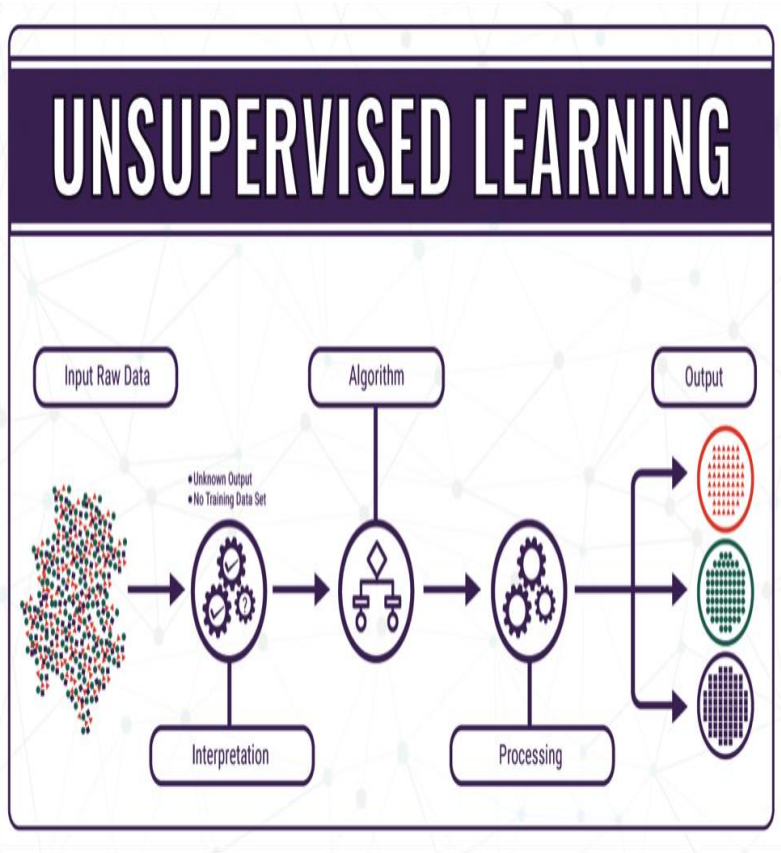
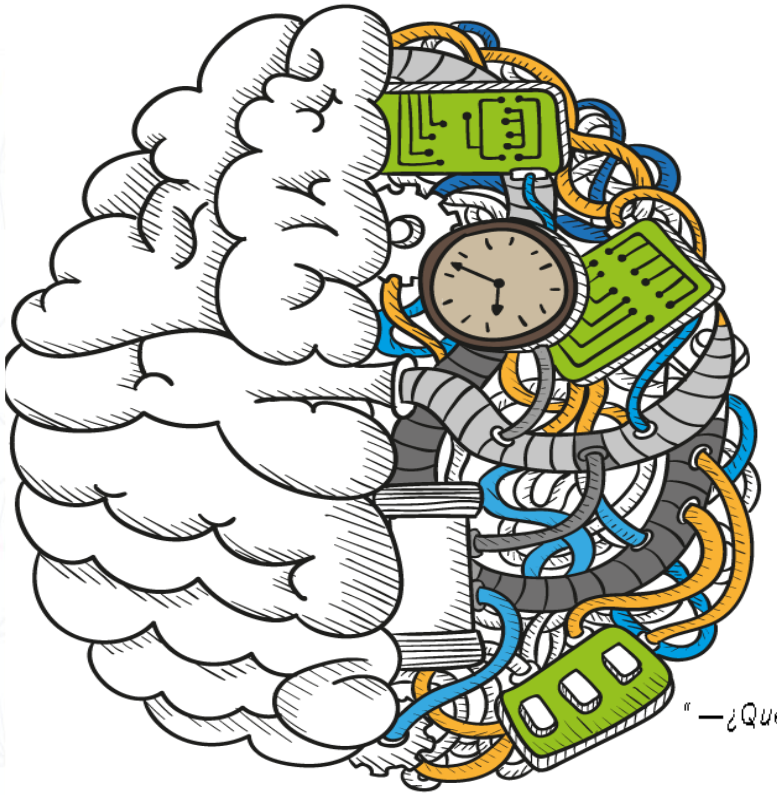
INTRODUCCIÓN AL APRENDIZAJE AUTOMÁTICO

APRENDIZAJE NO SUPERVISADO – AGRUPAMIENTO
(K – means, HAC, SOM)

LAURA DIAZ DÁVILA – FRANCISCO TAMARIT

ML:MODELOS Y MÁS MODELOS





" —¿Qué sabes de este asunto?— preguntó el Rey a Alicia.

