Informe del Cuarto Mini Proyecto de Estadística. Curso 2019-2020

Daniel Alberto García Pérez D.GARCIA@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Leonel Alejandro García López L.GARCIA3@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Roberto Marti Cedeño R.MARTI@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Tutor(es):
Msc. Dalia Diaz Sistachs, *Facultad de Matemática y Computación, Universidad de La Habana*

Tema: Estadística, Técnicas de Clasificación.

1. Ejercicios

Mostramos a continuación los enunciados de los ejercicios propuestos. Todos los códigos relacionados con la respuesta a los ejercicios propuestos se encuentran dentro de la carpeta Code adjunta al proyecto.

1.1 Ejercicio 1

Trabajando para encontrar una sintomatología estándar para comportamientos de anorexia y bulimia se realice un estudio de 55 adolescentes con desórdenes alimenticios conocidos. En cada observación las pacientes fueron valoradas en síntomas diferentes.

1. Realice un análisis de componentes principales. Calcule las componentes, interprete sus valores.
2. Agrupe los datos utilizando técnicas de clúster jerárquico.
3. Realice un agrupamiento utilizando el algoritmo k-means.
4. Diga si es posible construir un árbol de clasificación usando CART tomando como variable respuesta el diagnostico. De ser posible constrúyalo. De no ser posible existe otra variable que se pueda usar como dependiente. Explique su respuesta.

2. Solución

Comenzaremos analizando las variables presentes en los datos. Debido a la gran cantidad de variables presentes en la muestra, y al tamaño de sus gráficas pasaremos directamente a analizar la matriz de correlación de forma gráfica.

Como podemos apreciar en el gráfico (Figura 1), la matriz no se encuentra altamente correlacionada. Por lo que podemos decir que la mayoría de sus variables son independientes. Pasamos entonces al análisis de sus componentes.

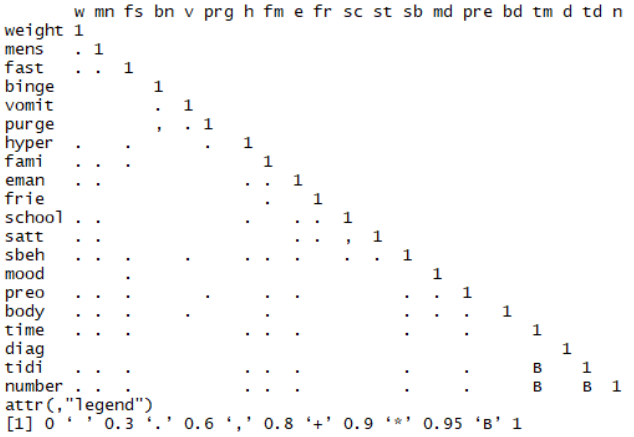

Figure 1: Matriz de correlación variante gráfica.

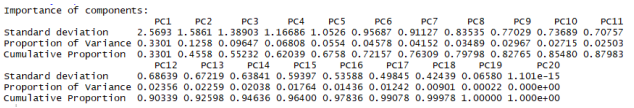

Figure 2: Importancia de las componentes.

Siguiendo el criterio de Kaiser en la importancia de los componentes (Figura 2), podemos llegar a la conclusión de que las componentes principales son las 5 primeras. Sin embargo, incumplimos el criterio del porcentaje dado que hasta la 5ta componente acumulamos solo un 68% de los datos aproximadamente. Se acordó entonces incluir una sexta componente dado que su valor propio es aproximadamente 1 y aglomeramos mas del 70% de los datos de la muestra.

Entonces nuestra primera componente se caracteriza por pacientes que presentan un alto valor de peso corporal, menstruación, que suelen restringirse el consumo de comida (fasting), son hiperactivos, mantiene buenas relaciones familiares, son independientes, poseen diver-

sos amigos, tienen un buen desempeño en la escuela o trabajo, poseen una actitud y comportamiento sexual positivo, también un buen estado de ánimo así como se preocupan por su percepción física, su peso y alimentos que ingieren. Por ultimo, su entrevista fue larga, interactuaron bastante durante la misma y se caracterizan por numero de paciente elevado.

Nuestra segunda componente se caracteriza por pacientes que no suelen darse atracones de comida, no suelen vomitar ni suelen provocarse el vómito. Poseen pocos amigos. Poseen una baja actitud y comportamiento sexual. Por otro lado, su tiempo de entrevista fue largo e interactuaron bastante durante la misma, su diagnóstico fue elevado y suelen pertenecer al ultimo grupo de pacientes.

La tercera componente se caracteriza por pacientes que, no se atracan de comida, no suelen provocarse pérdidas de comida y cuya entrevista fue de poca duración. Por otro lado, poseen buenas relaciones de amistad, un alto desempeño escolar o laboral y poseen una conducta sexual alta.

