4. Segmentación

1. Introducción

En el contexto de la elaboración de un modelo de calificación crediticia, el clustering no supervisado puede ser utilizado para agrupar a los clientes en diferentes grupos basados en sus características. Esto puede ayudar a los analistas a entender mejor cómo diferentes factores afectan la calificación crediticia de los clientes y a desarrollar un modelo de calificación crediticia más preciso.

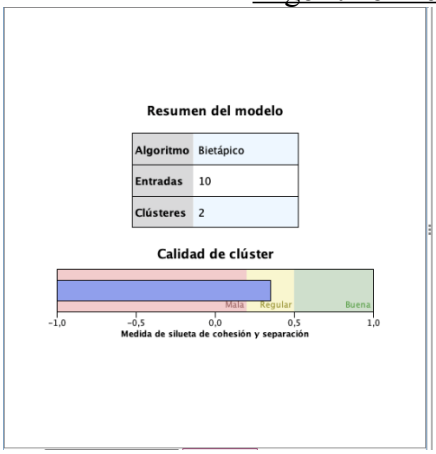
El clustering no supervisado es una técnica de aprendizaje automático que se utiliza para agrupar objetos similares en grupos o "clústeres" basados en sus características. El objetivo del clustering es encontrar patrones o estructuras en los datos que de otra manera serían difíciles de detectar.

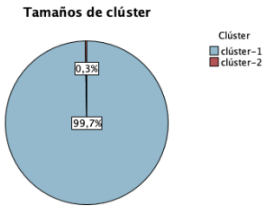
Hay varios algoritmos de clustering no supervisado disponibles, cada uno con sus propias ventajas y desventajas. Algunos ejemplos populares de algoritmos de clustering incluyen k-medias, Redes de Kohonen y Método bietápico, todos estos son los que utilizaremos en nuestro modelo. Tras estos paso se realizará una Detección de Anomalías. Es importante elegir el algoritmo adecuado para el conjunto de datos y el objetivo específico del modelo.

Además, una vez se han agrupado los clientes en diferentes clusters, se puede utilizar esta información para segmentar el mercado y desarrollar estrategias de marketing y riesgo específicas para cada grupo.

2. Algoritmo Bietápico / Two-Steps

Algoritmo Bietápico





Tamaño del clúster más pequeño	41 (0,3%)
Tamaño del clúster más grande	12845 (99,7%)
Cociente de tamaños: De clúster más grande a clúster más pequeño	313,29

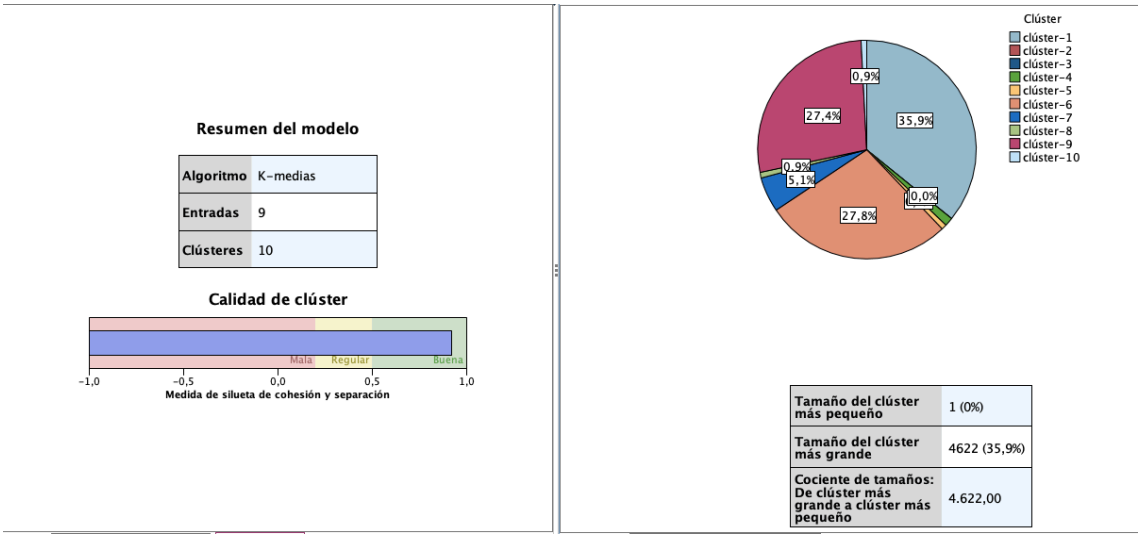
El primer algoritmo de agrupamiento ha sido Two-Steps o Bietápico. El algoritmo bietápico es un algoritmo de clustering no supervisado que se utiliza para agrupar objetos similares en clusters. Es una variante del algoritmo k-medias y se diferencia en que utiliza dos pasos para determinar los centroides de los clusters. En primer lugar, el algoritmo elige al azar un conjunto de puntos de inicio como representantes de los clusters y luego asigna cada objeto del conjunto de datos al cluster cuyo centroide está

