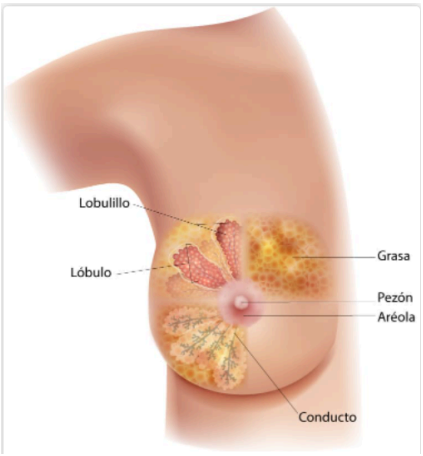
Fase I. Comprensión del problema

1.1. Cáncer de mama

El cáncer es una colección de enfermedades relacionadas que causa que algunas células del cuerpo comiencen a dividirse sin detenerse y se extiendan a tejidos cercanos, este puede generarse en casi cualquier parte del cuerpo. Específicamente se trata del resultado de mutaciones o cambios anormales en los genes que regulan el crecimiento de la célula. El ciclo normal sucede cuando una célula crece y se divide para formar otras nuevas según la necesidad del cuerpo, sin embargo cuando las células envejecen y resultan dañadas, estas mueren, y nuevas células las reemplazan, así las cosas cuando el cáncer se desarrolla, el proceso celular se descompone. Las mutaciones pueden "activar" ciertos genes y "desactivar" otras en la célula. La célula modificada adquiere la habilidad de dividirse sin ningún control u orden, lo que produce células idénticas y generan un tumor (Gallegos, Torres, Álvarez y Torres, 2017. Pág. 46) .

Uno de los casos más comunes de cáncer es el de mama y se origina en unas glándulas llamadas lóbulos o en los tubos delgados denominados ductos que cumplen con la función de transportar la leche hasta el pezón. Como se puede observar en la figura 1. "Las mamas constan de tres partes principales: lobulillos, conductos y tejido conectivo. Los lobulillos son las glándulas que producen leche. Los conductos son los tubos que transportan la leche al pezón. El tejido conectivo (formado por tejido fibroso y adiposo) rodea y sostiene todas las partes de la mama. La mayoría de los cánceres de mama comienzan en los conductos o en los lobulillos" (CDC, 2021).

Imagen 1. tejidos de las mamas



Fuente: Tomado de: CDC, 2021

Los tipos de cáncer de mama más comunes son:

- Carcinoma ductal infiltrante. Las células cancerosas se originan en los conductos y después salen de ellos y se multiplican en otros tejidos mamarios. Estas células cancerosas invasoras también pueden diseminarse, o formar metástasis, en otras partes del cuerpo (CDC, 2021).

- Carcinoma lobulillar infiltrante. Las células cancerosas se originan en los lobulillos y después se diseminan de los lobulillos a los tejidos mamarios cercanos. Estas células cancerosas invasoras también pueden diseminarse a otras partes del cuerpo (CDC, 2021.
- Enfermedad de paget: es un tipo de cáncer poco común que afecta la piel del pezón y, por lo general, el círculo más oscuro de piel de su derredor el cual se llama areola. Las células de Paget a menudo tienen una apariencia grande, redonda, al microscopio; se pueden encontrar como células aisladas o como grupos pequeños de células en el interior de la epidermis (Instituto Nacional del Cáncer, 2012).
- Cáncer de mama inflamatorio: El cáncer inflamatorio de seno es una enfermedad poco común y muy agresiva en la que las células cancerosas bloquean los vasos linfáticos en la piel del seno. Este tipo de cáncer de seno se llama "inflamatorio" porque muchas veces el seno se ve hinchado y enrojecido, o "inflamado". La mayoría de los cánceres inflamatorios de seno son carcinomas ductales invasivos, lo que significa que se formaron de células que revisten los conductos de leche del seno y luego se diseminaron más allá de los conductos (Instituto Nacional del Cáncer, 2012).

