ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

Laboratorio de: Inteligencia de negocios
Práctica No.: 4 de Datamining
Tema: Clustering

Nombre: Díaz Padilla Danny Sebastián **Fecha:** 26/01/2020

1. Objetivos:

1.1. Objetivo General

Agrupar conjuntos de datos utilizando la herramienta WEKA y una hoja de cálculo para resolver un ejercicio de Datamaning por Bramer.

1.2. Objetivos Específicos

Realizar el proceso de iteraciones para determinar los centroides de clústers manualmente.

Calcular la distancia objetivo usando la suma de las últimas distancias al calcular los centroides.

2. Marco teórico:

WEKA

Es un software programado en Java que está orientado a la extracción de conocimientos desde bases de datos con grandes cantidades de información.[1]

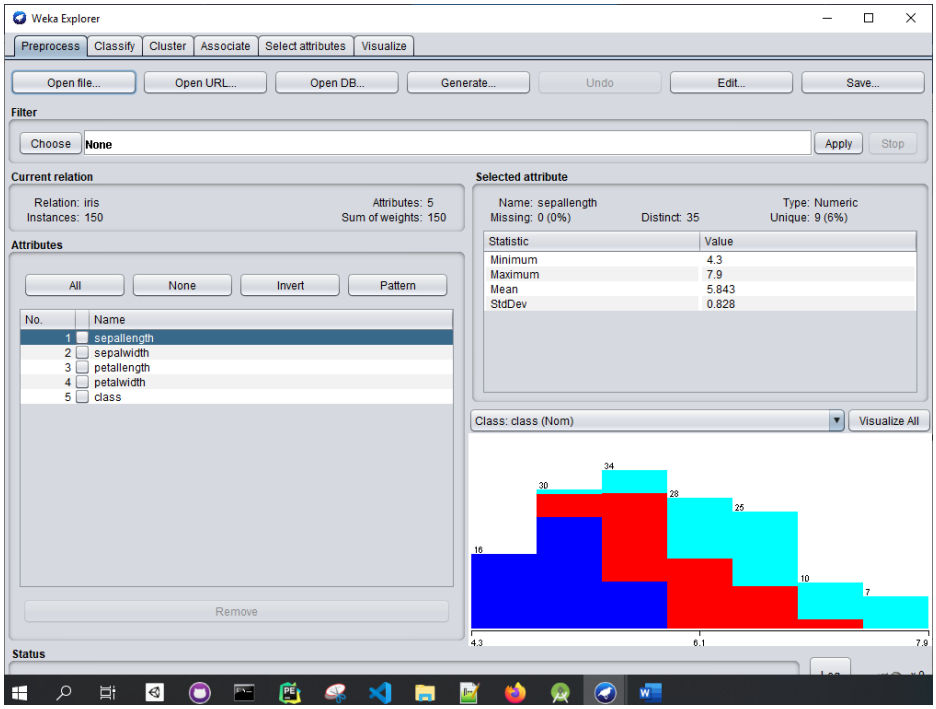
Clustering

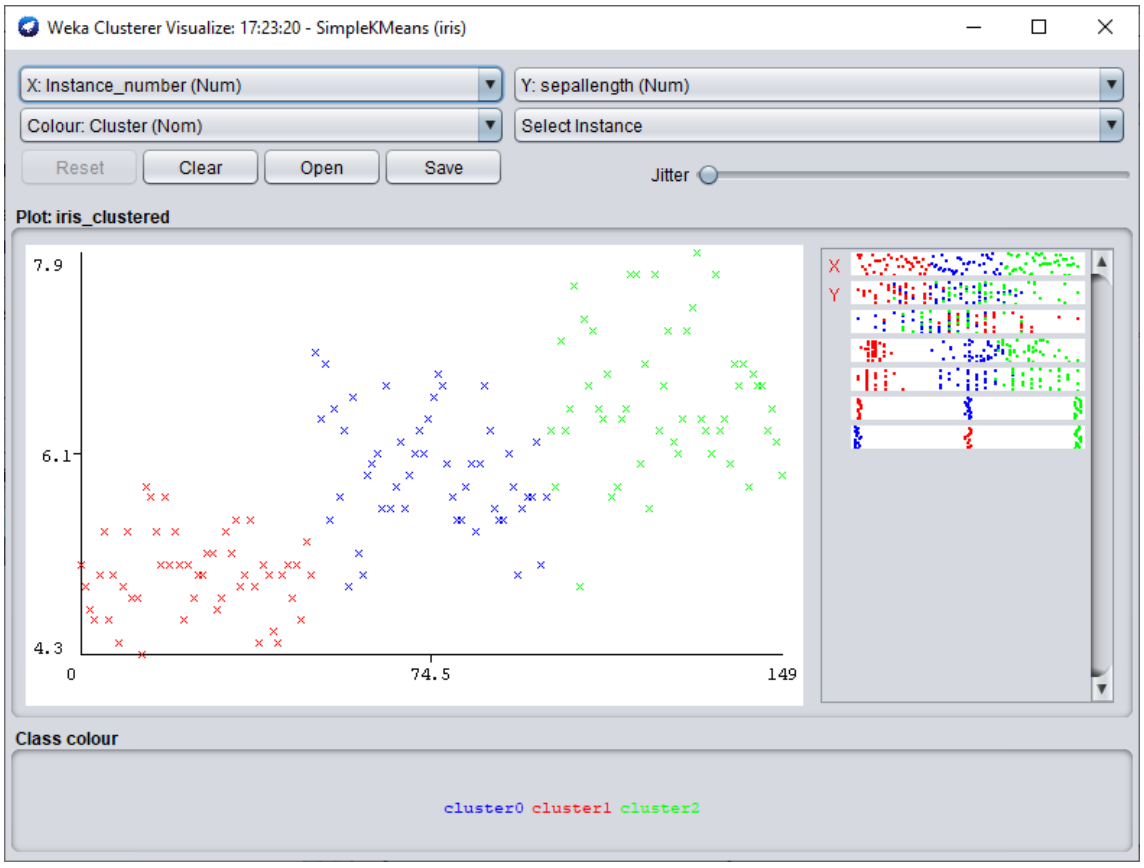
La agrupación es la agrupación de objetos específicos en función de sus características y similitudes. En cuanto a la minería de datos, esta metodología divide los datos que mejor se adaptan al análisis deseado utilizando un algoritmo de unión especial. Este análisis permite que un objeto no sea parte o estrictamente parte de un clúster, lo que se denomina partición dura de este tipo. Sin embargo, las particiones suaves sugieren que cada objeto en el mismo grado pertenece a un grupo. Se pueden crear divisiones más específicas como objetos de múltiples grupos, se puede obligar a un solo grupo a participar o incluso se pueden construir árboles jerárquicos en las relaciones grupales. [2]

3. Desarrollo de la práctica:

Primer ejercicio

- Utilizando el archivo iris.arff generar 3 clusters usando el algoritmo Kmeans.





- Escoger como medida de distancia la distancia euclidiana

- Indicar el número de instancias en cada cluster

```
Clustered Instances

0       50 ( 33%)
1       50 ( 33%)
2       50 ( 33%)
```

- Los centroides iniciales

```
Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor
Cluster 2: 6.9,3.1,5.1,2.3,Iris-virginica
```

- Los centroides finales

```
Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0              1              2
                          (150.0)          (50.0)         (50.0)         (50.0)
================================================================================
sepallength                5.8433           5.936          5.006          6.588
sepalwidth                  3.054            2.77          3.418          2.974
petallength                3.7587            4.26          1.464          5.552
petalwidth                 1.1987           1.326          0.244          2.026
class                 Iris-setosa Iris-versicolor     Iris-setosa  Iris-virginica
```

- La suma de los cuadrados de los errores

```
Within cluster sum of squared errors: 7.817456892309574
```

