A continuación se van a exponer las técnicas de CLUSTERING llevadas a cabo en el proyecto IBERINFORM sobre CREDIT SCORING desarrollado por nuestro equipo en el IBM SPSS MODELLER para la cual hemos estimado los siguientes algoritmos de clusterización:

* KOHONEN
* K-MEDIAS
* CLUSTER BIETÁPICO

Una vez escogidas las diferentes opciones o técnicas de CLUSTERING, vamos a comentar por que se han escogido y la calidad de las mismas teniendo en cuenta la METADATA que es objeto de nuestro estudio IBERINFORM sobre CREDIT SCORING atendiendo a las distintas variables.

* INTRODUCCIÓN CLUSTERING

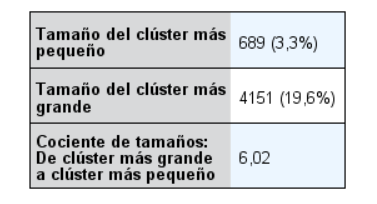
También conocido como agrupamiento, es una de las técnicas de [minería de datos](https://www.ecured.cu/Miner%C3%ADa_de_datos) que consiste en la división de los datos de manera **NO SUPERVISADA** en grupos de objetos con características similares. Cuando se representa la información obtenida a través de clusters se pierden algunos detalles de los datos, pero a la vez se simplifica dicha información.

De forma general, las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se colocan en clases que son muy similares internamente (entre los miembros de la misma clase) y a la vez diferente entre los miembros de las diferentes clases.

* ***KOHONEN***

Los [mapas autoorganizativos de ***Kohonen*** son un algoritmo](http://en.wikipedia.org/wiki/Kohonen), a veces agrupado dentro de las [redes neuronales](http://es.wikipedia.org/wiki/Redes_Neuronales), que a partir de un proceso iterativo de comparación con un conjunto de datos y cambios para aproximarse a los mismos, crea un modelo de esos mismos datos que puede servir para agruparlos por criterios de similitud; adicionalmente, este agrupamiento se produce de forma que la proyección de estos datos sobre el mapa distribuya sus características de una forma gradual. El Mapa de Kohonen se conoce como **SOM** (*self-organizing map*, mapa autoorganizativo) o **SOFM** (*self-organizing feature map*, mapa autoorganizado de características).

Como podemos observar en los resultados obtenidos en nuestro Diagrama SPSS Modeller, este nos aporta un Modelo con 12 clusters y buena silueta por lo que tiene buena calidad y nos dice que la variable “***Años o Fecha de Constitución***” es la que más repercute o tiene mayor importancia predictora hacia nuestro Target binario de agrupamiento (NO SUPERVISADO), sin embargo tiene poca homogeneidad en la distribución, siendo el tamaño del clúster más pequeño un 3,3% mientras que el mayor ocupa un 19,6% y de ahí esa deficiencia en la homogeneidad.



Por lo tanto, a continuación para proseguir nuestro análisis vamos a probar con un cluster no jerárquico como puede ser el K-Means o K-Medias y ver que diferencias ofrece.

* ***Simple K-Means***

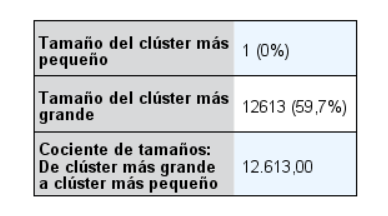
Este algoritmo debe definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar. Los pasos básicos para aplicar el algoritmo son muy simples. En primer lugar se determina la cantidad de clusters en los que se quiere agrupar la información, en este caso las simulaciones. Luego se asume de forma aleatoria los centros por cada clusters. Una vez encontrados los primeros centroides el algoritmo hará los tres pasos siguientes:

1. Determina las coordenadas del centroide.
2. Determina la distancia de cada objeto a los centroides.
3. Agrupa los objetos basados en la menor distancia.

Finalmente quedarán agrupados por clusters, los grupos de simulaciones según la cantidad de clusters que el investigador definió en el momento de ejecutar el algoritmo.

En nuestro estudio hemos observado que este algoritmo se ha usado buscando una buena calidad pero las similitudes de los porcentajes en cada clúster son muy desiguales por lo tanto no va a ser nuestro clúster óptimo al no particionar muy bien.

