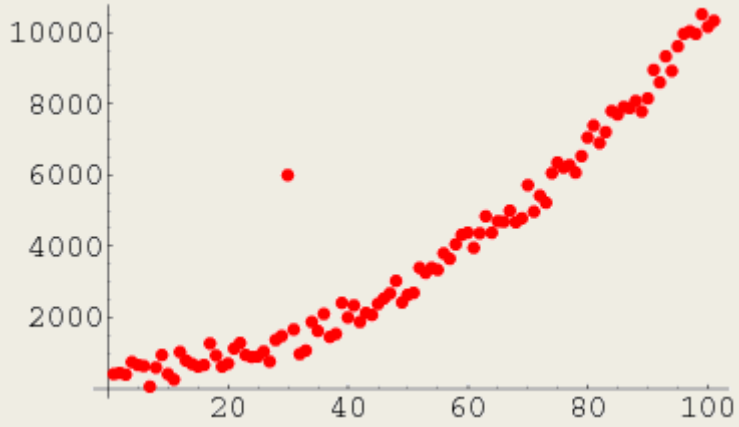
OUTLIERS

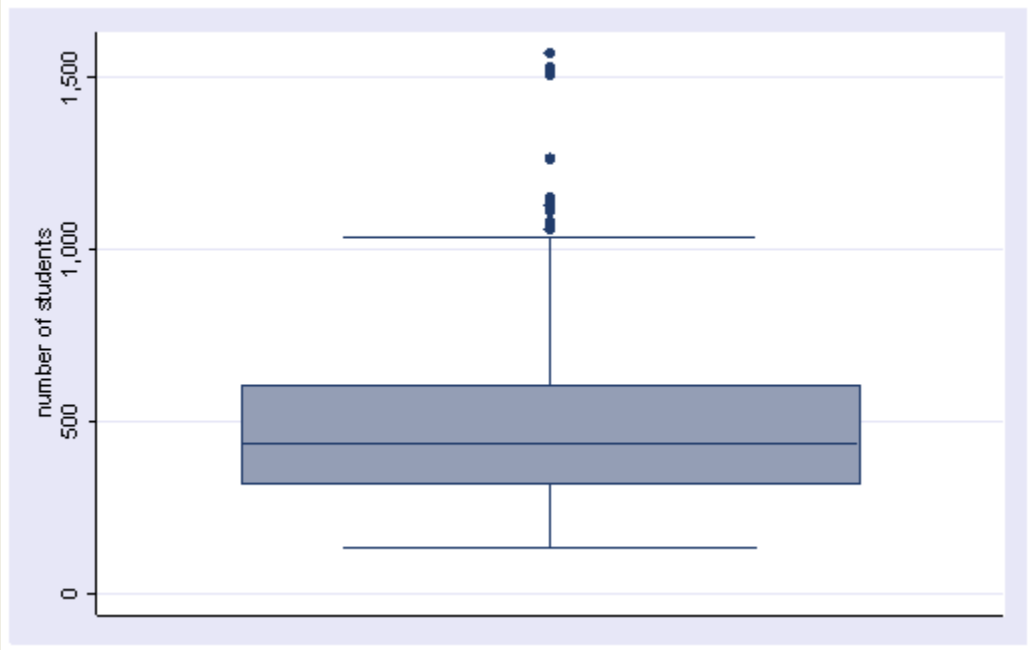
Minería de Datos Grupo 012
Edgar Bladimir López Alonzo 1753141
Cristian Antonio Jaramillo Arriaga 1680776

Introducción

El análisis de la calidad de los datos es de gran importancia para las organizaciones, ya que datos con problemas pueden conducir a decisiones erróneas con consecuencias como pérdida de dinero, tiempo y credibilidad. Entre los posibles problemas que pueden presentar los datos, se encuentran los conocidos como valores atípicos o “Outliers”.