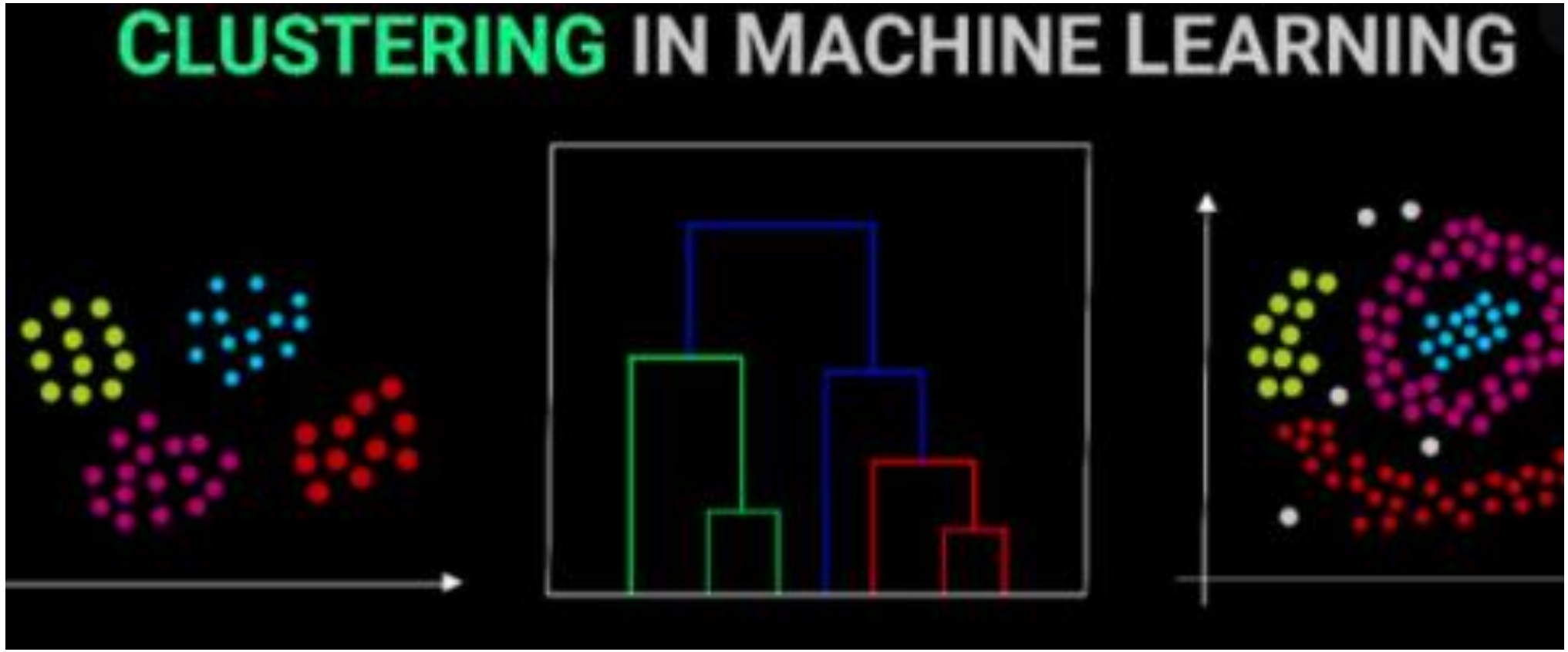
—Nada— dijo Alicia.

—¿Absolutamente nada?— insistió el Rey.

—Absolutamente nada— dijo Alicia.

—Esto es importante— dijo el Rey, volviéndose hacia los jurados."

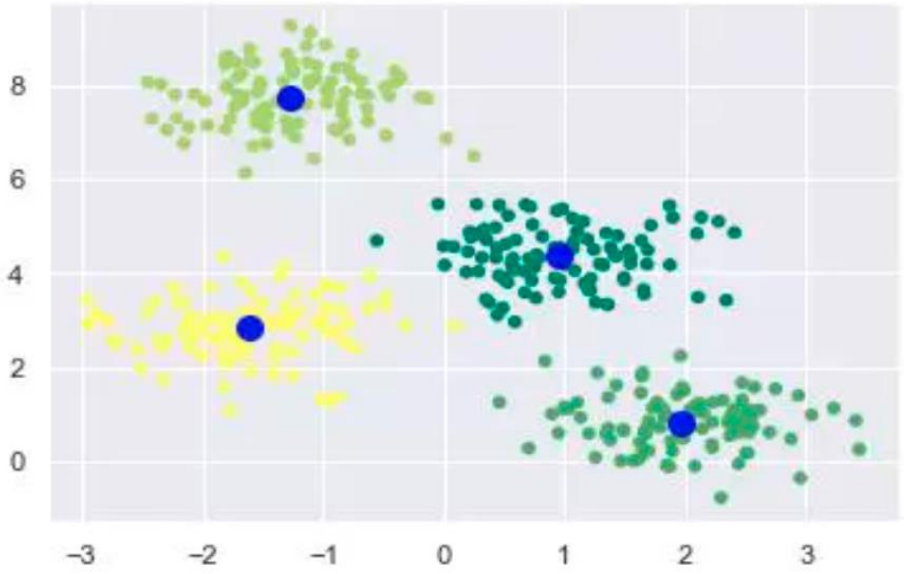
Lewis Carroll, Alicia en el país de la maravillas, en capítulo XII, La Declaración de Alicia (1865)


CLUSTERING IN MACHINE LEARNING

CLUSTERING CON K- MEANS

OBTENDREMOS CENTROIDES

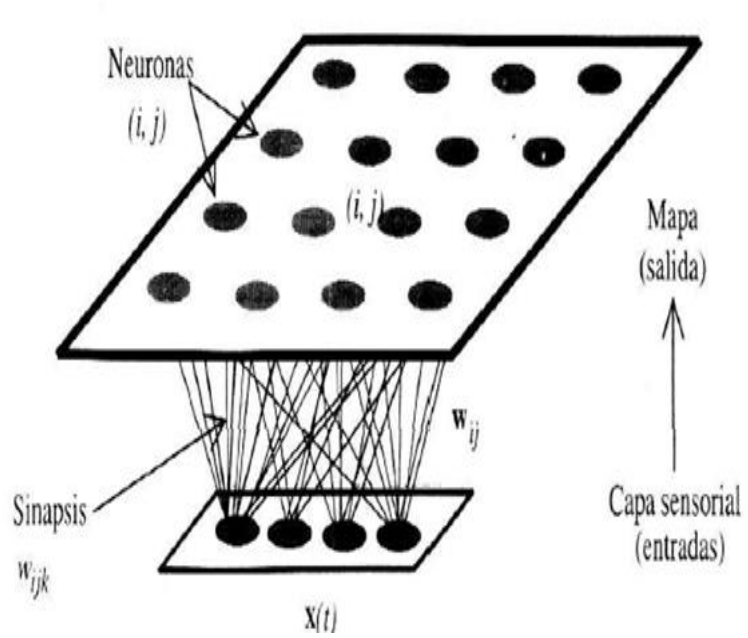
Y CONJUNTOS O AGRUPAMIENTOS DE LOS EJEMPLOS A LOS QUE SE ENFRENTE EL ALGORITMO



AL FINALIZAR, EL ALGORITMO HA ETIQUETADO AUTOMÁTICAMENTE LOS DATOS ASIGNÁNDOLES UNA PARTICIÓN O CLUSTER

MODELO DE MAPAS AUTOORGANIZADOS (SOM) KOHONEN – R.N.A.