Síntomas asociados:

1. Un bulto nuevo en la mama o la axila (debajo del brazo)
2. Aumento del grosor o hinchazón de una parte de la mama.
3. Irritación o hundimientos en la piel de la mama.
4. Enrojecimiento o descamación en la zona del pezón o la mama.
5. Hundimiento del pezón o dolor en esa zona.
6. Secreción del pezón, que no sea leche, incluso de sangre.
7. Cualquier cambio en el tamaño o la forma de la mama.
8. Dolor en cualquier parte de la mama.
9. Algunas personas no presentan ningún síntoma

Factores de riesgo que no se pueden cambiar

1. Hacerse mayor. El riesgo de cáncer de mama aumenta con la edad. La mayoría de los cánceres de mama se diagnostican después de los 50 años de edad.
2. Mutaciones genéticas. Las mujeres que han heredado cambios (mutaciones) heredados en ciertos genes, tales como en el BRCA1 y el BRCA2, tienen mayor riesgo de presentar cáncer de mama y de ovario.
3. Historial reproductivo. El comienzo de la menstruación antes de los 12 años de edad y de la menopausia después de los 55 años de edad exponen a las mujeres a hormonas por más tiempo, lo cual aumenta el riesgo de cáncer de mama.
4. Tener mamas densas. Las mamas densas tienen más tejido conjuntivo que tejido adiposo, lo cual, a veces, puede hacer difícil la detección de tumores en una mamografía. Las mujeres con mamas densas tienen más probabilidades de tener cáncer de mama.
5. Antecedentes personales de cáncer de mama o ciertas enfermedades de las mamas que no son cancerosas. Las mujeres que han tenido cáncer de mama tienen mayores probabilidades de tener esta enfermedad por segunda vez. Algunas enfermedades de las mamas que no son cancerosas, como la hiperplasia atípica o el carcinoma lobulillar *in situ,* están asociadas a un mayor riesgo de tener cáncer de mama.

6. Antecedentes familiares de cáncer de mama o cáncer de ovario. El riesgo de una mujer de tener cáncer de mama es mayor si su madre, una hermana o una hija (parientes de primer grado) o varios integrantes de la familia por el lado paterno o materno han tenido cáncer de mama o cáncer de ovario. Tener un pariente de primer grado de sexo masculino con cáncer de mama también aumenta el riesgo para la mujer.
7. Tratamientos previos con radioterapia. Las mujeres que han recibido radioterapia en el pecho o las mamas antes de los 30 años de edad (por ejemplo, para el tratamiento del linfoma de Hodgkin) tienen un riesgo mayor de presentar cáncer de mama más adelante en la vida.
8. Exposición al medicamento dietilestilbestrol. Dietilestilbestrol se administró a algunas mujeres embarazadas en los Estados Unidos entre los años 1940 y 1971 para prevenir el aborto espontáneo. Las mujeres que tomaron dietilestilbestrol, o cuyas madres tomaron dietilestilbestrol cuando estaban embarazadas de ellas, tienen un mayor riesgo de tener cáncer de mama.

Factores de riesgo que si se pueden cambiar

1. No mantenerse físicamente activa. Las mujeres que no se mantienen físicamente activas tienen un mayor riesgo de tener cáncer de mama.
2. Tener sobrepeso u obesidad después de la menopausia. Las mujeres mayores que tienen sobrepeso u obesidad tienen mayor riesgo de tener cáncer de mama que las que tienen un peso normal.
3. Tomar hormonas. Algunas formas de terapia de reemplazo hormonal (aquellas que incluyen tanto estrógeno como progesterona) que se toman durante la menopausia pueden aumentar el riesgo de cáncer de mama si se toman por más de cinco años.
4. Ciertos anticonceptivos orales (píldoras anticonceptivas) aumentan el riesgo de cáncer de mama también.
5. Historial reproductivo. Quedar embarazada por primera vez después de los 30 años de edad, no amamantando y nunca tener un embarazo que llegue a término puede aumentar el riesgo de cáncer de mama.
6. Tomar alcohol. Algunos estudios muestran que el riesgo de la mujer de tener cáncer de mama aumenta cuanto mayor sea la cantidad de alcohol que tome.
7. Tabaquismo y exposición a sustancias químicas.

Tasas de supervivencia

de acuerdo con datos tomados de de las publicaciones *Cancer Facts & Figures 2022* y *Cancer Facts & Figures 2020* de la Sociedad Americana contra el cáncer, las tasas de supervivencia según la etapa de detección y estado de avance de la enfermedad son:

1. La tasa de sobrevivencia promedio a 5 años para las mujeres con cáncer de mama invasivo no metastásico es del 90%. La tasa de sobrevivencia promedio a 10 años para las mujeres con cáncer de mama invasivo no metastásico es del 84%.
2. Si el cáncer de mama invasivo se encuentra solo en la mama, la tasa de supervivencia a 5 años para mujeres con esta enfermedad es 99%.

