

A large, thick red stylized letter 'U' that curves around the central text, starting from the top left and ending at the bottom right.

IU Digital
de Antioquia

**PRONÓSTICO DE TRATAMIENTO PARA PACIENTE CON CÁNCER DE MAMA - DATOS
MULTIVARIADOS: RELACIONES, COMPORTAMIENTO Y PREDICCIÓN**

**OSSMAN MEJÍA GUZMÁN
GRUPO PREICA2401B010075**

**INSTITUCIÓN UNIVERSITARIA DIGITAL DE ANTIOQUIA
INGENIERIA DE SOFTWARE Y DATOS
ESTADISTICA II
MARIA CLAUDIA NEGRET LOPEZ
MEDELLÍN
MAYO DE 2024**

TABLA DE CONTENIDO

RESUMEN	4
PROBLEMA	6
1. ANÁLISIS EXPLORATORIO DE DATOS	7
2. CÁLCULO DE MEDIDAS DE TENDENCIA CENTRAL Y DE DISPERSIÓN Y GRÁFICOS ESTADÍSTICOS PARA LA MUESTRA.....	11
3. CORRELACIÓN PARA VARIABLES CUANTITATIVAS (TODAS ENTRE SI)	1
4. A PARTIR DE LO ANTERIOR, IDENTIFIQUE (SI ES POSIBLE) LA O LAS VARIABLES QUE PUEDAN SER MÁS RELEVANTES RELACIONADAS CON LA MALIGNIDAD Y CON LA BENIGNIDAD.....	3
5. REALIZAR CAJAS Y BIGOTES PARA LOS DOS GRUPOS. Y DISCUTIR RESULTADOS. 3	
6. PLANTEE UNA HIPÓTESIS NULA Y ALTERNA PARA LAS VARIABLES QUE USTED CONSIDERÓ MÁS RELEVANTES PARA LA PRESENCIA DE MALIGNIDAD Y BENIGNIDAD ..3	
7. REALICE LA TABLA Y ESTABLEZCA PROBABILIDADES PARA LOS ERRORES DE TIPO I Y II. ¿CUÁL ES EL ERROR MÁS GRAVE?	5
8. REALICE ALGUNAS TABLAS DE CONTINGENCIA PARA LAS VARIABLES MÁS RELEVANTES QUE USTED ENCONTRÓ.....	8
9. CONCLUSIONES.....	10
10. ANEXOS.....	12

RESUMEN

En este trabajo se realizó un análisis exploratorio de datos sobre un conjunto de datos de pacientes con cáncer de mama, con el objetivo de investigar las relaciones entre las variables y su capacidad para predecir el diagnóstico (maligno o benigno). Para ello, se utilizó una muestra aleatoria representativa y se llevaron a cabo diversas técnicas de análisis estadístico y visualización.

Se instalaron y cargaron las librerías necesarias para el análisis de datos y visualización, incluyendo dplyr, ggplot2, GGally, tidyr, corrplot y knitr. Posteriormente, se calculó el tamaño de muestra necesario para asegurar un nivel de confianza del 99% y un margen de error del 5%. Luego, se creó una muestra aleatoria simple a partir del conjunto de datos original. La representatividad de la muestra se validó comparando las proporciones del diagnóstico (maligno vs. benigno) entre el conjunto de datos original y la muestra.

A continuación, se calcularon estadísticas descriptivas (media, mediana, desviación estándar, rango intercuartílico, mínimo y máximo) para cada variable, agrupadas por el diagnóstico (maligno o benigno). Estos resultados se reorganizaron para facilitar su interpretación.

Luego de esto, se crearon varias visualizaciones para entender mejor el comportamiento de las variables: Histogramas para observar la distribución de las variables por grupo de diagnóstico. Boxplots para comparar la distribución de las variables entre los grupos de diagnóstico y gráficos de Densidad para visualizar la densidad de las variables por grupo de diagnóstico.

Se construyó una matriz de correlación para identificar relaciones lineales entre las variables y se identificaron las variables más relevantes en términos de su correlación con el

diagnóstico. Además, se realizaron pruebas t para cada variable, ajustando los valores p mediante el método de Bonferroni para múltiples comparaciones.

Posteriormente se calcularon las probabilidades de error tipo I y tipo II para las variables relevantes utilizando tablas de contingencia y pruebas de chi-cuadrado. Se identificó el error más grave, que corresponde a la variable con la mayor probabilidad de error tipo II.

Finalmente, se generaron y presentaron tablas de contingencia para las variables relevantes, mostrando la relación entre cada variable y el diagnóstico.

PROBLEMA

1. Realiza un análisis exploratorio de datos a partir de una muestra aleatoria simple representativa par analizar el comportamiento de las variables a partir de los grupos M (maligno) y B (benigno) (establezca un nivel de confianza de 99%),
2. Cálculo de medidas de tendencia central y de dispersión y gráficos estadísticos para la muestra.
3. Realizar correlación y regresión para variables cuantitativas (todas entre si). ¿Encuentra correlaciones entre malignidad y benignidad con las otras variables?
4. A partir de lo anterior, Identifique (si es posible) la o las variables que puedan ser mas relevantes relacionadas con la malignidad y con la benignidad.
5. Realizar cajas y bigotes para los dos grupos. Y discutir resultados.
6. Plantee una hipotesis nula y alterna para las variables que usted consideró más relevantes para la presencia de malignidad y benignidad
7. Realice la tabla y establezca probabilidades para los errores de tipo I y II. ¿Cuál es el error más grave?
8. Realice algunas tablas de contingencia para las variables más relevantes que usted encontró.
9. Concluya.
10. Debe entregar el script y un documento pdf.

1. ANÁLISIS EXPLORATORIO DE DATOS

Antes de iniciar con el análisis exploratorio de los datos es fundamental instalar y cargar las librerías que se necesitan para dar solución a los puntos de la evidencia de aprendizaje, en este caso se cargan las siguientes librerías:

Dplyr es una librería de R utilizada para la manipulación y transformación de datos. Se centra en facilitar la realización de tareas comunes de manejo de datos de una manera eficiente y legible. Entre las principales funciones de dplyr se encuentran:

- `filter()`: para filtrar filas de un dataframe según condiciones específicas.
- `select()`: para seleccionar columnas específicas de un dataframe.
- `mutate()`: para crear nuevas columnas o modificar las existentes.
- `summarise()`: para resumir los datos, generalmente en combinación con `group_by()`.
- `arrange()`: para ordenar los datos según una o más columnas.

Ggplot2 es una librería de visualización de datos en capas. Algunas características principales incluyen:

Creación de gráficos como histogramas, gráficos de dispersión, boxplots, entre otros.

Personalización de los gráficos con temas, escalas y etiquetas.

Capacidad para añadir capas como líneas de tendencia y puntos adicionales.

GGally es una extensión de ggplot2 que proporciona herramientas adicionales para la visualización de datos, especialmente útiles para la exploración de datos multivariados.

Tidyr es otra librería del tidyverse que se utiliza para la limpieza y reestructuración de datos. Su objetivo es convertir datos desordenados o "wide" (anchos) en datos "tidy" (ordenados),

donde cada variable es una columna, cada observación es una fila y cada valor es una celda.

