Análisis de Componentes Principales usando datos de pacientes de Cancer de mama

Alfredo Barra, Felipe Altamirano

2023-07-21

Abstract Según la OMS, el cancer de mama es una enfermedad que afectó a aproximadamente 2.3 millones de mujeres en 2020, siendo una de las principales causas de mortalidad en mujeres en todo el mundo. Sin embargo, se indica que una de las principales formas de combate a esta enfermedad está en el diagnostico temprano. Teniendo esto en cuenta, ¿cuales son los parámetros de los signos vitales de una persona más relevantes para poder determinar si está afectada por el cancer de mama?. A continuación se analizarán los datos de 116 pacientes, de las cuales 64 poseen cancer de mama.

1 Descripción de los datos

Los datos que usaremos a continuación corresponden a 116 pacientes mujeres, de los cuales 64 han sido diagnosticados con cancer de mama. Las variables de estos datos corresponden a:

- Age (years): Edad en años
- BMI (kg/m2): Índice de Masa Corporal
- Glucose (mg/dL): Glucosa en sangre
- Insulin (µU/mL): Insulina
- HOMA: Índice de resistencia a la insulina
- Leptin (ng/mL): Leptina
- Adiponectin (µg/mL): Adiponectina
- Resistin (ng/mL): Resistina
- MCP-1(pg/dL): Proteína quimiotáctica de monocitos 1
- Classification: como etiquetado si el paciente está enfermo (2) o está sano (1)

2 Preparación de los datos

Dado que los datos presentan distintas escalas, será necesario normalizar los datos. Para ello se utilizará la siguiente función.

```
normalize_min_max <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}
```

Y posteriormente crearemos un nuevo *DataFrame* con los datos ya normalizados en una escala de 0 a 1.

```
columnas_a_normalizar = c(
  "Age",
  "BMI",
  "Glucose",
  "Insulin",
  "HOMA",
  "Leptin",
```

```
  "Adiponectin",
  "Resistin",
  "MCP.1",
  "Classification"
)

data_normalizada = data

data_normalizada[columnas_a_normalizar] = lapply(data_normalizada[columnas_a_normalizar], normalize_min
```

Ahora tenemos una data normalizada como la siguiente.

```
head(data_normalizada)
```

```
##         Age       BMI    Glucose    Insulin        HOMA     Leptin Adiponectin
## 1 0.3692308 0.2538503 0.07092199 0.00490826 0.000000000 0.05229908  0.22115173
## 2 0.9076923 0.1148262 0.22695035 0.01219033 0.009742070 0.05272598  0.10370677
## 3 0.8923077 0.2352777 0.21985816 0.03687442 0.022057677 0.15852575  0.57102109
## 4 0.6769231 0.1483278 0.12056738 0.01417149 0.005911266 0.06481057  0.15153757
## 5 0.9538462 0.1356398 0.22695035 0.01993646 0.013748471 0.02778211  0.08693991
## 6 0.3846154 0.2219066 0.22695035 0.01417149 0.010766749 0.02932103  0.33046769
##     Resistin     MCP.1 Classification
## 1 0.06066485 0.2246591              0
## 2 0.01082583 0.2559263              0
## 3 0.07690645 0.3079117              0
## 4 0.12113069 0.5339336              0
## 5 0.09337495 0.4405654              0
## 6 0.09009507 0.2932155              0
```

2Análisis de los datos

2.1 Matriz de covarianza

```
## Warning: package 'ggplot2' was built under R version 4.3.1
```

```
## Warning: package 'corrplot' was built under R version 4.3.1
```

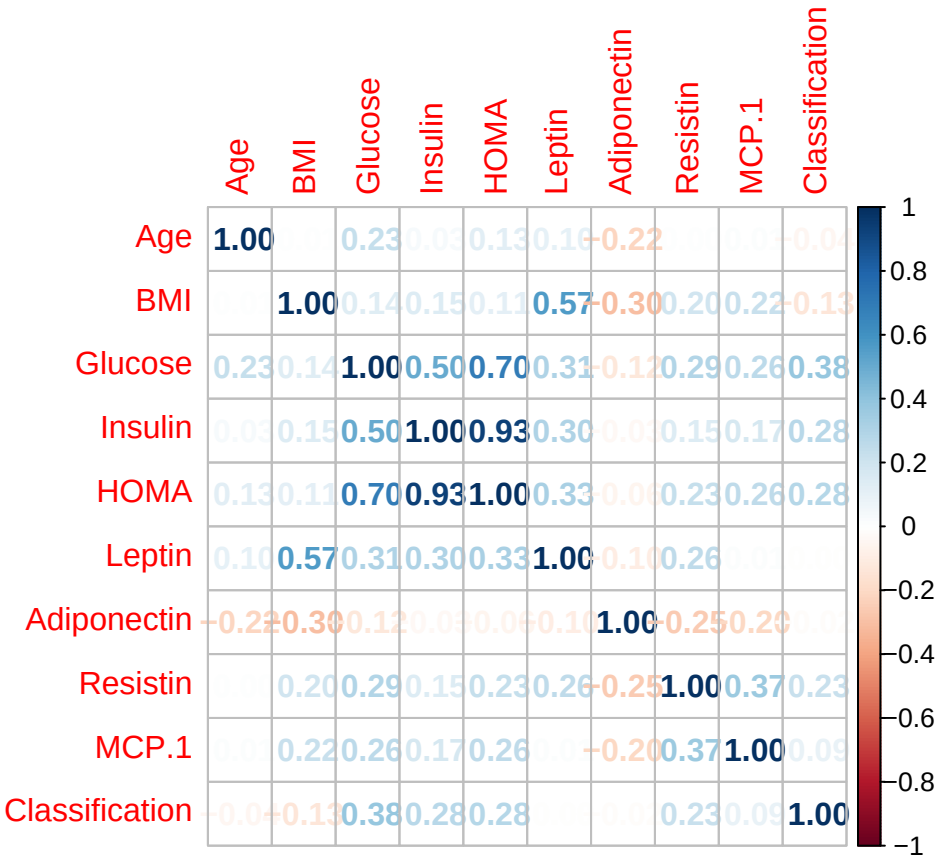
```
## corrplot 0.92 loaded
```

Con los datos ya normalizados, será necesario generar una matriz de covarianza para averiguar qué tan correlacionados se encuentran los datos.

```
cov_matrix = cor(data_normalizada)
```

Para mejorar la visualización de estos datos, usaremos un gráfico de mapa de calor (*heatmap*).

```
corrplot(cov_matrix, method = 'number')
```



Si nos enfocamos en la clasificación del paciente, podemos observar que existe cierta correlación con sus parámetros de Glucosa, Insulina, Indice de resistencia a la insulina (HOMA) y Resistina, mientras que existe una leve correlación inversa con la edad y el índice de masa corporal.