3. Si el cáncer se ha diseminado hacia los ganglios linfáticos regionales, la tasa de supervivencia a 5 años es del 86%.
4. Si el cáncer se ha diseminado a una parte distante del cuerpo, la tasa de supervivencia a 5 años es del 29%. Las tasas de sobrevivencia son 10% más bajas en las mujeres negras en comparación con las mujeres blancas en los Estados Unidos.

1.2. Contexto social

De acuerdo con la Organización Panamericana de la Salud (OPS, 2020) el pronóstico después de un diagnóstico de cáncer de mama ha mejorado en los últimos 40 años, lo cual se evidencia en una reducción del 40% en la tasa de mortalidad en 2020 frente a 1980 gracias a los programas de detección temprana y protocolos de tratamiento estandarizados. Sin embargo, aún hay brechas en el tratamiento de la enfermedad, por ejemplo, la incidencia de esta patología es mucho mayor en los países en vías de desarrollo frente a los países del primer mundo.

> "Las Américas representarán casi una cuarta parte de los nuevos casos de cáncer de mama en 2020. En América Latina y el Caribe, la proporción de mujeres afectadas por la enfermedad antes de los 50 años (32%) es mucho mayor que en América del Norte (19%)" (OPS, 2021). Además, el mayor porcentaje de muertes por cáncer de mama (50%) ocurre en mujeres menores de 65 años en países de ALC en comparación con las mujeres que residen en América del norte (37%).

Tan solo en 2020 a nivel mundial se diagnosticaron 2.261.419 nuevos casos y se estima que de ellos alrededor de 684.996 mujeres murieron. De esta cantidad, aproximadamente 210.000 nuevos casos y 68.000 muertes se presentaron en América Latina, de las cuales 15.509 casos se diagnosticaron en Colombia y 4.411 fallecieron.

Al respecto la OMS indica que la detección temprana y el acceso a tratamiento efectivo siguen siendo un reto para países con recursos limitados, a pesar de que existen intervenciones probadas y rentables, en contextos donde hay suficientes recursos se recomienda el tamizaje de mamografía cada dos años para mujeres entre los 50 y los 69 años, por su parte, para entornos de pocos recursos donde los programas de detección temprana no arrojan resultados eficientes se recomienda el examen clínico, pues el tratamiento puede ser eficaz, especialmente cuando se detecta a tiempo. Por lo general, implica cirugía con o sin radiación y medicamentos. La efectividad del tratamiento depende de someterse al curso completo del tratamiento.

1.3. Objetivos

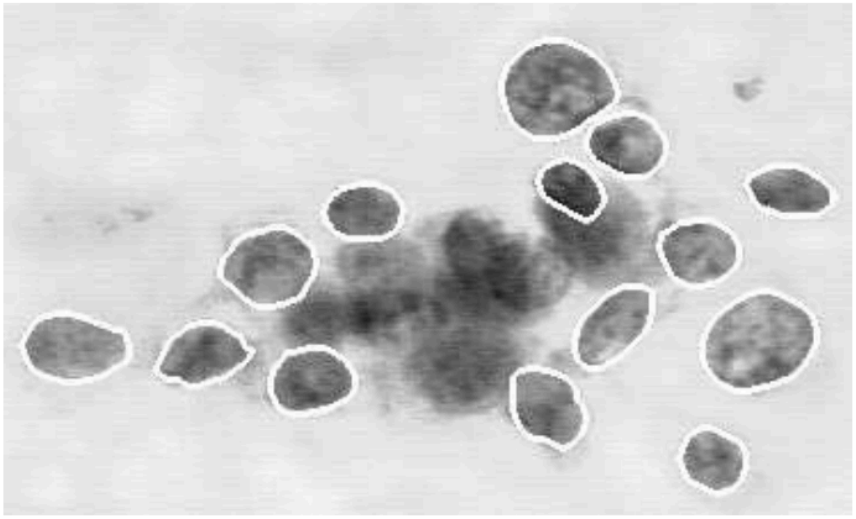
- Implementar modelos de machine learning para predecir el tipo de diagnóstico de cáncer de mama ante nuevos registros.
- Analizar a través de validación cruzada cuál de los modelos entrenados realiza una mejor clasificación del diagnóstico de cáncer de mama

Fase II. Estudio y comprensión de los datos

2.1. Diccionario variables

La información que se usará en este trabajo corresponde a la base de datos de diagnóstico de cáncer de mama de Wisconsin de la Universidad de California. Este conjunto de datos se compone de 32 variables que brindan información relevante de las características más importantes para analizar esta patología. El doctor Wolberg en los años 90 junto a dos de sus estudiantes identificó nueve características evaluadas visualmente mediante muestra de FNA[1] (masas mamarias basadas únicamente en una aspiración con aguja fina) que consideró relevantes para el diagnóstico del cáncer de mama. Posteriormente, con estas nueve características se diagnosticaron con éxito el 97 % de los casos nuevos de cáncer.