Como podemos observar en los resultados obtenidos en nuestro Diagrama SPSS Modeller, este K-MEANS nos aporta un Modelo con 8 clusters y buena silueta por lo que también tiene buena calidad y también nos dice que la variable “***Años o Fecha de Constitución***” es de las que tiene mayor importancia predictora hacia nuestro Target binario de agrupamiento pero no la diferencia del resto de variables como “***Fondos Propios, Endeudamiento a corto plazo o Ventas***” , cosa que si hizo el anterior algoritmo de Kohonen, y por lo tanto mucha menos homogeneidad en la distribución que el modelo anterior, siendo el tamaño del clúster más pequeño un 0,0% mientras que el mayor ocupa un 59,7% y de ahí esa deficiencia demasiado significativa en la homogeneidad que nos concluye que tampoco este algoritmo va a ser el óptimo en nuestro análisis.



Para finalizar el análisis clúster de este trabajo y buscando siempre la optimización de las variables, vamos a llevar a cabo el estudio con un clúster Bietápico:

* ***BIETÁPICO***

Los métodos anteriormente descritos presentan problemas cuando hay un número de individuos muy grande para ser clasificados como es el caso de nuestro estudio de minería de datos.

Como alternativa surge el Método “Two-Step” o Bietápico” en dos fases que ha conseguido hacerse muy popular porque es único y además esta implementado en el SPSS de IBM.

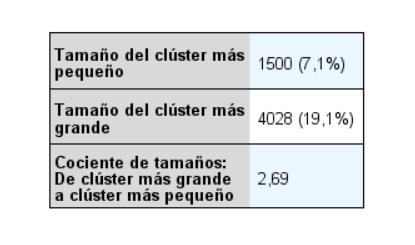
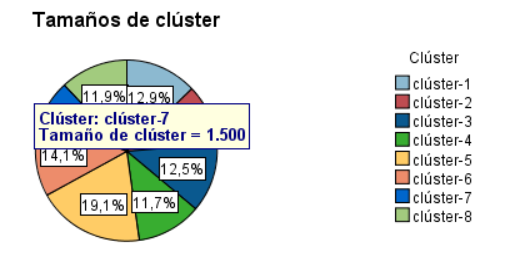
Los tres aspectos fundamentales y diferenciadores del resto de técnicas de clúster son: permite trabajar con información mixta (variables categóricas y continuas), realiza una selección automática del numero de clusteres y permite el análisis de grandes volúmenes de datos.

El algoritmo va a determinar por tanto de forma automática el numero óptimo de clusters.

Dicho ello, vamos a ver cuales han sido los resultados arrojados por este algoritmo Bietápico en nuestro estudio IBERINFORM de CREDIT SCORIN en IBM SPSS Modeller.

El modelo nos da un número óptimo de 8 clusters con buena calidad y silueta, en el cual la mayor importancia predictora hacia el Target de agrupamiento la soporta la variable “***Año o Fecha de constitución***” como fue el caso de KOHONEN.

La principal diferencia con los otros 2 algoritmos es que en este caso la distribución parece muy buena y homogénea, siendo el tamaño del clúster más pequeño un 7,1% mientras que el mayor ocupa un 19,1%, por lo que es más eficiente también en término de homogeneidad.



Una vez llevada a cabo la comparación entre los distintos algoritmos, decidiendo que vamos a utilizar el Bietápico, procedemos a analizar los datos con el mismo y vamos a sacar las siguientes conclusiones de como vamos a agrupar las observaciones, según sus características más similares.

Como se puede observar en los resultados arrojados por el algoritmo Bietápico, este ha agrupado las cerca de 21.000 observaciones y sus variables y para ellos vamos a tener en cuentas características tales como el número de EMPLEADOS, el volumen de VENTAS de la compañía, reclamaciones y otro tipo de incidencias, si se encuentran en consurso de Acreedores o no, para las cuales se crearon una seri de dummies binarias, y por supuesto la variable con mayor poder predictivo como es el ***“Año o Fecha de Constitución”*** de la compañía.

**CLUSTER 1**: (12,9%) Está conformado por un total de 2722 observaciones y lo que más concentra las similitudes en este grupo es que el ***Año de constitución*** de la compañía se comprende entre el 10 y el 12, por lo que se crearon en un espacio temporal casi idéntico. En cuanto a reclamaciones e incidencias también son similares ya que hay muchos más ceros que unos, lo que las hace ser empresas más eficaces en su actividad y jurídicamente saneadas, ya que hacen frente a menos casos de impago.

**CLUSTER 2**: Este es uno de los grupos más pequeño que nos ha proporcionado el Bietápico con un total de 2282 observaciones, no muy lejano al anterior grupo en todo caso. Se constituyen en años más tempranos que el anterior pero no tan concentradas. Son similares en volumen por ventas y podemos encontrar empresas de las 19 categorías CNAE que hemos descrito según sea su sector y actividad laboral, tales como Agricultura o Ganadería, Industria Manufacturera y la Construcción.

