Contents

Carátula

MAESTRÍA EN CIENCIA DE DATOS

Rosario

Cohorte 2021 - 2022

Ing. Emiliano Olivares

36227254

emiliano.olivares@unc.edu.ar[1]

Presentado a fin de cumplimentar con el Final Integrador

Materia: Data Mining

Fecha: 2021-09-15

Realizado utilizando R Studio Versión: 4.1.0

[1] mailto:emiliano.olivares@unc.edu.ar

Introducción

Exploracion sencilla del dataset

Realizamos una exploración aleatoria de los datos, tomando un sampleo de 5 individuos.

Table 0.1: Muestra Dataset

id	sexo	imc	perimetro_abdo	hto	glicemia	ct	hdl	tgd
50	1	29.34803	99	44.9	1.55	159	38	1.62
43	0	36.84049	104	47.8	1.30	191	70	0.74
6	0	31.40496	106	40.6	0.99	222	58	1.09
9	0	27.62618	93	43.9	0.96	220	57	1.44
28	1	24.25867	100	52.2	1.10	226	60	1.15

Ademas, se suma la metadata asociada a nuestra base de datos, correspondiente a la definición precisa de variable que comprende nuestro estudio clínico, ademas de un elemento de suma importancia: unidades de medida.

Table 0.2: Metadata: Descripcion de variables

Variable	Descripcion
id	Identificacion anonimizada de usuario
sexo	Sexo del paciente (0: Femenino, 1: Masculino)
imc	Índice de masa corporal: 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)

Analisis univariado y bivariado

Table 0.3: Medidas de tendencia central y dispersion univariadas para hombres

	imc	perimetro_abdo	hto	glicemia	ct	hdl	tgd
Min	22.22	78.5	35.4	0.78	154	24	0.7
Q1	26.35	94	40.7	0.9	184	41	1.01
Mediana	29.31	102	43.7	0.96	202	49	1.18
Q3	30.97	108	47.6	1.06	227	58	1.68
Max	45.54	139	52.2	1.65	293	93	5.21
Promedio	29.32	103.44	43.85	1.01	208.24	51.06	1.47
DesvEst	4.91	14.55	4.65	0.19	33.91	15.02	0.84
Valores normales	20 a 25	Menor a 102	38.3 a 48.6	70 a 100	Menor a 200	Menor a 40	Menor a 1.7

Table 0.4: Medidas de tendencia central y dispersion univariadas para mujeres

	imc	perimetro_abdo	hto	glicemia	ct	hdl	tgd
Min	21.91	75	28.2	0.68	143	34	0.54
Q1	25.56	88.5	38.1	0.86	193	56.5	1
Mediana	27.73	94	40.6	0.95	220	62	1.11
Q3	31.35	105.75	41.65	1.1	242	69	1.31
Max	43.12	127	48.9	1.33	296	89	2.75
Promedio	29.23	97.53	40.19	0.98	217.37	62.83	1.25
DesvEst	5.26	12.73	3.98	0.17	40.82	12.34	0.5
Valores normales	20 a 24	Menor a 88	35.5 a 44.9	70 a 100	Menor a 200	Menor a 50	Menor a 1.7

Se obtuvo valor de referencia de Mayo Foundation for Medical Education and Research (MFMER). Se separaron las tablas al haber valor de referencia que poseen diferencias entre sexo. La tabla es util para identificar si la media/promedio de cada variable, diferenciada segun sexo, se aproxima a los valores normales de referencia obtenidos.

Ademas, deberemos prestar especial atencion a aquellos valores maximos o minimos, ya que podrian tratarse de outliers que estan empujando nuestro promedio (mas aun, al tratarse de pocos individuos). En nuestra tabla, al ser un analisis univariado, no podemos detectar si los valores extremos en cada una de las variables corresponden al mismo individuo. Por ejemplo, para la tabla de mujeres: 296 en ct (límite normal es 200) y 127 cm en perimetro_abdo (límite normal 88), ¿corresponden al mismo individuo? Podemos aplicar un filtro sobre el individuo con valor extremo máximo en perimetro_abdo y ver todas las realizaciones de variables para ese individuo (ver tabla a continuación).

Además, se detecta que las variables glicemia, hdl y tgd se encuentran escaladas (se desconece puntualmente el escalamiento utilizado), por lo que no podremos compararlos directamente con el valor de referencia. En caso de conocer el escalamiento podríamos aplicarlo sobre el valor de referencia normal y comparar. Esta

información (qué escalamiento se aplicó) no se encuentra en la metadata del problema por lo que el análisis quedará pendiente.

