Modelos no Supervisados

Sebastian Jaremczuk

2020-04-17

carga de datos

Reducción de Dimensión

Análisis de Componentes Principales

Se aplica a técnica de Componentes Principales para reducir las variables predictoras pero que tengan un gran porcentaje de la variabilidad total.

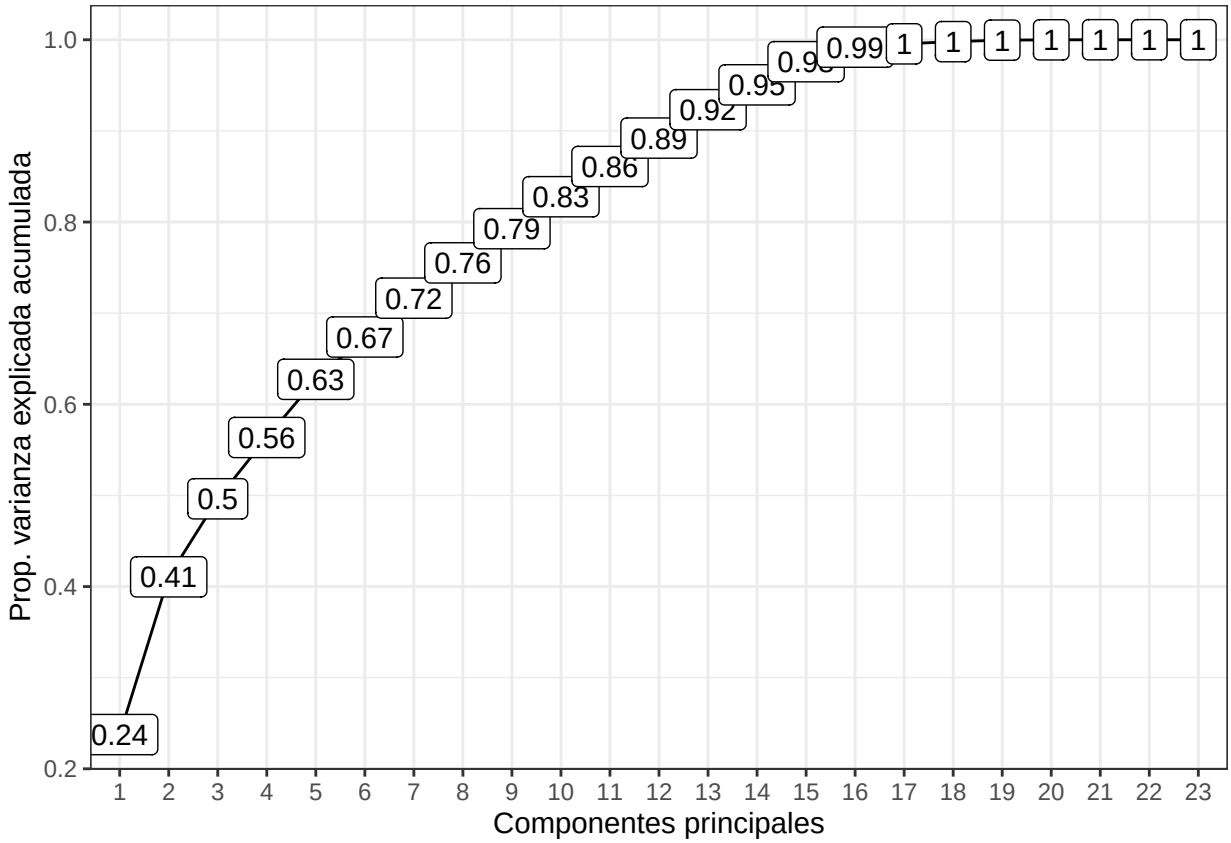
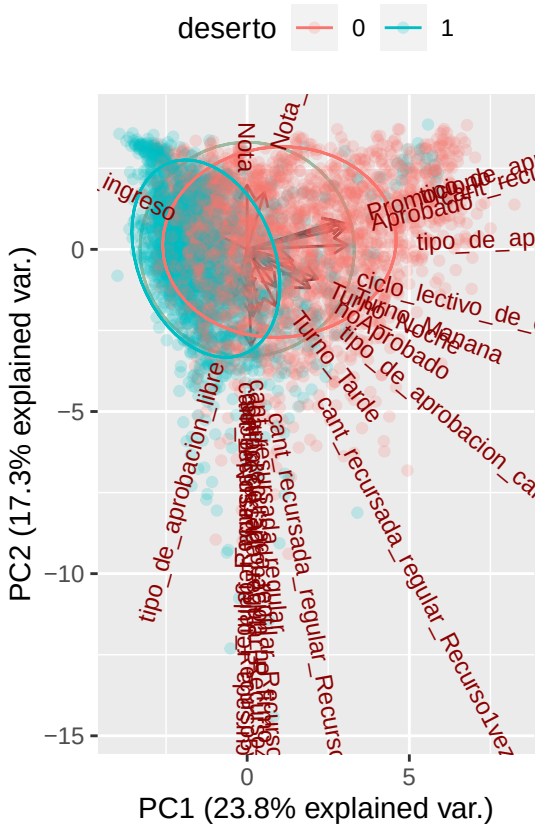