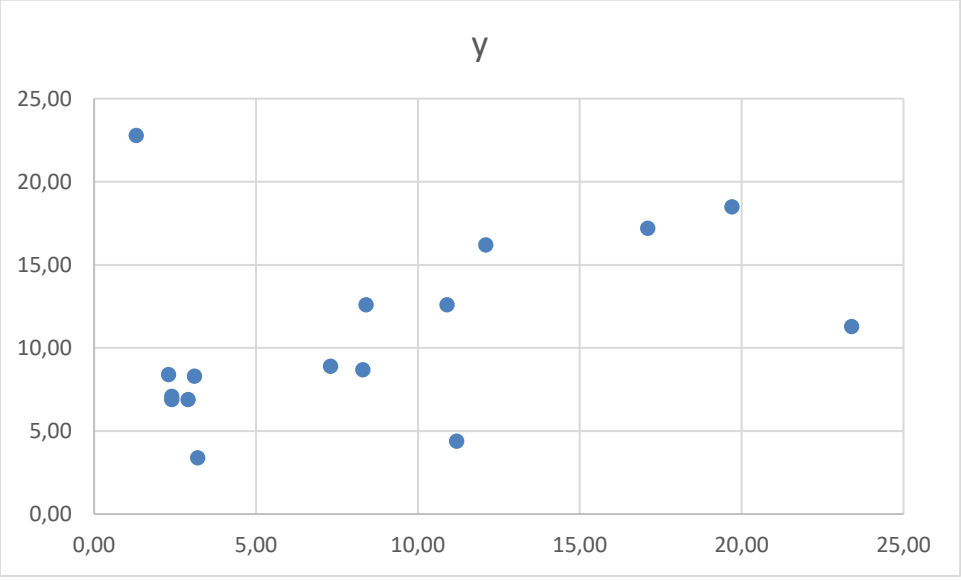
Ejercicio 1 Capítulo 19, Principles of Datamining, Bramer

1. Usando el método que se muestra en la Sección 19.2, agrupe los siguientes datos en tres grupos, utilizando el método k-means.

Datos

Punto	x	y	Cluster final	Cluster final	Cluster final
1	10,90	12,60	2	2	2
2	2,30	8,40	1	1	1
3	8,40	12,60	1	2	2
4	12,10	16,20	3	3	3
5	7,30	8,90	1	2	2
6	23,40	11,30	3	3	3
7	19,70	18,50	3	3	3
8	17,10	17,20	3	3	3
9	3,20	3,40	1	1	1
10	1,30	22,80	1	1	1
11	2,40	6,90	1	1	1
12	2,40	7,10	1	1	1
13	3,10	8,30	1	1	1
14	2,90	6,90	1	1	1
15	11,20	4,40	2	2	2
16	8,30	8,70	2	2	2

Representación



Centroides de las 3 iteraciones

Centroide 1	
2,40	6,90

Centroide 2	
11,20	4,40

Centroide 3	
19,70	18,50

Nuevo centroide 1	
3,70	9,48

Nuevo centroide 2	
10,13	8,57

Nuevo centroide 3	
18,08	15,80

Nuevo centroide 1	
2,51	9,11

Nuevo centroide 2	
9,22	9,44

Nuevo centroide 3	
18,08	15,80

Iteración 1

Distancia 1	Distancia 2	Distancia 3	Cluster final
10,23	8,21	10,59	2
1,50	9,76	20,12	1
8,28	8,66	12,75	1
13,44	11,83	7,94	3
5,29	5,95	15,68	1
21,46	14,02	8,10	3
20,83	16,46	0,00	3
17,95	14,09	2,91	3
3,59	8,06	22,37	1
15,94	20,89	18,90	1
0,00	9,15	20,83	1
0,20	9,20	20,72	1
1,57	8,99	19,48	1
0,50	8,67	20,42	1
9,15	0,00	16,46	2
6,17	5,19	15,03	2