Imagen 2. Muestra FNA capturada con Microscopio y procesada con el Software Xcyt



Fuente: Millán y Robles, 2020. Pág 31.

La variable **diagnóstico** es el resultado de la evaluación de las características de la célula y solo tiene dos posibles diagnósticos Bueno (B) o Malo (M).

1. **Radio:** fue medido promediando la longitud de los segmentos de líneas radiales definidos por el centroide de la célula y los puntos individuales en el límite de la célula
2. **Textura:** fue medida encontrando la varianza en intensidades de escala de grises en los pixeles de la computadora

[1] "Se toma una FNA de la masa mamaria. Este material se monta en la platina de muestras del microscopio y se remite para resaltar los núcleos celulares. Una parte de la platina en la que las células están bien diferenciadas se escanea con una cámara digital y una placa de captura de fotogramas. El usuario luego aísla los núcleos individuales usando Xcyt . Usando un puntero del mouse, el usuario dibuja el límite aproximado de cada núcleo. Usando un enfoque de visión por computadora conocido como snakes, estas aproximaciones luego convergen a los límites nucleares exactos. Este proceso interactivo lleva entre dos y cinco minutos por muestra.Una vez que todos (o la mayoría) de los núcleos han sido aislados, el programa calcula los valores para cada una de las diez características de cada núcleo, midiendo el tamaño, la forma y la textura. Se calculan la media, el error estándar y los valores extremos de estas características, lo que da como resultado un total de 30 características nucleares para cada muestra" (Millán & Robles, 2020). Pág. 32).

3. **Perímetro:** Es definido como la distancia total entre puntos individuales llamados puntos serpientes. Estos puntos individuales comprenden las líneas blancas en el perímetro de las células
4. **Área:** se obtiene contando el número de píxeles en el interior de la línea blanca añadiendo la mitad de los píxeles en el perímetro.
5. **Suavidad del núcleo de la célula:** se calcula midiendo la diferencia entre la longitud de una línea radial y la longitud principal que la rodea. Básicamente, la suavidad es la variación local en las longitudes de radio.
6. **Compacidad:** para obtener está medida el perímetro y el área son combinados para calcular la compacidad; la cual es una medida de forma que sigue la siguiente fórmula; c= perímetro^2 /área.
7. **Concavidad:** analiza las irregularidades de forma en el núcleo de la célula.
8. **Puntos cóncavos:** usan una medida similar a la concavidad, pero ésta característica solo mide el número, más que la magnitud, de las concavidades del contorno
9. **Simetría:** se obtiene encontrando la línea más larga que pase por el centro. Entonces, se trazan líneas perpendiculares a dicha línea para medir la diferencia de longitudes en las dos direcciones de la línea central.
10. **Dimensión fractal:** es una característica de forma, es decir, a mayor valor corresponde a un menor contorno y por tanto a una mayor probabilidad malignidad

Nota: las definiciones fueron tomadas del artículo de Gallegos et.al

De cada una de estas variables se cuenta con datos de la media, el error estándar y los valores extremos de estas características.