Table 1: Loadings de PCA en Tablon

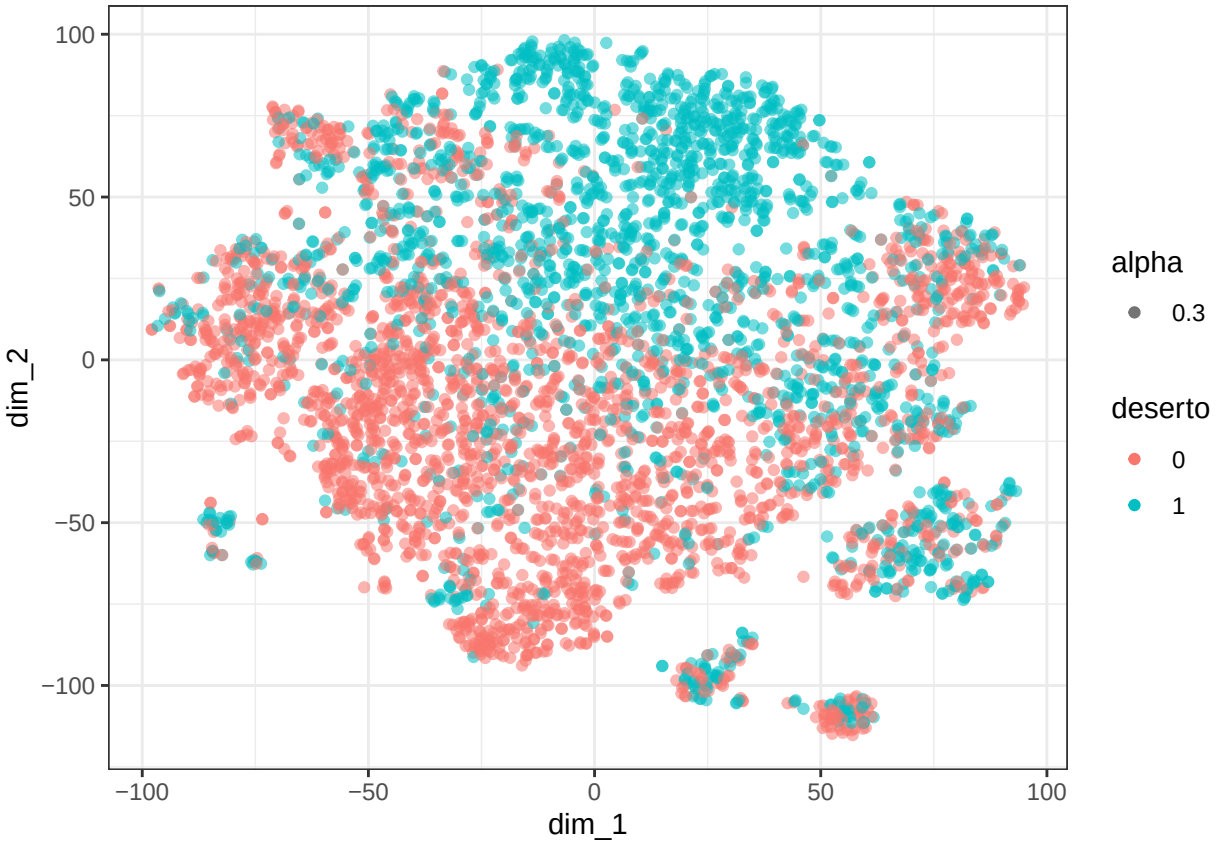
variable	PC1	PC2	PC3	PC4
Turno_Manana	0.2813063	-0.1375132	0.0171629	0.0891187
tipo_de_aprobacion_libre	-0.0516192	-0.2204644	-0.1151088	-0.3721061
tipo_de_aprobacion_cambio_curso	0.0985927	-0.0960438	0.0259816	-0.0003118
tipo_de_aprobacion_promociono	0.3581733	0.1308245	-0.0808955	0.0804953
edad_al_ingreso	-0.1269356	0.0733137	-0.2349852	-0.2976394
tipo_de_aprobacion_no_firmo	0.0156176	-0.4488391	-0.1126781	-0.1126948
ciclo_lectivo_de_cursada	0.1818683	-0.0532719	-0.1144249	-0.0532596
tipo_de_aprobacion_firmo	0.3996032	0.0206986	0.0690403	-0.0132481
cant_resursada_regular	0.0227835	-0.3127946	-0.4101361	0.3674317
cant_recursada_regular_No_Recurso	0.3869691	0.1219610	0.0468956	0.0352279
cant_recursada_regular_Recurso1vez	0.1188652	-0.2681558	0.0654967	-0.1891977
cant_recursada_regular_Recurso2vez	0.0451665	-0.2939066	-0.0071695	-0.1971879
cant_recursada_regular_Recurso3vez	0.0120116	-0.2503402	-0.0867234	-0.1612458
cant_recursada_regular_Recurso4vez	0.0044083	-0.1971933	-0.0845225	-0.1281708
cant_recursada_regular_Recurso5vez	0.0006389	-0.1464459	-0.1411299	0.0048529
cant_recursada_regular_RecursoNveces	-0.0021967	-0.1955233	-0.4257671	0.4879239
Turno_Tarde	0.1153361	-0.1754377	0.0127343	0.0859205
Turno_Noche	0.2032766	-0.1098708	-0.0798642	-0.3755512
Aprobado	0.3918783	0.1006967	-0.0242947	-0.0374767
Promociono	0.3589365	0.1289055	-0.0812155	0.0798013
noAprobado	0.2504186	-0.1806504	0.2037648	-0.0519875
Nota	-0.0001346	0.3003970	-0.4732939	-0.2012817
Nota_max_prom	0.0688392	0.2670847	-0.4728596	-0.2279385