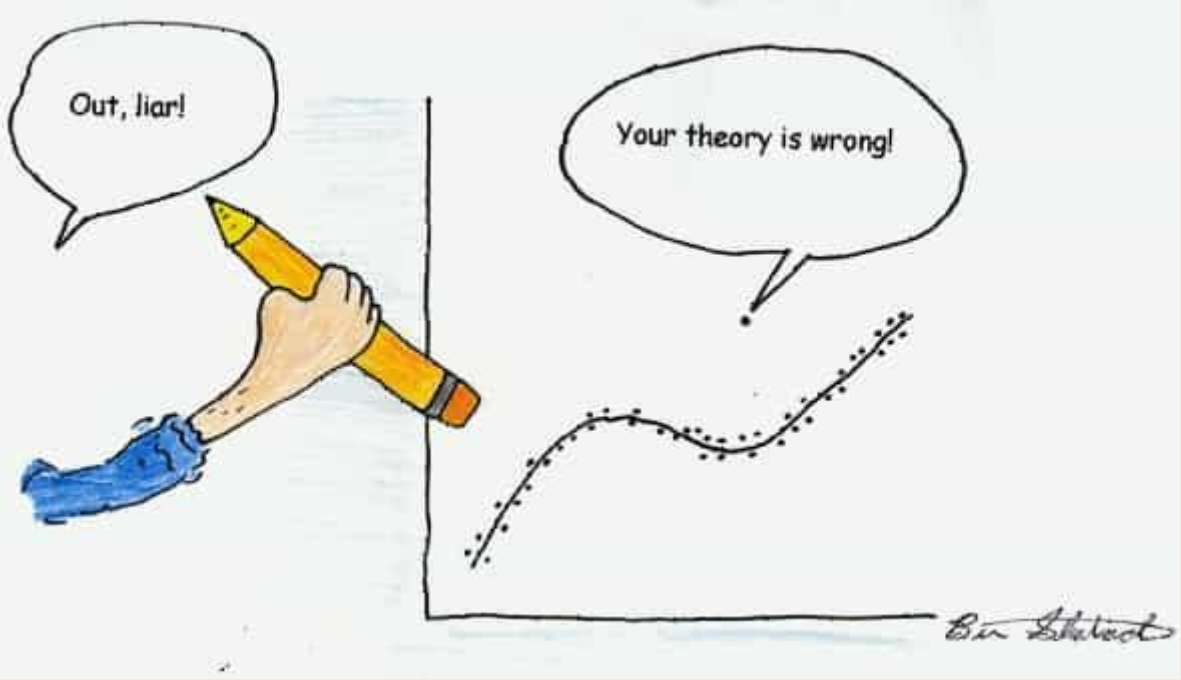


Un **outlier** es una observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente.