MODELO SOM DE KOHONEN: *APRENDIZAJE NO SUPERIVSADO*. ***APRENDE A AGRUPAR LOS EJEMPLOS SEGÚN LAS SIMILITUDES QUE DESCUBRE ENTRE ELLOS. CLUSTERING***



EJEMPLO

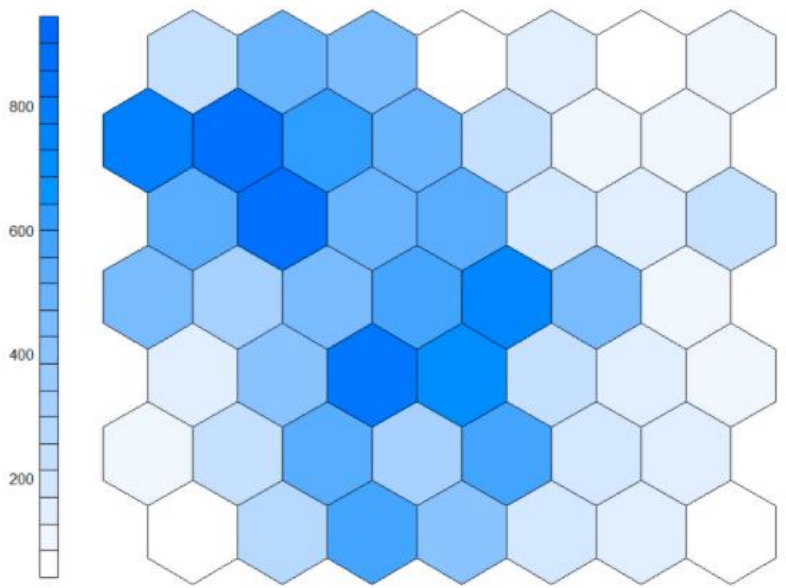
DE REPOSITORIO UCI

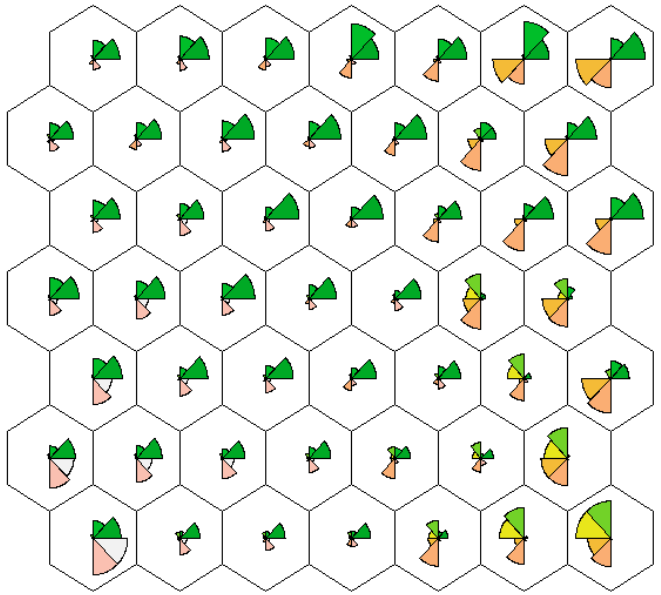
HTRU2 Data Set

17897 observaciones con 8 variables

X140.5625	X55.68378214	X.0.234571412	X.0.699648398	X3.199832776	X19.11042633	X7.975531794	X74.24222492	X0
102.50781	58.88243	0.465318154	-0.515087909	1.6772575	14.860146	10.5764867	127.3935796	0
103.01562	39.34165	0.323328365	1.051164429	3.1212375	21.744669	7.7358220	63.1719091	0
136.75000	57.17845	-0.068414638	-0.636238369	3.6429766	20.959280	6.8964989	53.5936607	0
88.72656	40.67223	0.600866079	1.123491692	1.1789298	11.468720	14.2695728	252.5673058	0
93.57031	46.69811	0.531904850	0.416721117	1.6362876	14.545074	10.6217484	131.3940043	0
119.48438	48.76506	0.031460220	-0.112167573	0.9991639	9.279612	19.2062302	479.7565669	0
130.38281	39.84406	-0.158322759	0.389540448	1.2207358	14.378941	13.5394560	198.2364565	0

