Exámen Interciclo - Inteligencia Artificial

Daniela Cuesta

```
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(caret)
```

```
Loading required package: lattice
```

```
library(e1071)
library(ggstatsplot)
```

```
You can cite this package as:
     Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach.
     Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167
```

```r
library(psych)
```

```
Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

    %+%, alpha
```

Descripción del conjunto de datos.

```r
datos <- read.csv("./datos/cancer.csv")
```

```r
str(datos)
```

```
'data.frame':   569 obs. of  31 variables:
 $ diagnostico            : chr  "M" "M" "M" "M" ...
 $ mean_radius            : num  18 20.6 19.7 11.4 20.3 ...
 $ mean_texture           : num  10.4 17.8 21.2 20.4 14.3 ...
 $ mean_perimeter         : num  122.8 132.9 130 77.6 135.1 ...
 $ mean_area              : num  1001 1326 1203 386 1297 ...
 $ mean_smoothnes         : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
 $ mean_compactness       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
 $ mean_concavity         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
 $ mean_concave_points    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
 $ mean_simmetry          : num  0.242 0.181 0.207 0.26 0.181 ...
 $ mean_fractal_dimension : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
 $ se_radius              : num  1.095 0.543 0.746 0.496 0.757 ...
 $ se_texture             : num  0.905 0.734 0.787 1.156 0.781 ...
 $ se_perimeter           : num  8.59 3.4 4.58 3.44 5.44 ...
 $ se_area                : num  153.4 74.1 94 27.2 94.4 ...
 $ se_smoothnes           : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
 $ se_compactness         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
 $ se_concavity           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
 $ se_concave_points      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
 $ se_simmetry            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
 $ se_fractal_dimension   : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
 $ worst_radius           : num  25.4 25 23.6 14.9 22.5 ...
```

```
$ worst_texture          : num  17.3 23.4 25.5 26.5 16.7 ...
$ worst_perimeter        : num  184.6 158.8 152.5 98.9 152.2 ...
$ worst_area             : num  2019 1956 1709 568 1575 ...
$ worst_smoothnes        : num  0.162 0.124 0.144 0.21 0.137 ...
$ worst_compactness      : num  0.666 0.187 0.424 0.866 0.205 ...
$ worst_concavity        : num  0.712 0.242 0.45 0.687 0.4 ...
$ worst_concave_points   : num  0.265 0.186 0.243 0.258 0.163 ...
$ worst_simmetry         : num  0.46 0.275 0.361 0.664 0.236 ...
$ worst_fractal_dimension: num  0.1189 0.089 0.0876 0.173 0.0768 ...
```

1. ***Realizar una estadística descriptiva numérica de los datos***

```
summary(datos)  # Resumen estadístico básico
```

```
diagnostico         mean_radius      mean_texture   mean_perimeter
Length:569         Min.   : 6.981   Min.   : 9.71   Min.   : 43.79
Class :character   1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17
Mode  :character   Median :13.370   Median :18.84   Median : 86.24
                   Mean   :14.127   Mean   :19.29   Mean   : 91.97
                   3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10
                   Max.   :28.110   Max.   :39.28   Max.   :188.50
  mean_area      mean_smoothnes    mean_compactness  mean_concavity
Min.   : 143.5   Min.   :0.05263   Min.   :0.01938   Min.   :0.00000
1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492   1st Qu.:0.02956
Median : 551.1   Median :0.09587   Median :0.09263   Median :0.06154
Mean   : 654.9   Mean   :0.09636   Mean   :0.10434   Mean   :0.08880
3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040   3rd Qu.:0.13070
Max.   :2501.0   Max.   :0.16340   Max.   :0.34540   Max.   :0.42680
mean_concave_points mean_simmetry    mean_fractal_dimension   se_radius
Min.   :0.00000     Min.   :0.1060   Min.   :0.04996        Min.   :0.1115
1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770        1st Qu.:0.2324
Median :0.03350     Median :0.1792   Median :0.06154        Median :0.3242
Mean   :0.04892     Mean   :0.1812   Mean   :0.06280        Mean   :0.4052
3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612        3rd Qu.:0.4789
Max.   :0.20120     Max.   :0.3040   Max.   :0.09744        Max.   :2.8730
  se_texture      se_perimeter       se_area         se_smoothnes
Min.   :0.3602   Min.   : 0.757   Min.   :  6.802   Min.   :0.001713
1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850   1st Qu.:0.005169
Median :1.1080   Median : 2.287   Median : 24.530   Median :0.006380
Mean   :1.2169   Mean   : 2.866   Mean   : 40.337   Mean   :0.007041
3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190   3rd Qu.:0.008146
Max.   :4.8850   Max.   :21.980   Max.   :542.200   Max.   :0.031130
```

```
se_compactness      se_concavity     se_concave_points   se_simmetry
Min.   :0.002252   Min.   :0.00000   Min.   :0.000000   Min.   :0.007882
1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638   1st Qu.:0.015160
Median :0.020450   Median :0.02589   Median :0.010930   Median :0.018730
Mean   :0.025478   Mean   :0.03189   Mean   :0.011796   Mean   :0.020542
3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710   3rd Qu.:0.023480
Max.   :0.135400   Max.   :0.39600   Max.   :0.052790   Max.   :0.078950
se_fractal_dimension  worst_radius   worst_texture   worst_perimeter
Min.   :0.0008948    Min.   : 7.93   Min.   :12.02   Min.   : 50.41
1st Qu.:0.0022480    1st Qu.:13.01   1st Qu.:21.08   1st Qu.: 84.11
Median :0.0031870    Median :14.97   Median :25.41   Median : 97.66
Mean   :0.0037949    Mean   :16.27   Mean   :25.68   Mean   :107.26
3rd Qu.:0.0045580    3rd Qu.:18.79   3rd Qu.:29.72   3rd Qu.:125.40
Max.   :0.0298400    Max.   :36.04   Max.   :49.54   Max.   :251.20
  worst_area     worst_smoothnes   worst_compactness worst_concavity
Min.   : 185.2   Min.   :0.07117   Min.   :0.02729   Min.   :0.0000
1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720   1st Qu.:0.1145
Median : 686.5   Median :0.13130   Median :0.21190   Median :0.2267
Mean   : 880.6   Mean   :0.13237   Mean   :0.25427   Mean   :0.2722
3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910   3rd Qu.:0.3829
Max.   :4254.0   Max.   :0.22260   Max.   :1.05800   Max.   :1.2520
worst_concave_points worst_simmetry   worst_fractal_dimension
Min.   :0.00000      Min.   :0.1565   Min.   :0.05504
1st Qu.:0.06493      1st Qu.:0.2504   1st Qu.:0.07146
Median :0.09993      Median :0.2822   Median :0.08004
Mean   :0.11461      Mean   :0.2901   Mean   :0.08395
3rd Qu.:0.16140      3rd Qu.:0.3179   3rd Qu.:0.09208
Max.   :0.29100      Max.   :0.6638   Max.   :0.20750
```