La cuarta componente se caracteriza por pacientes que, poseen poca menstruación, se limitan a la hora de comer, suelen estar deprimidos, no se preocupan por la comida ni por su peso, suelen ser delgados, y su diagnóstico es bajo. Por otro lado, son hiperactivos, son independientes, poseen una buen desempeño escolar o laboral, presentan una actitud sexual positiva, su entrevista fue de larga duración e interactuaron bastante en la misma. Son principalmente del ultimo grupo de pacientes.

La quinta componente se caracteriza por pacientes que, no se atracan de comida, tienen relaciones familiares pobres, dependen de su familia, tienen relaciones de amistad pobres, y su diagnóstico es bajo. Por otro lado, presentan un alto peso corporal, se restringen la ingestión de alimentos y poseen un alto autoestima.

La ultima componente, se caracteriza por pacientes que no suelen vomitar, no son hiperactivos, de complexión delgada, que se atracan de comida, poseen un buen desempeño laboral o educativo, presentan una actitud sexual positiva y un estado de ánimo elevado.

Siguiendo los resultados obtenidos del análisis de las componentes principales, realizamos un análisis de tipo clúster (Figuras 3 y 4). Podemos destacar lo acertado de considerar 6 componentes principales en lugar de 5 dado que, en el caso de 5 componentes se pierde información sobre una rama considerable de la jerarquía.

Después de la ejecución del algoritmo kmeans, con 6 componentes como hemos mantenido hasta ahora, (Figura 5) obtenemos una similitud por componentes, de un 44% resultado directo de la desproporción presente en la 2da de las componentes, lo que significa que en los datos propuestos abundan los pacientes con las características propias de esa componente. Sucesivos intentos de corridas del algoritmo con mas componentes incrementan dicha similitud, pero incumplen con los análisis determinados anteriormente, este comportamiento se debe posiblemente a que las observaciones de las variables aleatorias no superan un rango determinado entre 1 y 4.

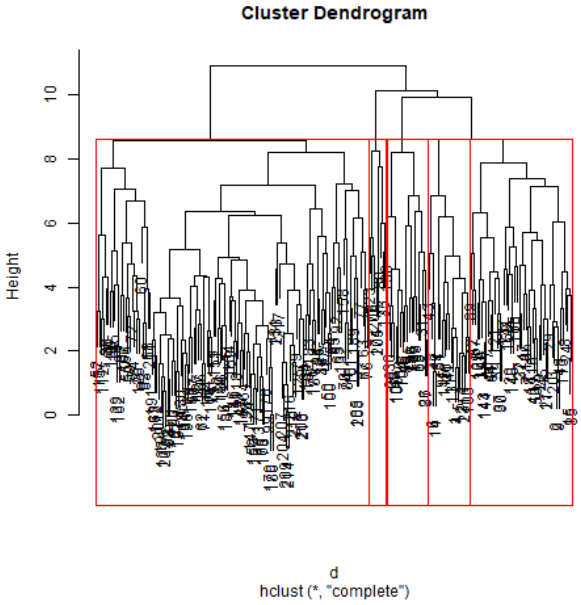

Figure 3: Clúster jerárquico de 5 componentes.