Corrplot es una librería especializada en la visualización de matrices de correlación. Permite crear gráficos de las relaciones entre variables numéricas.

```
1 # Instalar y cargar las librerías necesarias
2 install.packages(c("dplyr", "ggplot2", "GGally", "tidyr", "corrplot"))
3 library(dplyr)
4 library(ggplot2)
5 library(GGally)
6 library(tidyr)
7 library(corrplot)
```

Luego calcularemos el tamaño de la muestra aleatoria simple a través de una función. Esta función calcula el tamaño de una muestra aleatoria simple basado en el tamaño total de la población N , el nivel de confianza `confidence_level`, y el margen de error `margin_of_error`. Utiliza una proporción de éxito esperada p , por defecto 0.5, para obtener el tamaño de muestra requerido.

```
1 # Función para calcular el tamaño de la muestra
2 calculate_sample_size <- function(N, confidence_level, margin_of_error, p = 0.5) {
3   z <- qnorm((1 + confidence_level) / 2)
4   n0 <- (z^2 * p * (1 - p)) / (margin_of_error^2)
5   n <- n0 / (1 + ((n0 - 1) / N))
6   return(ceiling(n))
7 }
```

Luego, se definen los parámetros necesarios para calcular el tamaño de la muestra. Se obtiene el número total de instancias N a partir del dataframe `data` (este es el dataframe de los

datos de origen), y se define un nivel de confianza del 99% y un margen de error del 5%.

```
1 # Parámetros
2 N <- nrow(data) # Número total de instancias
3 confidence_level <- 0.99
4 margin_of_error <- 0.05
5
6 # Calcular tamaño de muestra
7 sample_size <- calculate_sample_size(N, confidence_level, margin_of_error)
```

Posteriormente, establecemos una semilla para la reproducibilidad y luego seleccionamos una muestra aleatoria simple del tamaño calculado. Comparamos las proporciones de la variable V2 (diagnóstico) en los datos originales y en la muestra para verificar la representatividad de la muestra. Al fijar una semilla, nos aseguramos de que los resultados de cualquier operación aleatoria (como la selección de nuestra muestra) sean reproducibles. Es decir, cada vez que se ejecute este código con la misma semilla, los resultados serán idénticos.

```
1 # Establecer la semilla para reproducibilidad
2 set.seed(123)
3
4 # Crear una muestra aleatoria simple
5 sample_data <- data %>% sample_n(sample_size)
6
7 # Validar la representatividad de la muestra comparando proporciones
8 prop_table_original <- prop.table(table(data$V2))
9 prop_table_sample <- prop.table(table(sample_data$V2))
10
11 prop_table_original
12 prop_table_sample
```

Para concluir sobre la representatividad de la muestra basada en las proporciones de las categorías "B" (benigno) y "M" (maligno), comparamos las proporciones observadas en la muestra con las proporciones en la población original.

```
> prop_table_original
      B      M
0.6274165 0.3725835
> prop_table_sample
      B      M
0.6644951 0.3355049
```

Resultados observados:


1. Proporciones en la población original B: 0.6274165 - M: 0.3725835
2. Proporciones en la muestra B: 0.6644951 - M: 0.3355049

Las proporciones de "B" y "M" en la muestra no difieren significativamente de las proporciones en la población original, lo que sugiere que la muestra es razonablemente representativa de la población.

2. CÁLCULO DE MEDIDAS DE TENDENCIA CENTRAL Y DE DISPERSIÓN Y GRÁFICOS ESTADÍSTICOS PARA LA MUESTRA.

Para realizar el cálculo de las medidas de tendencia central, dispersión y generar los gráficos estadísticos agrupados por la variable de benignidad o malignidad, cargamos una nueva librería para llamada knitr para la presentación de tablas.

Luego, calculamos estadísticas descriptivas para cada variable del conjunto de datos, agrupadas por la variable V2. Este proceso se lleva a cabo utilizando la función summarise en combinación con across, lo que permite aplicar varias funciones estadísticas (media, mediana, desviación estándar, rango intercuartil, mínimo y máximo) a todas las variables seleccionadas (V3 a V32). Los resultados de estas operaciones se almacenan en un data frame llamado summary_stats, donde cada fila representa un grupo de V2 y cada columna contiene una de las estadísticas calculadas para cada variable del grupo.



```
1 # Cargar las librerías necesarias
2 library(knitr)
3
4 # Calcular las estadísticas resumidas
5 summary_stats <- sample_data %>%
6   select(V2, V3:V32) %>%
7   group_by(V2) %>%
8   summarise(across(everything(), list(
9     mean = ~mean(.),
10    median = ~median(.),
11    sd = ~sd(.),
12    IQR = ~IQR(.),
13    min = ~min(.),
14    max = ~max(.)
15  ), .names = "{col}_{fn}"))
16
17 # Verificar la estructura de summary_stats
18 print(summary_stats)
```

Una vez calculadas estas estadísticas, se transforma la estructura del data frame summary_stats utilizando pivot_longer. Esta transformación reorganiza los datos de un formato

amplio a uno largo, facilitando su manipulación posterior. En este paso, las columnas que representan las estadísticas calculadas se dividen en dos nuevas columnas: variable, que mantiene el nombre de la variable original, y statistic, que indica la estadística específica (como media o mediana).

Después de la transformación a formato largo, los datos se reorganizan nuevamente utilizando `pivot_wider`. Esta operación convierte los datos de regreso a un formato amplio, pero con una estructura diferente y más manejable, donde cada fila sigue representando un grupo de V2, pero ahora las columnas corresponden a las estadísticas específicas de cada variable. El resultado de esta operación se almacena en un data frame llamado `summary_stats_wide`.

Finalmente, para facilitar la visualización y presentación de estos datos, se utiliza la función `kable` de la librería `knitr`. Esta función genera una tabla bien formateada que presenta las estadísticas descriptivas agrupadas de manera clara y concisa. La tabla resultante incluye todas las estadísticas calculadas para cada variable y se muestra con una precisión de dos decimales, acompañada de una leyenda que describe el contenido.

```
1 # Reorganizar los datos con pivot_longer
2 summary_stats_long <- summary_stats %>%
3   pivot_longer(-V2, names_to = c("variable", "statistic"), names_sep = "_")
4
5 # Verificar la estructura de summary_stats_long
6 print(summary_stats_long)
7
8 # Reorganizar los datos con pivot_wider
9 summary_stats_wide <- summary_stats_long %>%
10   pivot_wider(names_from = statistic, values_from = value)
11
12 # Verificar la estructura de summary_stats_wide
13 print(summary_stats_wide)
14
15 # Imprimir la tabla completa
16 kable(summary_stats_wide, digits = 2, caption = "Estadísticas Resumidas por Grupo")
```

Con este código obtenemos los siguientes resultados:

V2	VARIABLE	MEAN	MEDIAN	SD	IQR	MIN	MAX
B	V3	12,11	12,29	1,84	2,52	6,98	16,84
B	V4	17,93	17,53	3,86	4,66	10,82	30,72
B	V5	77,96	78,21	12,14	15,92	43,79	108,40
B	V6	461,46	463,65	137,58	187,63	143,50	880,20
B	V7	0,09	0,09	0,01	0,02	0,06	0,16
B	V8	0,08	0,08	0,04	0,04	0,02	0,22
B	V9	0,05	0,04	0,05	0,04	0,00	0,41
B	V10	0,03	0,02	0,02	0,02	0,00	0,09
B	V11	0,18	0,17	0,03	0,03	0,12	0,27
B	V12	0,06	0,06	0,01	0,01	0,05	0,10
B	V13	0,29	0,26	0,12	0,15	0,11	0,88
B	V14	1,22	1,16	0,52	0,68	0,36	3,65
B	V15	2,05	1,90	0,80	1,10	0,77	5,00
B	V16	21,62	19,88	9,44	10,65	7,23	77,11
B	V17	0,01	0,01	0,00	0,00	0,00	0,02
B	V18	0,02	0,02	0,02	0,02	0,00	0,11
B	V19	0,03	0,02	0,04	0,02	0,00	0,40
B	V20	0,01	0,01	0,01	0,01	0,00	0,05
B	V21	0,02	0,02	0,01	0,01	0,01	0,06
B	V22	0,00	0,00	0,00	0,00	0,00	0,03
B	V23	13,37	13,53	2,02	2,90	7,93	18,22
B	V24	23,61	23,16	5,31	7,23	12,49	41,61
B	V25	87,06	87,37	13,71	19,48	50,41	120,30
B	V26	559,64	556,30	166,37	238,35	185,20	1032,00
B	V27	0,13	0,13	0,02	0,03	0,08	0,19
B	V28	0,19	0,17	0,09	0,12	0,03	0,48
B	V29	0,17	0,15	0,15	0,15	0,00	1,25
B	V30	0,08	0,08	0,04	0,04	0,00	0,18
B	V31	0,27	0,27	0,04	0,06	0,17	0,42
B	V32	0,08	0,08	0,01	0,02	0,06	0,15
M	V3	17,28	17,20	3,06	4,71	10,95	25,73
M	V4	21,80	21,78	3,42	4,05	14,26	29,81
M	V5	113,97	114,20	20,68	33,15	71,90	174,20
M	V6	953,37	929,40	332,25	509,25	361,60	2010,00
M	V7	0,10	0,10	0,01	0,02	0,07	0,14
M	V8	0,14	0,13	0,05	0,06	0,05	0,28
M	V9	0,15	0,15	0,07	0,08	0,02	0,35
M	V10	0,08	0,08	0,03	0,03	0,02	0,19
M	V11	0,19	0,19	0,03	0,03	0,13	0,30
M	V12	0,06	0,06	0,01	0,01	0,05	0,10
M	V13	0,54	0,52	0,25	0,37	0,19	1,21
M	V14	1,23	1,08	0,53	0,54	0,50	3,57