Table 0.5: Individuos con ct elevado, sexo femenino

	imc	perimetro_abdo	hto	glicemia	ct	hdl	tgd
Max	43.12	127	48.9	1.33	296	89	2.75

Distribucion univariada

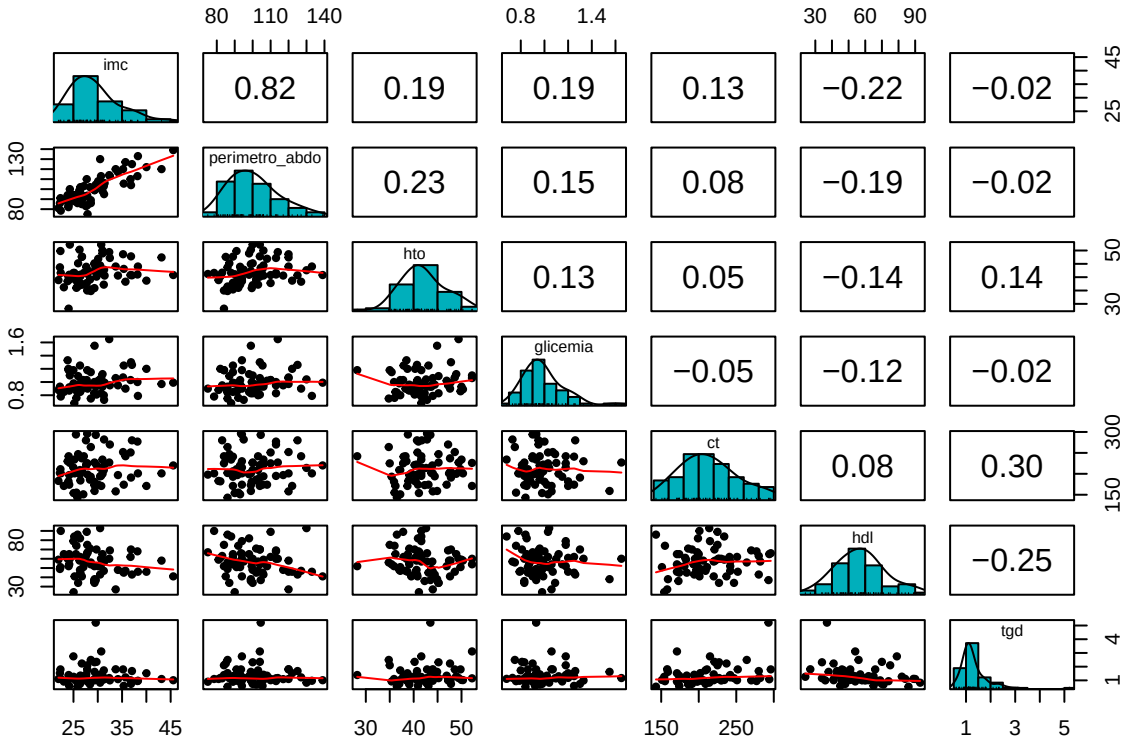

Figure 0.1: Densidad univariada y correlacion lineal bivariada

Se descartó el ID y el Sexo como variable ya que no corresponden al análisis de distribución (Sexo es binaria). Además, la primera, nos servirá como identificación anonimizada de individuos y la segunda para implicar el sexo en el análisis de clusterización/aglomeración.

Se observan, en general, distribuciones normales con asimetría hacia la izquierda. Es decir, contamos con individuos con valores distintos por encima del promedio. Además, observando el gráfico de tgd, se evidencia la posible aparición de outliers (observemos como la distribución llega hasta el valor escalado 5). Respecto al analisis bivariado, la correlación lineal es en general baja en magnitud, excepto para imc y el perímetro abdominal que resulta de magnitud importante. Es decir, que existe una correlación positiva entre el índice de masa corporal y el perímetro abdominal. Sin implicar causalidad, un perímetro abdominal

por encima del promedio se ve en individuos que poseen un imc por encima del promedio, en nuestro set de datos.

Correlación: otro forma de visualizar

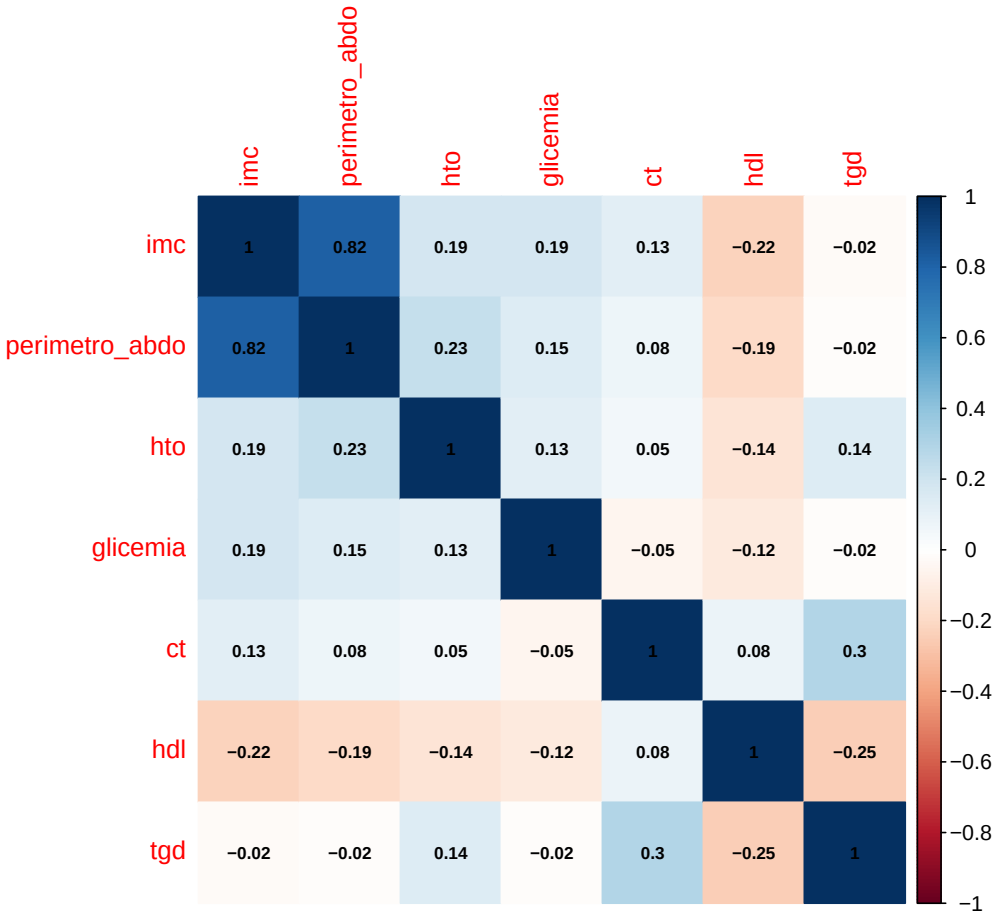

Figure 0.2: Matriz de correlacion lineal

El gráfico muestra la matriz de correlaciones lineales de manera clara. Se observa rápidamente por la escala de colores que solamente la correlación entre perimetro_abdo y imc es atendible.
Para reforzar la visualización anterior, podemos estudiar con que grado de significación podemos aseverar la correlación lineal. Se muestra a continuación.