Interpretación

Cuando se aplica **`summary()`** a un conjunto de datos, se calculan varias estadísticas descriptivas para cada columna. Estas estadísticas pueden incluir:

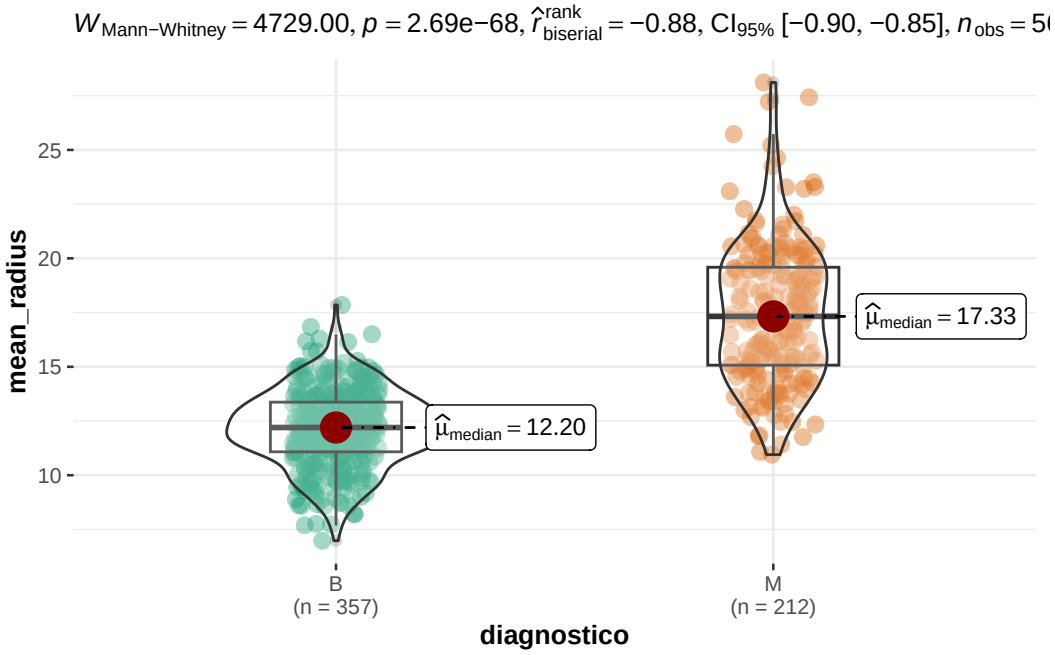
Para variables numéricas:

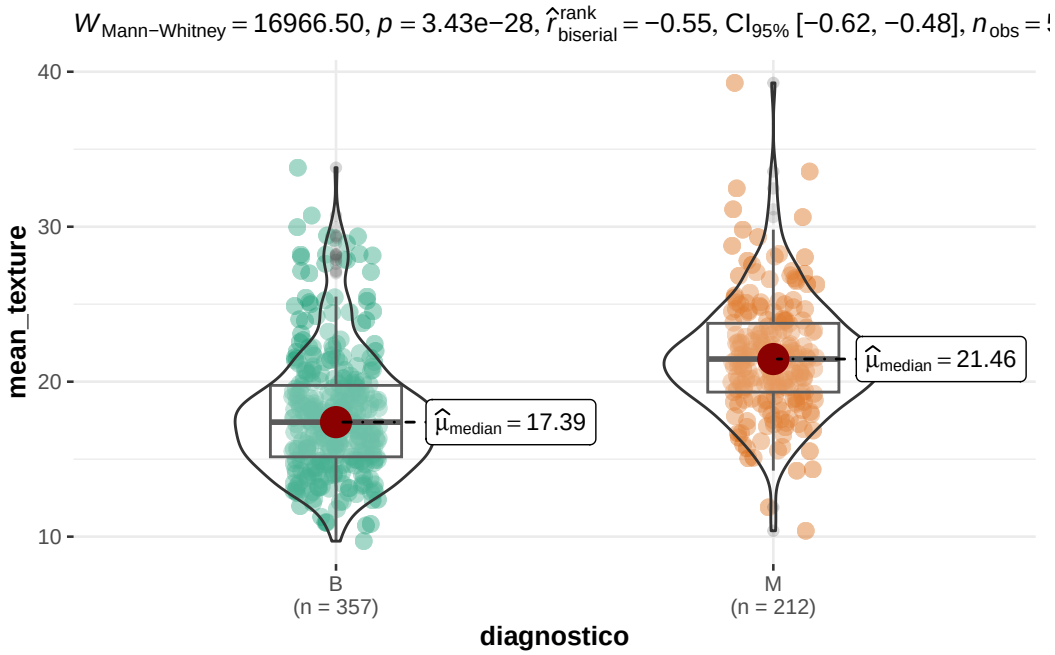
- Mínimo (valor mínimo)
- Cuartil 1 (Q1, el valor que divide el conjunto de datos en el 25% inferior)
- Mediana (Q2, el valor que divide el conjunto de datos en el 50%)
- Media (promedio aritmético)
- Cuartil 3 (Q3, el valor que divide el conjunto de datos en el 75% inferior)