más cerca. En el segundo paso, se recalculan los centroides y se vuelve a asignar los objetos a los clústeres. El proceso se repite varias veces hasta que no se producen cambios significativos en la asignación de objetos a los clusters. Es una buena opción para trabajar con conjuntos de datos grandes y requerir alta precisión, pero puede ser más lento que otros algoritmos.

Este algoritmo permite especificar un rango para el número de clusters deseados, se ha establecido un rango máximo de 15 clusters y un mínimo de 2 clusters. Sin embargo, después de aplicar el algoritmo, se obtuvo un modelo con una calidad de 0,3 , lo cual se considera una calidad regular. Además, el modelo solo generó dos clusters, uno de ellos con un 99,7% de los datos y el otro con solo un 0,03%. Dado que el modelo no cumple con los estándares deseados de calidad, se ha decidido descartarlo y buscar un algoritmo alternativo para continuar con el análisis.

3. Algoritmo K-Medias

Algoritmo K-Medias



K-means es un algoritmo de aprendizaje automático no supervisado utilizado para la agrupación de datos. Funciona dividiendo un conjunto de datos en k grupos (donde k es un número especificado previamente, en nuestro caso 10) de manera que los puntos en cada grupo son similares entre sí. El algoritmo se inicializa asignando cada punto de datos a un grupo al azar, luego calcula el centroide de cada grupo y reasigna cada punto al grupo cuyo centroide es el más cercano. Este proceso se repite varias veces hasta que no se producen más cambios en la asignación de puntos a grupos.

Con este algoritmo obtenemos un 10 clusters con una precisión del 0,9, considerándolo como bueno. Todas las variables del algoritmo son muy importantes, tienen un nivel 1 que es lo máximo, menos la variable “Incidencias Totales”. Los clúster 2 y 3 tienen un tamaño de 0%, estos puntos de datos serán excluidos del análisis debido a que no contribuyen al cálculo de los centroides de los clúster. Esto puede ser debido a que

estos puntos de datos se encuentran muy alejados de los demás puntos de datos en el conjunto de datos o pueden ser valores atípicos.

Clústers

Clústeres

Importancia de entrada (predictor)
1,0 0,8 0,6 0,4 0,2 0,0

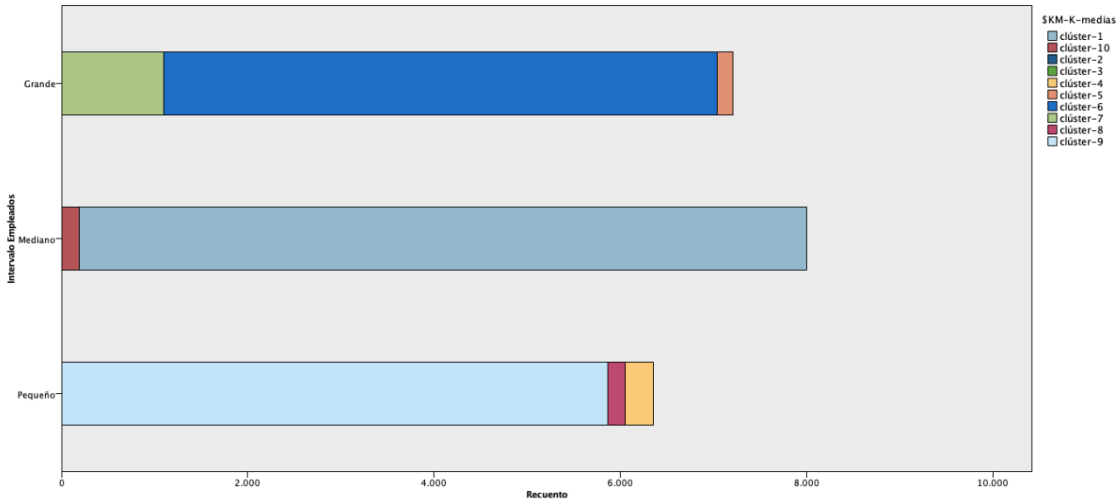
Clúster	clúster-1	clúster-6	clúster-9	clúster-7	clúster-4	clúster-8	clúster-10	clúster-5	clúster-2	clúster-3
Tamaño	35,9%	27,8%	27,4%	5,1%	1,3%	0,9%	0,9%	0,7%	0,0%	0,0%
Entradas	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO
	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL
	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO
	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA
	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA
	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados
	Pasivo	Pasivo	Pasivo	Pasivo	Pasivo	Pasivo	Pasivo	Pasivo	Pasivo	Pasivo
	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO
	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales

En el resumen del modelo podemos obtener la información acerca de los datos de entrenamiento, el proceso de estimación y los clústeres definidos por el modelo. Se muestra el número de clústeres y el historial de iteración. Este nos cuenta que el algoritmo necesita solo 3 iteraciones para ajustarse, la primera iteración tiene un error del 0,765 la segunda del 0,051, y la tercera de 0. Cuando la última iteración es 0, significa que el algoritmo de k-means ha completado su proceso y ha asignado todos los puntos de datos a un grupo. Esto significa que se ha alcanzado una solución óptima o una solución estable para el conjunto de datos en cuestión.

También podemos ver los campos utilizados en el modelo, vemos que se utilizan todos menos “Provincia” y “Riesgo”.

Hemos separado nuestros clusters en 3 grupos, correspondientes al tamaño de la empresa. En el primer grupo, “Grande”, obtenemos, los clusters 5, 6 y 7. El segundo, “Mediano”, los clústers 1 y 10. En el último, “Pequeño”, los clusters 4, 8 y 9. Como podemos ver, no selecciona los clusters de tamaño 0%.

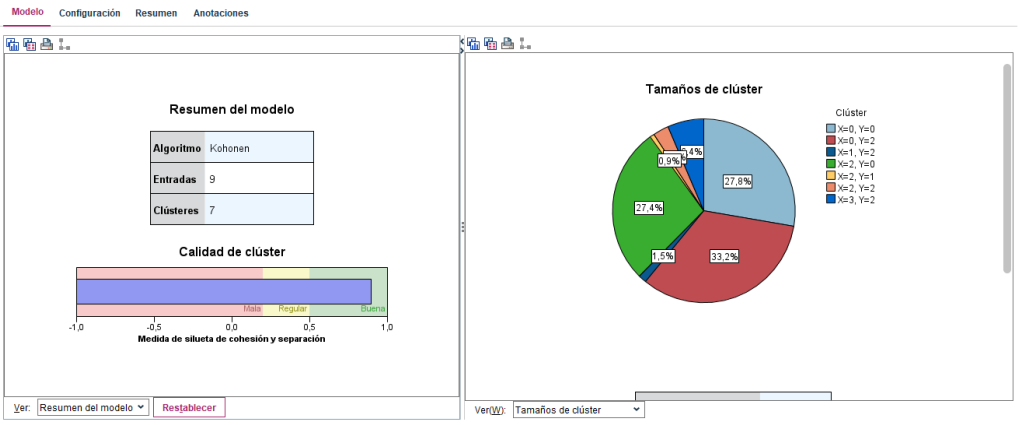
Clústers por tamaño de la empresa



Consideramos este clúster para nuestra selección final y seguimos probando el resto para ver si hay uno mejor.

4. Algoritmo Kohonen

Algoritmo Kohonen

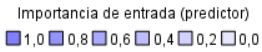


Las redes de Kohonen son un algoritmo de aprendizaje no supervisado utilizado para la agrupación de datos en clústeres mediante un mapa autoorganizativo. Utiliza una estructura de dos capas (entrada y salida) con unidades básicas llamadas neuronas. Estas neuronas se conectan entre sí formando una estructura topológica, que se asemeja a la organización de las células en el cerebro. El algoritmo se inicializa asignando valores aleatorios a cada nodo, luego se entrena en base a un conjunto de datos de entrada, para asignar cada punto de datos a un grupo de acuerdo con su similitud.