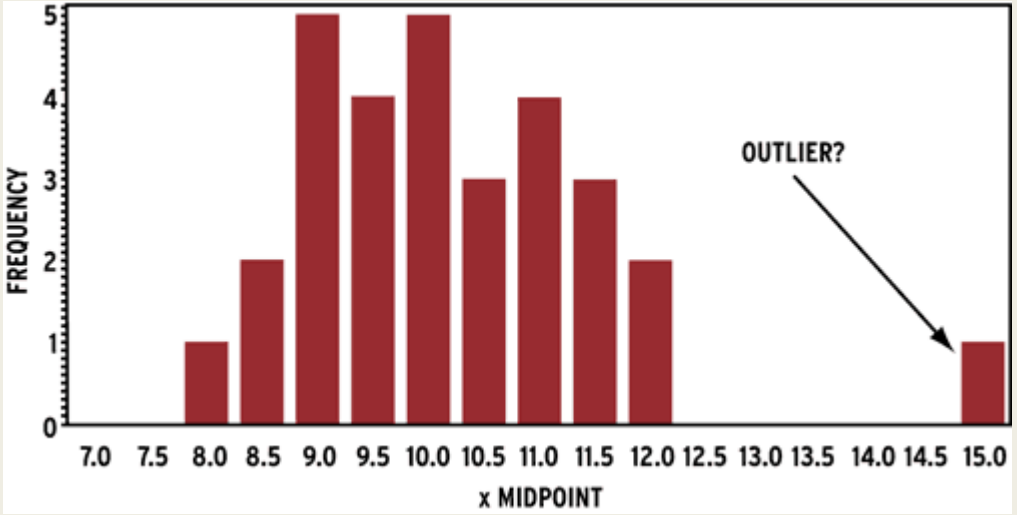


Tipos de outliers

1. Casos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificarse como datos ausentes.

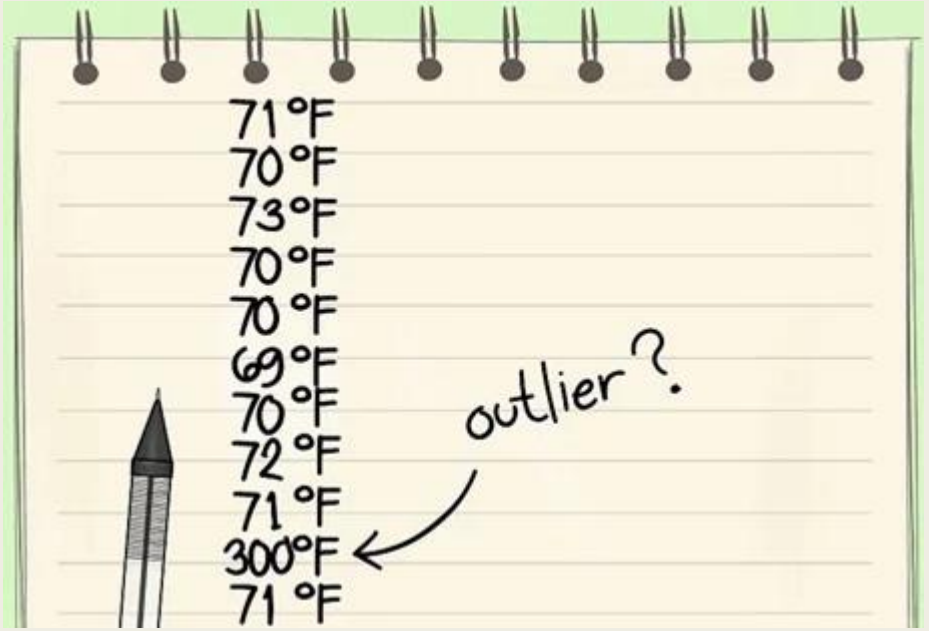


2. Observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento valido de la población y puede ser eliminado del análisis.



3. Observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables. Estas observaciones deberían ser retenidas en el análisis pero estudiando que influencia ejercen en los procesos de estimación de los modelos considerados.

4. Datos extraordinarios para las que el investigador no tiene explicación. En estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista debería reportarlo en sus conclusiones y debería averiguar el por que de dichas observaciones.