Se puede observar que si bien se puede reducir la cantidad de variables predictoras y mantener una alta variabilidad de la información explicada, los diagramas de biplot en este caso no nos servirían de mucha ayuda ya que en las primeras 2 componentes solo se explica el 41% y en las primeras 4 componentes solo el 56%. Además, los loadings de dichos componentes no tienen una clara identificación de la proyección que quieren significar, por lo que sería complicado explicar el modelo que se queira desarrollar según estas nuevas variables.

Reducción por t-SNE

Al igual que PCA, existen otro algoritmos que pueden realizar ruducción de dimensionalidad. Unos de esos casos es el método no lineal t-distributed stochastic neighbor embedding (t-SNE), que en ciertos casos es ventajoso respecto a PCA que aplica reducción pero utilizando combinaciones lineales de las variables originales.



Utilizando el método T-SNE, si bien no se arman grupos bien definidos, al identificar cada observación con un color en el gráfico según el target puede observarse que están mas separadas y hay menos solapamiento entre ellas que con el método PCA. Este resultado puede insinuar que es posible clasificar un gran porcentaje de los casos correctamente.

Clusters

Se utilizarán técnicas de clusters sobre la base del tablón y las reducciones de dimensión calculadas anteriormente.

Criterio: * Como son solamente 2 las variables categóricas (EsTecnico y sexo), en vez de calcular las distancias numéricas por un lado (euclídea, manhattan, correlación, etc), las distancias categóricas por otras (SMC, Jaccard, etc) y tratar de transformar esas matrices de distancias en una nueva matriz unificada con criterios, se decide transformar los datos categóricos en numéricos aprovechndo que ambos campos tiene solamente

2 valores por lo que estarán en los extremos tomando una normalización entre 0 y 1. En el caso de las observaciones que con valor nulo en la variable EsTecnico, se imputará con el valor 0.5 (mitad entre extremos)

```
## [1] 0.8893004
```