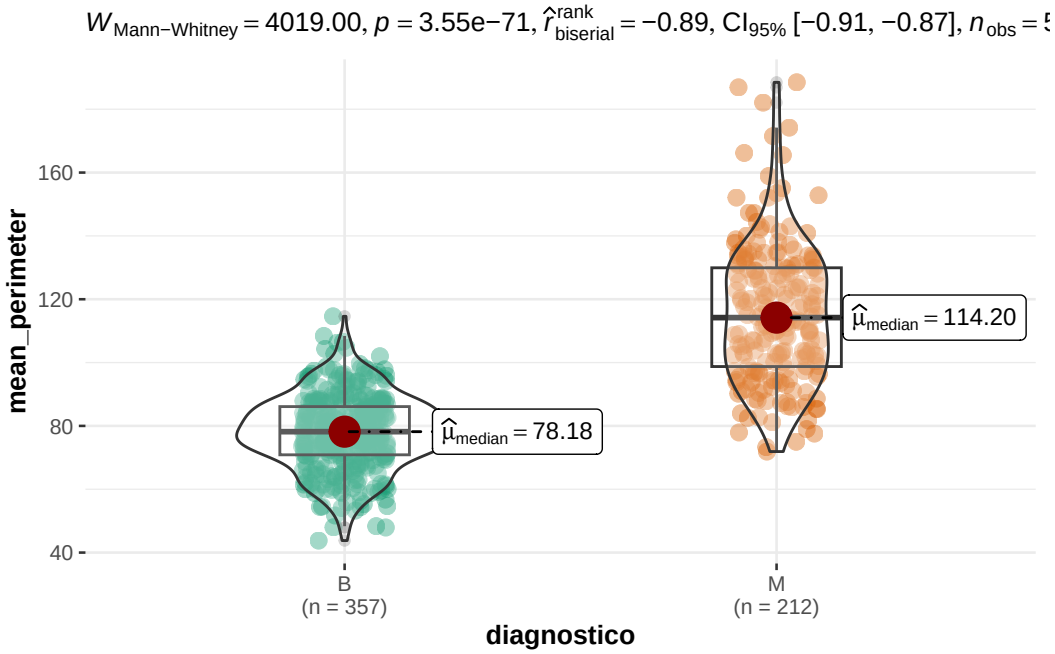
- Máximo (valor máximo)

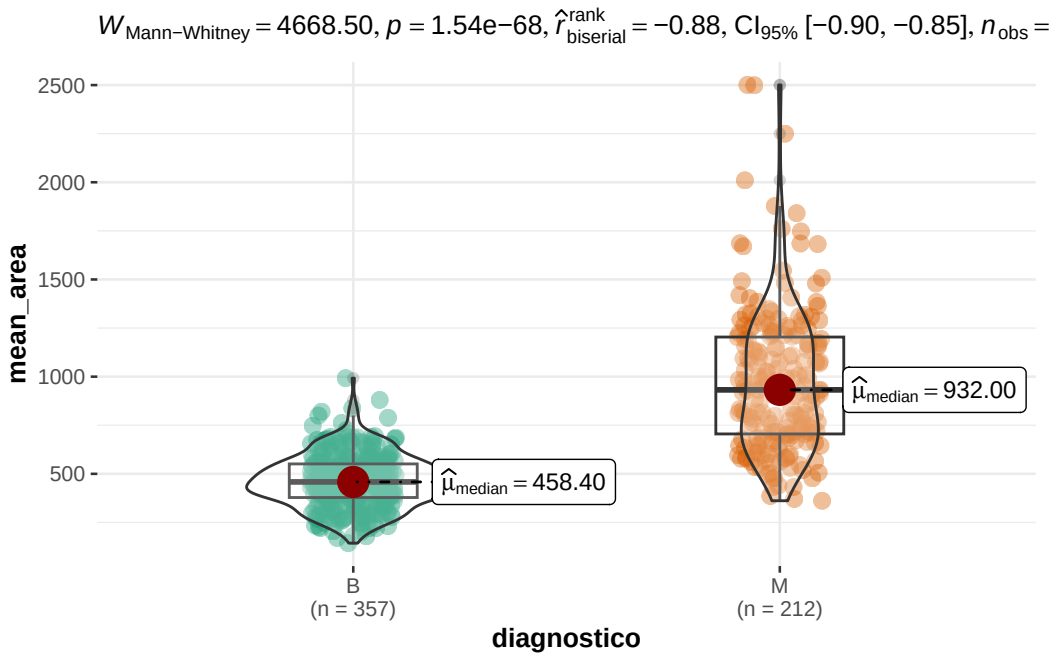
2. ***Realizar estadística descritptiva univariante inferencial para las 10 primeras columnas. Po***

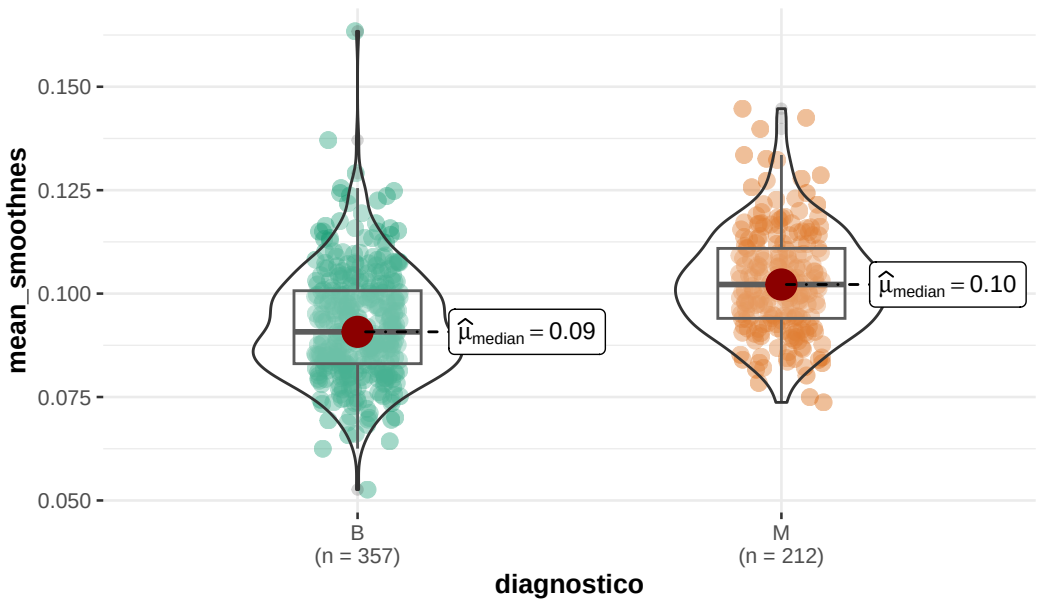
```
columnas2a11 <-c("mean_radius", "mean_texture", "mean_perimeter", "mean_area", "mean_smoot
for (ccol in columnas2a11){
  simbolo <-as.name(ccol)#col representa el nombre de una columna en columnas2a11. Al util
  plot <- ggbetweenstats(data = datos, x = diagnostico, y = !!simbolo, type="np")
  print(plot)
}
```