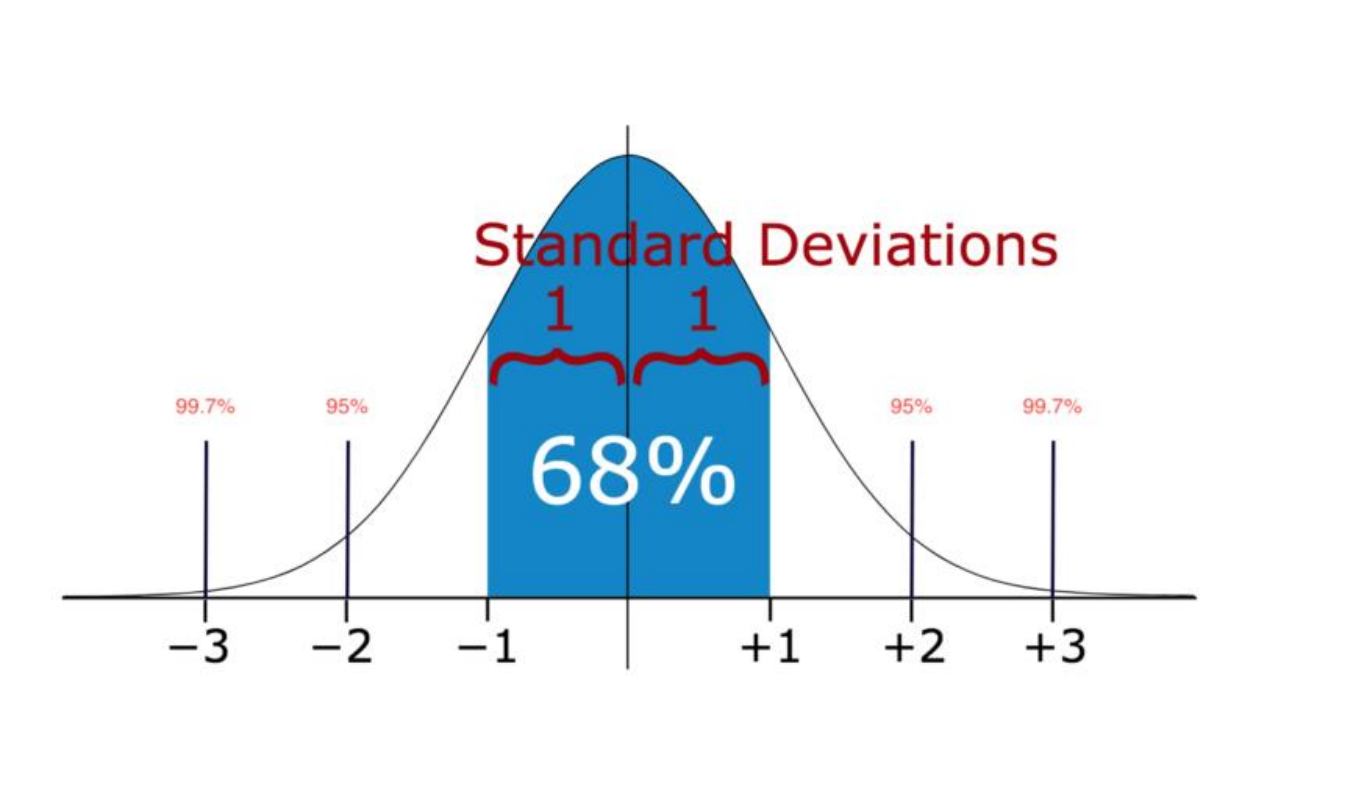
Los métodos de detección

Se pueden dividir en univariados y multivariados. Los outliers multivariantes son observaciones que se consideran extrañas no por el valor que toman en una determinada variable, sino en el conjunto de aquellas. Son más difíciles de identificar que los outliers unidimensionales, dado que no pueden considerarse “valores extremos”, como sucede cuando se tiene una única variable bajo estudio

Método de detección: Desviación estándar

En estadística, si una distribución de datos es aproximadamente normal, aproximadamente el 68% de los valores de los datos se encuentran dentro de una desviación estándar de la media y aproximadamente el 95% están dentro de dos desviaciones estándar, y aproximadamente el 99,7% se encuentran dentro de tres desviaciones estándar.

Método de detección: Desviación estándar



Si se tiene algún punto de datos que sea más de 3 veces la desviación estándar, es muy probable que esos puntos sean anómalos o atípicos.

Ejemplo

```
import numpy as np
np.random.seed(1)

valoresAleatorios = np.random.randn(50000) * 20 + 20

def encontrarAnomalias(vector):
    anomalias = []

    desviacionEstandar = np.std(vector)
    media = np.mean(vector)
    anomaly_cut_off = desviacionEstandar * 3

    limiteInferior = media - anomaly_cut_off
    limiteSuperior = media + anomaly_cut_off

    print('Desviación estandar: ', desviacionEstandar)
    print('Media: ', media)
    print('Limite inferior: ', limiteInferior)
    print('Limite superior: ', limiteSuperior, '\n')
    # Generate outliers
    for outlier in vector:
        if outlier > limiteSuperior or outlier < limiteInferior:
            anomalias.append(outlier)
    return anomalias

valoresAtipicos = encontrarAnomalias(valoresAleatorios)
print('Valores atipicos: ', valoresAtipicos)
```

