Universidad Austral

Maestría en Ciencia de Datos

Introducción al Data Mining

INTRODUCCIÓN A LA MINERÍA DE DATOS

TRABAJO PRÁCTICO FINAL

Alumno

Mg. Arnaldo Viera

Docentes

Mg. Leandro Kovalevski
Mg. Pablo Beltramone

Universidad Austral
Maestría en Ciencia de Datos
Introducción al Data Mining

Trabajo Práctico Final

Fecha de entregra:

1. Introducción

ESTUDIO DE PREVENCIÓN CARDIOVASCULAR EN EL ADULTO MAYOR

El incremento de la expectativa de vida ha generado un aumento en la incidencia de enfermedades cardiovasculares y neurológicas. En los adultos mayores es donde se conjugan diversas patologías con alta morbi-mortalidad requiriendo gran cantidad de recursos materiales y humanos. Dado que, la determinación en análisis sanguíneos de rutina permite la detección de alteraciones que en determinadas circunstancias podrían progresar a patologías definidas con serias repercusiones cardiovasculares en este grupo poblacional, se estudió la frecuencia en adultos mayores de diversas patologías subclínicas y la asociación entre los factores de riesgo cardiovascular y dichas patologías subclínicas.

Se cuenta con información de 68 personas de ambos sexos mayores de 60 años, a quienes se les midieron las siguientes variables (disponibles en el archivo 'cardio.xls'):

sexo: Sexo del paciente (0: Femenino, 1: Masculino).

imc: Índice de masa corporal (es el cociente del peso en kg y la estatura al cuadrado en metros).

perimetro_abdo: Perímetro abdominal (en centímetros).

hto: Hematocrito (porcentaje del volumen de eritrocitos en el volumen de sangre).

glicemia: Glicemia (en mg/dL).

ct: Colesterol Total (en mg/dL).

hdl: Colesterol HDL (en mg/dL).

tgd: Triglicéridos (en mg/dL).

2. Consignas

a. Describa la distribución univariada de las variables presente en el conjunto de datos. ¿Se evidencian *outliers* en alguna de ellas?
b. Calcule e interprete la matriz de correlaciones.
c. Realice un análisis de componentes principales. ¿Qué porcentaje de la variabilidad total logran explicar las dos primeras componentes? ¿Es posible realizar una interpretación sobre los componentes? ¿Cuál? ¿Logran esas componentes diferenciar a los pacientes según el sexo?

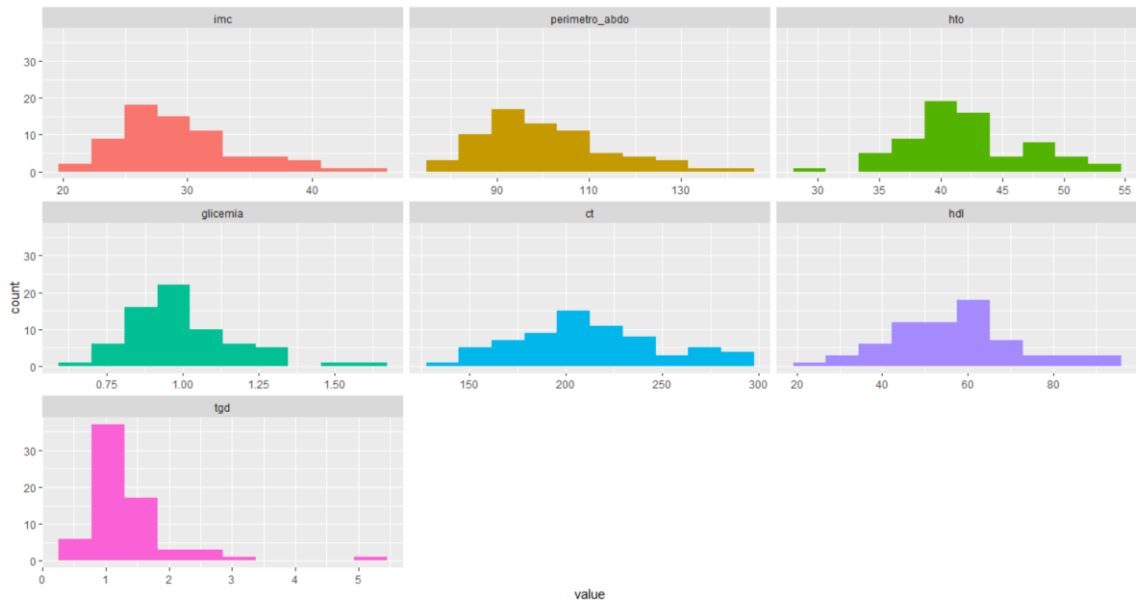
d. ¿Existen distintos subgrupos de pacientes en los datos? ¿Cuántos logra identificar? ¿Qué características tienen? Explique la metodología utilizada.

e. Construya la variable dicotómica 'obesidad' (índice de masa corporal mayor a 30) y construya un modelo predictivo utilizando sólo el sexo del paciente y las variables de los resultados de las pruebas de laboratorio (hematocrito, glicemia, colesterol Total, colesterol HDL y triglicéridos). ¿Qué capacidad predictiva tiene ese modelo?