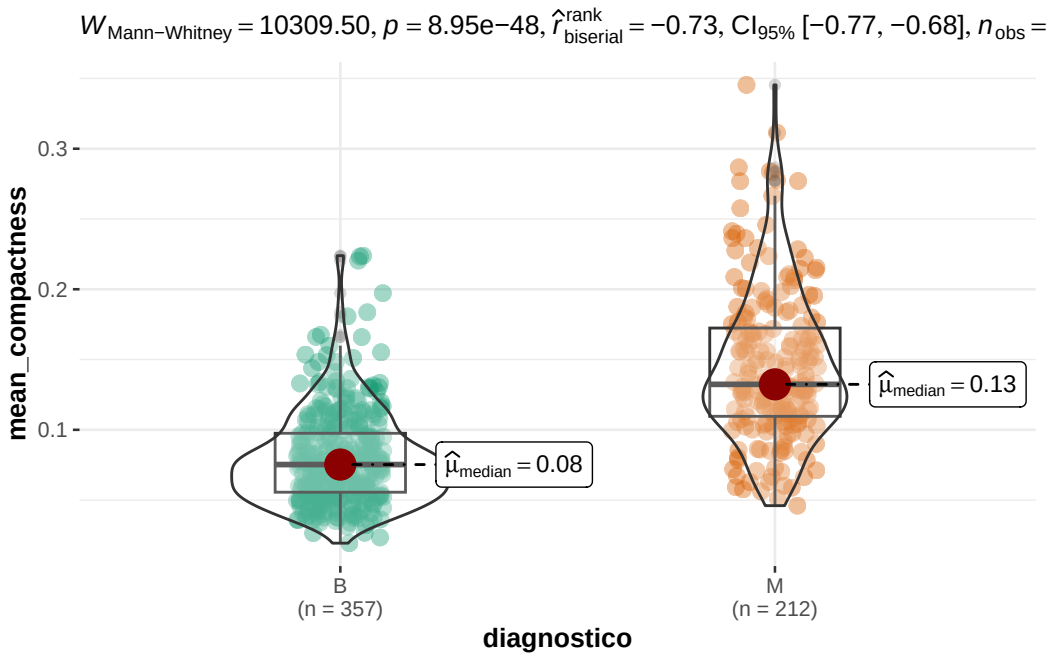


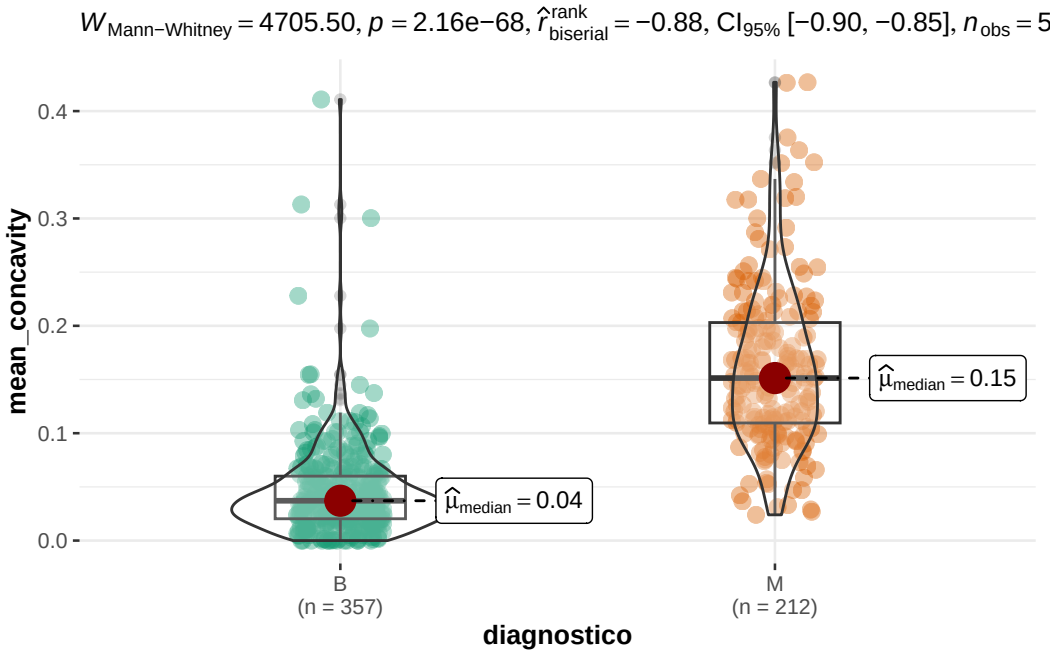

$W_{\text{Mann-Whitney}} = 16966.50$, $p = 3.43e{-}28$, $\hat{r}^{\text{rank}}_{\text{biserial}} = -0.55$, $\text{CI}_{95\%}$ [−0.62, −0.48], $n_{\text{obs}} =$
mean_texture
40
30
20
10
$\hat{\mu}_{\text{median}} = 17.39$
$\hat{\mu}_{\text{median}} = 21.46$
B
(n = 357)
M
(n = 212)
diagnostico

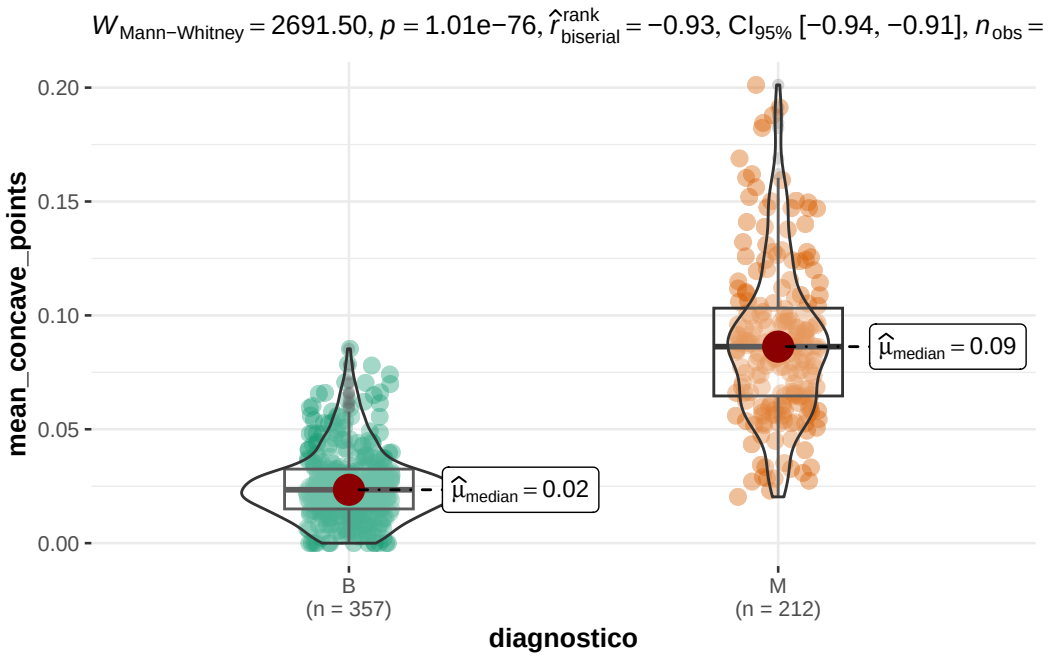

$W_{\text{Mann-Whitney}} = 4019.00$, $p = 3.55e{-}71$, $\hat{r}^{\text{rank}}_{\text{biserial}} = -0.89$, $\text{CI}_{95\%}$ [−0.91, −0.87], $n_{\text{obs}} =$
mean_perimeter
160
120
80
40
$\hat{\mu}_{\text{median}} = 78.18$
$\hat{\mu}_{\text{median}} = 114.20$
B
(n = 357)
M
(n = 212)
diagnostico