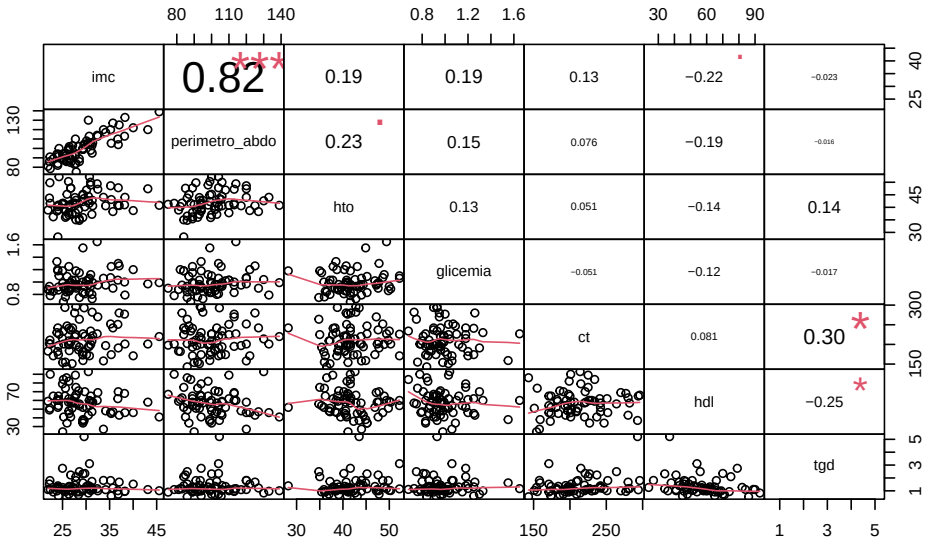

Figure 0.3: Correlacion

El tamaño de la letra nuevamente simboliza magnitud. Los tres asteriscos rojos representan que podemos aseverar esa correlación lineal con significancia estadística.

Boxplots y outliers

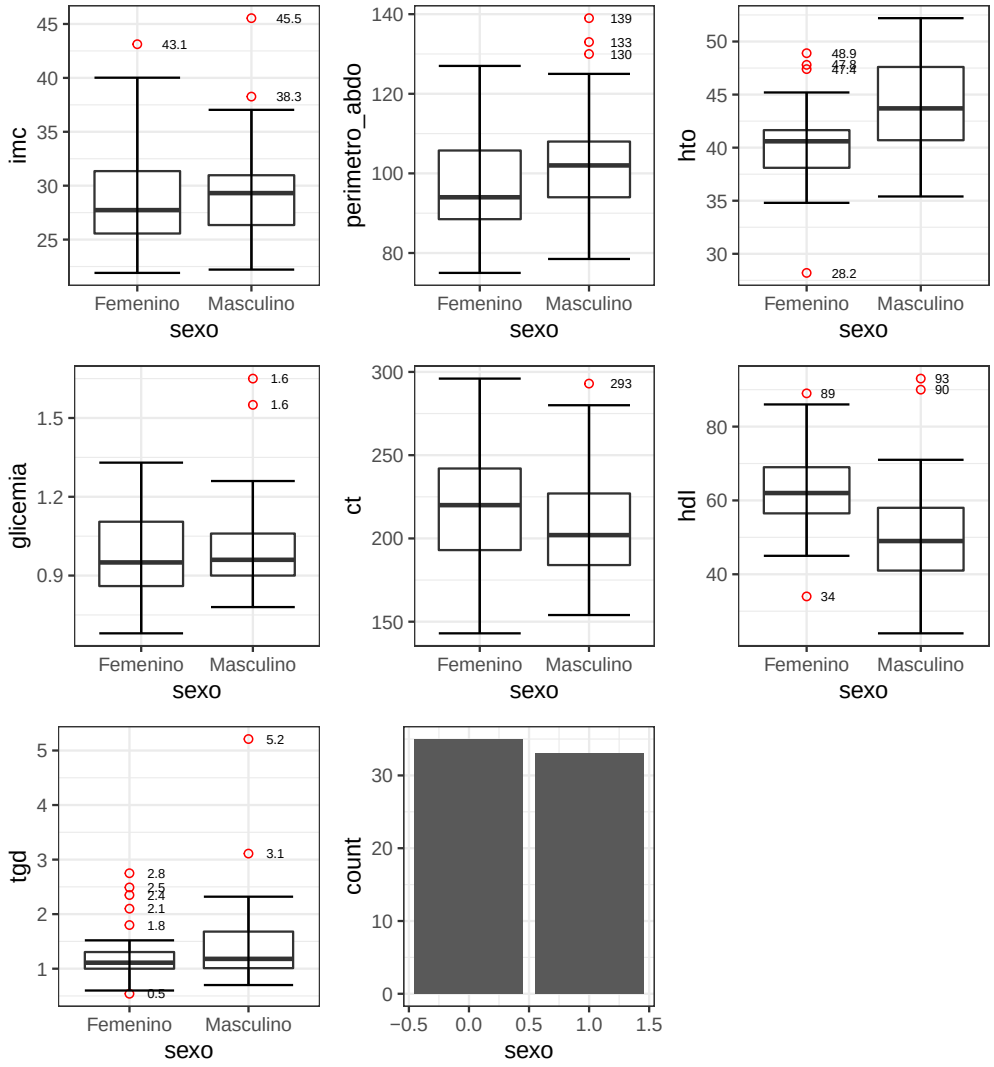

Figure 0.4: Boxplots segun sexo para cada variable numérica continua

Se observa que existen realizaciones muy por arriba del promedio para la mayoría de las variables en el caso de individuos masculinos. Por ejemplo, existen individuos con un imc muy alto y su correspondiente perímetro abdominal elevado. ¿Son los mismos individuos?

Table 0.6: Individuos extremos, masculinos

id	sexo	imc	perimetro_abdo	hto	glicemia	ct	hdl	tgd
32	1	37.03561	125	40.7	1.26	280	43	1.68
68	1	35.20818	120	48.6	1.07	252	46	1.80
42	1	32.49008	113	47.2	0.95	270	69	1.00

Un experto de dominio nos aconsejaría precaución: ya detectamos que no existe fuerte correlación entre la mayoría de nuestras variables. Mientras un observador no experto tendería a pensar que las realizaciones

extremas en nuestro boxplots corresponden a los mismos individuos, una simple tabulacion nos detalla como un individuo con un altísimo valor de ct se corresponde con un valor promedio de tgd. El individuo 32 puede tener una alimentación o una condición hormonal que cause esa relación lipídica a priori no esperable. Por ejemplo ciertos deportistas pueden tener valores elevados de ct y a la vez tgd bajos.
El analisis multivariado para la identificación de "individuos outliers" no es tan superficial como puede serlo el detectar realizaciones outliers para la distribución univariada (que quedan expresamente marcados en nuestros boxplots).
Para el caso de individuos femeninos, la distribución de outliers en tgd resulta interesante. Podria, nuevamente, consultarse con un experto de dominio para estudiar a que puede deberse esta situacion.