Durante el entrenamiento se actualizan las ponderaciones de las conexiones entre las neuronas para mejorar la representación de los datos. Al finalizar el entrenamiento, los registros similares se encuentran cercanos en el mapa de resultados mientras que los diferentes se encuentran alejados. Es utilizado para problemas de agrupamiento, visualización de datos y reducción de dimensionalidad.

Con este algoritmo obtenemos 7 clusters con una precisión del 0,9, considerándolo como bueno. Las variables más importantes son “Forma Jurídica” e “Intervalo de Empleados” , tienen un nivel 1 que es lo máximo. El resto de las variables tienen un nivel muy bajo de importancia que esta entre 0,1 y 0. La menos importante es “Fondo de Maniobra".

clústers



Clúster	X=0, Y=2	X=0, Y=0	X=2, Y=0	X=3, Y=2	X=2, Y=2	X=1, Y=2	X=2, Y=1
Tamaño	33,2%	27,8%	27,4%	6,4%	2,7%	1,5%	0,9%
Entradas	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA	FORMA_JURIDICA
	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados	Intervalo Empleados
	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO	ENDEUDAMIENTO_CORTO_PLAZO
	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO	ACTIVO
	Pasivo	Pasivo	Pasivo	Pasivo	Pasivo	Pasivo	Pasivo
	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO	PATRIMONIO_NETO
	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL	CAPITAL_SOCIAL
	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales	Incidencias totales
	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA	FONDO_MANIOBRA

En este caso todos los clústers tienen un tamaño mayor a 0%, por lo que todos ellos serán incluidos en el analisis ya que todas las neuronas contribuyen al cálculo de los clusters.

El resumen del Nuggets indica los siguientes detalles sobre el análisis:

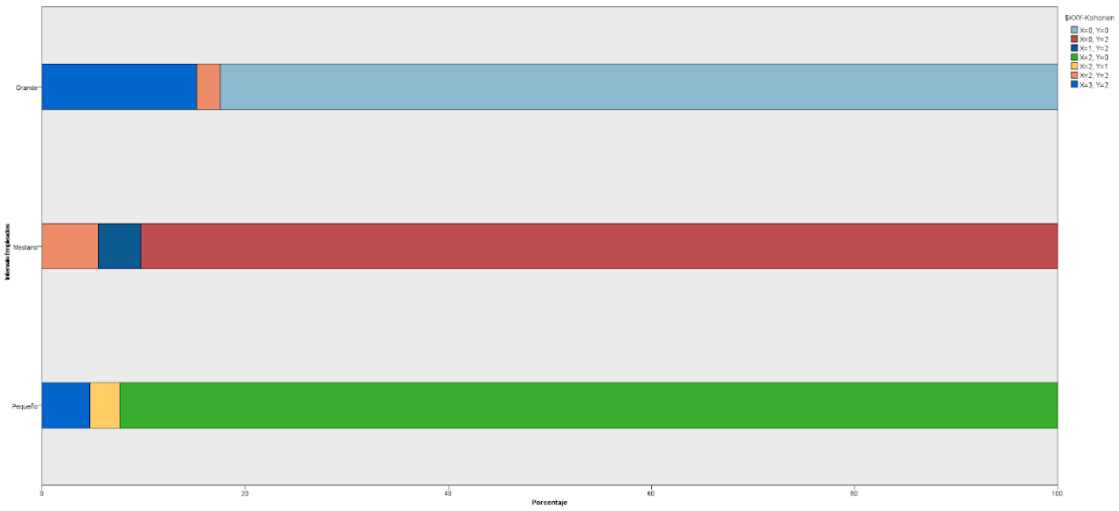
- $KX-Kohonen: 4: indica que el análisis ha sido realizado utilizando 4 dimensiones en el eje X.
- $KY-Kohonen: 3: indica que el análisis ha sido realizado utilizando 3 dimensiones en el eje Y.
- Capa de entrada: 26 neuronas: indica que la capa de entrada del análisis contiene 26 neuronas. La capa de entrada es la primera capa de la red neuronal, donde se introducen los datos para ser analizados.
- Capa de salida: 12 neuronas: indica que la capa de salida del análisis contiene 12 neuronas. La capa de salida es la última capa de la red neuronal, donde se generan las salidas del análisis.

En resumen, el análisis de Kohonen se realizó utilizando 4 dimensiones en el eje X, 3 dimensiones en el eje Y, con una capa de entrada de 26 neuronas y una capa de salida de 12 neuronas. Estos detalles son importantes para entender la configuración utilizada en el análisis y pueden ser útiles para interpretar los resultados obtenidos.