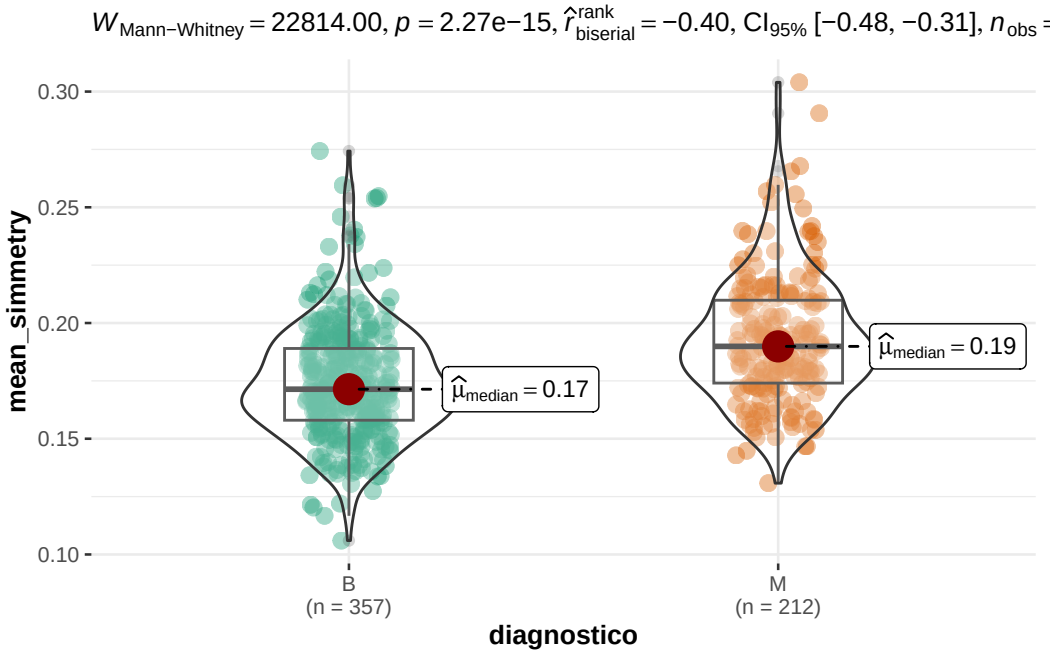

$W_{\text{Mann–Whitney}} = 4668.50$, $p = 1.54e{-}68$, $\hat{r}^{\text{rank}}_{\text{biserial}} = -0.88$, $\text{CI}_{95\%}$ [−0.90, −0.85], n_{obs} =
mean_area
2500
2000
1500
1000
500
$\hat{\mu}_{\text{median}} = 458.40$
$\hat{\mu}_{\text{median}} = 932.00$
B
(n = 357)
M
(n = 212)
diagnostico

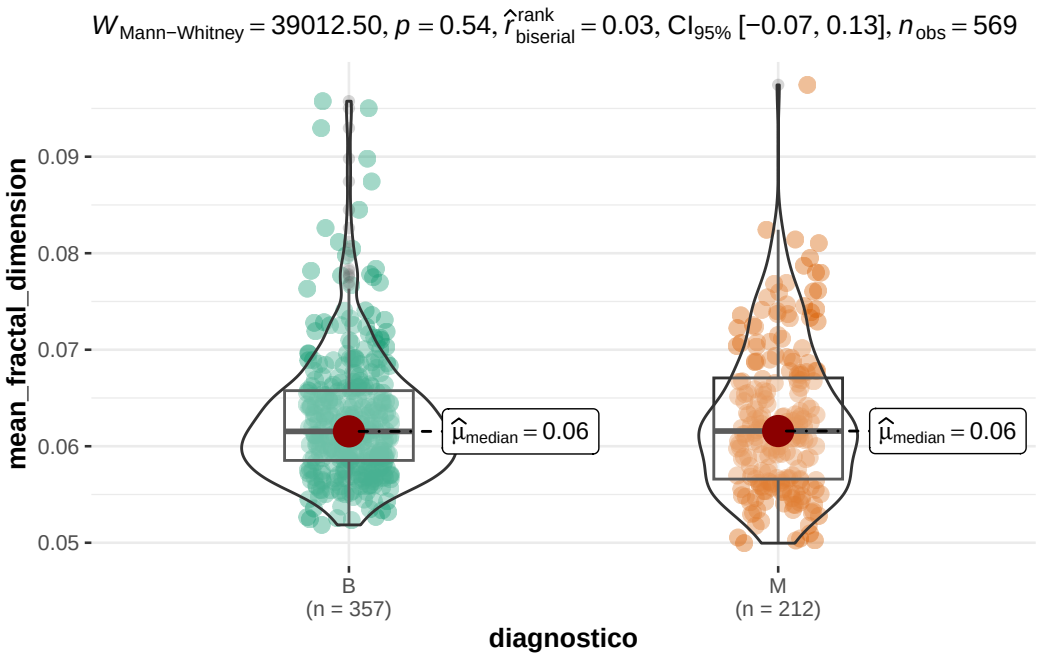

$W_{\text{Mann–Whitney}} = 21037.00$, $p = 7.79e{-}19$, $\hat{r}^{\text{rank}}_{\text{biserial}} = -0.44$, $\text{CI}_{95\%}$ [−0.52, −0.36], n_{obs}
mean_smoothnes
0.150
0.125
0.100
0.075
0.050
$\hat{\mu}_{\text{median}} = 0.09$
$\hat{\mu}_{\text{median}} = 0.10$
B
(n = 357)
M
(n = 212)
diagnostico