Analisis de Componentes Principales

Table 0.7: Autovalores

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.1147820	30.211171	30.21117
Dim.2	1.3370175	19.100250	49.31142
Dim.3	1.1047924	15.782748	65.09417
Dim.4	0.9094237	12.991768	78.08594
Dim.5	0.8398588	11.997983	90.08392
Dim.6	0.5185336	7.407622	97.49154
Dim.7	0.1755921	2.508459	100.00000

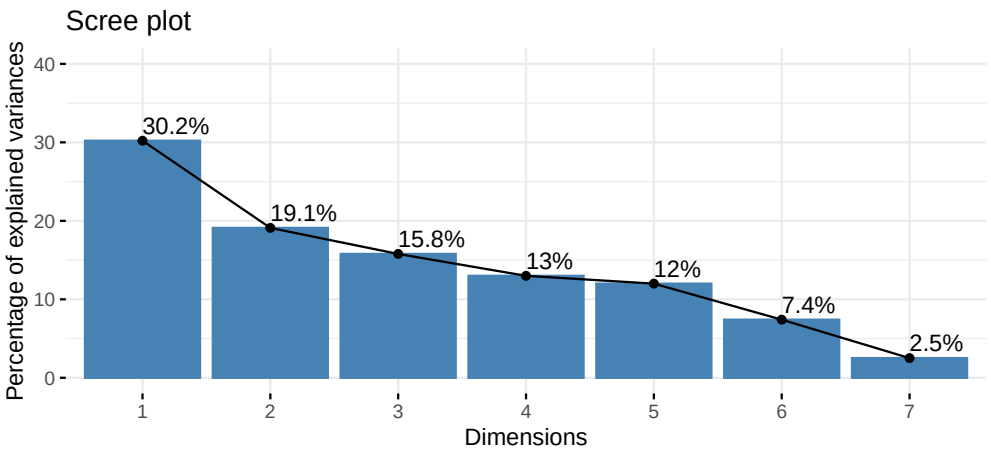

Figure 0.5: Scree Plot de PCA

En las visualizaciones anteriores (tabla y scree plot) se observa una situación esperable, en función de la correlación lineal bivariada de magnitud baja, las primeras dimensiones resumen una porción reducida de la variabilidad total de los datos.

Con las dos primeras dimensiones logramos explicar un 49.3 % de la variabilidad total, porcentaje honestamente bajo.

Siguiendo el enunciado, se realizará un análisis más profundo tomando estos dos primeros ejes factoriales.

Análisis de las dos primeras componentes

Table 0.8: Variabilidad

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
imc	0.8785296	-0.2120159	0.2616912	-0.1261235	-0.0753029	0.0640351	0.2984534
perimetro_abdo	0.8705989	-0.2262081	0.2534029	-0.1400010	0.0216003	0.1493705	-0.2903379
hto	0.4549259	0.2063210	-0.2457950	0.3821048	0.7320817	-0.0881391	0.0185137
glicemia	0.3465680	-0.1953584	-0.4114817	0.6694029	-0.4729356	0.0185045	-0.0172601
ct	0.1783759	0.6241889	0.5831280	0.2454863	-0.1933952	-0.3741520	-0.0296148
hdl	-0.4444783	-0.2306512	0.6125108	0.4644292	0.1520541	0.3671702	0.0209957
tgd	0.1690544	0.8469661	-0.1647940	-0.0620816	-0.1165672	0.4573914	0.0162183

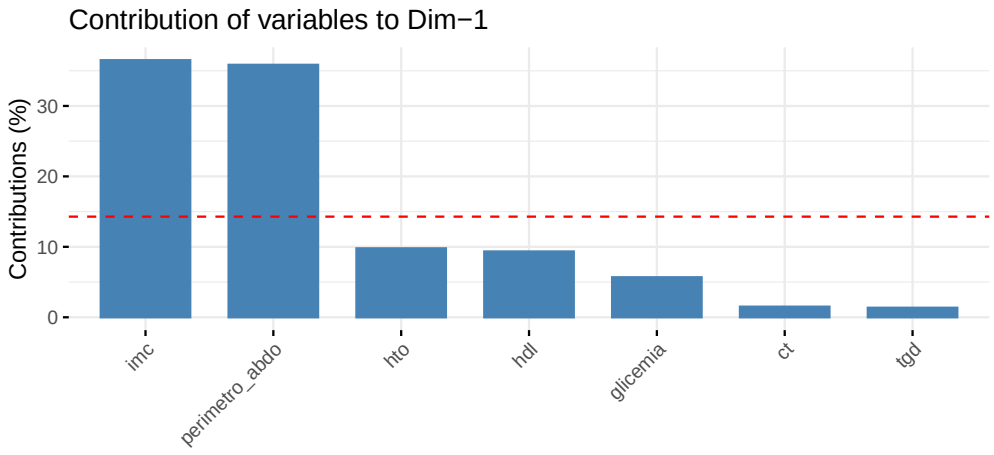

Figure 0.6: Fviz - Contribucion de variables a la Dimension 1

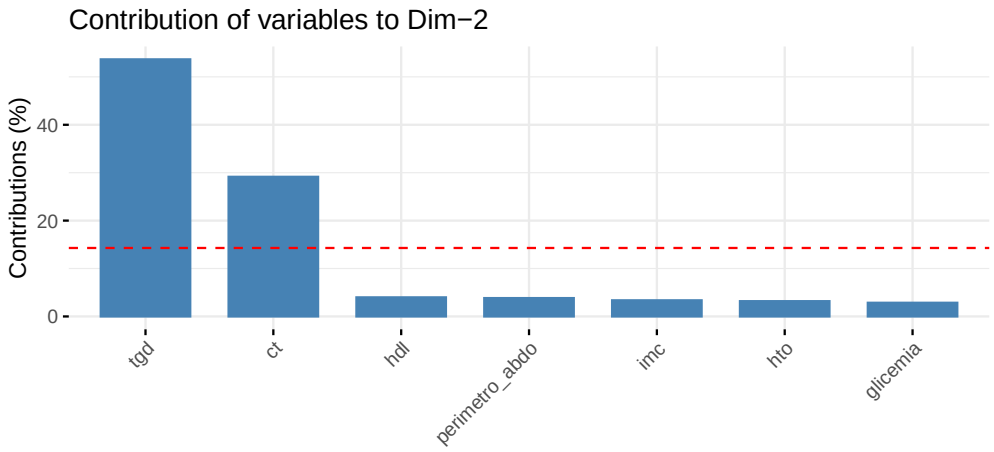

Figure 0.7: Fviz - Contribucion de variables a la Dimension 2

Observamos otro ordenamiento lógico en nuestros ejes factoriales. La primera dimensión muestra que la contribución de las variables más fuertemente correlaciones -perímetro abdominal e indice de masa corporal- son las que superan el umbral de aporte significativo (línea de puntos roja, que se corresponde con el valor esperado si las contribuciones fueran perfectamente uniformes).