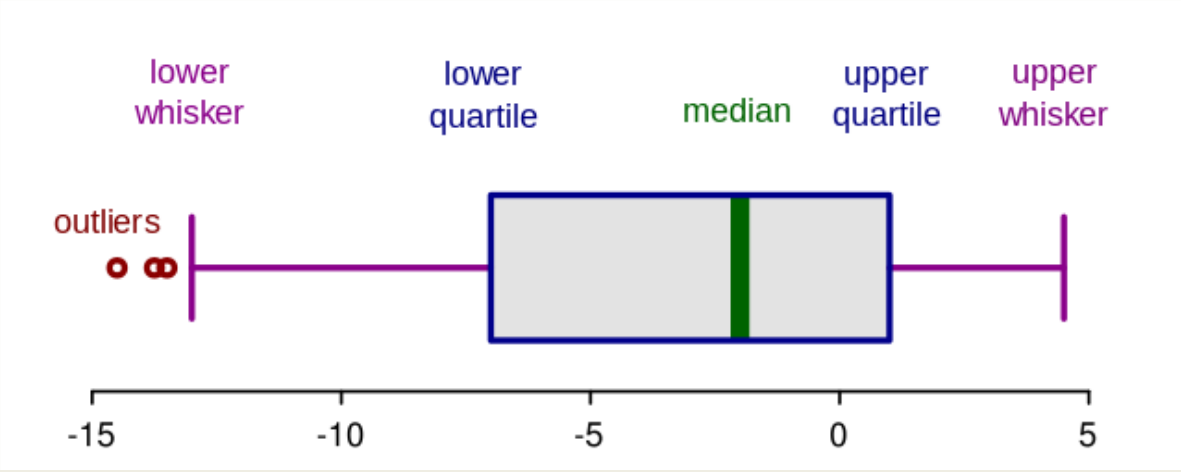
Ejemplo

```
Desviación estandar:  20.013530816925808
Media:  20.07576979467771
Limite inferior:  -39.96482265609971
Limite superior:  80.11636224545514

Valores atipicos:  [80.61714224744061, 99.17205408075927, 86.4215751234073, -41.0752876085261, -43.06714900382097, -40.32063970418597, 88.65326
863591119, -41.28282712210201, 94.80497807409178, 82.70094680175814, 84.76686393504752, -45.06068469953237, 82.64059456281011, 80.2550913647446
6, -43.81232693648161, -45.21230191553079, 88.08604551497943, 82.369591815651, 85.9708104656157, 82.97968050439259, -45.8971681339703, -41.1126
6480828613, 100.53698089094755, -44.38264211247683, 82.23783662033449, 92.26554014166295, 87.96231313004853, 91.21746641411097, -53.12880198509
5905, -45.641575926212354, 91.21225295344836, -43.46923274683489, 82.13498657145581, 82.90737927394908, -46.216851244091984, -41.9565310384887,
-45.6065519410858, 80.5243698860871, -43.30420249353775, -42.25263324372853, 81.59215546151304, 103.36235355910189, 96.68762041821407, -40.5868
79558462776, -48.71851620008829, -45.18412137279293, 81.54158108441533, 84.5587224070789, -40.7688517191754, -44.08128900692253, -42.0432978738
81315, 87.71430311862296, 88.9391203157937, -40.52113314693459, -49.02805812510347, 86.72050502651298, -44.452278374625166, -58.55028132123577,
87.83885146048503, -41.810065776968784, -42.38237155225082, 91.19937116220078, 93.19531683513023, -41.0172829464469, 82.94710559091223, 86.8410
257826826, -47.358945718270846, -47.27471768342224, -44.05283959700523, 81.39224116690686, -64.66329594864382, -41.91924535511986, 86.982934021
21955, -42.78324538163221, -43.44070544512959, 83.14254869231635, 80.29722046012112, 80.97664481662183, -49.135306899574985, 90.95359817188209,
85.7504914101956, 90.63353726783707, -52.25624570621119, -48.427273980448774, 92.45148786479089, -40.60136533430394, 90.66197938200516, -40.652
000332079666, 85.47272253055641, -40.59878231029168, -53.957837173283735, -40.80036485428288, 82.4740392321337, 82.69560613707037, -46.18148126
636443, 86.9908487614501, 80.1457297060347, 82.24721888631221, -45.832275955612076, 84.176266941649, 84.4193809675114, -48.12080165824392, -47.
66016305888587, 83.03299706691035, -46.34572895708007, -43.94127801481702, -44.888893484963376, 82.73490952329395, 83.57335544403156, 80.772741
3101268, -52.50413207267792, -49.69745467233162, 80.22051938742219, -61.76977168084022, -41.57467213920625, -58.86284640187061, -51.19738385716
6216, -56.11928395638613, -48.85042290718998, -52.64159591899764, -40.858015925037485, 90.00512514161373, -44.150803760611865, 81.075750841652
4, 88.42082610751223, -48.97515466922074, -41.87410601743178, 81.91272478059072, 84.69303765209298, -40.93984641793596, -52.615615036883995, -4
9.39834453038375, 84.48319182306491, 80.41393381216217, 89.04319371214211, 83.11545348720793, 86.3948526034166, -54.34260433053814, -53.8840895
53553295, -60.83360878272147, -45.502951473453436, -42.84117089742673]
```