Iteración 2

Distancia 1	Distancia 2	Distancia 3	Respecto al anterior	
7,847819545	4,105551797	7,856247514	2	Igual
1,766806424	7,835106182	17,4244261	1	Igual
5,642541236	4,390013921	10,19046736	2	Diferente
10,75863707	7,882610961	5,988374153	3	Igual
3,646070098	2,852873795	12,79494529	2	Diferente
19,78409699	13,54531489	6,971773447	3	Igual
18,36846466	13,79101962	3,151289419	3	Igual
15,46585646	11,09364182	1,706055392	3	Igual

6,098309825	8,646707787	19,36557835	1	Igual
13,53667629	16,75158367	18,17692562	1	Igual
2,887029316	7,910892631	18,02541608	1	Igual
2,70994966	7,871185143	17,92751028	1	Igual
1,321801987	7,03838681	16,74815288	1	Igual
2,69906248	7,42286258	17,59234564	1	Igual
9,057252738	4,301033468	13,3126115	2	Igual
4,665290802	1,838175424	12,08141651	2	Igual

Iteración 3

Distancia 1	Distancia 2	Distancia 3	Cluster final	Respecto al anterior
9,081321939	3,578826623	7,856247514	2	Igual
0,745736179	6,997713912	17,4244261	1	Igual
6,840455886	3,264659247	10,19046736	2	Igual
11,92028797	7,347924877	5,988374153	3	Igual
4,790509325	1,994492417	12,79494529	2	Igual
20,99977162	14,30146846	6,971773447	3	Igual
19,58163446	13,85330286	3,151289419	3	Igual
16,67698524	11,05947558	1,706055392	3	Igual
5,755281514	8,527719508	19,36557835	1	Igual
13,73947835	15,53113003	18,17692562	1	Igual
2,217233061	7,277636979	18,02541608	1	Igual
2,017525257	7,210270453	17,92751028	1	Igual
1,003056553	6,225271078	16,74815288	1	Igual
2,247629136	6,811314117	17,59234564	1	Igual
9,882617186	5,414979224	13,3126115	2	Igual
5,800527773	1,180677771	12,08141651	2	Igual

función objetivo 1	función objetivo 2	función objetivo
27,72594005	15,43363528	17,817492

Clústers finales para los centroides:

Inicial

Centroide 1		Centroide 2		Centroide 3	
2,40	6,90	11,20	4,40	19,70	18,50

Final

Nuevo centroide 1		Nuevo centroide 2		Nuevo centroide 3	
2,51	9,11	9,22	9,44	18,08	15,80

Agrupando los datos de los puntos

Cluster		
1	2	3
2,9,10,11,12,13,14	1,3,5,15,16	4,6,7,8

4. Análisis de resultados:

La segunda iteración reestructuró bastante los datos sin embargo la tercera iteración dio los mismos resultados y por lo tanto el algoritmo termina y se limita.
Las agrupaciones finales dependen bastante de los centroides iniciales por lo que es importante realizar la gráfica de forma que se pueda iniciar en puntos que otorguen relevancia e información.

5. Conclusiones y recomendaciones:

- Se realizó la tarea de clustering usando WEKA.
- Con Excel se realizó las iteraciones necesarias para obtener una agrupación de datos coherente, por medio de hojas de cálculo.
- Se calculó una distancia objetivo de los tres clústers generados en el ejercicio de Bramer.
- Datamining utiliza mucho el cálculo de los errores cuadrados para reestructurar sus modelos en el tiempo.

- Se recomienda utilizar pestañas distintas en Excel para ubicar: Datos, Cálculos y gráficas.
- Para calcular la pertenencia de un punto a un clúster es útil usar operaciones lógicas (programación en Excel).

6. Bibliografía:

[1]"Introducción a la minería de datos con Weka", *Locualo.net*, 2007. [Online]. Available: http://www.locualo.net/programacion/introduccion-mineria-datos-weka/00000018.aspx. [Accessed: 21- Jan- 2020].

[2]"What is Clustering in Data Mining? | Application of clustering in data mining", *EDUCBA*. [Online]. Available: https://www.educba.com/what-is-clustering-in-data-mining/. [Accessed: 21- Jan- 2020].