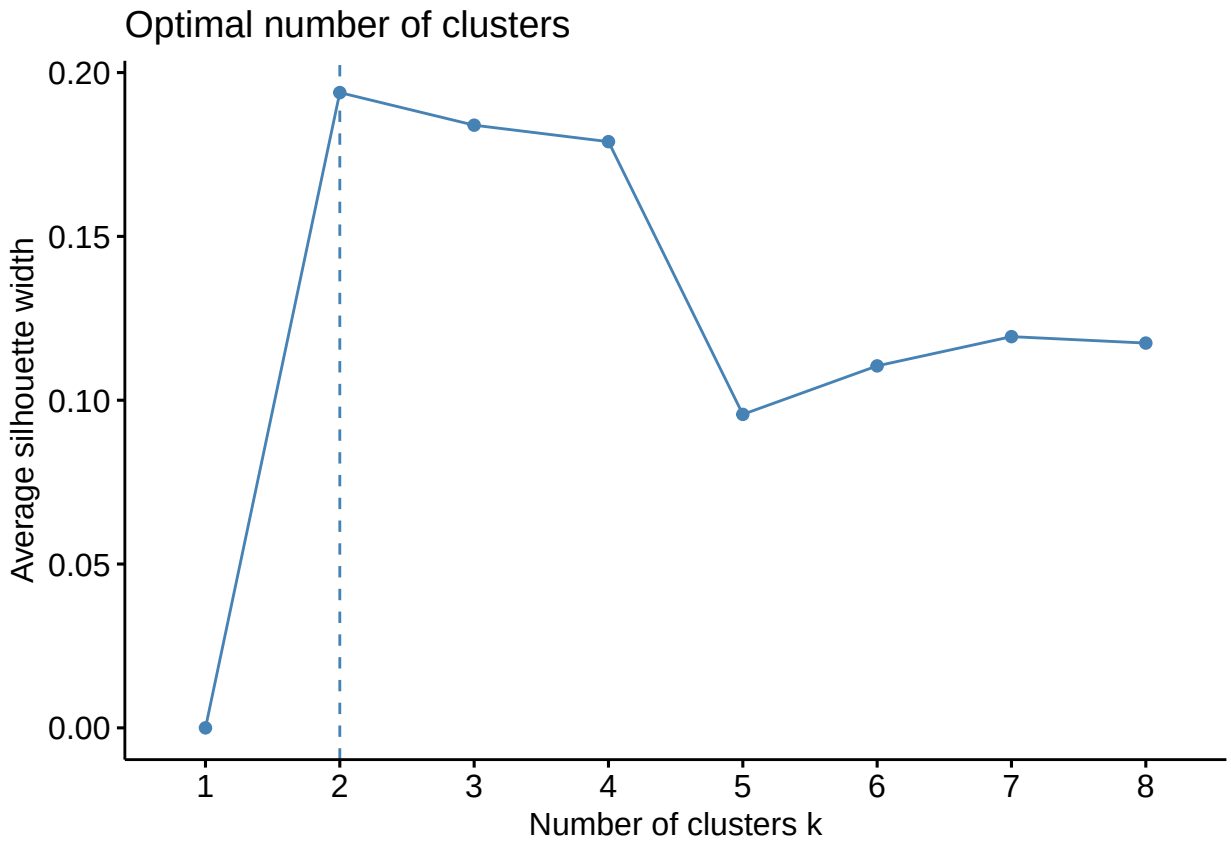
El estadístico de Hopkins sobre el tablon orifginal transformado las 2 variable cateóricas en numéricas da 0.8893004. Es un valor cercano a 1 por lo que tiene mucha tendencia a ser clusterizado.