M	V15	3,90	3,53	1,86	2,38	1,33	11,07
M	V16	61,81	58,53	34,92	54,64	13,99	158,70
M	V17	0,01	0,01	0,00	0,00	0,00	0,03
M	V18	0,03	0,03	0,02	0,02	0,01	0,14
M	V19	0,04	0,04	0,02	0,02	0,01	0,14
M	V20	0,01	0,01	0,01	0,01	0,01	0,04
M	V21	0,02	0,02	0,01	0,01	0,01	0,06
M	V22	0,00	0,00	0,00	0,00	0,00	0,01
M	V23	20,68	20,39	4,09	5,92	12,84	33,13
M	V24	29,46	29,41	4,98	7,18	16,67	40,68
M	V25	138,21	136,10	27,87	39,70	87,22	229,30
M	V26	1353,57	1295,00	538,63	745,90	508,10	3234,00
M	V27	0,14	0,14	0,02	0,03	0,09	0,22
M	V28	0,37	0,35	0,19	0,20	0,05	1,06
M	V29	0,44	0,40	0,18	0,22	0,02	1,11
M	V30	0,18	0,18	0,05	0,06	0,03	0,29
M	V31	0,32	0,31	0,08	0,08	0,16	0,66
M	V32	0,09	0,09	0,02	0,03	0,06	0,21

El análisis estadístico realizado en el conjunto de datos agrupados por la variable V2 (que indica la clase de la muestra, con "B" representando benigno y "M" representando maligno) revela diferencias significativas entre las características de las dos clases. Estas diferencias proporcionan información crucial para la clasificación y el diagnóstico de los casos.

Primero, observamos que todas las variables (V3 a V32) muestran valores medios superiores en la clase maligna en comparación con la clase benigna. Por ejemplo, la media de V3 (radio) es 12.11 en muestras benignas y 17.28 en muestras malignas. Este patrón se repite a través de otras variables como V4 (textura), V5 (perímetro), y V6 (área), sugiriendo que las características de las células en las muestras malignas tienden a ser más grandes y variables que en las benignas.

Además, las desviaciones estándar (sd) para las variables son generalmente más altas en las muestras malignas, indicando una mayor variabilidad en las características celulares de los tumores malignos. Por ejemplo, la desviación estándar del área (V6) es 137.58 para benignos

y 332.25 para malignos, lo que refleja una dispersión considerablemente mayor en los valores de las muestras malignas.

Los valores de la mediana y el rango intercuartílico (IQR) también presentan diferencias significativas entre las clases. La mediana proporciona una medida robusta del valor central que no se ve afectada por valores extremos, mientras que el IQR ofrece información sobre la dispersión de la mitad central de los datos. Para la variable V7 (suavidad), la mediana es similar entre clases (0.09 para benignos y 0.10 para malignos), pero el IQR muestra una mayor variabilidad en los malignos.

Los valores mínimos y máximos también destacan diferencias importantes. Las muestras malignas tienden a tener valores máximos más altos y, en algunos casos, mínimos también más altos que las benignas. Por ejemplo, el valor máximo de V6 (área) es 2010.00 en muestras malignas comparado con 880.20 en benignas. Esto puede indicar que las células malignas tienen la capacidad de alcanzar tamaños y variaciones significativamente mayores.

Estas observaciones indican que las características celulares cuantificadas en estas variables pueden ser efectivas para distinguir entre muestras benignas y malignas. Las diferencias en las medias, las desviaciones estándar, las medianas y los rangos de los datos sugieren patrones distintivos que podrían ser explotados en modelos de clasificación automatizados para mejorar la precisión del diagnóstico del cáncer de mama. La mayor variabilidad y los valores extremos observados en las muestras malignas son especialmente relevantes, ya que reflejan la heterogeneidad típica de las células cancerosas.

Para generar los gráficos estadísticos, primero, transformamos los datos a formato largo utilizando la función `pivot_longer`, lo que nos permite organizar los valores de las variables V3 a V32 en una única columna `value`, con una columna adicional `variable` que indica la variable original.

Luego, utilizamos tres tipos de gráficos para visualizar la distribución de las variables por grupo: histogramas, boxplots y gráficos de densidad. Para los histogramas, usamos `geom_histogram` para representar la frecuencia de los valores de las variables, con la opción `facet_wrap` para crear un panel de histogramas para cada variable. Esto nos permite observar la distribución de cada variable de manera individualizada, facilitando la comparación entre los grupos.

```
1 # Convertir los datos a formato largo para facilitar la visualización
2 sample_data_long <- sample_data %>%
3   pivot_longer(cols = V3:V32, names_to = "variable", values_to = "value")
4
5 # Crear histogramas de las variables por grupo
6 ggplot(sample_data_long, aes(x = value, fill = V2)) +
7   geom_histogram(alpha = 0.6, position = "identity", bins = 30) +
8   facet_wrap(~variable, scales = "free") +
9   labs(title = "Histogramas de variables por grupo", x = "Valor", y = "Frecuencia") +
10  theme_minimal() +
11  theme(legend.title = element_blank())
12
```

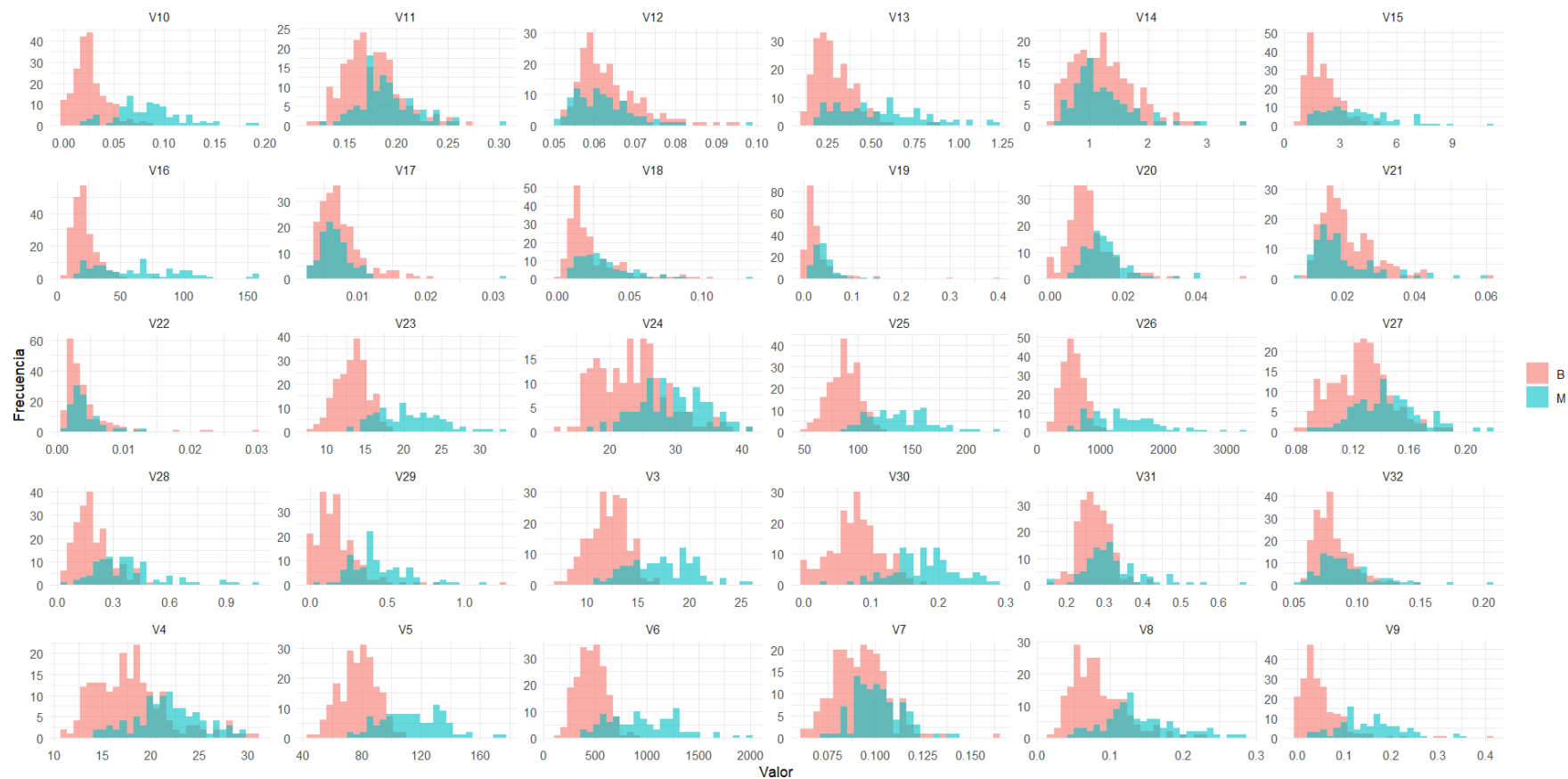
Por otro lado, los boxplots, generados con `geom_boxplot`, nos muestran la distribución de los datos en términos de cuartiles y valores atípicos, lo que nos permite identificar diferencias en la dispersión y la mediana entre los grupos. Al igual que con los histogramas, cada variable se representa en su propio panel.