**CLUSTER 3**: (12,5%) 2629 registros en este caso. Encontramos empresas con Años de constitución no muy similares aquí pero si que podemos observar que son pequeñas o medianas empresas, debido a que gran parte de este grupo se caracteriza en que el número de empleados es muy bajo en la mayoría de sus datos, lo que nos puede a llevar a pensar que se trate de trabajadores por cuenta propia o autónomos, los cuales se rigen en un ámbito fiscal diferente y esto puede ser un condicionante a la hora de solicitar préstamos, no incluyente de cualquier modo en el grupo de Riesgo de impago o default.

**CLUSTER 4**: (11,7%) 2463 registros. Años de constitución muy similares de nuevo, concentrándose entre el 16 y el 18. El número de reclamaciones de Organismos Públicos motivadas por el impago a Seguridad Social, Hacienda Pública o Administraciones Locales es bastante alto en este grupo, lo que nos hace pensar que se pueden encontrar en concurso de Acreedores cuando una persona física o jurídica deviene en una situación de insolvencia en la que no puede hacer frente a la totalidad de los pagos que adeuda. El concurso de acreedores abarca las situaciones de quiebra y las de suspensión de pagos.

**CLUSTER 5** : (19,1%) Este es nuestro cluster más grande, con un total de 4028 registros. Los años de constitución en este grupo de empresas vuelven a ser dispares y diversos en el tiempo.

Tanto los Fondos Propios, así como su fondo de maniobra, es decir, el Activo circulante disponible en la empresa para hacer frente a los pasivos más inmediatos tienen medias positivas muy altas con respecto a otros clusters, lo que nos lleva a pensar que son compañias muy bien estructuradas y saneadas, con un número irrisorio de reclamaciones por impagos.

**CLUSTER 6** : (14,1%) Encontramos 2908 registros, lo que lo sitúa en el 2º cluster por tamaño de observaciones pero muy cercano tal y como estamos comprobando, lo que hace al bietápico un algoritmo muy óptimo. Volvemos a econtrar empresas o compañias pequeñas por numero de empleados, pero sin embargo, con una media alta en cuanto al volumen de ventas anuales o facturacion (902.379 euros), lo que las convierte en empresas solventes pero con un endudamiento a corto plazo (obligaciones con terceros por explotación) que tiene una media de 343.284 euros anuales.

En todo caso su apalancamiento o leverage financiero es bueno.

**CLUSTER 7** : (7,1%) Este es el cluster más pequeño en número de observaciones que nos ha proporcionado el bietápico con un total de 1500 registros. El año de constitucion de las empresas de este grupo de situa entre el 19 y el 22, por lo que se agrupan en torno a la variable predictora.

Albergan un número muy bajo de reclamaciones o Incidencias derivadas de demandas judiciales y entre sus sectores por actividad destacan comercio al por mayor y por menor, Transporte y almacenamiento.

Aunque su media en Fondo de maniobra para hacer frente al activo circulante, es más bajo que otros grupos, lo compensa con la media en Fondos Propios, la cual está por encima de otros, lo que las hace estar equilibradas financieramente.

**CLUSTER 8** : (11,7%) Aquí tenemos los 2513 registros restantes que completan nuestro dataset sobre IBERINFORM CREDIT SCORING. El año de constitucion de estas compañias se comprende ahora entre el 19 y el 21, por lo tanto diríamos que correlacionan mucho en esta variable predictora.

Son empresas muy repartidas geográficamente en el territorio nacional, al igual que en los clusters anteriores y con un numero promedio de empleados de 8,64.

Superan el millón de euros de facturación anuales con un endeudamiento medio ligeramente mayor que otros grupos, pero en ningun caso factor de riesgo en el scoring ya que disponen de cifras altas en cuanto a recursos propios. Tampoco hacen frente a impagos en la mayoría de casos en este grupo por lo que no se detectan grande anomalías o puntos diferenciales con respecto a otros factores desarrollados en este estudio.

* CONCLUSIONES

Concluimos que el algoritmo más eficiente en términos de calidad, optimización y distribución homogénea, va a ser el llamado TWO STEP O Bietápico tanto para variables continuas y discretas.

Posteriormente, trás haber realizado todo el estudio, el resultado es considerado como bastante óptimo. Se han conseguido ubicar las 21.127 empresas de características bastante similares en muchos casos, en 8 clusters o agrupaciones, de manera que en cada cluster están las empresas o compañías con una mayor similitud, tanto a nivel técnico financiero como a nivel de actividad laboral. A pesar de poder haber sido ubicados en únicamente 8 clusters, durante todo el proceso se ha intentado seguir un criterio basado en el rating o Credit Scoring que pueden obtener estas compañías, añadido a otros criterios ya mencionados anteriormente.