Por su parte, la segunda dimensión se explica principalmente por el aporte de tgd y ct, que poseen una correlación baja pero de mayor magnitud que el resto de las correlaciones lineales bivariadas (aparte, obviamente, de perimetro_abdo-imc). Esto es relativamente lógico en tanto el colesterol total se computa como la suma de los componentes clínicos de colesterol (entre ellos tgd).

En un análisis gráfico que muestre las dos primeras dimensiones obtendremos alineados al eje D1 hacia los valores positivos aquellos individuos con un alto valor de imc y perímetro abdominal. Respecto al eje D2 hacia los valores positivos del eje observaremos los individuos con altos valores de ct y tgd.

Calidad de representación de las variables

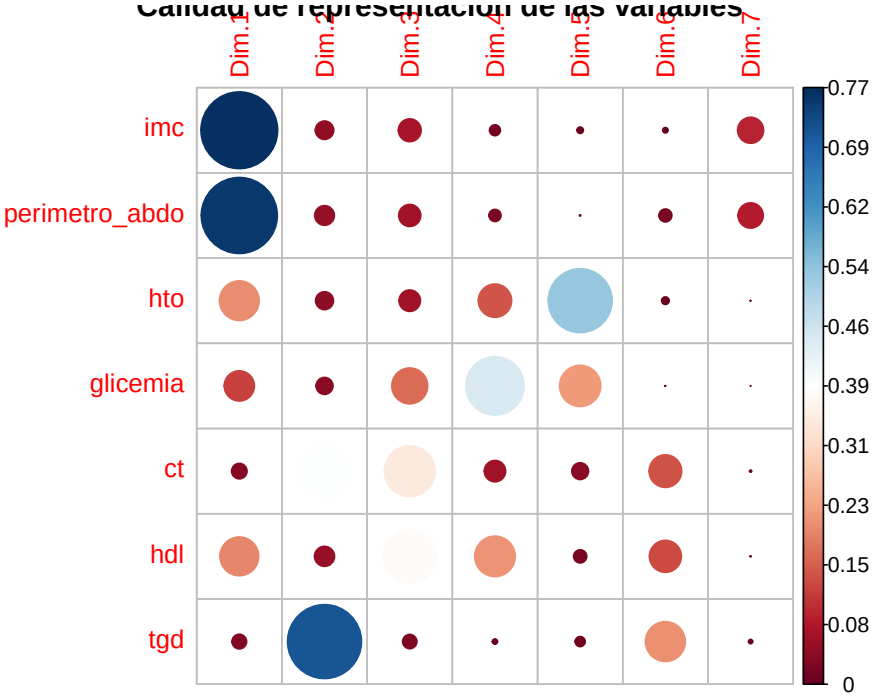

Figure 0.8: Calidad de representacion de las variables, cos2

Sabemos que un valor cercano a 1, significa un alto porcentaje de explicación de esa variable en esa dimensión. Observamos que esto se cumple para imc y perimetro_abdo en la primera dimensión y tgd principalmente en la segunda.
Hay una alta dispersión de la explicación de la varianza de los datos en el resto de las dimensiones. Esto se debe también a que nuestro dataset cuenta con una estructura de correlación baja.

Círculos de correlación

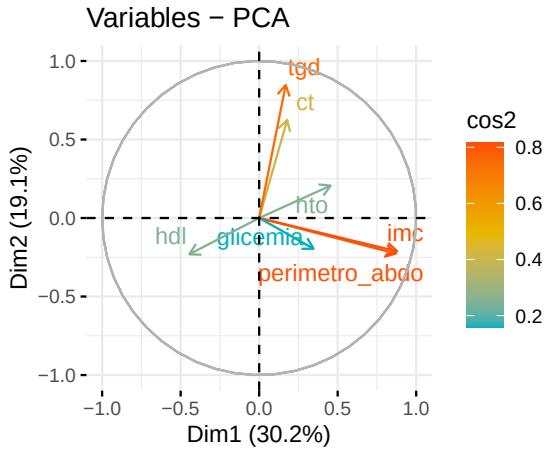

Figure 0.9: Circulos de correlacion: Dimension 1 y 2

Se observa en el círculo de correlación los elementos mencionados anteriormente y esperados: las flechas que representan a perímetro_abdo e imc se ubican hacia los valores positivos de la dimensión 2, por su

parte tgd y ct se ubican hacia los valores positivos de la dimensión 2. La cercanía con los ejes mencionados también era esperable, una cercanía al eje respresenta un valor de cos2 igual a uno y un fuerte componente de explicación de la varianza en esa dimensión.

