K-MODES

El algoritmo K-MODES fue diseñado para agrupar grandes conjuntos de datos categóricos, y tiene como objetivo obtener las k modas que representan al conjunto. Permite extender el k-means, a partir del cálculo de una medida de disimilitud que permita comparar observaciones categóricas y la utilización de modas en lugar de medias para calcular los clusters.

El primer paso será seleccionar k número de modas. Queremos realizar un análisis de agrupamiento utilizando el algoritmo de K-Mode con 5 clusters. Como ya hemos detectado en el clustering jerárquico, con k = 5 conseguimos un mejor corte y un perfilamiento de grupos más detallado.

La manera de medir la distancia entre dos vectores de variables categóricas es la cantidad de valores que son diferentes en la misma variable entre clusters.

Función para ejecutar K-MODES

Cuadro 33: Obtención de los Parámetros de los Clústers

CODE_GENDER	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS
M	Working	Secondary / secondary special	Married
F	Commercial associate	Secondary / secondary special	Married
F	Working	Higher education	Married
M	Working	Secondary / secondary special	Married
F	Pensioner	Secondary / secondary special	Married

OCCUPATION_TYPE	ORGANIZATION_TYPE	REGION_RATING_CLIENT	TARGET
Unknown	Business and bank	2	1
Mid skill laborers	Business and bank	2	1
Mid-high skill laborers	Education	2	0
Low-mid skill laborers	Business and bank	2	0
Unknown	Unknown	2	0

Con esta tabla podemos ver estos 5 clusters. El primer cluster, y por lo tanto la clase mayoritaria, corresponde a hombres casados, que trabajan en empresas o bancos y educación secundaria o secundaria especial. Pertenecen al grupo de clientes sin dificultad de pago.

El segundo cluster está formado por mujeres casadas trabajadoras como asociadas en empresas o bancos con habilidades medianas y educación secundaria o secundaria especial. Pertenecen al grupo de clientes sin dificultad de pago.

En el tercer cluster tenemos a mujeres casadas, que trabajan en educación con habilidades medias o altas y educación alta. Pertenecen al grupo de clientes con dificultad de pago.

El cuarto está compuesto por hombres casados, que trabajan en empresas o bancos con habilidades bajas o medias y educación secundaria o secundaria especial. Pertenecen al grupo de clientes con dificultad de pago.

Y finalmente, el quinto y último cluster formado por mujeres casadas y pensionistas que han tenido educación secundaria o secundaria especial. Pertenecen al grupo de clientes con dificultad de pago.

Para poder calcular las distancias entre el primer y segundo cluster miramos la separación que existe usando las diferencias entre la primera fila y la segunda. Obtiene un valor de 3, ya que ni code_gender ni name_income_type ni occupation_type coinciden. Las dos variables que coincide en los cinco clusters son name_family_status y region_rating_client.