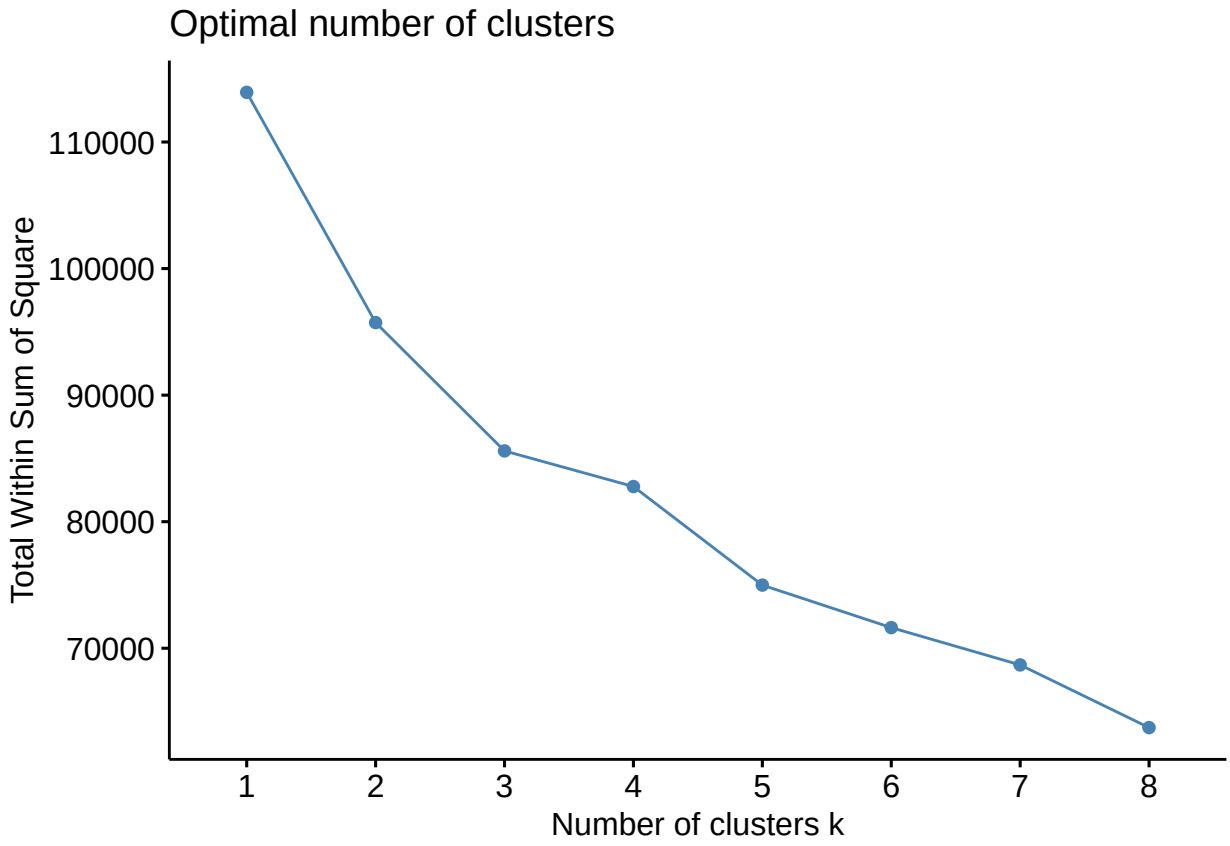
Matriz distancia entre variables

Matriz

Estudio del número óptimo de clusters

Por la cantidad de datos que hay en el dataset y el poder computacional





Verificación de las funciones anteriores de número óptio de clsuters pero calculadas manualmente

```
##
## Clustering Methods:
##  kmeans hierarchical
##
## Cluster sizes:
##  2 3 4 5
##
## Validation Measures:
##                                  2         3         4         5
##
## kmeans       Connectivity     0.0000    3.1667  730.2742 1316.3881
##              Dunn             0.4693    0.5735    0.0471    0.0345
##              Silhouette       0.5937    0.5336    0.2070    0.1796
## hierarchical Connectivity     3.1667    3.1667    9.2357   12.1647
##              Dunn             0.5713    0.5735    0.2744    0.2744
##              Silhouette       0.7329    0.5336    0.5198    0.4879
##
## Optimal Scores:
##
##              Score  Method       Clusters
## Connectivity 0.0000 kmeans       2
## Dunn         0.5735 kmeans       3
## Silhouette   0.7329 hierarchical 2
```

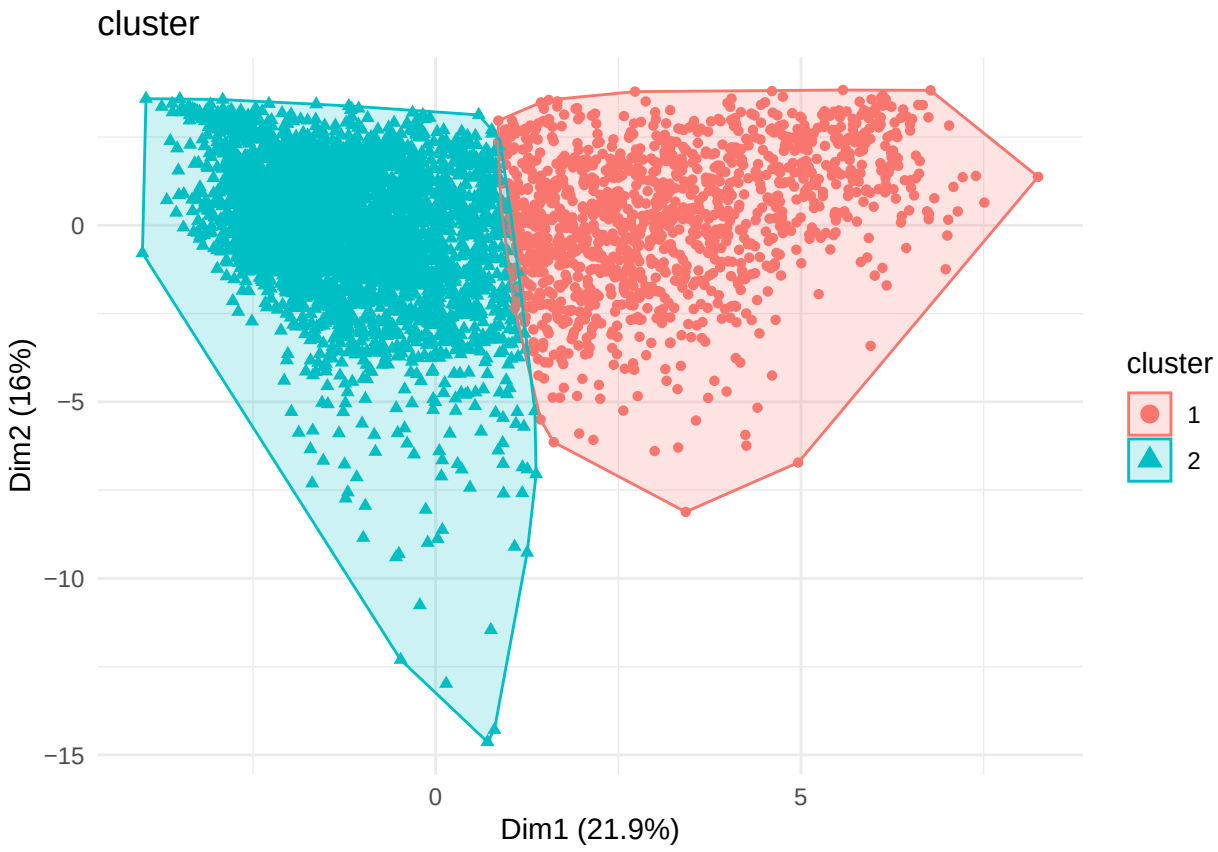
- Método Kmeans: Los resultados anteriores de validaciones internas, nuevamente nos da que la òptima solución son 2 clusters en las métricas de Siloutte y Connectivity mientras que para la métrica de Dunn el óptimo número de clusters sería 3.
- Método Jerárquico: E esta validación se agregó el método jerárquico en el cual las métricas Connectivity y Dunn dan resultados iguales mientras que en silhouette es mejor el resultado con 2 clusters.