Método de detección: Boxplots

Los diagramas de caja son una representación gráfica de datos numéricos a través de cuantiles. Es una forma muy simple pero efectiva de visualizar valores atípicos. Los bigotes inferiores y superiores pueden verse como los límites de la distribución de datos. Cualquier punto de datos que se muestre por encima o por debajo de los bigotes, puede considerarse atípico o anómalo.

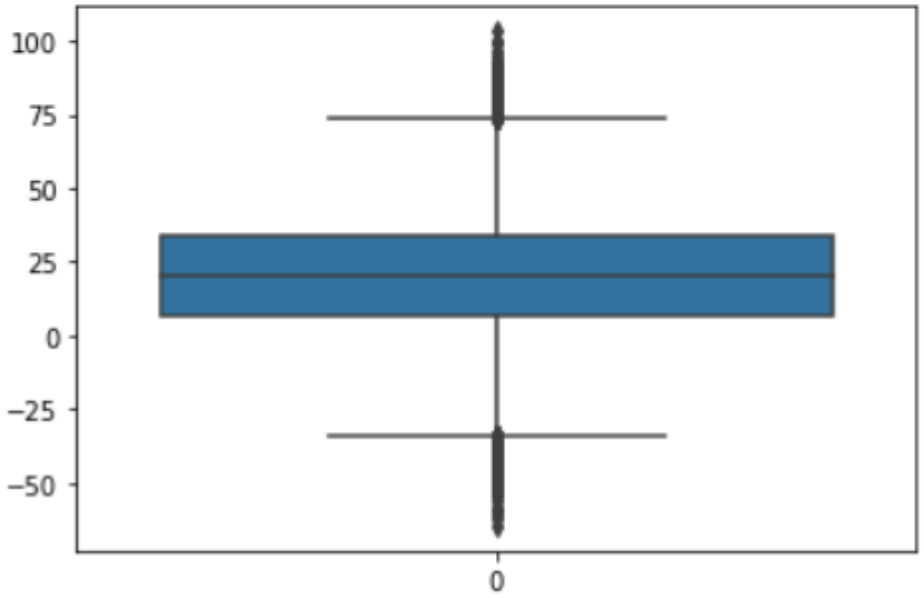


Ejemplo

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
np.random.seed(1)

valoresAleatorios = np.random.randn(50000) * 20 + 20
print('\n\033[95mBoxplot: ')
sns.boxplot(data=valoresAleatorios)
```