2.2. Revisión de los datos

Tabla 1. Estructura de los datos

```
> (str(data)) ### Se puede ver que inicialmente contamos con un cojunto de 32 datos incluyendo el id
spec_tbl_df [569 x 32] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ id                    : num [1:569] 842302 842517 84300903 84348301 84358402 ...
 $ diagnosis             : chr [1:569] "M" "M" "M" "M" ...
 $ radius_mean           : num [1:569] 1799 2057 1969 1142 2029 ...
 $ texture_mean          : num [1:569] 1038 1777 2125 2038 1434 ...
 $ perimeter_mean        : num [1:569] 1228 1329 130 7758 1351 ...
 $ area_mean             : num [1:569] 1001 1326 1203 3861 1297 ...
 $ smoothness_mean       : chr [1:569] "0.1184" "0.08474" "0.1096" "0.1425" ...
 $ compactness_mean      : chr [1:569] "0.2776" "0.07864" "0.1599" "0.2839" ...
 $ concavity_mean        : chr [1:569] "0.3001" "0.0869" "0.1974" "0.2414" ...
 $ concave points_mean   : chr [1:569] "0.1471" "0.07017" "0.1279" "0.1052" ...
 $ symmetry_mean         : chr [1:569] "0.2419" "0.1812" "0.2069" "0.2597" ...
 $ fractal_dimension_mean: chr [1:569] "0.07871" "0.05667" "0.05999" "0.09744" ...
 $ radius_se             : chr [1:569] "1.095" "0.5435" "0.7456" "0.4956" ...
 $ texture_se            : chr [1:569] "0.9053" "0.7339" "0.7869" "1.156" ...
 $ perimeter_se          : chr [1:569] "8.589" "3.398" "4.585" "3.445" ...
 $ area_se               : num [1:569] 1534 7408 9403 2723 9444 ...
 $ smoothness_se         : chr [1:569] "0.006399" "0.005225" "0.00615" "0.00911" ...
 $ compactness_se        : chr [1:569] "0.04904" "0.01308" "0.04006" "0.07458" ...
 $ concavity_se          : chr [1:569] "0.05373" "0.0186" "0.03832" "0.05661" ...
 $ concave points_se     : chr [1:569] "0.01587" "0.0134" "0.02058" "0.01867" ...
 $ symmetry_se           : chr [1:569] "0.03003" "0.01389" "0.0225" "0.05963" ...
 $ fractal_dimension_se  : chr [1:569] "0.006193" "0.003532" "0.004571" "0.009208" ...
 $ radius_worst          : num [1:569] 2538 2499 2357 1491 2254 ...
 $ texture_worst         : num [1:569] 1733 2341 2553 265 1667 ...
 $ perimeter_worst       : num [1:569] 1846 1588 1525 9887 1522 ...
 $ area_worst            : num [1:569] 2019 1956 1709 5677 1575 ...
 $ smoothness_worst      : chr [1:569] "0.1622" "0.1238" "0.1444" "0.2098" ...
 $ compactness_worst     : chr [1:569] "0.6656" "0.1866" "0.4245" "0.8663" ...
 $ concavity_worst       : chr [1:569] "0.7119" "0.2416" "0.4504" "0.6869" ...
```

Fuente: elaboración propia

En el dataset original se cuenta con un total de 569 observaciones y 32 variables, dentro de las cuales se encuentra incluido el id, el cual no será necesario para el análisis. También se identifica que a pesar de que todo el conjunto de datos es numérico, 22 variables están almacenadas como

caracteres por lo que más adelante será necesario realizar una transformación del tipo de los datos. Al observar los primeros 6 registros del dataset, se puede ver que efectivamente se trata de datos numéricos.

Tabla 2. Visualización de los primeros 6 registros de cada variable

```
> head(data)
# A tibble: 6 x 32
      id diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean concavity_mean `concave point~` symmetry_mean fractal_dimensi~ radius_se
   <dbl> <chr>          <dbl>        <dbl>          <dbl>     <dbl>           <dbl> <chr>            <chr>           <chr>           <chr>          <chr>           <chr>
1 8.42e5 M               1799         1038           1228      1001 0.1184          0.2776           0.3001          0.1471          0.2419         0.07871         1.095
2 8.43e5 M               2057         1777           1329      1326 0.08474         0.07864          0.0869          0.07017         0.1812         0.05667         0.5435
3 8.43e7 M               1969         2125            130      1203 0.1096          0.1599           0.1974          0.1279          0.2069         0.05999         0.7456
4 8.43e7 M               1142         2038           7758      3861 0.1425          0.2839           0.2414          0.1052          0.2597         0.09744         0.4956
5 8.44e7 M               2029         1434           1351      1297 0.1003          0.1328           0.198           0.1043          0.1809         0.05883         0.7572
6 8.44e5 M               1245          157           8257      4771 0.1278          0.17             0.1578          0.08089         0.2087         0.07613         0.3345
# ... with 19 more variables: texture_se <chr>, perimeter_se <chr>, area_se <dbl>, smoothness_se <chr>, compactness_se <chr>, concavity_se <chr>, `concave points_se` <chr>,
#   symmetry_se <chr>, fractal_dimension_se <chr>, radius_worst <dbl>, texture_worst <dbl>, perimeter_worst <dbl>, area_worst <dbl>, smoothness_worst <chr>,
#   compactness_worst <chr>, concavity_worst <chr>, `concave points_worst` <chr>, symmetry_worst <chr>, fractal_dimension_worst <chr>
```

Fuente: elaboración propia

A continuación en la tabla 3, se puede apreciar que en el conjunto de datos no hay variables con datos faltantes.