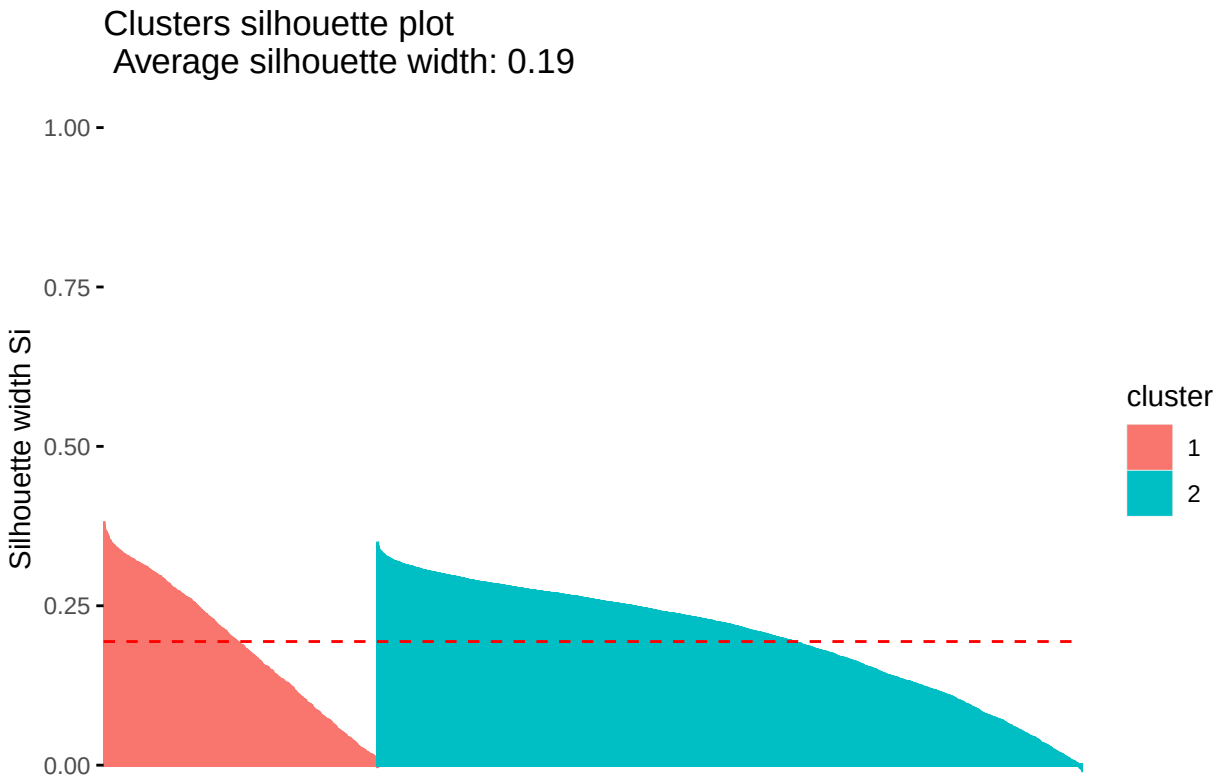
La conslución que podes sacar es que pareciera ser que la mejor solución es hacer clusters de 2 grupos ya sea por el método Kmeans o Jerárquico.

Cluster Kmeans

Tomando como referencia las validaciones anteriores, se realizará un cluster kmeans con 2 centroides. Este método se repetirá 25 veces distintas se elegirá el mejor.

```
##   cluster size ave.sil.width
## 1       1 1275          0.19
## 2       2 3283          0.20
```