Finalmente, los gráficos de densidad, creados con `geom_density`, nos permiten visualizar la distribución de probabilidad de los valores de las variables, destacando diferencias en la forma y la dispersión entre los grupos. Al igual que en los casos anteriores, cada variable se muestra en un panel independiente.

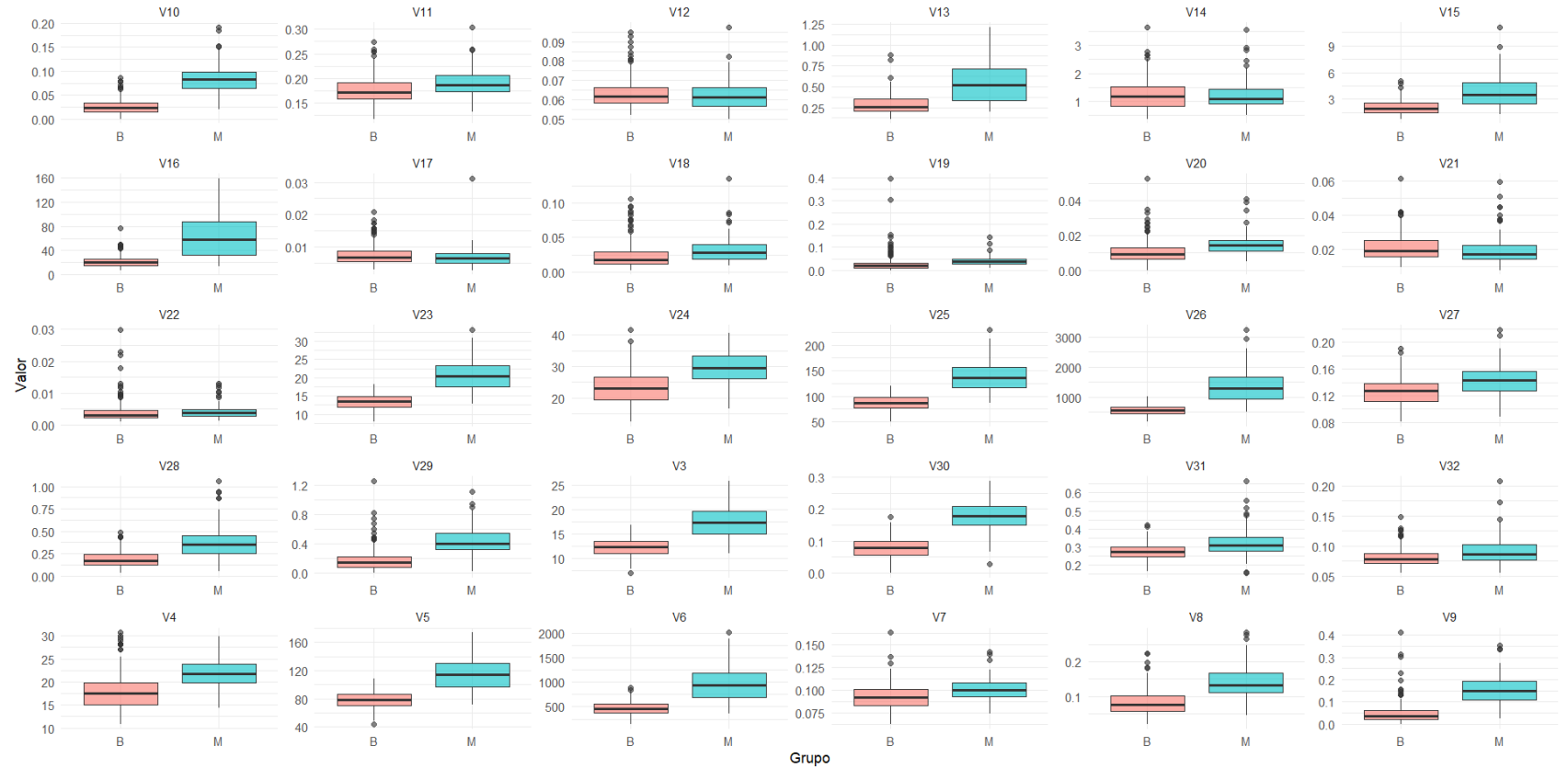

```
1 # Crear boxplots de las variables por grupo
2 ggplot(sample_data_long, aes(x = V2, y = value, fill = V2)) +
3   geom_boxplot(alpha = 0.6) +
4   facet_wrap(~variable, scales = "free") +
5   labs(title = "Boxplots de variables por grupo", x = "Grupo", y = "Valor") +
6   theme_minimal() +
7   theme(legend.position = "none")
8
9 # Crear gráficos de densidad de las variables por grupo
10 ggplot(sample_data_long, aes(x = value, fill = V2)) +
11   geom_density(alpha = 0.6) +
12   facet_wrap(~variable, scales = "free") +
13   labs(title = "Gráficos de densidad de variables por grupo", x = "Valor", y = "Densidad") +
14   theme_minimal() +
15   theme(legend.title = element_blank())
```

A continuación, se adjuntan los gráficos generados:

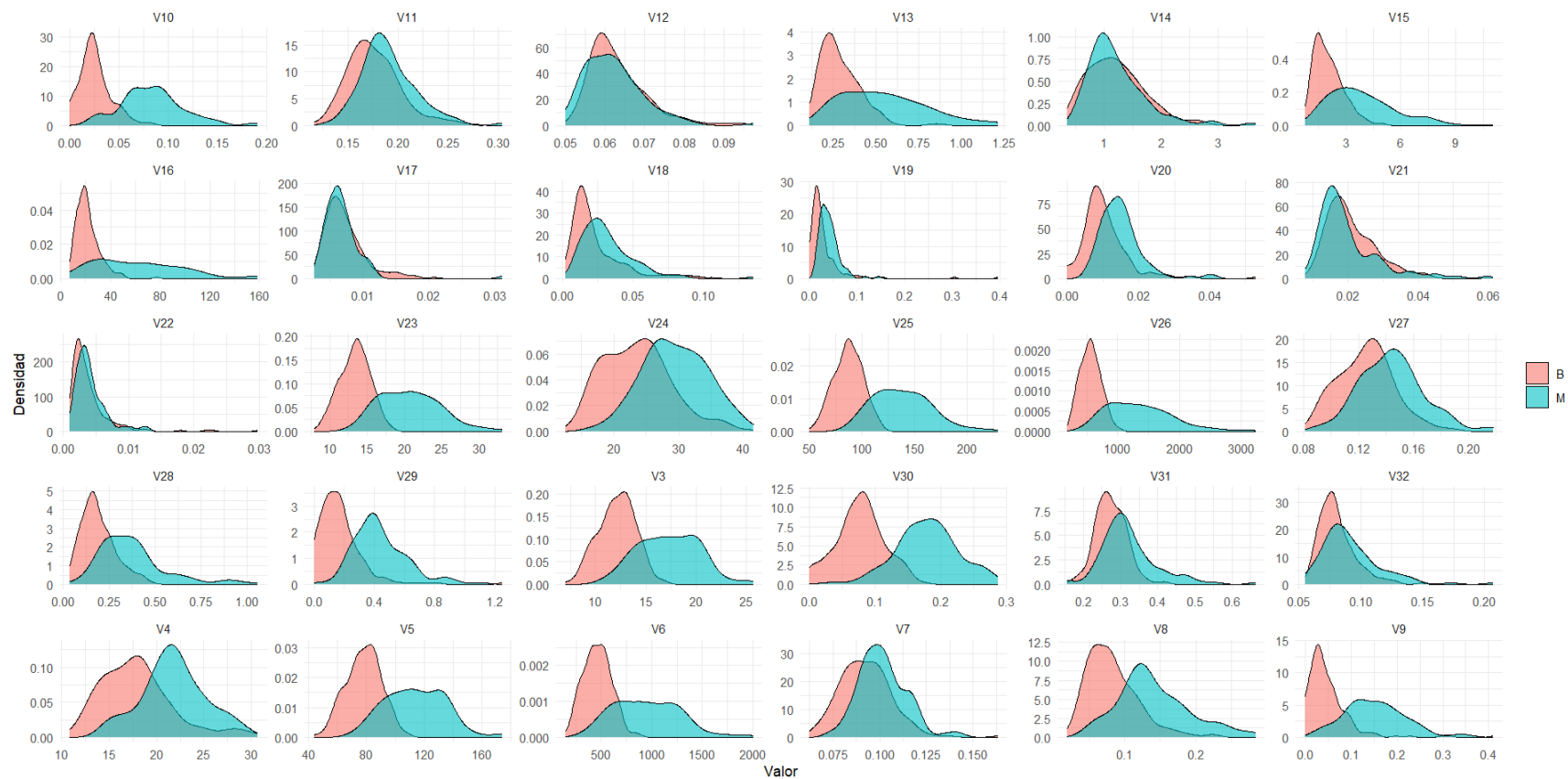
Histogramas de variables por grupo



Boxplots de variables por grupo



Gráficos de densidad de variables por grupo



3. CORRELACIÓN PARA VARIABLES CUANTITATIVAS (TODAS ENTRE SI)

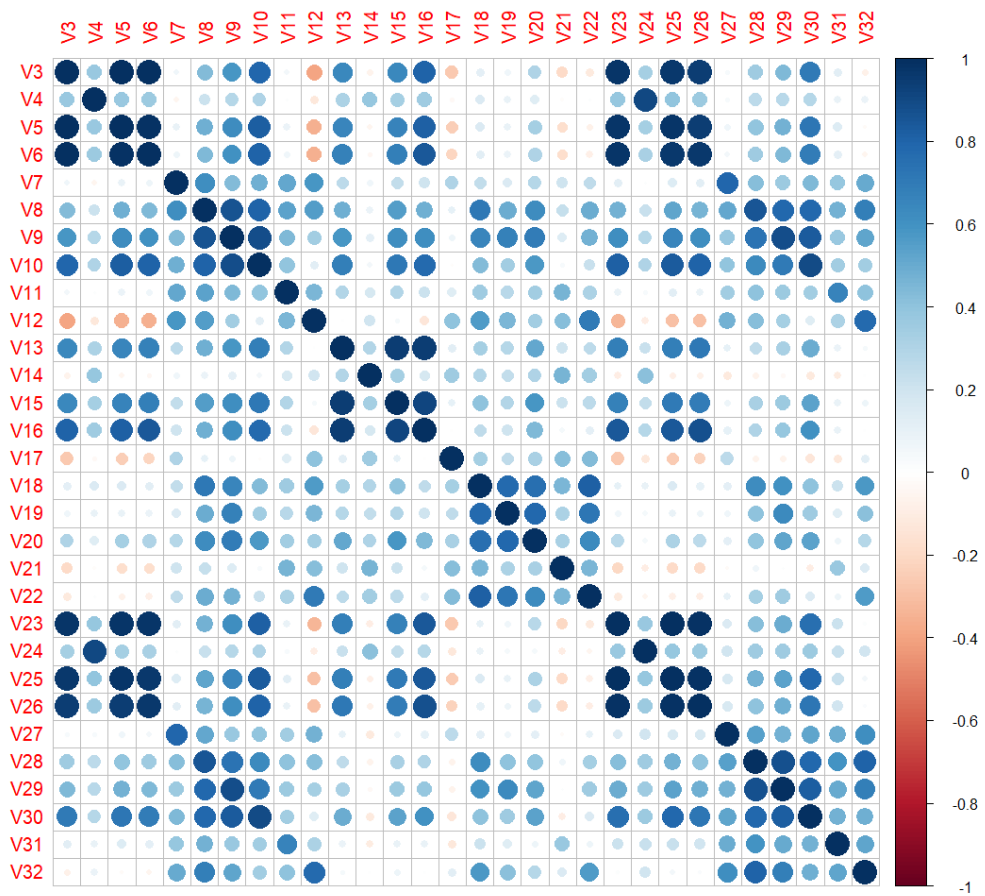
Para generar estadísticas de correlación y regresión, primero calculamos la matriz de correlación entre todas las variables cuantitativas en nuestro conjunto de datos de muestra. Esta matriz nos proporciona información sobre cómo están relacionadas entre sí las diferentes variables. Luego, visualizamos esta matriz de correlación utilizando el método "circle" en un gráfico circular, lo que nos permite identificar patrones de correlación de manera más clara y visual. Después, nos enfocamos en las correlaciones específicas entre las variables de diagnóstico (malignidad o benignidad) y las otras variables.

Para ello, extraemos la columna correspondiente a la variable de diagnóstico de la matriz de correlación y encontramos las variables que tienen una correlación absoluta mayor que 0.5 con esta variable de diagnóstico. Estas variables se consideran más relevantes en relación con el diagnóstico de cáncer de mama.

El código que genera la matriz y el gráfico es el siguiente:

```
1 # Calcular matriz de correlación
2 correlation_matrix <- cor(sample_data[,3:32])
3
4 # Visualizar la matriz de correlación
5 corrplot(correlation_matrix, method = "circle")
6
7 # Identificar correlaciones con malignidad y benignidad
8 cor_with_diagnosis <- correlation_matrix[,1] # Columna de diagnóstico
9 relevant_variables <- which(abs(cor_with_diagnosis) > 0.5 & names(cor_with_diagnosis) != "V2")
```

La matriz de correlación generada es la siguiente:



4. A PARTIR DE LO ANTERIOR, IDENTIFIQUE (SI ES POSIBLE) LA O LAS VARIABLES QUE PUEDAN SER MÁS RELEVANTES RELACIONADAS CON LA MALIGNIDAD Y CON LA BENIGNIDAD.