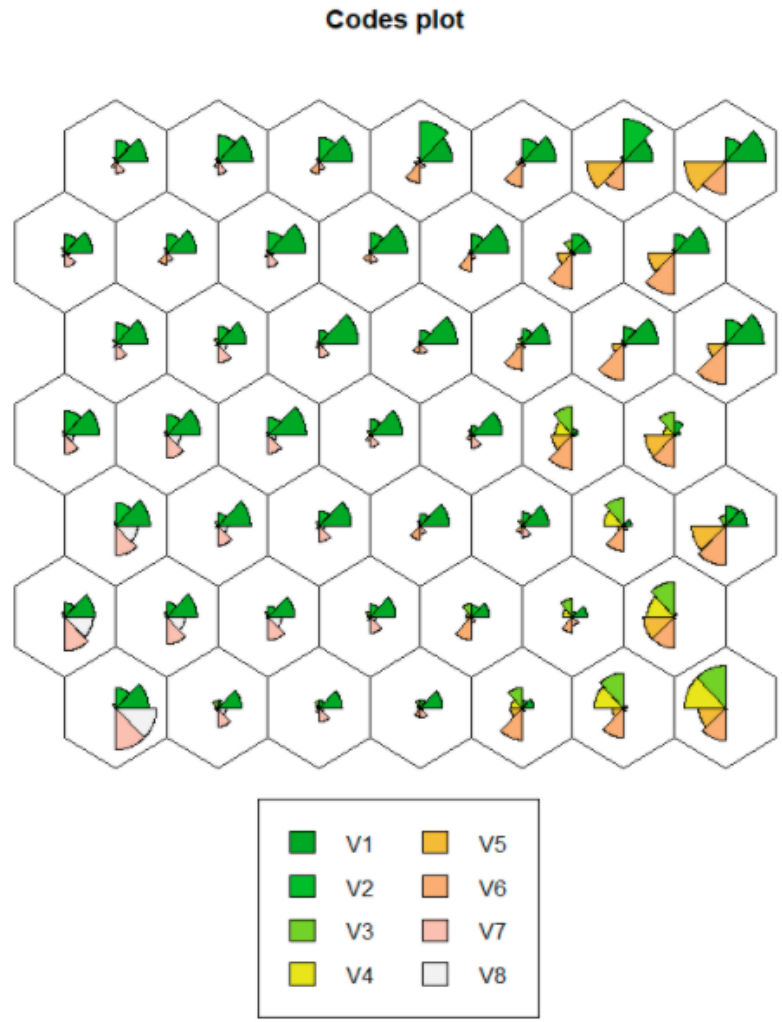
VISUALIZACIÓN DE LOS CLUSTERS





Ejemplo de uso de un Mapa Auto-Organizado (SOM) de Kohonen en R

```
library(kohonen)
library(dplyr)
library(plot3D)
library(plot3Drgl)
```



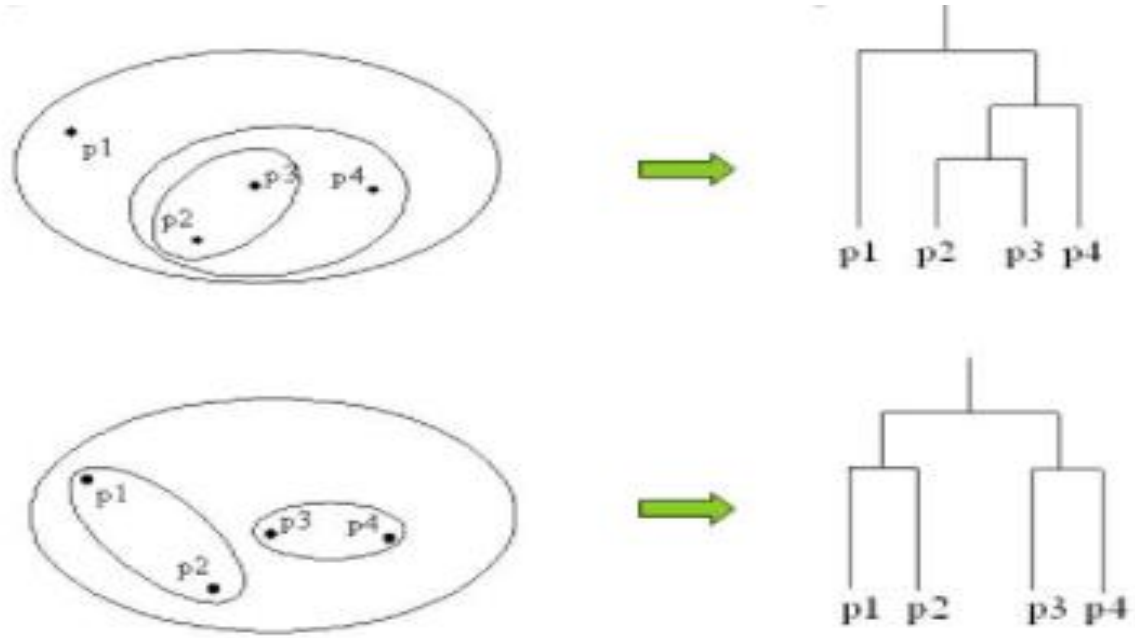
http://exponentis.es/ejemplo-de-uso-de-un-mapa-auto-organizad