Segregación por sexo

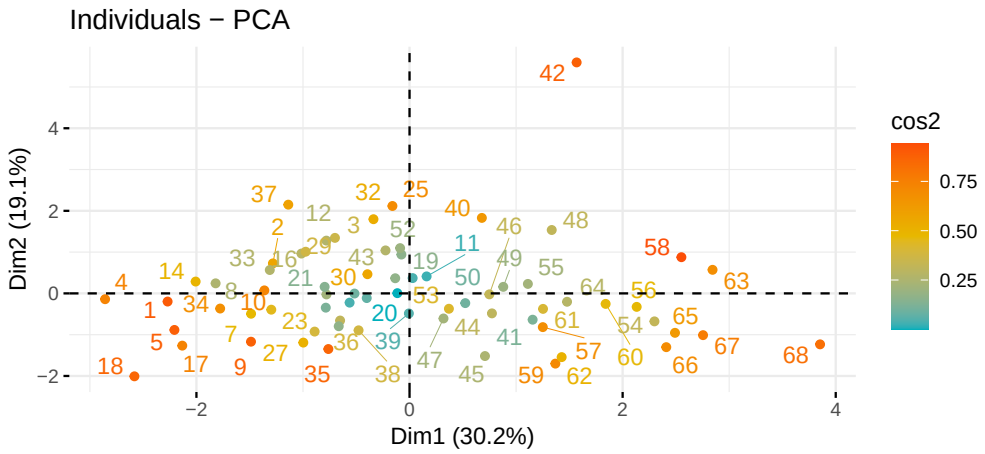

Figure 0.10: Individuos y su ubicacion en los dos primeros ejes factoriales

En principio, se aprecia una leve tendencia de la dimensión uno a separar los individuos según sexo. Pero podría deberse a una casualidad en función de la cantidad relativamente baja de la población y que, proporcionalmente, tengamos una mayor cantidad de individuos masculos con alto imc. Es decir, nuestro PCA parece guiarse más por ello que por sexo.
Se acompañan dos tablas, una que identifica los individuos hacia los valores positivos de la dimensión uno y otra hacia los valores negativos. Además, se tomó aquellos cuyo cos2 según la codificación en colores es mayor a 0.6.

Table 0.9: Individuos sobre valores positivo de la dim 1, con cos2 mayor a 0.6

sexo	imc	perimetro_abdo	hto	glicemia	ct	hdl	tgd
Masculino	32.35996	114	49.4	1.65	227	60	1.17
Masculino	33.69005	120	44.0	1.06	158	27	1.81
Masculino	35.20818	120	48.6	1.07	252	46	1.80
Masculino	37.03561	125	40.7	1.26	280	43	1.68
Masculino	38.26531	133	44.0	0.79	201	46	0.97
Femenino	40.01763	122	38.7	1.20	171	46	1.52
Femenino	43.12500	120	47.4	0.97	201	58	1.10
Masculino	45.54111	139	40.8	0.99	220	41	1.01

Table 0.10: Individuos sobre valores negativo de la dim 1, con cos2 mayor a 0.6

sexo	imc	perimetro_abdo	hto	glicemia	ct	hdl	tgd
Femenino	21.90758	81.0	38.9	0.88	183	59	1.10
Masculino	22.38835	78.5	41.6	0.78	201	90	1.10
Femenino	23.02632	89.0	36.2	0.85	171	65	0.91
Femenino	24.12879	94.0	40.9	0.92	174	65	0.60
Femenino	24.13138	91.5	39.1	0.87	197	54	1.30
Femenino	25.61291	92.0	36.0	0.76	155	74	0.94
Femenino	25.83968	86.0	36.5	0.84	143	86	0.54
Femenino	27.91111	75.0	41.2	0.90	198	67	0.90

Aglomeracion

Los métodos de clusterización enfrentan el problema de la colinealidad (alta magnitud de correlación lineal entre dos variables): cada variable tiene asignado un peso para influir en la clusterización de cada individuo, pero si dos variables están muy correlacionadas implican, en nuestro análisis, lo mismo. Ese decir, ese peso se “duplica” y “tira” el método hacia lo que indiquen esas variables. Bajo múltiples colinealidades, se presenta el problema de que las variables más correlacionadas ocultan la información de las que no lo están.

En función de la argumentación anterior se ha dropear la variable perimetro_abdo y mantener la variable imc que contiene mayor información intrínseca y es un número con mayor expansión de uso clínico. Además, se usará k-means con un cálculo de distancia euclidiana, solucionado el problema de correlación.
Se llevará adelante un proceso de normalización, para evitar problemas de escala y one hot encoding sobre variable sexo.
Se aclara, además, que se probó -no se muestra en este trabajo ya que no se consideró que aporte demasiado- una implementación con one-hot-encoding sobre la variable sexo, pero ese método consiguió unificar a todos los hombres en un mismo cluster. Esto puede deberse a que las distancias respecto a la dimension sexo utilizada como variable dummy puede superar el resto de los aportes (provenientes de otras dimensiones), dejándonos con sexo como variable dominante a la hora de la clusterizacion. Tambien, nuevamente, puede estar impactando el numero bajo de individuos. Otro argumento fue el siguiente: ¿por qué incluir una variable que nos genera una clusterización tal que incluye un grupo para solamente diferenciar por sexo cuando esa diferenciación podemos verla en los datos originales de maenra sencilla y, a priori, no parece realizar un aporte sustancial?
Por esto, se decidió dejar de lado la variable sexo para la implementación del método kmeans.

Silhoutte y SCE

Implementamos silhoutte y SCE para identificar el numero de clusters recomendados.