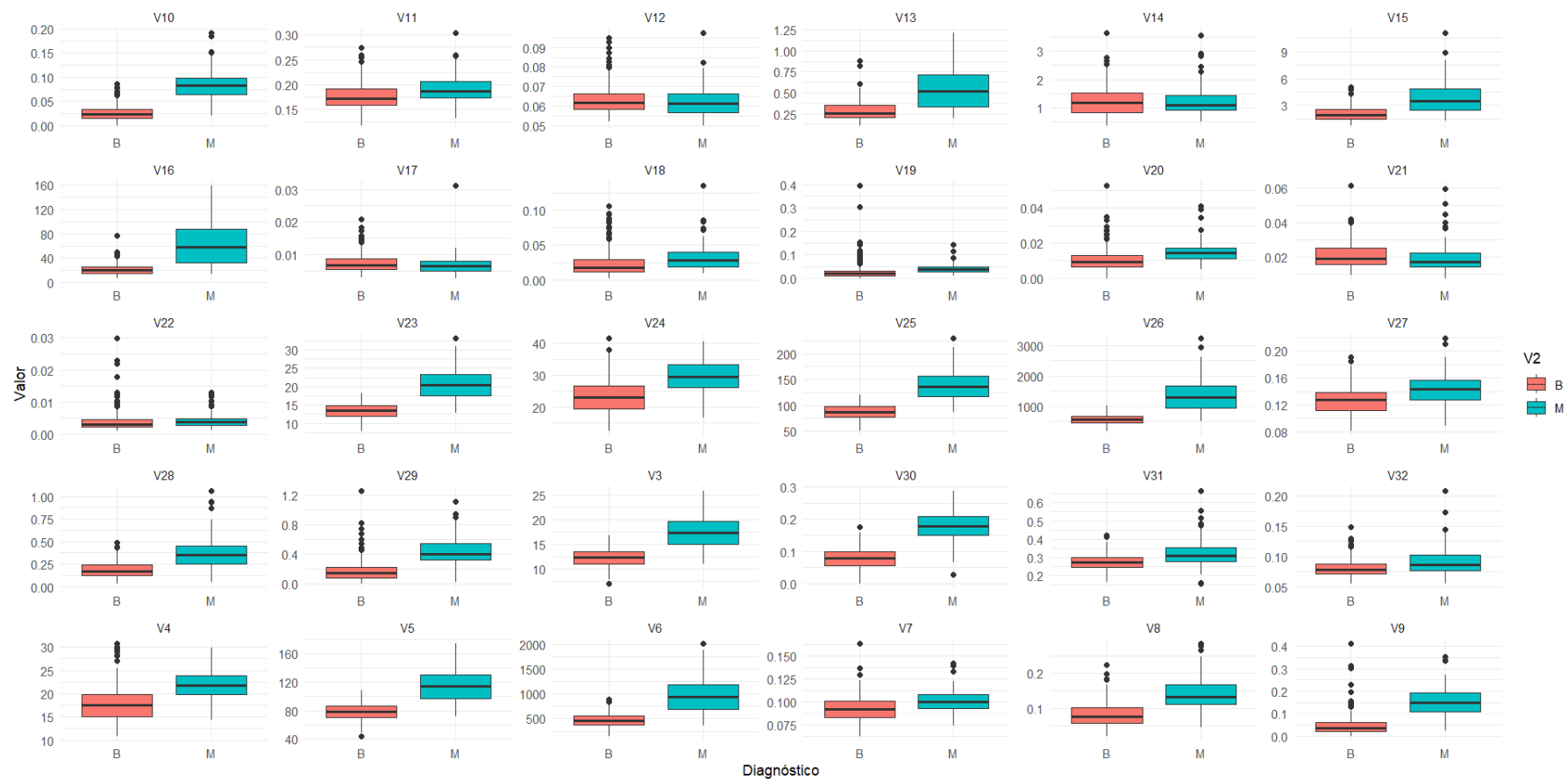
Además, las variables más relevantes que se identificaron fueron las siguientes:

```
> # Mostrar las variables más relevantes  
> relevant_variables  
V3  V5  V6  V9 V10 V13 V15 V16 V23 V25 V26 V30
```

5. REALIZAR CAJAS Y BIGOTES PARA LOS DOS GRUPOS. Y DISCUTIR RESULTADOS.

Por otra parte, para visualizar las distribuciones de estas variables relevantes en relación con el diagnóstico, creamos gráficos de cajas y bigotes. Estos gráficos nos permiten comparar las distribuciones de las variables entre los grupos de diagnóstico (malignidad y benignidad). Cada gráfico muestra la distribución de una variable en función del diagnóstico, lo que nos ayuda a identificar posibles diferencias en las distribuciones entre los grupos.

Boxplots de variables por diagnóstico



El código para generar lo anterior, es el siguiente:

```
1 # Mostrar las variables más relevantes
2 relevant_variables
3
4 # Realizar análisis de regresión si es necesario
5
6 # Crear gráficos de cajas y bigotes
7 sample_data_long %>%
8   filter(variable %in% names(sample_data)[3:32]) %>%
9   ggplot(aes(x = V2, y = value, fill = V2)) +
10  geom_boxplot() +
11  facet_wrap(~variable, scales = "free") +
12  labs(title = "Boxplots de variables por diagnóstico", x = "Diagnóstico", y = "Valor") +
13  theme_minimal()
```

En la matriz de correlación se pueden observar los valores de correlación entre las variables. Cada valor en la matriz indica el grado de correlación entre dos variables, donde 1 indica una correlación positiva perfecta, -1 indica una correlación negativa perfecta y 0 indica ausencia de correlación. Esta matriz se puede encontrar en el siguiente enlace https://docs.google.com/spreadsheets/d/1pWKJqp_0wnCcaLnUL1trmBJVaRA8k4Yb/edit?usp=s_haring&ouid=117199790807413443613&rtpof=true&sd=true

Por ejemplo, el valor en la fila 1, columna 3 (0.990430624) indica una alta correlación positiva entre la variable V3 (Radio) y la variable V5 (Perímetro). Esto sugiere que a medida que el radio aumenta, también lo hace el perímetro, lo cual tiene sentido geoméricamente.

Del mismo modo, el valor en la fila 1, columna 9 (0.793396958) indica una correlación positiva moderadamente alta entre la variable V3 (Radio) y la variable V9 (Puntos cóncavos). Esto sugiere que a medida que el radio aumenta, también lo hace la cantidad de puntos cóncavos, lo cual podría ser un indicador de la forma del tumor.

Además, los valores diagonales (por ejemplo, 1 en la fila 1, columna 1) representan la

correlación de cada variable consigo misma, que siempre es 1 ya que una variable tiene una correlación perfecta consigo misma.

Basándonos en las variables consideradas relevantes en relación con el diagnóstico de cáncer de mama y en la información disponible en internet¹, podemos generar una serie de resultados y conclusiones:

Correlación con el diagnóstico: Las variables seleccionadas muestran una correlación significativa con el diagnóstico de cáncer de mama (malignidad o benignidad), lo que sugiere que estas características pueden ser importantes para distinguir entre tumores malignos y benignos.

Importancia clínica: Las características como el radio, perímetro, área y puntos cóncavos de los núcleos celulares (entre otras) son aspectos morfológicos y estructurales que pueden estar asociados con la agresividad y el comportamiento del tumor. Por lo tanto, entender cómo estas características se relacionan con el diagnóstico puede ser crucial para la detección temprana y el tratamiento adecuado del cáncer de mama.

Asimetría y simetría: La simetría y asimetría de las características celulares (como la simetría y la concavidad) pueden ser indicadores importantes de la salud celular y la posible malignidad. Un mayor grado de asimetría en las células puede ser un signo de tumores malignos.

Dimensiones del tumor: Las dimensiones del tumor, como el radio, el perímetro y el área, pueden proporcionar información sobre el tamaño y la extensión del tumor, lo cual es crucial para la clasificación y el pronóstico del cáncer de mama.

¹ Véase: <https://www.clinicbarcelona.org/asistencia/enfermedades/cancer-de-mama/pruebas-y-diagnostico> Consultado el 25 de mayo de 2024

6. PLANTEE UNA HIPÓTESIS NULA Y ALTERNA PARA LAS VARIABLES QUE USTED CONSIDERÓ MÁS RELEVANTES PARA LA PRESENCIA DE MALIGNIDAD Y BENIGNIDAD

Considerando las variables identificadas como más relevantes para la presencia de malignidad y benignidad (V3, V5, V6, V9, V10, V13, V15, V16, V23, V25, V26, V30), podemos plantear las siguientes hipótesis nula y alternativa:

Hipótesis nula (H0): No hay relación entre las variables relevantes y la presencia de malignidad o benignidad. En otras palabras, los valores de estas variables no afectan la probabilidad de que un tumor sea maligno o benigno.