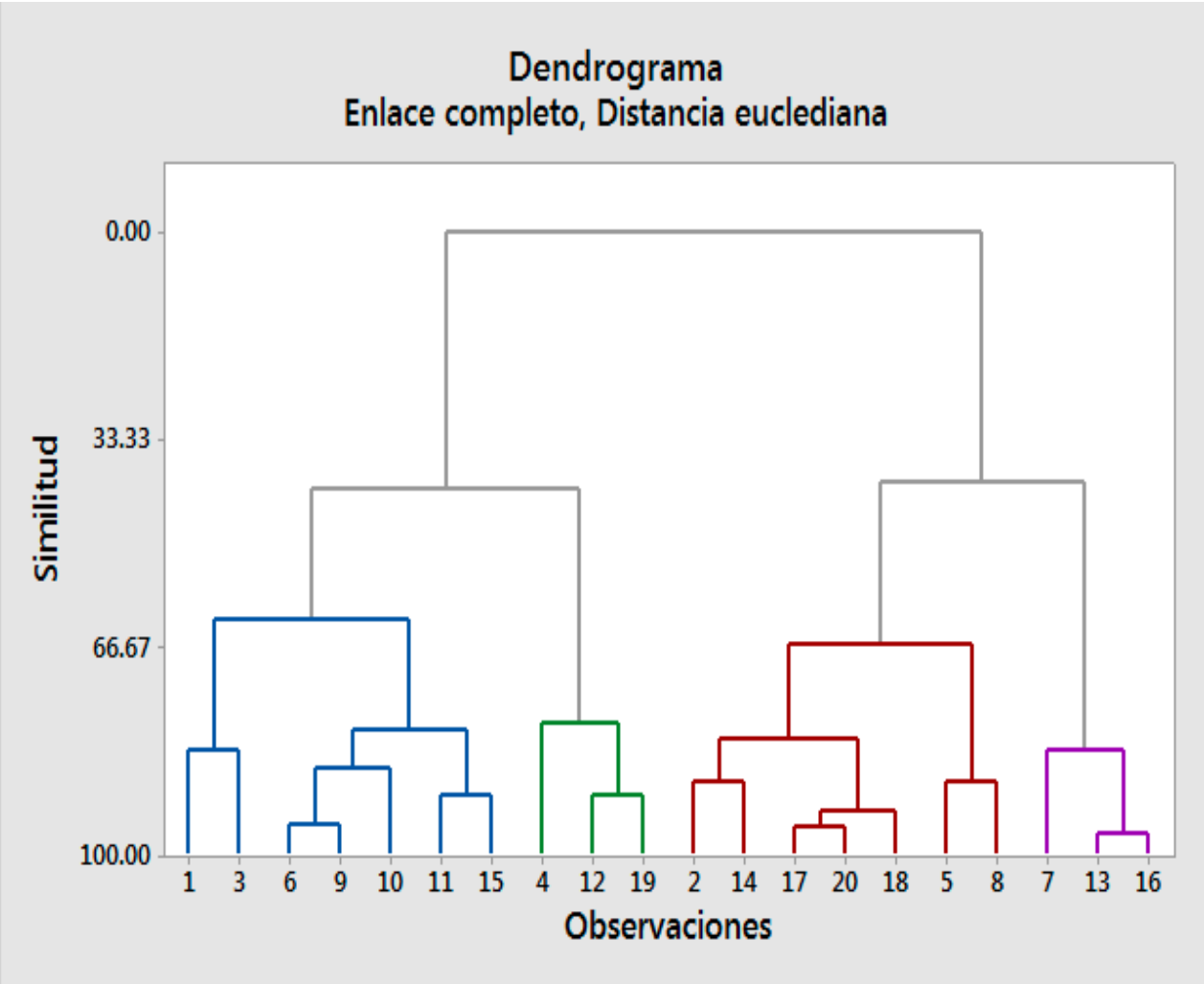
CLUSTERING JERÁRQUICO: H.A.C.

AGRUPACIÓN JERÁRQUICA AGLOMERATIVA (DE ABAJO HACIA ARRIBA)

AGRUPACIÓN JERÁRQUICA DIVISIVA (DE ARRIBA HACIA ABAJO)

UN CONJUNTO DE CLUSTERS ANIDADOS, ORGANIZADOS COMO UN ÁRBOL JERÁRQUICO, CON ÚNICO CLUSTER ARRIBA, AGRUPANDO TODOS LOS INDIVIDUOS Y CLUSTERS CON UN SOLO ELEMENTO ABAJO



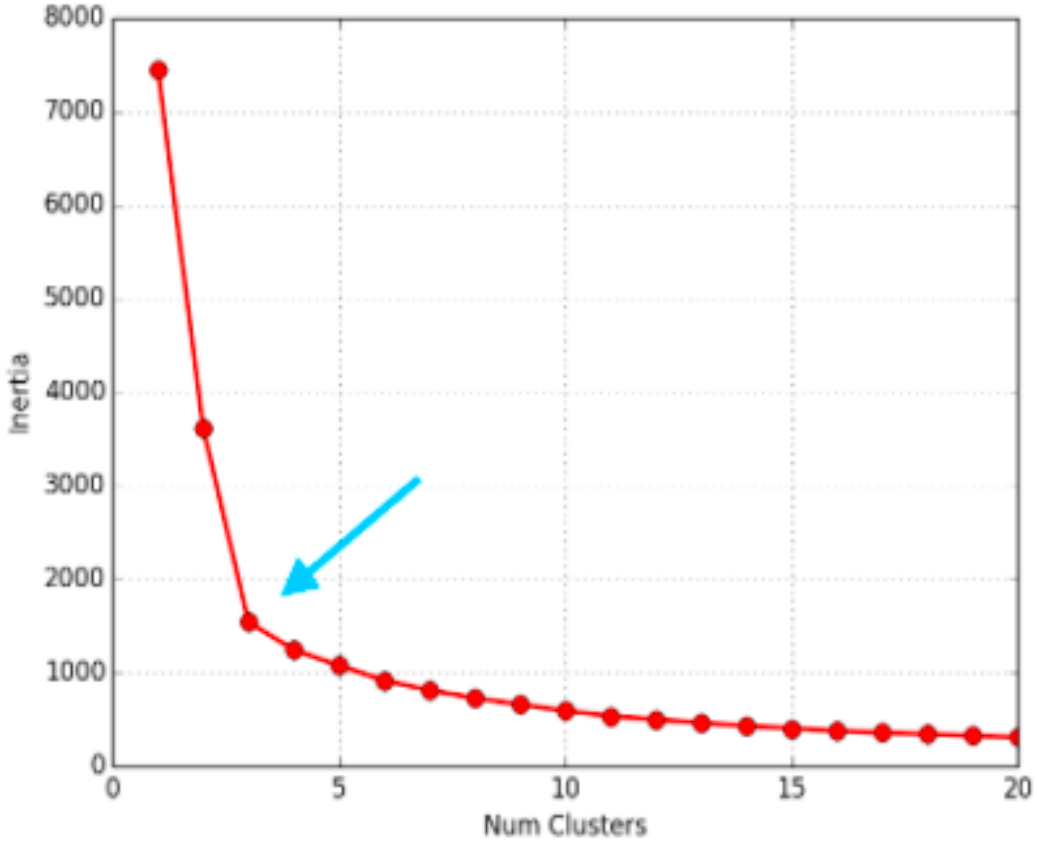


El algoritmo es comparativamente más lento y no escala bien para grandes conjuntos de datos

El algoritmo HAC es sensible a valores atípicos

https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-observations/interpret-the-results/all-statistics-and-graphs/dendrogram/