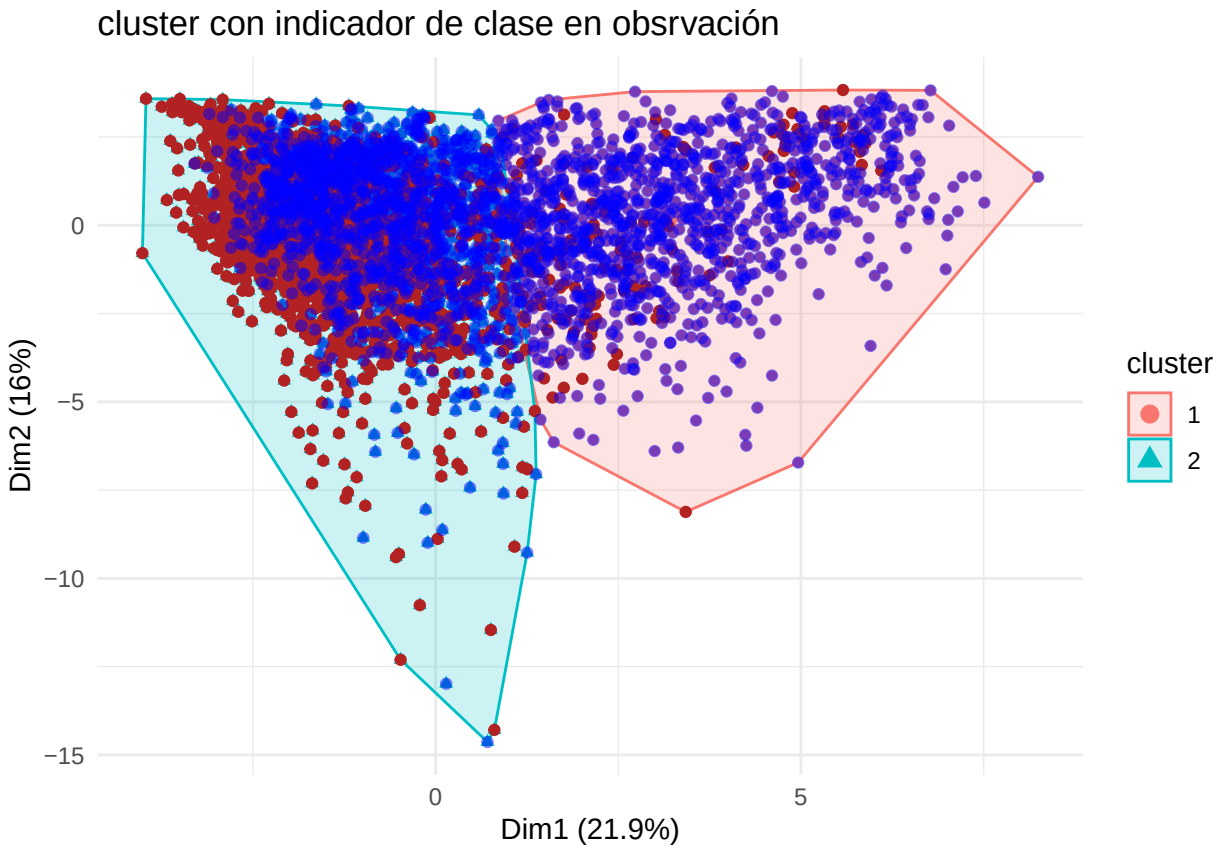




El cluster parece ser muy bueno. Por lo tanto el próximo paso es saber en que cluster cae cada observación según su target

Table 2: Resumen composición de cluster Kmeans según clase desertor

grupo	cant_integrantes	cant_desertores	cant_desertores_pct	cant_no_desertores	cant_no_dese
1	1275	142	11.13725	1133	
2	3283	1858	56.59458	1425	



Puede observarse que un grupo contiene solamente un 11,13% de datos erroneos, pero el otro grupo està muy balanceado. Por lo tanto, el cluster de 2 grupos a pesar de que tenga muy buenos valores en las validaciones realizadas, al verificar con el target real donde queda cada obsevación no es un buen resultado. En el gráfico se observa lo expresado en la tabla.

cluster jerárquico

Siguiendo lo sugerido se realiza el cluster jerárquico. Es de destacar que esta vez no puede mostrarse el dendograma completo ya que la cantidad de observaciones son muchas. Por lo cual, según lo sugerido en las validaciones, se hará el cote en 2 grupos y se verificará según la clase si la composición de los grupos se relaciona bien con e target.

Se realizan 4 cluster jerarquicos, cada uno utilizando las medidas de distancias “complete”, “average”, “single”, “ward”.

Table 3: Coef. cophenetic por cada tipo de cluster jerárquico

metodo	coeficiente_cofenetico
complete	0.7199776
average	0.8142721
single	0.7217462
ward	0.3765082

El tipo de distancia del cluster jerárquico que arroja el mayor coeficiente cophenetico es “average” con un valor de 0.8142721.

se puede observar que forma un grupo muy numeroso y otro muy chico y ambos estan muy mezclados en

Table 4: Composición de clusters según la clase deserto

grupo	cant_integrantes	cant_desertores	cant_desertores_pct	cant_no_desertores	cant_no_desertores_pct
1	4550	1996	43.86813	2554	56.13187
2	8	4	50.00000	4	50.00000

función el target. Por lo tanto, el método jerárquico no es adecuado.