```
Boxplot:
<matplotlib.axes._subplots.AxesSubplot at 0x15fd2056790>
```



Método de detección: DBScan Clustering

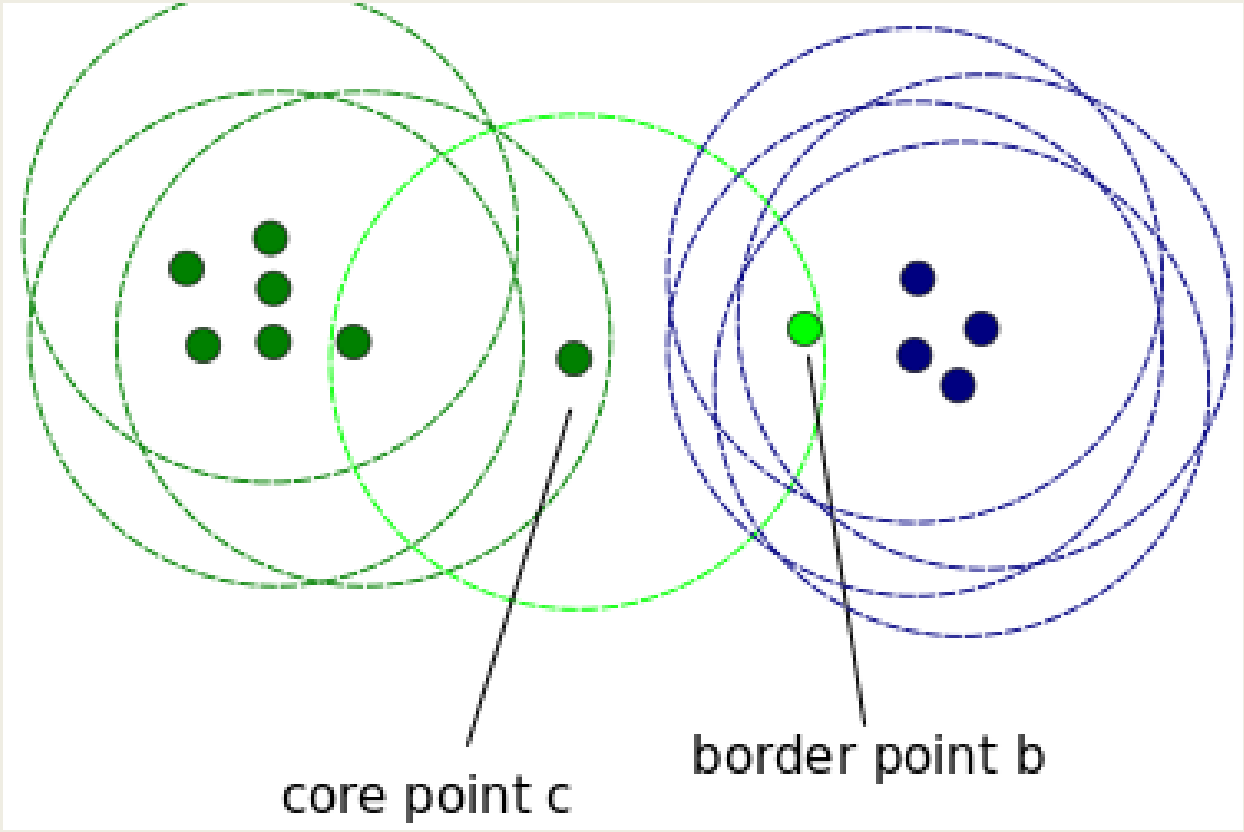
DBScan es un algoritmo de agrupación en clústeres que utiliza datos agrupados en grupos. También se utiliza como un método de detección de anomalías basado en la densidad con datos unidimensionales o multidimensionales. También se pueden utilizar otros algoritmos de agrupación como k-medias y agrupación jerárquica para detectar valores atípicos.

Método de detección: DBScan Clustering

Core Points: para comprender el concepto de puntos centrales, se debe revisar algunos de los hiperparámetros utilizados para definir el trabajo DBScan. El primer hiperparámetro (HP) es **min_samples**. Este es simplemente el número mínimo de puntos centrales necesarios para formar un grupo. El segundo HP importante es el **eps**. **eps** es la distancia máxima entre dos muestras para que se consideren como en el mismo grupo.

Border Points: se encuentran en el mismo grupo que los puntos centrales, pero mucho más lejos del centro del grupo.

Método de detección: DBScan Clustering



Método de detección: DBScan Clustering

Todo lo demás se denomina **Puntos de ruido (Noise Points)**, son puntos de datos que no pertenecen a ningún grupo. Pueden ser anómalos o no anómalos y necesitan más investigación.

Ejemplo

```
import numpy as np
from sklearn.cluster import DBSCAN
np.random.seed(1)
valoresAleatorios = np.random.randn(50000,2)  * 20 + 20

outlierDetection = DBSCAN(min_samples = 2, eps = 3)
clusters = outlierDetection.fit_predict(valoresAleatorios)
numeroDePosiblesAnomalias = list(clusters).count(-1)
print('\033[95mNumero de posibles anomalias: \033[0m', numeroDePosiblesAnomalias)
```

```
Numero de posibles anomalias:  94
```

Bibliografía

- https://core.ac.uk/download/pdf/297178308.pdf
- https://www.ugr.es/~fmocan/MATERIALES%20DOCTORADO/Tratamiento%20de%20outliers%20y%20missing.pdf