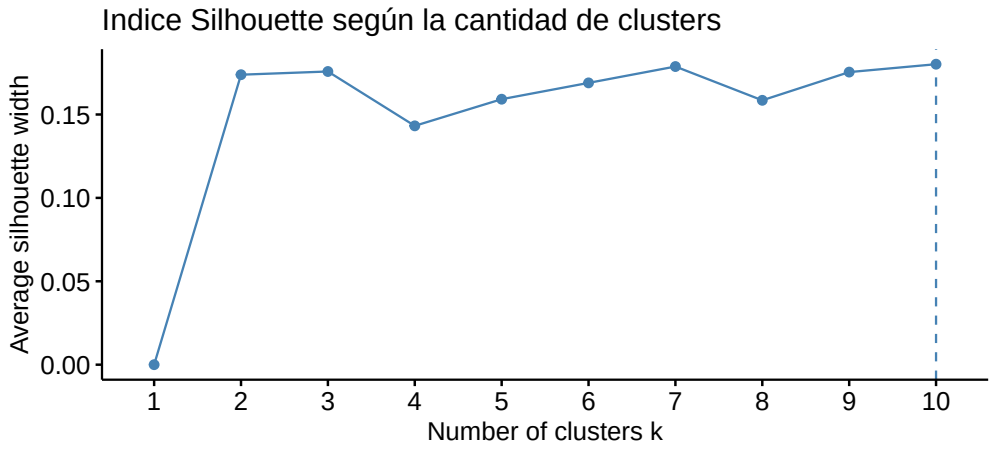

Figure 0.11: Indice de Silhouette segun cantidad de clusters

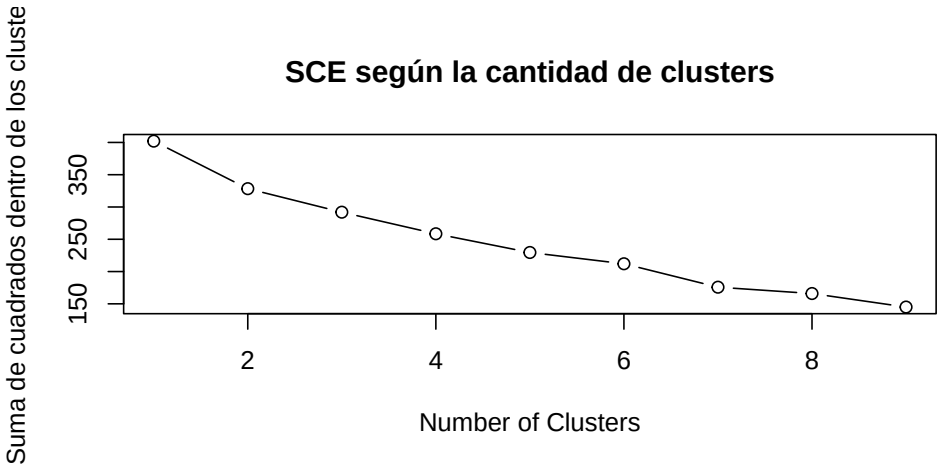

Figure 0.12: SCE segun cantidad de clusters

Puede tomarse 4 clusters como un numero adecuado para el análisis.

Boxplots para visualizar clusters

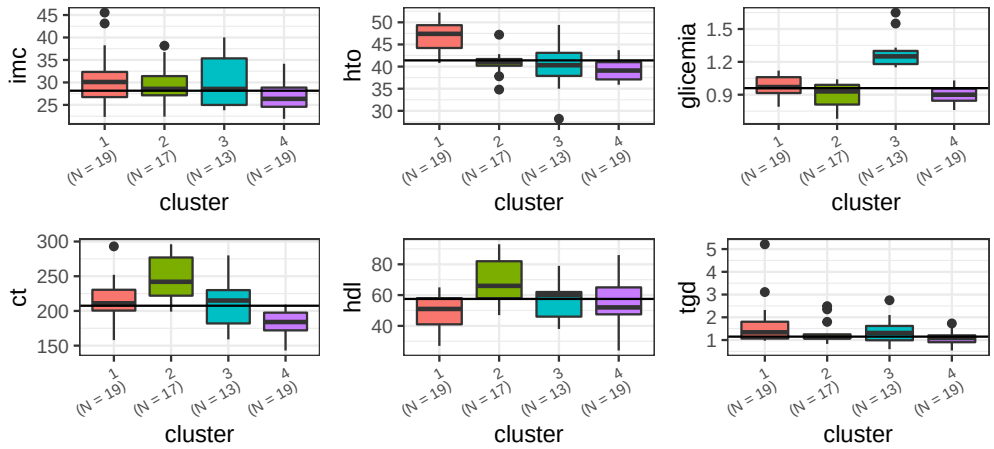

Figure 0.13: Composicion de clusters segun variables

Cluster numero uno Identifica pacientes con un valores promedios en la gran mayoría de variables pero con un componente de hto elevado.

Cluster numero dos Presenta pacientes con valores elevados de colesterol total y de hdl (componente de ct). Llamativamente estos mismos individuos poseen un icm relativamente promedio.

Cluster numero tres Identifica pacientes con un valores promedios en la gran mayoría de variables y valores de colesterol promedio tendiendo a bajo. Ademas de un icm con mayor spread. Identifica a los pacientes con glicemia elevada (¿diabéticos?).

Cluster numero cuatro Cluster que aglomera a los individuos "promedio" o al menos cuyos boxplots se aproximan al promedio en todas las variables de estudio.

Conclusiones El método, sin incluir la variable sexo mediante mecanismos de encodeado, no separa a los pacientes por sexo, lo que puede resultar valioso si estamos tratando de identificar población patológica y normal -en sentido clínico de normalidad-. Pero, también, los valores de referencia no son los mismos para estos grupos. Deberemos incluir un experto de dominio para tomar una decisión final sobre este punto.
El método implementado logra identificar pacientes con valores por encima del promedio en algunas variables y aglomerarlos -tal es el caso del cluster número uno y número tres, por ejemplo-. Se observa ademas un cierto spread de los pacientes con valores promedio entre los clusters. Podría pulirse el método para tratar de separar con algún criterio a pacientes con valores más cercanos al promedio pero alguna diferenciación particular.

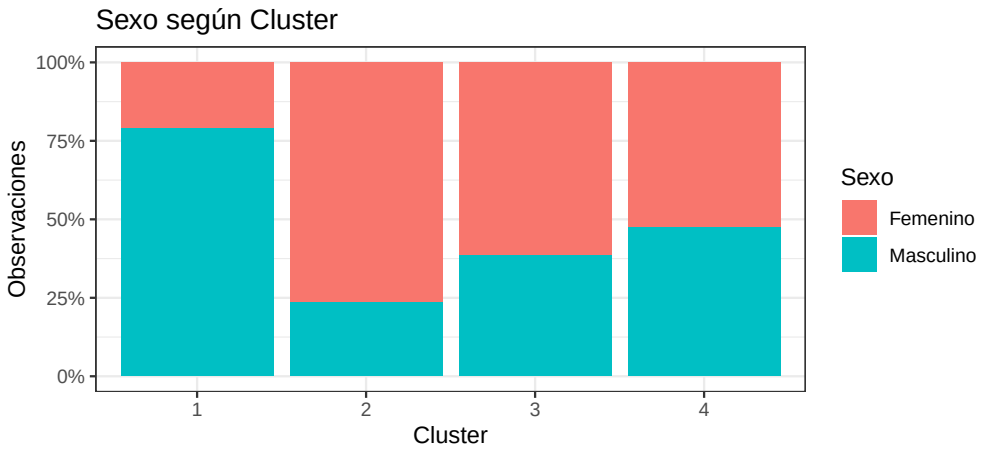

Figure 0.14: Distribucion de sexo en clusters

La distribucion por sexo presenta un spread significativo. Situacion a analizar en profundidad con un experto de dominio segun lo que se busque estudiar.