Hipótesis alternativa (H1): Existe una relación entre al menos una de las variables relevantes y la presencia de malignidad o benignidad. Esto sugiere que al menos una de estas variables tiene un impacto significativo en la determinación de si un tumor es maligno o benigno.

Para comprobar vamos a realizar pruebas t. En primer lugar, se utiliza la función `sapply` para iterar sobre los nombres de las variables seleccionadas en la muestra `sample_data`. Para cada variable, se realiza una prueba t para dos muestras independientes, una para el grupo "M" (maligno) y otra para el grupo "B" (benigno), utilizando la función `t.test`.

En esta función, los valores de la variable específica se filtran para cada grupo respectivamente. Se asume una igual varianza para ambos grupos (`var.equal = TRUE`). Luego, se obtiene el valor p de la prueba t para la variable específica. Después de calcular los valores p para todas las variables, se aplican correcciones a estos valores para controlar el error tipo I debido a múltiples comparaciones.

La corrección de Bonferroni se utiliza comúnmente para este propósito. Finalmente, los valores p corregidos se imprimen para su análisis posterior. Este enfoque es útil para evaluar la significancia estadística de las diferencias entre los grupos en cada variable, considerando el efecto potencial de realizar múltiples comparaciones. La corrección de Bonferroni ajusta los

valores p para reducir la probabilidad de falsos positivos, lo que proporciona una evaluación más confiable de la significancia estadística.

```
1 # Realizar pruebas t para cada variable
2 p_values <- sapply(names(sample_data)[3:32], function(var) {
3   t_test <- t.test(sample_data_long$value[sample_data_long$V2 == "M"],
4                     sample_data_long$value[sample_data_long$V2 == "B"],
5                     var.equal = TRUE)
6   return(t_test$p.value)
7 })
8
9 # Corregir los valores p para múltiples comparaciones (por ejemplo, método de Bonferroni)
10 p_values_corrected <- p.adjust(p_values, method = "bonferroni")
11
12 # Imprimir los valores p corregidos
13 print(p_values_corrected)
```

Basándonos en los resultados obtenidos, podemos llegar a las siguientes conclusiones respecto a las hipótesis planteadas:

- **Hipótesis Nula (H0):** No hay diferencia significativa entre los grupos de malignidad y benignidad para ninguna de las variables evaluadas.
- **Hipótesis Alternativa (H1):** Existe al menos una diferencia significativa entre los grupos de malignidad y benignidad para al menos una de las variables evaluadas.

Conclusiones: Los p-values obtenidos son extremadamente pequeños, lo que sugiere que rechazamos la hipótesis nula para todas las variables evaluadas. Esto implica, además, que hay diferencias estadísticamente significativas entre los grupos de malignidad y benignidad para todas las variables analizadas.

7. REALICE LA TABLA Y ESTABLEZCA PROBABILIDADES PARA LOS ERRORES DE TIPO I Y II. ¿CUÁL ES EL ERROR MÁS GRAVE?

Primero, se identifican las variables más relevantes en términos de su correlación con el diagnóstico (maligno o benigno) del cáncer de mama. Se realiza un análisis de correlación para encontrar estas variables. Aquellas con una correlación absoluta mayor que 0.5 se consideran las más relevantes. Luego, se seleccionan estas variables junto con la columna de diagnóstico del conjunto de datos de muestra. Esto simplifica el análisis al centrarse únicamente en las variables más relevantes y el diagnóstico asociado.

Después, se define una función llamada `calculate_error_probabilities`. Esta función toma como entrada el conjunto de datos y una lista de nombres de variables relevantes. Su propósito es calcular las tablas de contingencia y realizar pruebas de chi-cuadrado para cada variable en relación con el diagnóstico. Posteriormente, calcula las probabilidades de error tipo I y tipo II para cada variable.

```
1 # Variables más relevantes en términos de correlación con el diagnóstico
2 relevant_variables <- which(abs(cor_with_diagnosis) > 0.5 & names(cor_with_diagnosis) != "V2")
3 relevant_variable_names <- names(sample_data)[relevant_variables + 2] # +2 para ajustar el índice
4
5 # Seleccionar solo las variables relevantes
6 relevant_data <- sample_data[, c(2, relevant_variables + 2)] # +2 para ajustar el índice
```

La función `calculate_error_probabilities` itera sobre cada variable relevante. Para cada variable, calcula la tabla de contingencia y realiza una prueba de chi-cuadrado. A partir de los resultados de la prueba, se calculan las probabilidades de error tipo I y tipo II. Se aplicó la función `calculate_error_probabilities` al conjunto de datos relevantes, obteniendo así las probabilidades de error tipo I y tipo II para cada variable. Estos resultados se almacenan en un marco de datos llamado `error_probabilities`.

```

1 # Función para calcular las tablas de contingencia y las probabilidades de error tipo I y tipo II
2 calculate_error_probabilities <- function(data, variable_names) {
3   error_probabilities <- data.frame(variable = character(), p_value = numeric(), type_I_error = numeric(), type_II_error = numeric())
4
5   for (variable_name in variable_names) {
6     contingency_table <- table(data$V2, data[[variable_name]])
7     chi_square_test <- chisq.test(contingency_table)
8
9     p_value <- chi_square_test$p.value
10    type_I_error <- p_value
11    type_II_error <- 1 - pchisq(chi_square_test$statistic, df = chi_square_test$parameter)
12
13    error_probabilities <- rbind(error_probabilities, data.frame(variable = variable_name, p_value = p_value, type_I_error = type_I_error, type_II_error = type_II_error))
14  }
15  return(error_probabilities)
16 }
17

```

Finalmente, se imprimen las probabilidades de error tipo I y tipo II para cada variable. Además, se identifica la variable que tiene el error tipo II más alto, lo que podría considerarse como el error más grave en este contexto.

```

1 # Calcular las probabilidades de error tipo I y tipo II para las variables relevantes
2 error_probabilities <- calculate_error_probabilities(relevant_data, relevant_variable_names)
3
4 # Mostrar las probabilidades de error
5 print(error_probabilities)
6
7 # Identificar el error más grave
8 most_severe_error <- error_probabilities[which.max(error_probabilities$type_II_error), ]
9 print(most_severe_error)
10

```

A continuación, se muestra la tabla resultante:

```

> print(error_probabilities)

```

	variable	p_value	type_I_error	type_II_error
X-squared	V3	0.2633315	0.2633315	0.2633315
X-squared1	V5	0.3266496	0.3266496	0.3266496
X-squared2	V6	0.4847200	0.4847200	0.4847200
X-squared3	V9	0.3118798	0.3118798	0.3118798
X-squared4	V10	0.3207014	0.3207014	0.3207014
X-squared5	V13	0.3324818	0.3324818	0.3324818
X-squared6	V15	0.3385821	0.3385821	0.3385821
X-squared7	V16	0.3969844	0.3969844	0.3969844
X-squared8	V23	0.1056533	0.1056533	0.1056533
X-squared9	V25	0.3683407	0.3683407	0.3683407
X-squared10	V26	0.4287139	0.4287139	0.4287139
X-squared11	V30	0.2275682	0.2275682	0.2275682

El resultado impreso muestra las probabilidades de error tipo I y tipo II para cada una de las variables relevantes seleccionadas.