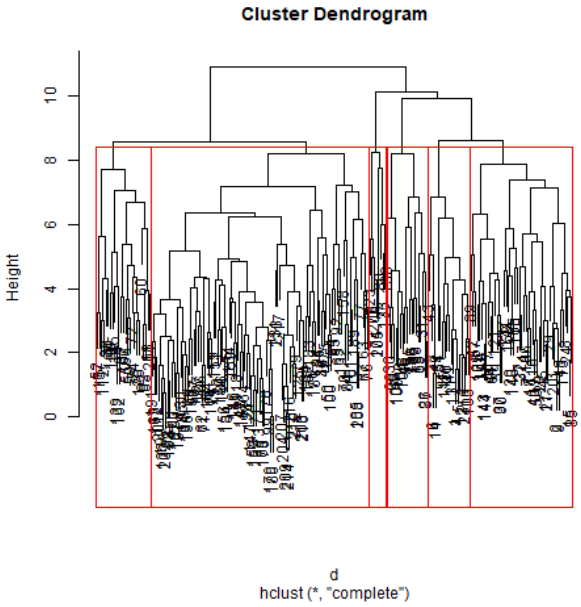

Figure 4: Clúster jerárquico de 6 componentes.

```
K-means clustering with 6 clusters of sizes 34, 38, 16, 68, 39, 22

Cluster means:
      weight       mens       fast     binge      vomit      purge      hyper       fami       eman       frie     school       satt
1 -0.8514008 -0.82843997 -0.9538750  0.4739157  0.2205333  0.3605424 -0.75915925 -1.42045624 -1.1334657 -0.7441224 -0.8912825 -0.84467763
2 -0.2516893  0.05373792  0.4190151  0.3530984  0.3238934  0.3618508  0.09065149  0.34120011 -0.2656432 -0.1448710 -0.8989233 -0.92066903
3  0.1247223  0.18064698 -0.4950644 -1.5346722 -1.6526883 -1.3704075 -0.48820932 -0.12565917 -0.1531967 -0.8712831 -0.5011269 -0.56944630
4  0.8423789  0.90605486  0.7590179  0.2578659  0.2722133  0.4351255  0.63429753  0.51207670  0.7969102  0.4687953  0.8326609  0.77963845
5 -0.3702839 -0.67963835 -0.3910673  0.3326523  0.3238934  0.2158384  0.15052926 -0.01614897  0.1864494  0.4206398  0.5074149  0.54177564
6 -0.2874741 -0.53960311 -0.5423358 -1.6129289 -1.1137515 -1.9133127 -0.85567574  0.14313859 -0.4717270 -0.1607837 -0.1786093 -0.06041189
       sbeh       mood       preo       body       time       diag       tidi     number
1 -0.86278801 -0.2942946 -0.7904799 -0.5745606 -1.0983333 -0.5148284 -1.1898999 -1.2151323
2 -0.09353829  0.2580393  0.5548859  0.3960238  0.5818180  0.0326910  0.5731208  0.5515753
3 -0.89080899 -0.3956120 -0.4541385 -0.7211593  0.8299598  0.4463062  0.9128018  0.8720003
4  0.84088774  0.4095944  0.7131642  0.5982288  0.7027120  0.1370137  0.7154506  0.7305315
5  0.13274721 -0.2538224 -0.3247764 -0.2562172 -0.5140477 -0.2944778 -0.5696591 -0.5480503
6 -0.69160528 -0.5192288 -1.0350919 -0.6664720 -1.1718938  0.5131230 -1.0163980 -0.9954341

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 5 1 1 1 5 5 1 5 1 6 6 1 6 1 6 6 6 5 1 1 5 5 1 5 1 1 1 1 5 5 6 2 5 2 5 1 2 6 3 2 2 5 3 2 2 5 5 5 4 5 2 2 2 3
 [70] 2 5 2 3 2 2 2 2 3 4 3 2 2 2 2 5 6 6 6 5 5 1 5 6 5 6 4 2 4 4 4 2 2 1 6 6 6 6 6 6 5 1 5 5 1 2 5 4 5 5 5 4 3 4 4 4 2 2 4 4 4 4 4 3 3 6 1 6
[139] 2 5 1 4 5 5 2 5 4 5 4 4 4 4 2 4 4 4 2 2 4 2 5 2 4 4 4 2 5 4 4 4 4 4 4 3 3 4 4 3 4 4 3 3 4 4 2 4 4 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 3 3 4
[208] 4 4 4 2 4 4 4 4 4 4

Within cluster sum of squares by cluster:
[1] 382.5128 519.4585 289.6231 476.0453 431.1327 324.7482
 (between_SS / total_SS =  43.9 %)
```

Figure 5: Resultado del algoritmo Kmeans con 6 componentes.

Finalmente se realizo el análisis del árbol de clasificación tomando como variable el diagnóstico de los pacientes. (Figura 6)

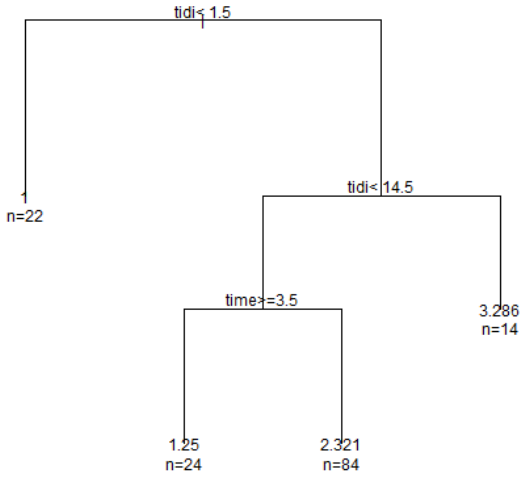

Figure 6: Árbol de Desición

Como resultado se obtuvo que para la predicción del diagnóstico empleando el árbol de decisión se toman en cuenta 2 variables, tidi y el tiempo de la entrevista. Si el paciente presenta un valor de tidi inferior a 1,5 podemos predecir que posee un diagnóstico de tipo 1. En cambio, si el paciente posee un valor de tidi superior a 14,5 su diagnóstico es de tipo 3. Finalmente si el tiempo de entrevista del paciente es superior a 3.5 se le diagnostica el trastorno alimenticio 2, eoc, su diagnóstico es de tipo 1.