$W_{\text{Mann–Whitney}} = 10309.50$, $p = 8.95\text{e}{-48}$, $\hat{r}^{\text{rank}}_{\text{biserial}} = -0.73$, $\text{CI}_{95\%}$ [−0.77, −0.68], n_{obs} =
mean_compactness
0.3
0.2
0.1
$\hat{\mu}_{\text{median}} = 0.08$
$\hat{\mu}_{\text{median}} = 0.13$
B
(n = 357)
M
(n = 212)
diagnostico


$W_{\text{Mann–Whitney}} = 4705.50$, $p = 2.16\text{e}{-68}$, $\hat{r}^{\text{rank}}_{\text{biserial}} = -0.88$, $\text{CI}_{95\%}$ [−0.90, −0.85], $n_{\text{obs}} = 5$
mean_concavity
0.4
0.3
0.2
0.1
0.0
$\hat{\mu}_{\text{median}} = 0.04$
$\hat{\mu}_{\text{median}} = 0.15$
B
(n = 357)
M
(n = 212)
diagnostico


$W_{\text{Mann–Whitney}} = 2691.50$, $p = 1.01\text{e}{-76}$, $\hat{r}^{\text{rank}}_{\text{biserial}} = -0.93$, $\text{CI}_{95\%}$ [−0.94, −0.91], n_{obs} =
mean_concave_points
0.20
0.15
0.10
0.05
0.00
$\hat{\mu}_{\text{median}} = 0.09$
$\hat{\mu}_{\text{median}} = 0.02$
B
(n = 357)
M
(n = 212)
diagnostico


$W_{\text{Mann–Whitney}} = 22814.00$, $p = 2.27\text{e}{-15}$, $\hat{r}^{\text{rank}}_{\text{biserial}} = -0.40$, $\text{CI}_{95\%}$ [−0.48, −0.31], n_{obs} =
mean_simmetry
0.30
0.25
0.20
0.15
0.10
$\hat{\mu}_{\text{median}} = 0.19$
$\hat{\mu}_{\text{median}} = 0.17$
B
(n = 357)
M
(n = 212)
diagnostico



Interpretación

Dentro de la interpretaciones que se van a realizar hay que tener claro que el número de tumores benignos incluidos en el análisis es mayor (n = 357) en comparación con el número de tumores malignos (n = 212). Además que en cada interpretacion la mediana cambia tanto para los tumores benignos como malignos.

Estas diferencias en las medianas podrían indicar una posible relación entre todas las características del tumor y la naturaleza benigna o maligna del mismo.

Mean_radius

El gráfico sugieren que, en promedio, los tumores benignos tienen un tamaño de radio medio menor (mediana de 12.20) en comparación con los tumores malignos (mediana de 17.33).

Mean_Texture

El gráfico sugieren que, en promedio, los tumores benignos tienen una textura media menor (mediana de 17.39) en comparación con los tumores malignos (mediana de 21.46).

Mean_Perimeter

El gráfico sugiere que, en promedio, los tumores benignos tienen un perímetro medio menor (mediana de 78.18) en comparación con los tumores malignos (mediana de 114.20).

Mean_Area

El gráfico sugiere que, en promedio, los tumores benignos tienen un área media menor (mediana de 458.40) en comparación con los tumores malignos (mediana de 932.00).

Mean_Smoothness

El gráfico indica que, en promedio, los tumores malignos tienen una suavidad media ligeramente mayor (mediana de 0.10) en comparación con los tumores benignos (mediana de 0.09).

Mean_Compactness

El gráfico indica que, en promedio, los tumores malignos tienen una compacidad media mayor (mediana de 0.13) en comparación con los tumores benignos (mediana de 0.06).

La compacidad de los tumores puede ser un factor relevante para distinguir entre tumores benignos y malignos.

Mean_Concavity

El gráfico indica que, en promedio, los tumores malignos tienen una concavidad media mayor (mediana de 0.15) en comparación con los tumores benignos (mediana de 0.04).

La concavidad de los tumores puede ser un factor relevante para distinguir entre tumores benignos y malignos.

Mean_Concave_Points

El gráfico indica que, en promedio, los tumores malignos tienen una mayor cantidad de puntos cóncavos medios (mediana de 0.09) en comparación con los tumores benignos (mediana de 0.02).

La presencia y cantidad de puntos cóncavos pueden ser características importantes para distinguir entre tumores benignos y malignos.

Mean_Simmetry

El gráfico indica que, en promedio, los tumores malignos tienen una simetría media ligeramente mayor (mediana de 0.19) en comparación con los tumores benignos (mediana de 0.17).

La simetría de los tumores puede ser una característica importante para distinguir entre tumores benignos y malignos.

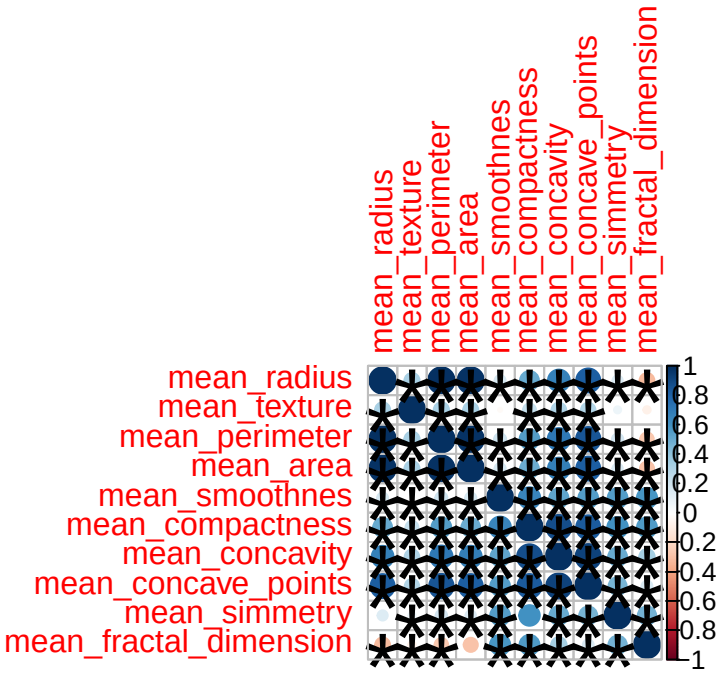
Mean_Fractal_Dimension

El gráfico indica que la mediana de la dimensión fractal media es similar tanto para los tumores benignos como para los tumores malignos (ambos con una mediana de 0.06).

La dimensión fractal puede ser una medida interesante para caracterizar los tumores, pero en este caso particular no parece haber una diferencia significativa entre los tumores benignos y malignos en términos de su dimensión fractal media.