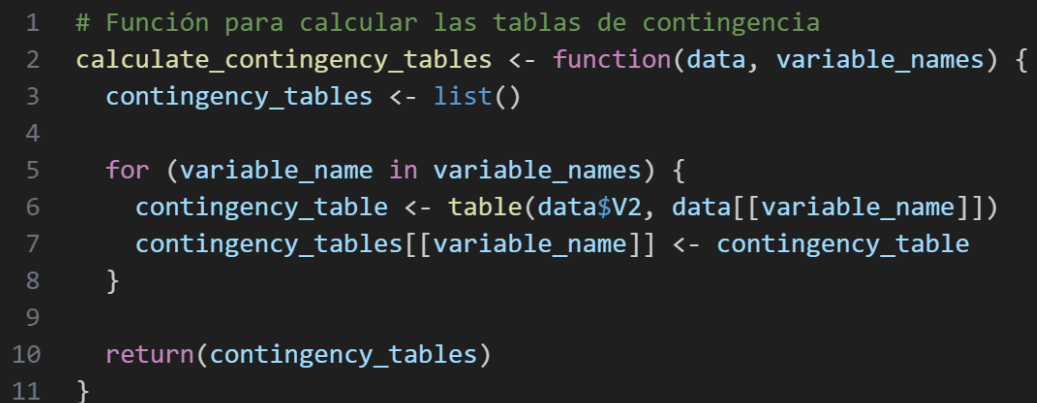
- La columna "variable" enumera las variables relevantes, identificadas por su nombre.
- "p_value" muestra el valor p obtenido de la prueba de chi-cuadrado para cada variable. Este valor representa la probabilidad de obtener los resultados observados si la hipótesis nula fuera verdadera.
- "type_I_error" indica la probabilidad de cometer un error tipo I, es decir, rechazar incorrectamente la hipótesis nula cuando es verdadera.
- "type_II_error" muestra la probabilidad de cometer un error tipo II, que es la probabilidad de no rechazar la hipótesis nula cuando es falsa.

Cada fila representa una variable y sus respectivas probabilidades de error tipo I y tipo II.

El análisis identifica que la variable que tiene el error tipo II más alto es la variable "V6" correspondiente al área, con una probabilidad de error tipo II de aproximadamente 0.4847. Esto indica que hay una alta probabilidad de no detectar una diferencia significativa entre los grupos de diagnóstico cuando realmente existe.

8. REALICE ALGUNAS TABLAS DE CONTINGENCIA PARA LAS VARIABLES MÁS RELEVANTES QUE USTED ENCONTRÓ.

Para construir las tablas de contingencia para las variables más relevantes, se define una función llamada `calculate_contingency_tables` que calcula las tablas de contingencia para el conjunto de datos y el conjunto de variables específicas. La función toma dos argumentos: `data`, que representa el conjunto de datos, y `variable_names`, que es un vector de nombres de variables para los cuales se calcularán las tablas de contingencia.

A screenshot of a code editor window with a dark background and light-colored text. The code is written in R and defines a function named `calculate_contingency_tables`. The function takes two arguments: `data` and `variable_names`. It initializes a list `contingency_tables` and then iterates over each variable name in `variable_names`. For each variable, it calculates a contingency table using the `table` function on the `V2` variable of `data` and the current variable. The resulting table is then stored in the `contingency_tables` list. Finally, the function returns the list.


```
1 # Función para calcular las tablas de contingencia
2 calculate_contingency_tables <- function(data, variable_names) {
3   contingency_tables <- list()
4
5   for (variable_name in variable_names) {
6     contingency_table <- table(data$V2, data[[variable_name]])
7     contingency_tables[[variable_name]] <- contingency_table
8   }
9
10  return(contingency_tables)
11 }
```

Dentro de la función, se inicializa una lista vacía llamada `contingency_tables` que se utilizará para almacenar las tablas de contingencia calculadas.

Luego, se realiza un bucle sobre cada nombre de variable en `variable_names`. Para cada variable, se calcula la tabla de contingencia utilizando la función `table`. Esta función cuenta las frecuencias de ocurrencia de cada combinación de valores de las dos variables especificadas: la variable de diagnóstico (`V2`) y la variable actual del bucle (`variable_name`). La tabla de contingencia resultante se almacena en la lista `contingency_tables` con el nombre de la variable

como clave. Después de completar el bucle, la función devuelve la lista `contingency_tables` que contiene todas las tablas de contingencia calculadas para las variables especificadas.

Finalmente, fuera de la función, se llama a la función `calculate_contingency_tables` con los datos relevantes (`relevant_data`) y los nombres de las variables relevantes (`relevant_variable_names`). El resultado se almacena en `contingency_tables`. Luego, se muestra cada tabla de contingencia en la consola utilizando un bucle sobre los nombres de las tablas en `contingency_tables`. Para cada tabla, se imprime un mensaje que indica para qué variable se calculó la tabla, seguido de la tabla de contingencia correspondiente. Esto se logra utilizando la función `cat` para imprimir mensajes sin saltos de línea y la función `print` para imprimir las tablas de contingencia.



```
1 # Calcular las tablas de contingencia para las variables relevantes
2 contingency_tables <- calculate_contingency_tables(relevant_data, relevant_variable_names)
3
4 # Mostrar las tablas de contingencia
5 for (variable_name in names(contingency_tables)) {
6   cat("Tabla de contingencia para", variable_name, ":\n")
7   print(contingency_tables[[variable_name]])
8   cat("\n")
9 }
```

Las tablas de contingencia pueden encontrarse en el siguiente archivo:
https://docs.google.com/spreadsheets/d/11IPMv6K0_dgMYKciUtKNGhFBsP66E8s6/edit?usp=s_haring&ouid=117199790807413443613&rtpof=true&sd=true

9. CONCLUSIONES

El análisis realizado sobre el conjunto de datos de pacientes con cáncer de mama ha proporcionado múltiples aprendizajes, que son valiosos tanto en el contexto académico de esta evidencia de aprendizaje como en el desarrollo de habilidades analíticas aplicadas.

Primero, el proceso de selección de una muestra representativa mediante el cálculo del tamaño de muestra y la validación de su representatividad demostró ser fundamental para asegurar la generalizabilidad de los resultados. Este paso es crucial en cualquier análisis estadístico y garantiza que las conclusiones obtenidas a partir de la muestra se puedan aplicar al conjunto completo de datos. Esta metodología robusta subraya la importancia de la rigurosidad en la selección de datos para obtener resultados fiables y válidos.

El análisis descriptivo de las variables, agrupadas por diagnóstico, permitió entender cómo se comportan diferentes características del tumor en función de su malignidad o benignidad. Este análisis no solo mejoró la comprensión de las diferencias fundamentales entre los tumores malignos y benignos, sino que también destacó la importancia de utilizar múltiples métodos estadísticos para obtener una visión completa y precisa de los datos. La realización de estadísticas descriptivas, junto con visualizaciones como histogramas, boxplots y gráficos de densidad, ilustró las distribuciones y diferencias entre los grupos, resaltando la relevancia de las visualizaciones en la interpretación de datos complejos.

El uso de la matriz de correlación para identificar relaciones entre variables fue un aprendizaje clave sobre cómo las características de los datos pueden interrelacionarse y afectar el diagnóstico. La identificación de variables con alta correlación con el diagnóstico es particularmente valiosa, ya que puede dirigir la atención hacia características cruciales que podrían mejorar la precisión de modelos predictivos.

La realización de pruebas t y el ajuste de valores p mediante el método de Bonferroni

enfaticaron la necesidad de un análisis estadístico riguroso para identificar variables significativas. Este proceso mostró la importancia de considerar la corrección por múltiples comparaciones para evitar conclusiones erróneas, lo que es fundamental en el análisis de grandes conjuntos de datos. Aprender a manejar y corregir los valores p adecuadamente es una habilidad esencial en la estadística.

Además, el análisis de las probabilidades de error tipo I y tipo II aportó una comprensión sobre las implicaciones de los errores estadísticos en el diagnóstico clínico. Identificar las variables con mayores probabilidades de errores tipo I y tipo II proporciona una perspectiva crítica sobre áreas donde el diagnóstico puede ser más susceptible a fallos.

Finalmente, la generación de tablas de contingencia para las variables relevantes proporcionó una visión clara y detallada de cómo las características del tumor se distribuyen entre diagnósticos malignos y benignos. Este ejercicio resaltó la utilidad de las tablas de contingencia en la comprensión de relaciones directas entre variables categóricas, una herramienta fundamental en el análisis estadístico.

En conjunto, esta evidencia de aprendizaje no solo ha ofrecido una profunda comprensión de las características del cáncer de mama y su diagnóstico, sino que también ha fortalecido habilidades clave en análisis de datos, estadística y visualización.

10.ANEXOS

Repositorio en GitHub que contiene los scripts de R que se utilizaron y las capturas de pantalla de los códigos.

https://github.com/ossmanmejia/MejiaOssman_EA1U3_Estadistica_II_Final