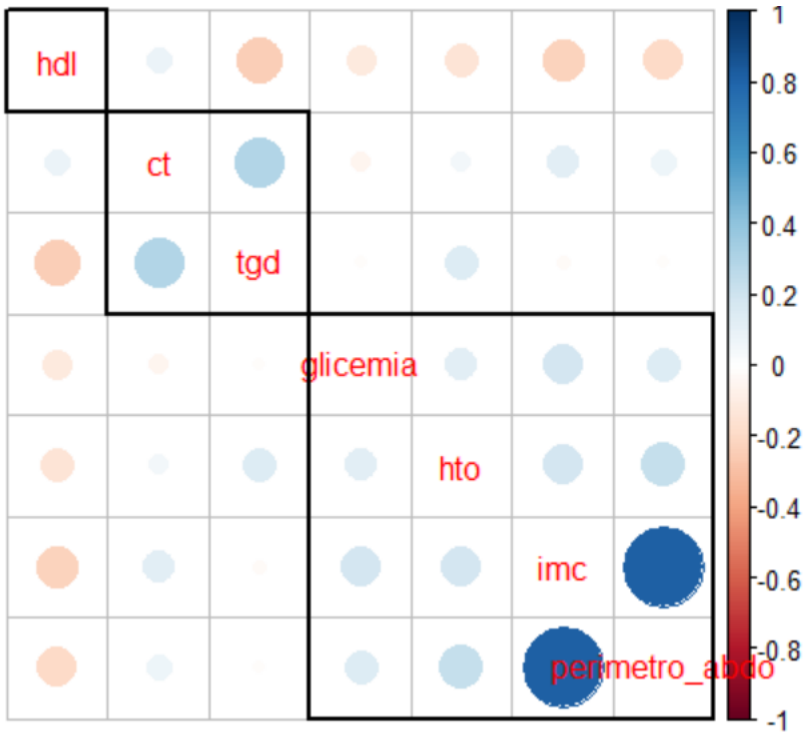
3. Respuestas

a. Se evidencian outliers en triglicéridos (tgd), glicemia, Hematorcrito (hto). No sólo son poco frecuentes, sino que en los gráficos se evidencia una interrupción de la curva.
Por otro lado, en cuando a las distribuciones todas tienen forma de campana invertida, con 1 pico y orientadas hacia la izquierda o derecha según el caso. El caso de hto tiene un segundo pico, pero podría atribuirse a la poca cantidad de observaciones.

Distribución de variables del dataset <<cardio>>



b. La matriz que se escogió para acompañar este párrafo sobre el dataset <<cardio>> es la que se logra con la función función **cor.plot(cor(cardio_cont))**



Se destaca una correlación fuerte (0,82) a simple vista: imc/perímetro_abdo. Entendiendo el origen de cada variable y cómo se construyen, tiene sentido que tengan una fuerte correlación positiva: a medida que la relación del peso y la altura suben, el perímetro abdominal también sube. Lo que quizás cabe destacar que las distintas variables tienen información redundante (lo veremos en el próximo punto).

El resto de las correlaciones no son significativas, sean estas positivas o negativas. Lo que llama la atención es que la variable hdl (que vendría a ser el colesterol "bueno") tiene una leve correlación negativa con el resto de las variables de forma bastante homogénea.

Por su parte, ct (colesterol total) tiene correlación casi nula, aun cuando el hdl forma parte de la medición del ct. En este sentido, podríamos inferir que la no-correlación entre hdl y ct se debe a una variable que también forma parte del ct, pero que no está incluida en el dataset <<cardio>>: el ldl o colesterol "malo".