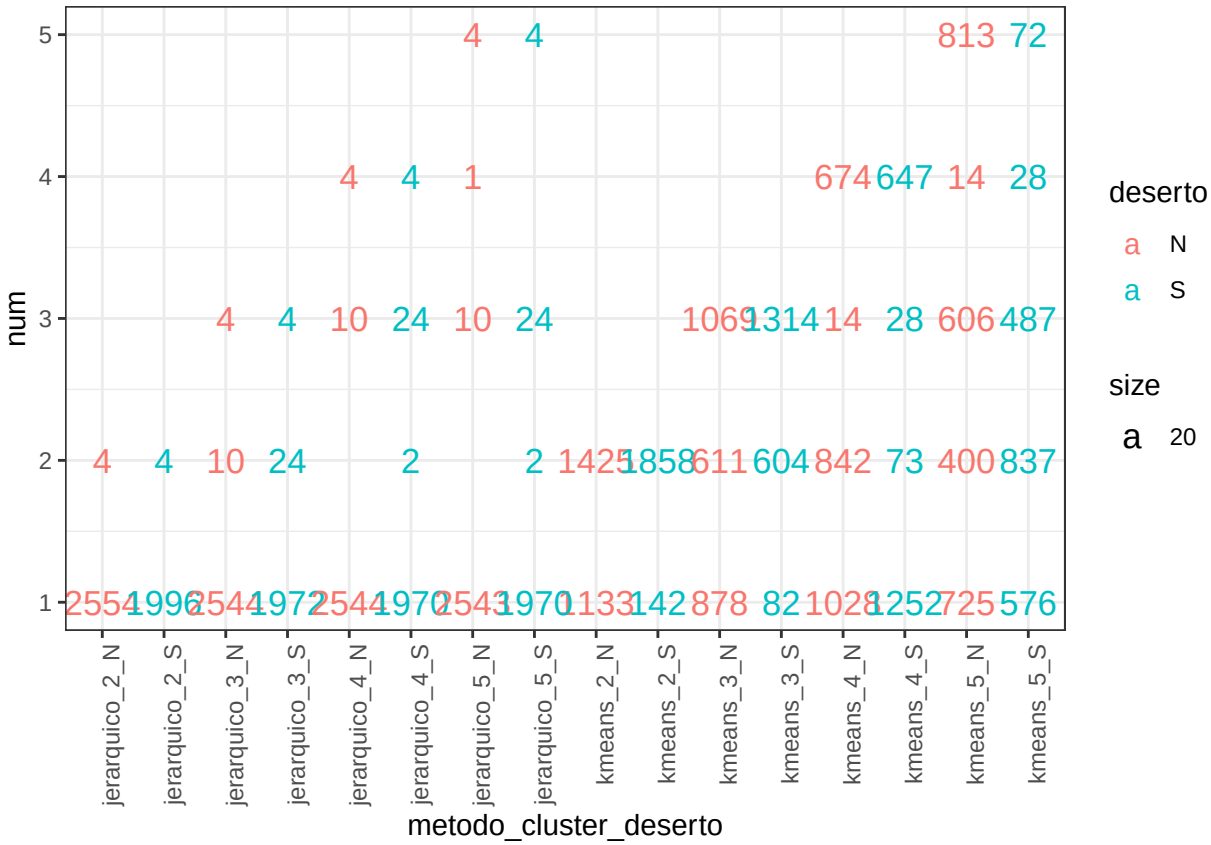
Extensión de Clusters

A pesar del estudio de validaciones y resultados anteriormente, igualmente queremos ver cuales son los resultados que nos arrojan hacer mas de 2 clusters. Como los resultados los referenciamos al target, campo que no se incluye en los datos para ahcer cluster al ser no supervisado, podría darse el caso de que en la situación real otro número de cluster sea óptima a la que arrojan las validaciones matemáticas.

Table 5: Resumen por tipo de cluster, cantidad de clusters y la composición de cada uno según la clase deserto

metodo	numero_clusters	1_N	1_S	2_N	2_S	3_N	3_S	4_N	4_S	5_N	5_S
jerarquico	2	2554	1996	4	4	NA	NA	NA	NA	NA	NA
jerarquico	3	2544	1972	10	24	4	4	NA	NA	NA	NA
jerarquico	4	2544	1970	NA	2	10	24	4	4	NA	NA
jerarquico	5	2543	1970	NA	2	10	24	1	NA	4	4
kmeans	2	1133	142	1425	1858	NA	NA	NA	NA	NA	NA
kmeans	3	878	82	611	604	1069	1314	NA	NA	NA	NA
kmeans	4	1028	1252	842	73	14	28	674	647	NA	NA
kmeans	5	725	576	400	837	606	487	14	28	813	72

A continuación mostramos un grafico resumen sobre



La figura muestra la misma información que el cuadro anterior. En el eje x indica que tipo de cluster es, cuantos clusters y la clase deserto ("S" y "N"). En el eje y se indica el numero de cluster, por lo que los cluster armados solo con 2 clusters, habrá información únicamente hasta esa altura. Por ejemplo, para el el caso de aplciar un método jerárquico de 2 clusters podemos observar que en el primer cluster tenemos 2554 casos Negatigos y 1996 casos positivos, mientras que el cluster número 2 esta conformado de 4 casos negativos y 4 casos positivos.