3. **Realizar un gráfico de correlaciones**

```r
# Filtrar las columnas relevantes en un nuevo data frame
datos_seleccionados <- datos[columnas2a11]

# Calcular la matriz de correlación y los valores p
obj.cor <- psych::corr.test(datos_seleccionados)
p.values <- obj.cor$p
p.values[upper.tri(p.values)] <- obj.cor$p.adj
p.values[lower.tri(p.values)] <- obj.cor$p.adj
diag(p.values) <- 1

# Visualizar el gráfico de correlaciones utilizando la función corrplot
corrplot::corrplot(corr = obj.cor$r, p.mat = p.values, sig.level = 0.05, insig = "label_si
```



- **Matriz de correlación**

  Los valores de correlación varían entre -1 y 1. Un valor de 1 indica una correlación positiva perfecta, -1 indica una correlación negativa perfecta y 0 indica una ausencia de correlación.

- **Valores p:**

  Indican si la correlación observada entre dos variables es estadísticamente significativa. Los valores p más bajos indican una mayor significancia estadística.

- **Gráfico de correlaciones:**

  El gráfico de correlaciones visualiza la matriz de correlación utilizando colores y símbolos.

  Las correlaciones positivas están representadas en tonos de azul, mientras que las correlaciones negativas están representadas en tonos de rojo.

  El color y la intensidad del cuadro indican la fuerza de la correlación. Tonos más intensos representan correlaciones positivas fuertes, mientras que tonos más claros representan correlaciones positivas más débiles.

- Las correlaciones negativas se representan en tonos de rojo.
- Los asteriscos (*) en las celdas indican la significancia estadística de las correlaciones. Los asteriscos más grandes indican una significancia estadística más fuerte.

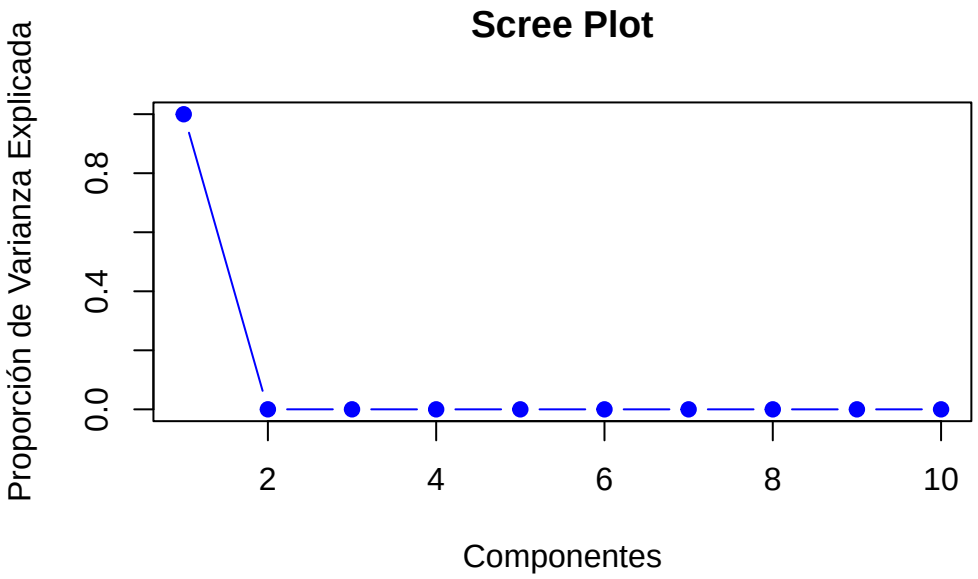
Si la celda está vacía, indica que la correlación no es estadísticamente significativa.

Interpretación

Según el análisis del gráfico de correlaciones se puede observar que la correlacion de mean_radius con mean_perimeter, y la correlacion entre mean_radius y mean_area es una correlación perfecta ya que su valor de p es bajo, esto quiere decir que tienen mayor significancia estadística

4. ***Realizar un PCA sobre las 10 primeras variables. Debe de contener: Scree plot y Biplot***

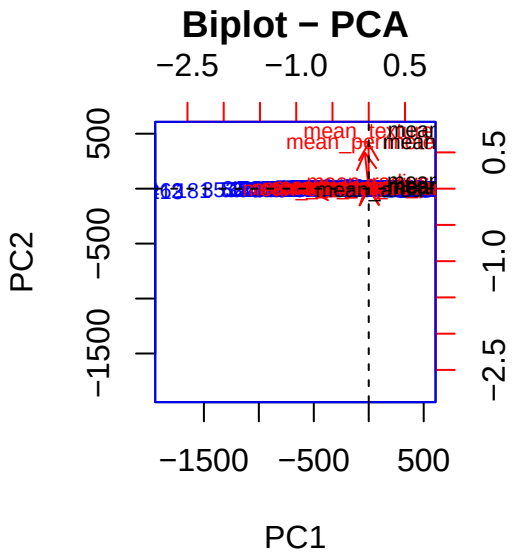
```r
# Calcular el PCA
pcx <- prcomp(datos_seleccionados, scale. = FALSE)

# Scree plot
scree <- pcx$sdev^2 / sum(pcx$sdev^2)
plot(1:length(scree), scree, type = "b", xlab = "Componentes", ylab = "Proporción de Varia
     main = "Scree Plot", col = "blue", pch = 19)
```




```
# Biplot
biplot(pcx, scale = 0, col = c("blue", "red"), cex = 0.7, main = "Biplot - PCA")
```

```
Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
arrow.len): zero-length arrow is of indeterminate angle and so skipped

Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
arrow.len): zero-length arrow is of indeterminate angle and so skipped

Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
arrow.len): zero-length arrow is of indeterminate angle and so skipped

Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
arrow.len): zero-length arrow is of indeterminate angle and so skipped
```

```
# Ajustar las etiquetas de los puntos
text(pcx$rotation[, 1], pcx$rotation[, 2], labels = colnames(datos_seleccionados), cex = 0

# Añadir líneas de referencia para los ejes
abline(h = 0, v = 0, lty = 2)
```



Interpretación

- **Scree Plot:**

  El Scree Plot muestra la proporción de varianza. Cada punto en el gráfico representa un componente principal, y en el eje y se muestra la proporción de varianza explicada. El eje x indica el número de componentes.

  En el Scree Plot, la línea trazada conectando los puntos muestra cómo se acumula la varianza explicada a medida que se agregan más componentes principales.

  El punto de inflexión en el gráfico, donde la curva se aplana o se desacelera, suele indicar el número óptimo de componentes principales a retener. Los componentes anteriores a este punto suelen explicar la mayor parte de la variabilidad en los datos, mientras que los componentes posteriores pueden contener información menos relevante o ruido.

