2. Research about the DBSCAN method, and answer the following questions:

a. ¿In which cases might it be more useful to apply?

In general, DBSCAN is an unsupervised clustering algorithm that should be used when you do not have a particular outcome variable you want to predict. Instead, you should have a set of features you want to use to identify patterns across your dataset. In general, DBSCAN is intended to be used in cases where all of your features are numeric.

When should you use DBSCAN over another clustering algorithm?

You suspect there may be irregularly shaped clusters: If you have reason to expect that the clusters in your dataset may be irregularly shaped, DBSCAN is a great option. DBSCAN will be able to identify clusters that are spherical or ellipsoidal as well as clusters that have more irregular shapes.

Data has outliers: DBSCAN is also a great option for cases where there are many outliers in your dataset. DBSCAN is able to detect outlying data point that do not belong to any clusters and exclude those data points from the the clusters.

Anomaly detection: Since DBSCAN automatically detects outliers and excludes them from all clusters, DBSCAN is also a good option in cases where you want to be able to detect outliers in your dataset.

When should you avoid using DBSCAN?

No drop in density between clusters: In general, DBSCAN requires there to be a drop in the density of data points in order to detect boundaries between clusters. That means that you should not use DBSCAN if you do not expect there to be much of a drop in density between different clusters. For example, if you expect many of your clusters overlap, multiple clusters might get grouped together into one large cluster.

Many categorical features: DBSCAN is generally intended to be used in scenarios where the majority of your features are numeric. That means that you should avoid using DBSCAN in cases where you have many categorial features. In these scenarios, you may be better off using hierarchical clustering with an appropriate

distance metric or an extension of k-means clustering like k-modes to k-prototypes.

b. What are the mathematical fundamentals of it?

DBSCAN is a simple algorithm that defines clusters by estimating local density. It can be divided into 4 stages:

- For each observation we look at the number of points at a maximum distance ε from it. This area is called the ε-neighborhood of the observation.
- If an observation has at least a certain number of neighbors, including itself, it is considered a central observation. In this case, a high-density observation has been detected.
- All observations in the neighborhood of a central observation belong to the same cluster. There may be central observations close to each other. Therefore, from one step to another, a long sequence of central observations is obtained that constitute a single cluster.
- Any observation that is not a central observation and does not have any central observations in its neighborhood is considered an anomaly.

Therefore, it is necessary to define two pieces of information before using DBSCAN:

What distance ε must be determined for each observation in the ε-neighborhood? What is the minimum number of neighbors needed to consider an observation as a central observation?

These two data are freely provided by the user. Unlike the k-means algorithm or bottom-up rank, the number of clusters does not need to be defined in advance, making the algorithm less rigid.

c. c. Is there any relation between DBSCAN and Spectral Clustering? If so, what is it?

The relations between dbscan and spectral clustering:

- Both needs numerical data to process
- Both works over irregular datasets
- Both works when there are no strong assumptions about cluster shape

“DBSCAN can be explained and rewritten under the framework of spectral clustering. Furthermore, eigenvectors are often approximately resolved, and k-

means is usually used in the final stage, often converges in local optimization. Hence, we rewrite spectral clustering by using nearest neighbor query instead of k-means to obtain exact result"[https://www.semanticscholar.org/paper/DBSCAN-Is-Semi-Spectral-Clustering*-Chen/978cd1ca9399eeffb8d908c2e71642b69daa876a]

References:

- https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm#:~:text=Distancia%20definida%20(DBSCAN)%3A%20utiliza,con%20todos%20los%20cl%C3%BAsteres%20potenciales.
- https://datascientest.com/es/machine-learning-clustering-dbscan
- https://anderfernandez.com/blog/dbscan-python/
- https://crunchingthedata.com/when-to-use-dbscan/#:~:text=In%20general%2C%20DBSCAN%20is%20an,identify%20patterns%20across%20your%20dataset.
- https://elutins.medium.com/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818