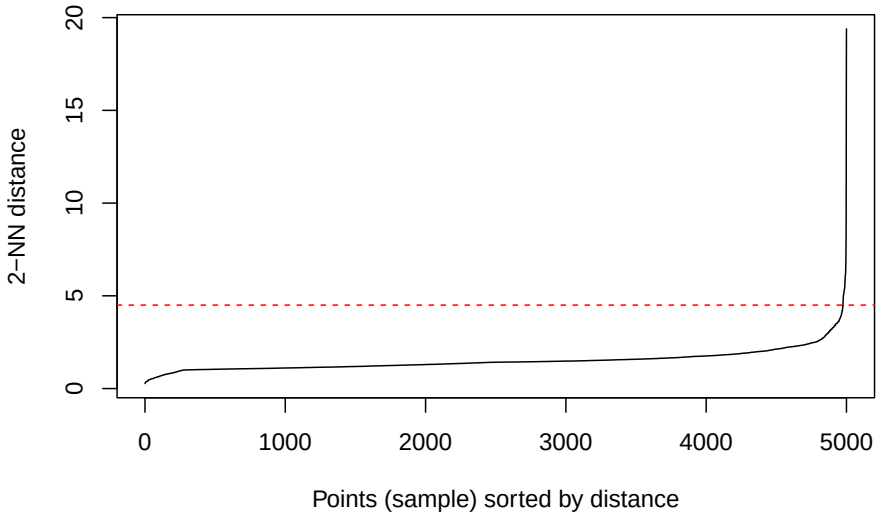
- El mínimo será de 2 para datos que sean muy pequeños
- El máximo será de 10 para datos con mucha información

Aplicando esto, se tiene que el número mínimo de puntos sería 11, pero lo limitamos a 10 en concordancia con la literatura.

Cálculo de épsilon

Se escogerá épsilon a partir del siguiente gráfico del codo, realizado con el método del k-NN. Como se han realizado otros métodos de clústering, se aplica el k-NN con el valor de K sacado de los métodos de clústerng jerárquico, que concluyen que el número de clústers k óptimo es 2. Estas k-distancias se trazan en orden ascendente con el objetivo es determinar la "codo", que corresponde al parámetro épsilon óptimo. A partir del siguiente gráfico del codo se puede observar el valor óptimo de épsilon.

Figura 63: Gráfico del Codo para el Valor Óptimo de Épsilon



El valor de épsilon se decide a partir de el corte en el máximo cambio de la pendiente. En el gráfico se observa que esto se da alrededor de épsilon = 4.5.

Resultado DBSCAN

Aplicamos el método DBSCAN con los valores extraídos: épsilon=4.5 y mínimo de puntos de 10.

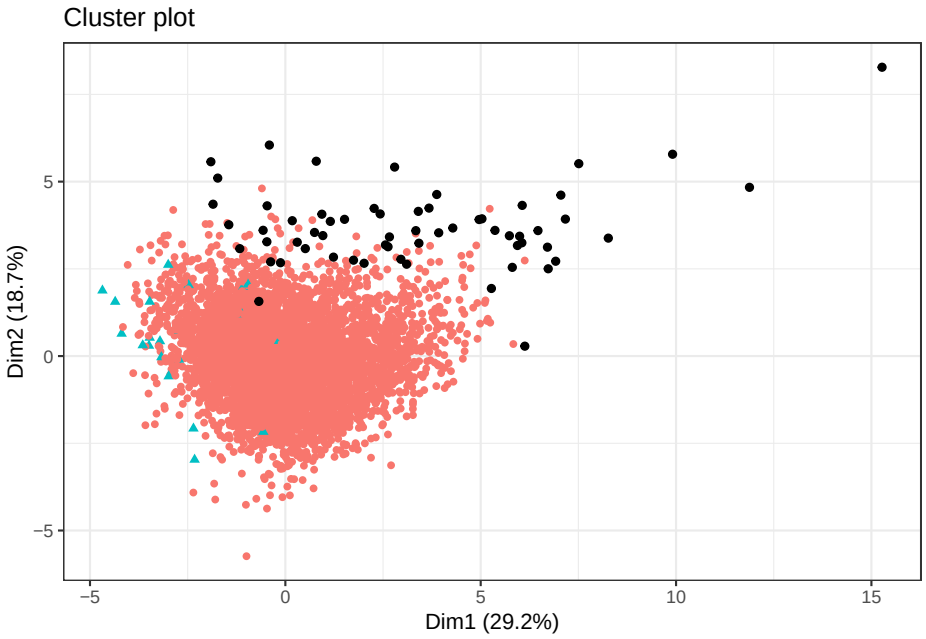
Cuadro 24: Resumen del número de puntos en cada clúster

Clúster	Frecuencia de puntos
0	61
1	4915
2	24

EL resultado del DBSCAN indica que agrupa los datos en 2 clústers, y devuelve 61 puntos como outliers.

Se presenta el gráfico de los clústers obtenidos con el DBSCAN:

Figura 64: Gráfico Clústers obtenidos con DBSCAN



Con estos resultados, ya se aprecia que el DBSCAN no realiza agrupaciones óptimas en estos datos, ya que la inmensa mayoría de datos se ubican en el primer clúster, y el segundo clúster contiene una proporción de datos ínfima. Esto puede ser debido a que el DBSCAN es un método que parte de las densidades, y en los datos que se agrupan con formas más simples y uniformes, como los que se tratan en este trabajo, puede no encontrar la solución ótpima, como se considera en este caso. A esta misma conclusión se llegará también con el método OPTICS, ya que también es un método de clústering basado en densidades.