Obesidad: variable dicotomica

Generaremos un modelo predictivo a partir de la construccion de una variable "obesidad" a ser utilizada como respuesta. Utilizaremos para tal fin las siguientes variables predictoras: hto, glicemia, ct, hdl y tgd./

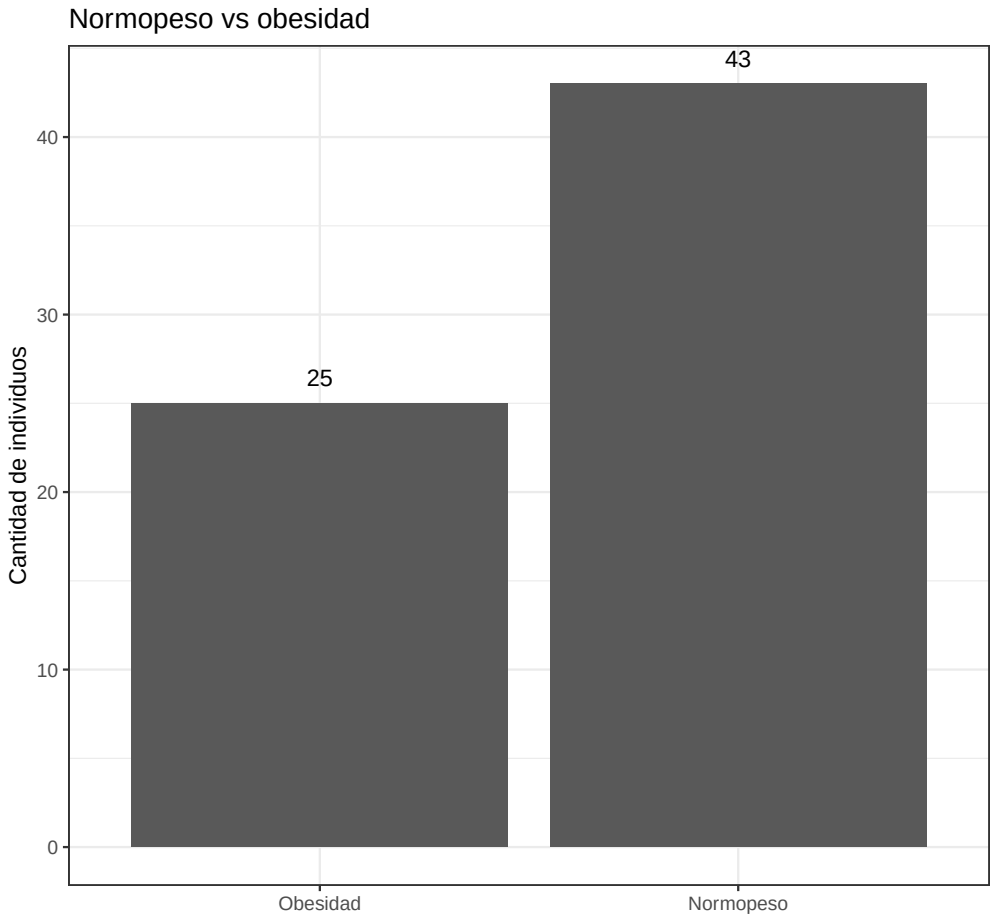

Figure 0.15: Distribucion de obesidad en nuestro dataset

Modelo: entrenamiento y validación

Se emplea el split tradicional, 80% en train y 20% en test. Usamos validación cruzada, con 10 sub-grupos o particiones y 3 repeticiones iterativas, para seleccionar el árbol que mejor ajusta a partir de la partición creada.

La librería caret toma como insumo los datasets particionados en test y train y selecciona el modelo optimo, balanceando CP y accuracy/kappa. Como se muestra a continuación, nos indica literalmente cual es el modelo elegido indicando según su CP.

```
## CART
##
## 55 samples
##  7 predictor
##  2 classes: 'Obesidad', 'Normopeso'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 50, 49, 49, 50, 50, 49, ...
## Resampling results across tuning parameters:
```

```
##
##    cp           Accuracy   Kappa
##    0.00000000  0.9877778  0.9705628
##    0.07142857  0.9877778  0.9705628
##    0.14285714  0.9877778  0.9705628
##    0.21428571  0.9877778  0.9705628
##    0.28571429  0.9877778  0.9705628
##    0.35714286  0.9877778  0.9705628
##    0.42857143  0.9877778  0.9705628
##    0.50000000  0.9877778  0.9705628
##    0.57142857  0.9877778  0.9705628
##    0.64285714  0.9877778  0.9705628
##    0.71428571  0.9877778  0.9705628
##    0.78571429  0.9877778  0.9705628
##    0.85714286  0.9877778  0.9705628
##    0.92857143  0.9877778  0.9705628
##    1.00000000  0.6333333  0.0000000
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.9285714.
```

Modelo: predicción y conclusiones

Se observa el arbol del mejor modelo seleccionado siguiendo el trabajo anterior.

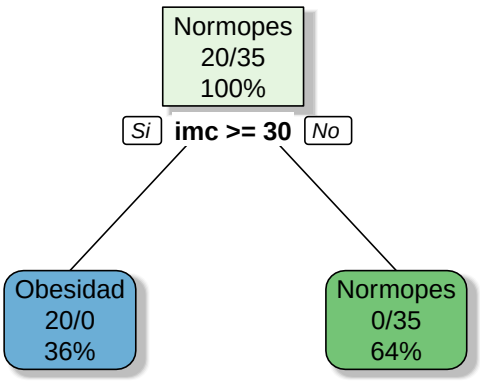

Figure 0.16: Mejor arbol

El modelo tiene una capacidad predictiva con accuracy igual a uno - como se muestra en la salida de la predicción a continuación-. Es decir, genera un criterio de clasificación casi idéntico al empleado para construir la variable (con diferencia del signo igual en la definición de obesidad en "imc $>$ 30"). Esto puede deberse a que según el criterio establecido, y empleado clinicamente, la obesidad tiene en cuenta únicamente una variable en su construcción, dejando el resto de las variables predictoras como secundarias.

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Obesidad Normopeso
##   Obesidad         5         0
##   Normopeso        0         8
##
##                Accuracy : 1
##                  95% CI : (0.7529, 1)
##     No Information Rate : 0.6154
##     P-Value [Acc > NIR] : 0.001815
##
```

```
##                  Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0000
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 1.0000
##              Prevalence : 0.3846
##          Detection Rate : 0.3846
##    Detection Prevalence : 0.3846
##       Balanced Accuracy : 1.0000
##
##        'Positive' Class : Obesidad
##
```