Al igual que en el modelo anterior, en la selección de campos no tiene en cuenta ni “Provincia”, ni “Riesgo”.

De nuevo hemos separado nuestros clústers en 3 grupos, correspondientes al tamaño de la empresa. El primero, “Grande”, obtenemos los clústers “X=3 Y=2”, “X=2 Y=2”, y “X=0 Y=0”. En el segundo, “Mediano”, los clústers “X=2 Y=2”, “X=1 Y=2”, y “X=0 Y=2”. En el último, “Pequeño”, los clústers “X=3 Y=2”, “X=2 Y=1” y “X=2 Y=0”.

Clústeres por tamaño de la empresa



También consideramos este clúster para nuestra selección final, pero antes analizamos las anomalías.

5. Detección de Anomalías

Un algoritmo de anomalía es una técnica utilizada para detectar patrones o valores atípicos en un conjunto de datos. Los algoritmos de anomalía se utilizan en una variedad de aplicaciones, como la detección de fraude financiero, la detección de intrusos en una red de computadoras y la monitorización de la salud de los equipos industriales.

El resultado de un algoritmo de anomalía es un conjunto de registros o valores que se consideran anómalos en comparación con el resto de los datos. Estos registros o valores pueden requerir una investigación adicional o una acción específica, dependiendo del contexto de la aplicación.

El objetivo de un algoritmo de anomalía es ayudar a los analistas a detectar patrones o valores atípicos en grandes cantidades de datos, para que puedan tomar decisiones informadas y tomar medidas para resolver cualquier problema detectado.

En SPSS Modeler, la detección de anomalías es una técnica utilizada para encontrar registros o patrones inusuales en los datos. Estos patrones inusuales pueden ser

indicativos de problemas, incidencias u oportunidades, dependiendo del contexto del análisis. La detección de anomalías se utiliza para identificar patrones o características que se desvían significativamente de lo que se considera "normal" o "esperado" en los datos.

La detección de anomalías en SPSS Modeler se realiza a través de un nugget específico, el cual permite configurar diferentes métodos y parámetros para analizar los datos. Dependiendo del método seleccionado, el análisis puede buscar patrones anómalos en una sola variable o en varias combinadas.

En el contexto de un análisis de detección de anomalías en SPSS Modeler, el término "grupo de homólogos" se refiere a un conjunto de registros que son similares entre sí en términos de una o varias características o variables. En nuestro resumen se indica que hay 3 grupos de homólogos, significa que el análisis ha dividido los datos en 3 grupos diferentes de registros que son similares entre sí.

Cada uno de estos grupos de homólogos se analizará de manera independiente y se buscarán registros que sean significativamente diferentes de los demás en el grupo, identificando así las anomalías. Al dividir los datos en varios grupos de homólogos se tiene una mayor precisión y flexibilidad en la detección de anomalías, especialmente si existen subgrupos o patrones específicos dentro de los datos.

En el primer homólogo, nuestro nugget de detección de anomalías está analizando un grupo compuesto por 4701 registros y ha encontrado 47 registros que se consideran anomalías en relación con el grupo.

La tabla que obtenemos presenta información sobre las características o variables que contribuyen a la detección de las anomalías. Cada fila representa una variable y se presenta la siguiente información:

- Contribución: es el porcentaje de contribución de la variable al cálculo de la anomalía.
- Recuento: es el número de registros en los que la variable contribuyó a la detección de la anomalía.
- Índice Promedio: es el promedio de la variable en los registros que contribuyeron a la detección de la anomalía.

De acuerdo con esta tabla, se puede concluir que las variables "Provincia" y "Forma jurídica" contribuyeron de manera significativa en las anomalías, con un 37,3% y un 46,8% respectivamente, mientras que la variable "Activo" tuvo una contribución menor con un 10,6%. El residuo del 0,51% significa que el 0,51% de los motivos que causan las anomalía no están incluidos en el informe.

En la segunda tabla, homologo 2, se puede concluir que las variables “Activo”, “Endeudamiento_corto_plazo” y “Patrimonio_Neto” contribuyeron de manera significativa en las anomalías, con un 14,5%, un 36,1% y un 21,5% respectivamente, mientras que la variable "Pasivo" tuvo una contribución menor con un 29,3%.