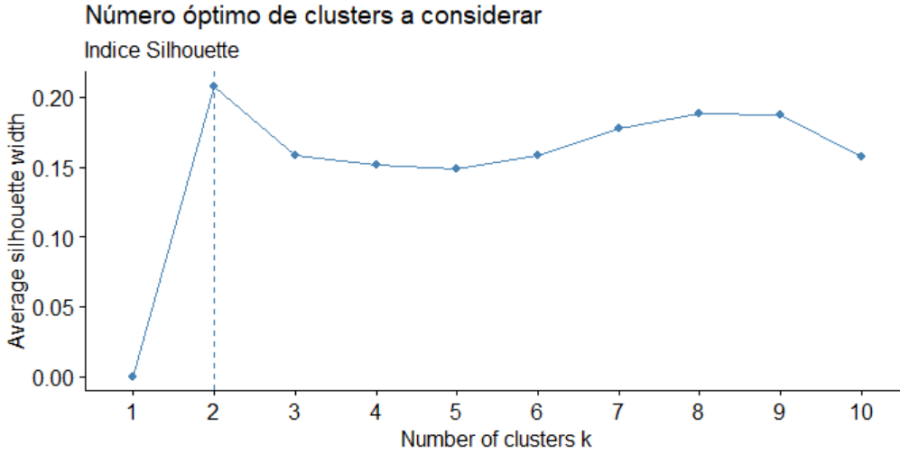
c. Continuando con el análisis anterior, al tener poca correlación entre las variables, no existen pocos componentes que nos expliquen los datos. Para llegar a un 80% o más de explicación tenemos que llegar a la 4ta (o 5ta) componente de las que estamos analizando.
Por otro lado, las primeras 2 componentes logran explicar un 49% de variabilidad. No logran diferenciar la variable sexo.

summary(xPCA)

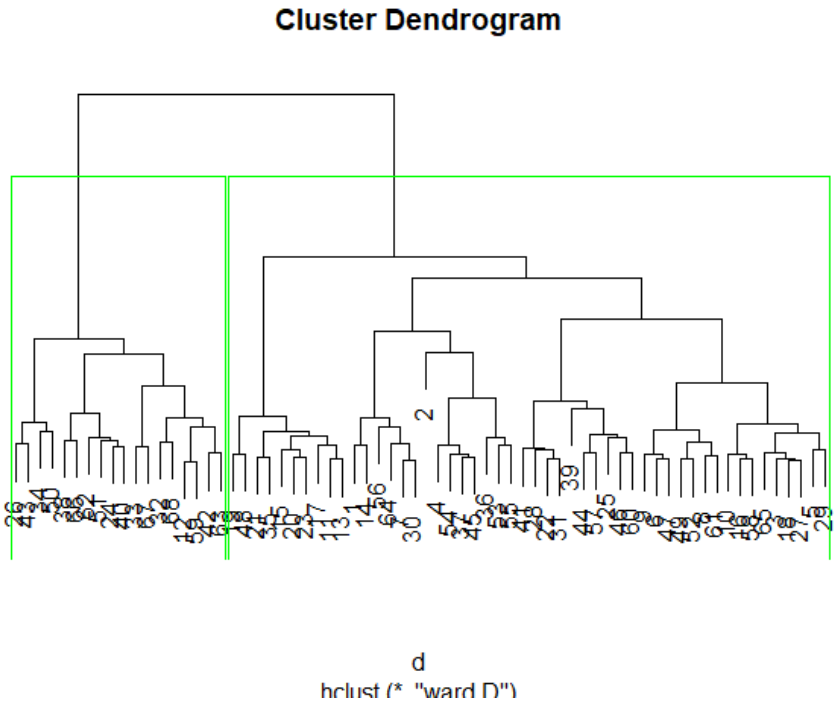
```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7
Standard deviation     1.4542 1.1563 1.0511 0.9536 0.9164 0.72009 0.41904
Proportion of Variance 0.3021 0.1910 0.1578 0.1299 0.1200 0.07408 0.02508
Cumulative Proportion  0.3021 0.4931 0.6509 0.7809 0.9008 0.97492 1.00000
```

d. Para detectar agrupaciones dentro del dataset se realizó un análisis de cluster jerárquico, aglomerativo y politético y se encontraron 2 grandes grupos de acuerdo a los siguientes criterios:

1° - Indice de Silhouette:



2°- Mediante la visualización del dendograma donde las distancias a la agrupación posterior eran “grandes”.



El grupo 1 tienen más bajas todas las variables, entre un 10 y un 20% menos que el grupo 2; excepto la variable el hdl que tiene valores levemente superiores en el grupo 1. Es la misma variable que en la matriz de correlaciones notábamos que tenía una leve correlación negativa con el resto.

e. Se creó la variable Obesidad según el criterio. Se realizó un modelo basado en ábroles de decisión (sin poda) y con los criterios descritos. La capacidad de predicción es 57.14% según el 30% de los datos que se dejaron para test.

4. Código

https://github.com/arnaldoviera/Individuales_TPAustral

⇨ **TPFinal_IDM**