- **Biplot:**

  El gráfico Biplot del PCA muestra la representación de los componentes principales en un espacio bidimensional, donde cada punto representa una variable original (en este caso, las columnas seleccionadas del conjunto de datos).

  El color de los puntos en el Biplot se utiliza para representar el diagnóstico de los casos (categoría “benigno” o “maligno”).

Las líneas de referencia que intersectan en el origen (punto [0,0]) representan los ejes de los componentes principales. Estas líneas proporcionan una referencia para la dirección y la orientación de las variables y las observaciones en el espacio del PCA.

Realizar una predicción del diagnóstico con Naive Bayes mediante el paquete e1071 (20 %)

1. Dividir el conjunto de datos en prueba y entrenamiento con la semilla de aleatorizaciónset.se

```
set.seed(123456)  # Semilla de aleatorización

# Dividir el conjunto de datos en entrenamiento y prueba
train_indices <- sample(1:nrow(datos), nrow(datos)*0.7)  # 70% de los datos para entrenami
train_data <- datos[train_indices, ]
test_data <- datos[-train_indices, ]
```

2. Entrenar y realizar la predicción del diagnóstico

```
# Definir la fórmula para el modelo Naive Bayes
formula <- as.formula(paste("diagnostico ~", paste(columnas2a11, collapse = "+")))

# Entrenar el modelo Naive Bayes
model <- naiveBayes(formula, data = train_data)

# Realizar la predicción en el conjunto de prueba
predictions <- predict(model, test_data)
```

3. Obtener la matriz de confusión. Obtener Accuracy, Specificity y Sensibility

```
# Obtener la matriz de confusión
confusion <- table(test_data$diagnostico, predictions)
print(confusion)
```

```
  predictions
     B   M
B 101   8
M   9  53
```

```
# Calcular Accuracy, Specificity y Sensibility
accuracy <- sum(diag(confusion)) / sum(confusion)
specificity <- confusion["B", "B"] / sum(confusion[, "B"])
```

```r
sensitivity <- confusion["M", "M"] / sum(confusion[, "M"])

# Imprimir los resultados
print(paste("Accuracy:", accuracy))
```

```
[1] "Accuracy: 0.900584795321637"
```

```r
print(paste("Specificity:", specificity))
```

```
[1] "Specificity: 0.918181818181818"
```

```r
print(paste("Sensitivity:", sensitivity))
```

```
[1] "Sensitivity: 0.868852459016393"
```

Interpretación

La matriz de confusión muestra las predicciones realizadas por el modelo. En este caso, hay 101 casos clasificados correctamente como "B" (benigno) y 53 casos clasificados correctamente como "M" (maligno). También hay 8 casos clasificados erróneamente como "M" cuando eran "B" y 9 casos clasificados erróneamente como "B" cuando eran "M", es decir se confundieron entre si eran malignos o beningnos

- Accuracy (Precisión): La precisión es una medida que indica qué tan bien el modelo predice correctamente todas las clases. En este caso, el modelo tiene una precisión del 90.06%.
- Specificity (Especificidad): La especificidad es una medida que indica cuántas instancias benignas son clasificadas correctamente. En este caso, el modelo clasifica correctamente el 91.82% de los casos benignos.
- Sensitivity (Sensibilidad o Tasa de Verdaderos Positivos): La sensibilidad es una medida que indica cuántas instancias malignas son clasificadas correctamente. En este caso, el modelo clasifica correctamente el 86.89% de los casos malignos.

Estos valores indican el rendimiento del modelo en términos de clasificar correctamente los casos en cada clase y pueden ser utilizados para evaluar su eficacia en la detección de casos malignos y benignos.

Realizar una extracción de las características más importantes

Realizar una regresión logística regularizada de LASSO

```
# Definir la cuadrícula de hiperparámetros para búsqueda
tuneGrid <- expand.grid(.alpha = 1, .lambda = seq(0, 1, by = 0.001))

# Definir el control de entrenamiento
trainControl <- trainControl(method = "repeatedcv", number = 10, repeats = 3, classProbs =
```

2. Entrenar y predecir el diagnóstico

```
# Ajustar el modelo de regresión logística con LASSO
model <- train(diagnostico ~ ., data = train_data, method = "glmnet", trControl = trainCon

# Realizar predicciones en el conjunto de prueba
predictions <- predict(model, test_data)
```

3. Obtener la matriz de confusión, Obtener Accuracy, Specificity y Sensibility

```
# Crear la matriz de confusión
confusion <- table(Actual = test_data$diagnostico, Predicted = predictions)

# Calcular Accuracy, Specificity y Sensitivity
accuracy <- sum(diag(confusion)) / sum(confusion)
specificity <- confusion["B", "B"] / sum(confusion[, "B"])
sensitivity <- confusion["M", "M"] / sum(confusion[, "M"])

# Imprimir la matriz de confusión
print("Matriz de Confusión:")
```

```
[1] "Matriz de Confusión:"
```

```
print(confusion)
```

```
      Predicted
Actual   B   M
     B 108   1
     M   3  59
```

```r
# Imprimir los resultados
print(paste("Accuracy:", accuracy))
```

```
[1] "Accuracy: 0.976608187134503"
```

```r
print(paste("Specificity:", specificity))
```

```
[1] "Specificity: 0.972972972972973"
```

```r
print(paste("Sensitivity:", sensitivity))
```

```
[1] "Sensitivity: 0.983333333333333"
```

En la matriz de confusión, las filas representan las clases reales (B para Benigno y M para Maligno), mientras que las columnas representan las clases predichas por el modelo. En este caso, se tienen los siguientes resultados:

- Para la clase Benigno (B):
  - Se predijeron correctamente 108 casos de la clase Benigno (verdaderos negativos).
  - Se predijo incorrectamente 1 caso de la clase Benigno como Maligno (falso positivo).
- Para la clase Maligno (M):
  - Se predijeron correctamente 59 casos de la clase Maligno (verdaderos positivos).
  - Se predijeron incorrectamente 3 casos de la clase Maligno como Benigno (falsos negativos).
- Precisión (Accuracy): En este caso, la precisión es de 0.9766, lo que indica que el modelo tiene un alto nivel de precisión en la clasificación general.
- Especificidad (Specificity): En este caso, la especificidad es de 0.9729, lo que indica que el modelo tiene una alta capacidad para identificar correctamente los casos Benigno.
- Sensibilidad (Sensitivity): En este caso, la sensibilidad es de 0.9833, lo que indica que el modelo tiene una alta capacidad para identificar correctamente los casos Maligno.