El residuo del 25,88% significa que el 25,88% de los motivos que causan las anomalía no están incluidos en el informe.

En la última tabla, homologo 3, se puede concluir que las variables "Provincia" y "Forma_juridica" contribuyeron de manera significativa en las anomalías, con un 35,5% y un 44,2% respectivamente, mientras que la variable "Pasivo" tuvo una contribución menor con un 26,6%.

El residuo del 5,88% significa que el 5,88% de los motivos que causan las anomalía no están incluidos en el informe.

Tras detectar estas anomalías, aplicaremos en el siguiente paso una preparación automática de datos para tratarlas.

Anomalías

Grupo de homólogos-1: 4701 registros

Anomalías: se han encontrado 47 registros de un total estimado de 4.701 registros.

Contribución	Recuento	Indice promedio
ACTIVO	1	0,106
PROVINCIA	47	0,373
FONDO_MANIOBRA	3	0,184
PATRIMONIO_NETO	3	0,312
CAPITAL_SOCIAL	2	0,507
FORMA_JURIDICA	44	0,468
Incidencias totales	41	0,147

Residuo de los motivos no incluidos en el informe: 0,51%

Perfil de grupo de homólogos

Grupo de homólogos-2: 623 registros

Anomalías: se han encontrado 5 registros de un total estimado de 623 registros.

Contribución	Recuento	Indice promedio
Pasivo	1	0,293
ENDEUDAMIENTO_CORTO_PLAZO	3	0,361
ACTIVO	4	0,145
FONDO_MANIOBRA	2	0,269
PATRIMONIO_NETO	3	0,215
CAPITAL_SOCIAL	2	0,285

Residuo de los motivos no incluidos en el informe: 25,88%

Perfil de grupo de homólogos

Grupo de homólogos-3: 7562 registros

Anomalías: se han encontrado 76 registros de un total estimado de 7.562 registros.

Contribución	Recuento	Indice promedio
Pasivo	1	0,266
ENDEUDAMIENTO_CORTO_PLAZO	3	0,236
ACTIVO	5	0,138
PROVINCIA	72	0,355
FONDO_MANIOBRA	7	0,415
PATRIMONIO_NETO	8	0,267
CAPITAL_SOCIAL	6	0,34
FORMA_JURIDICA	57	0,442
Incidencias totales	21	0,437
Intervalo Empleados	48	0,06

Residuo de los motivos no incluidos en el informe: 5,88%

6. Conclusión

Finalmente hemos podido ver que el algoritmo bietápico no se adaptaba a nuestro conjunto de datos. También hemos visto que nuestro conjunto tiene una serie de valores atípicos que tienen que ser arreglados en la siguiente fase del trabajo, antes de realizar el modelo.

Por último hemos visto que los modelos K-medias y Kohonen, ambos obtienen una evaluación buena, de un 90%.

K-means y Kohonen son dos algoritmos de agrupamiento diferentes con propósitos y usos específicos. K-means es un algoritmo de agrupamiento de centroide, mientras que Kohonen es un algoritmo de mapa auto-organizativo.

K-means es adecuado para agrupar datos basados en similitud y es ampliamente utilizado en una variedad de aplicaciones, como la segmentación de mercado y la clasificación de imágenes.
Por otro lado, Kohonen es utilizado para encontrar patrones no lineales y estructuras en datos, es especialmente útil en la minería de datos para encontrar patrones desconocidos.
En general, si el objetivo es agrupar datos basados en similitud y no hay interés en encontrar patrones no lineales, K-means sería la mejor opción. Si el objetivo es encontrar patrones no lineales, entonces Kohonen sería la mejor opción. Por lo que elegimos Kohonen.

Al elegir nuestro algoritmo, seleccionamos los clústeres que queremos y filtramos por tamaño de la empresa “Pequeña”, que es el intervalo que hemos elegido para nuestro modelo. Cómo podemos ver estos clústeres aparecen en varios de los tamaños de empresas por lo que realizaremos un filtro a través de operación con registros con el nodo seleccionar, y seleccionaremos solo los valores de la empresa “Pequeña”.

A continuación podemos ver los clústeres que componen nuestro modelo y el número total de registros que contiene cada uno.

Clústeres del modelo

	$KXY-Kohonen	Record_Count
1	X=3, Y=2	299
2	X=2, Y=0	5865
3	X=2, Y=1	188