¿Cuántos clusters?
Elbow Method



La inercia como métrica para determinar la cantidad óptima de clusters, en base a varias iteraciones de K-Means

$$Inercia = \sum_{i=0}^{N} \|x_i - \mu\|^2$$

https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_clustering_metrics.html

2.3. Clustering

2.3.1. Overview of clustering methods

2.3.2. K-means

2.3.3. Affinity Propagation

2.3.4. Mean Shift

2.3.5. Spectral clustering

2.3.6. Hierarchical clustering

2.3.7. DBSCAN

2.3.8. OPTICS

2.3.9. BIRCH

2.3.10. Clustering performance evaluation

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters, inductive	Distances between points
Ward hierarchical clustering	number of clusters or distance threshold	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, transductive	Distances between points
Agglomerative clustering	number of clusters or distance threshold, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances, transductive	Any pairwise distance

https://scikit-learn.org/stable/modules/clustering.html

2.3. Clustering

2.3.1. Overview of clustering methods

2.3.2. K-means

2.3.3. Affinity Propagation

2.3.4. Mean Shift

2.3.5. Spectral clustering

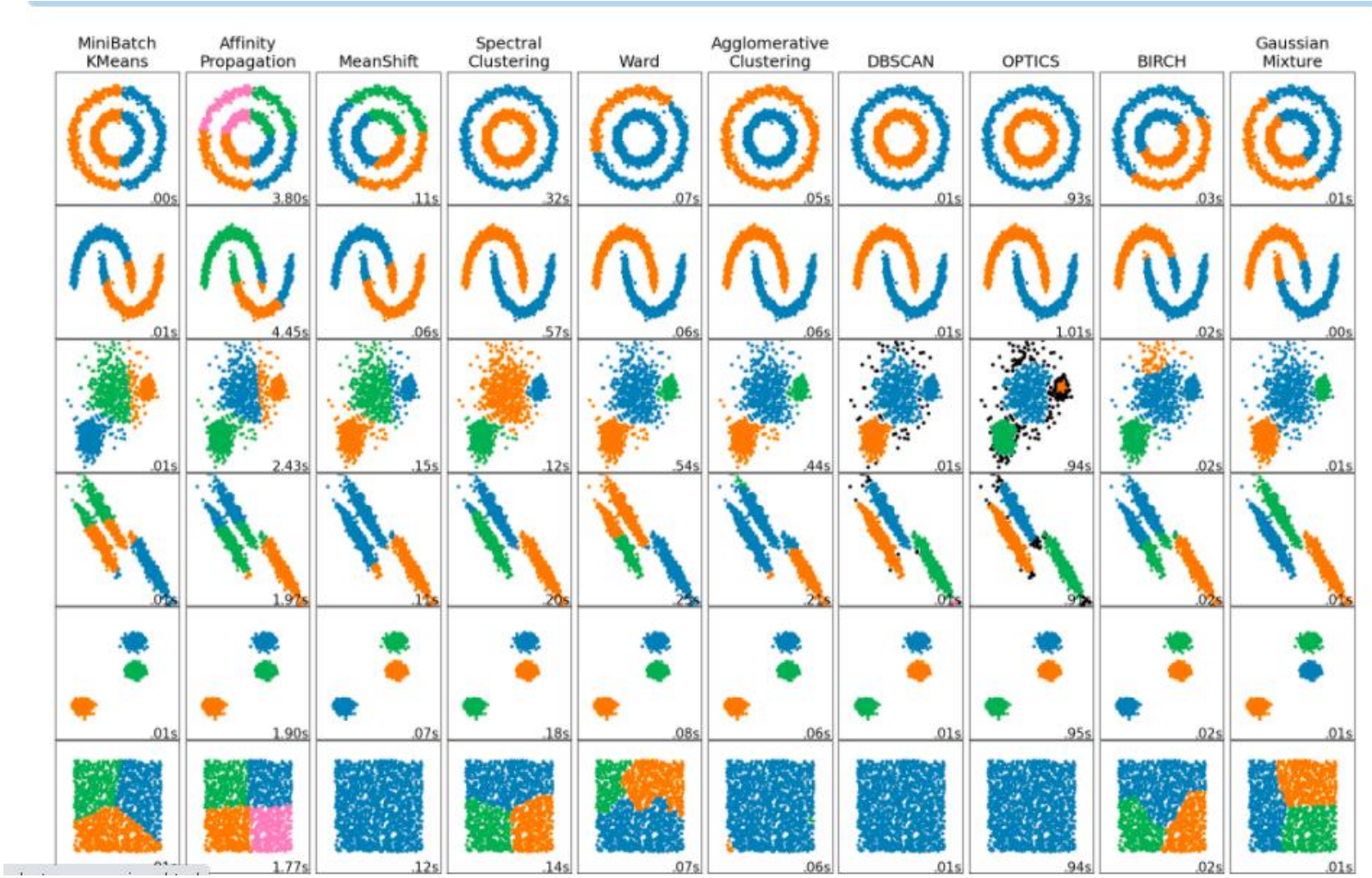
2.3.6. Hierarchical clustering

2.3.7. DBSCAN

2.3.8. OPTICS

2.3.9. BIRCH

2.3.10. Clustering performance evaluation



https://scikit-learn.org/stable/modules/clustering.html

DISTANCIAS GEOMÉTRICAS EN CLUSTERING CON AGLOMERACIÓN EN ADYACENCIAS POR AFINIDAD:

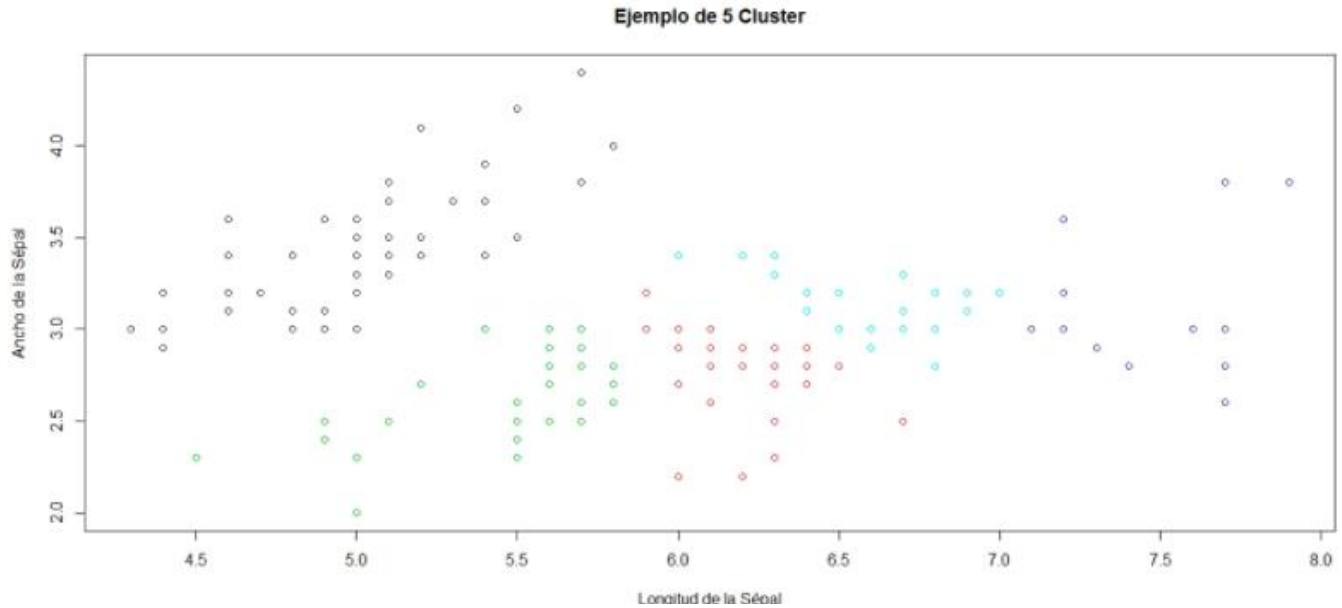
DISTANCIA COSENO

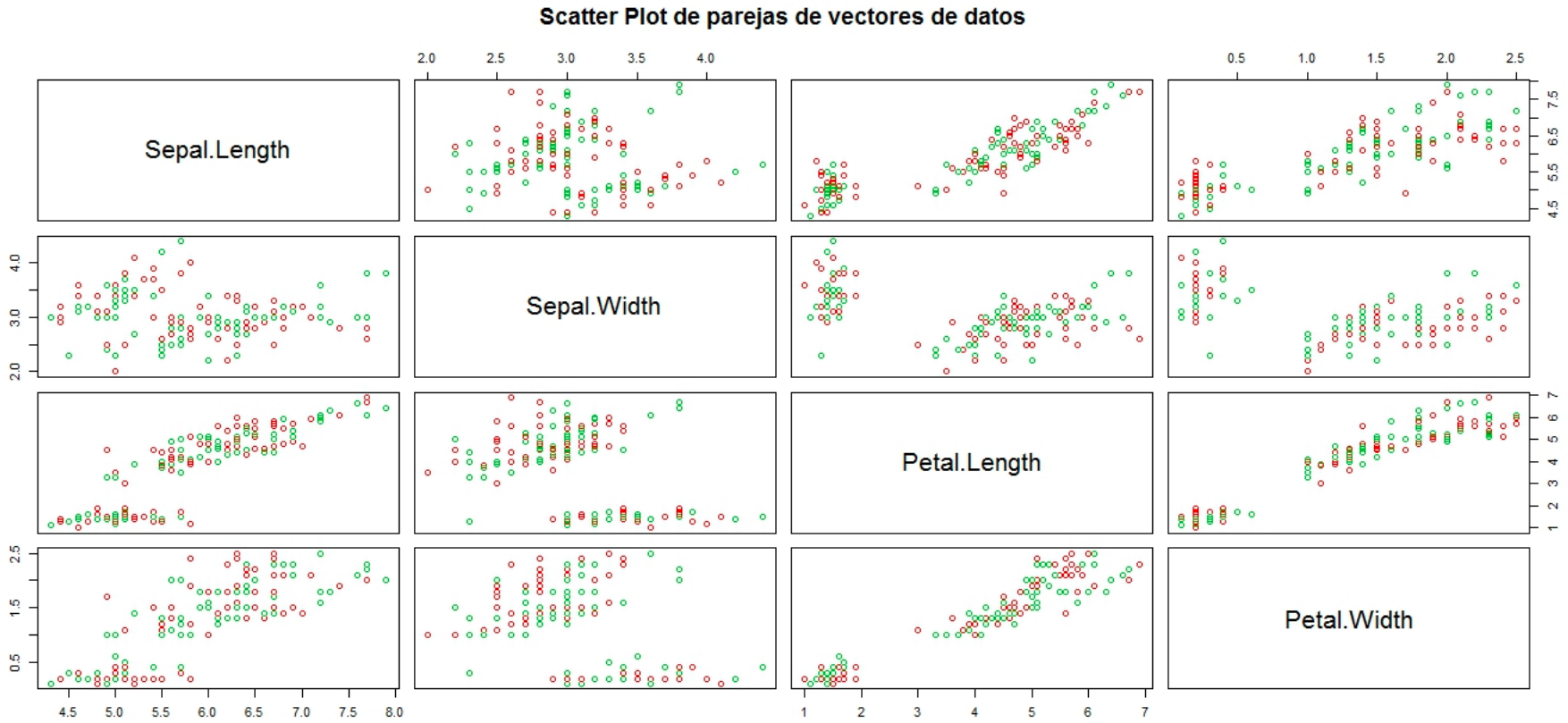
DISTANCIA ECUCLIDEA

Agglomerative clustering with different metrics

https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_clustering_metrics.html

```
#Scatter Plot
data(iris)
```





LAS ZONAS OSCURAS DE LAS INTERPRETACIONES EN LA VISUALIZACIÓN

EJEMPLO

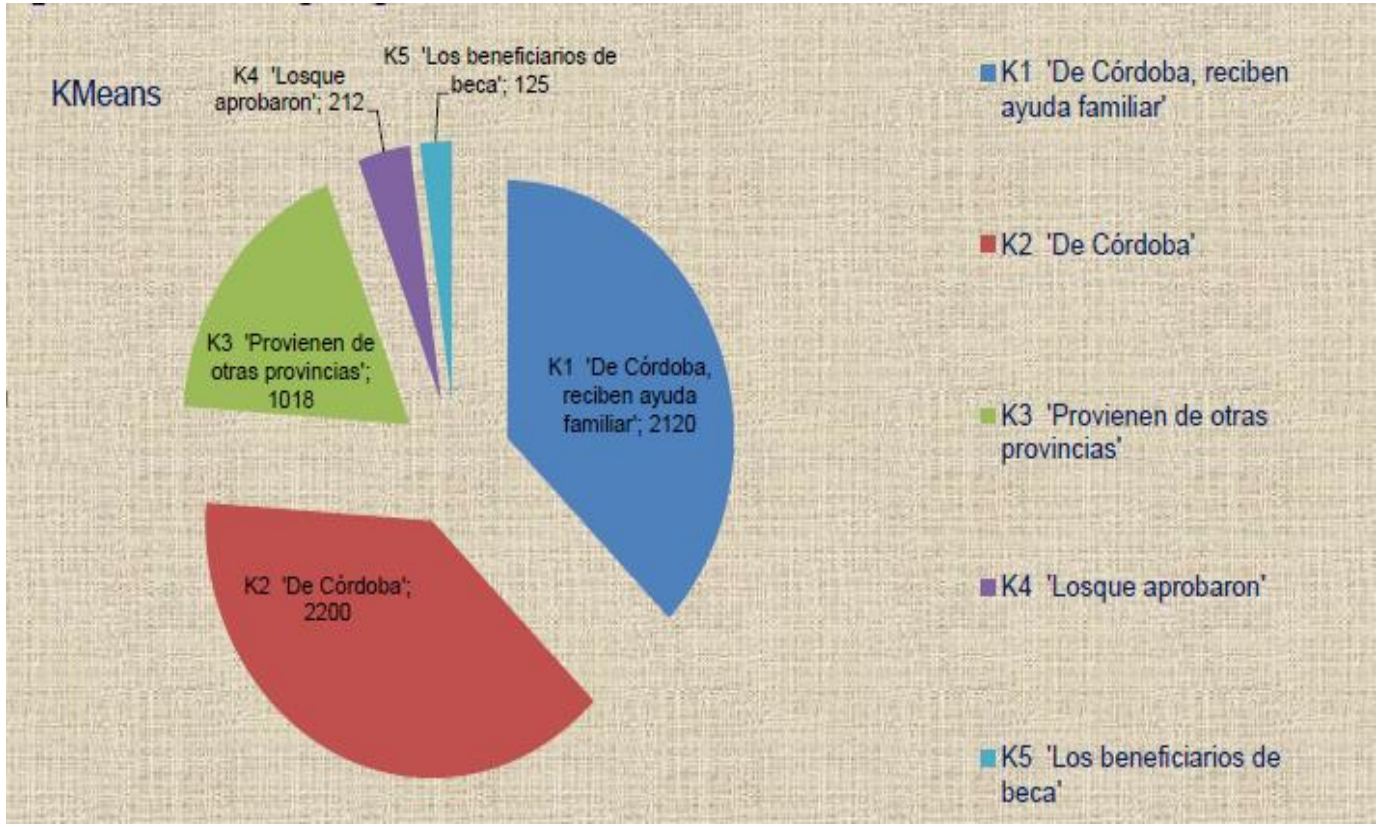
Explotación de Información Aplicada a la Caracterización de Patrones Socio-Económicos de la Población Estudiantil de Carreras de Ciencias Económicas

Base de Datos de SIU_Guaraní, 6500 registros de estudiantes que se inscribieron a la asignatura **Administración y Sistemas de Información Gubernamental** en los años 2012 a 2014

Variables:

- Procedencia del sustento económico del alumno: trabajo propio, familia y/o de beca. Dicotómicas.
- Últimos estudios formales alcanzados por su padre y madre, representados por una escala de 0 a 4, 0: no posee, 1: primario completos o secundario incompleto, 2: secundario completo o superior incompleto, 3: superior completo y 4: posee estudios de posgrado.
- Género del estudiante (1: Masculino, 0: Femenino).
- Ubicación de procedencia. Tres variables: Argentino, de la Provincia de Córdoba, y de Córdoba Capital. Booleanas.
- Aprobó la materia durante el mismo año que realizo la cursada (1 / 0).
- Cursa la asignatura acorde a lo establecido en el plan de estudios (1 / 0).
- Rendimiento académico en su primer año de ingreso (de 0 a 3 acorde a la cantidad de materias que cursó)
- Desempeño respecto al plan de estudios (de 0 a 4, materias aprobadas respecto del plan de estudios y del año en el que ingresaron).

¿Cómo se caracterizan los estudiantes de la carrera de Contador Público de la Facultad de Ciencias Económicas de la Universidad Nacional de Córdoba, tomando a la asignatura Administración y Sistemas de Información Gubernamental como eje para el análisis?



Para todos los grupos descubiertos:

el nivel de estudios de la madre es levemente superior al del padre,

el género de los estudiantes no parece tener mayor relevancia en su desempeño,

a los estudiantes que trabajan, importante cantidad, se les dificulta más sostener el plan de carrera, como así también a los de nivel económico bajo y a los que no proceden de Córdoba.

¿PREGUNTAS?

¡GRACIAS!