Tabla 3. Recuento de NA´s por variable

```
sapply(data, function(x) sum(is.na(x)))
                     id               diagnosis             radius_mean            texture_mean          perimeter_mean               area_mean         smoothness_mean
                      0                       0                       0                       0                       0                       0                       0
        compactness_mean         concavity_mean    concave points_mean          symmetry_mean fractal_dimension_mean               radius_se              texture_se
                      0                       0                       0                       0                       0                       0                       0
           perimeter_se                 area_se           smoothness_se          compactness_se            concavity_se       concave points_se            symmetry_se
                      0                       0                       0                       0                       0                       0                       0
  fractal_dimension_se            radius_worst           texture_worst         perimeter_worst              area_worst       smoothness_worst      compactness_worst
                      0                       0                       0                       0                       0                       0                       0
        concavity_worst    concave points_worst          symmetry_worst fractal_dimension_worst
                      0                       0                       0                       0
```

Fuente: elaboración propia

2.2. Preparación de los datos

En esta fase se inicia omitiendo la variable id, dado a que no aporta nada al análisis exploratorio ni en la parte de modelado para clasificar y predecir el diagnóstico de cáncer de mama. Posteriormente se procede a cambiar el tipo de las variables que se encontraban como caracteres a numéricas y la variable diagnóstico a factor.

Tabla 4. Resumen del conjunto de datos.

```
> summary(data) ## Observamos algunas medidas de tendencia central de los datos una vez se han convertido a numéricos
 diagnosis  radius_mean      texture_mean   perimeter_mean     area_mean      smoothness_mean   compactness_mean  concavity_mean      concave points_mean symmetry_mean
 B:357     Min.   :  9.00   Min.   : 17.0   Min.   :  63.0   Min.   : 181.0   Min.   :0.05263   Min.   :0.01938   Min.   :0.00000   Min.   :0.00000   Min.   :0.1060
 M:212     1st Qu.:11.54   1st Qu.:15.1   1st Qu.:  75.2   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492   1st Qu.:0.02956   1st Qu.:0.02031   1st Qu.:0.1619
           Median :13.37   Median :18.8   Median :  86.2   Median : 551.1   Median :0.09587   Median :0.09263   Median :0.06154   Median :0.03350   Median :0.1792
           Mean   :14.13   Mean   :19.3   Mean   :  91.9   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434   Mean   :0.08880   Mean   :0.04892   Mean   :0.1812
           3rd Qu.:15.78   3rd Qu.:21.5   3rd Qu.: 104.1   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040   3rd Qu.:0.13070   3rd Qu.:0.07400   3rd Qu.:0.1957
           Max.   :28.11   Max.   :39.3   Max.   : 188.5   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540   Max.   :0.42680   Max.   :0.20120   Max.   :0.3040
 fractal_dimension_mean    radius_se       texture_se     perimeter_se        area_se       smoothness_se       compactness_se      concavity_se     concave points_se
 Min.   :0.04996        Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  14.0   Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000
 1st Qu.:0.05770        1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.:1524   1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638
 Median :0.06154        Median :0.3242   Median :1.1080   Median : 2.287   Median :2157   Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930
 Mean   :0.06280        Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   :2960   Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796
 3rd Qu.:0.06612        3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.:3503   3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710
 Max.   :0.09744        Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :9904   Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790
  symmetry_se       fractal_dimension_se  radius_worst   texture_worst  perimeter_worst   area_worst    smoothness_worst  compactness_worst concavity_worst  concave points_worst
 Min.   :0.007882   Min.   :0.0008948    Min.   :  14   Min.   :  18   Min.   :  78    Min.   : 248   Min.   :0.07117   Min.   :0.02729   Min.   :0.0000   Min.   :0.00000
 1st Qu.:0.015160   1st Qu.:0.0022480    1st Qu.:1247   1st Qu.:1949   1st Qu.:1159    1st Qu.:1724   1st Qu.:0.11660   1st Qu.:0.14720   1st Qu.:0.1145   1st Qu.:0.06493
 Median :0.018730   Median :0.0031870    Median :1485   Median :2462   Median :1846    Median :4709   Median :0.13130   Median :0.21190   Median :0.2267   Median :0.09993
 Mean   :0.020542   Mean   :0.0037949    Mean   :1757   Mean   :2330   Mean   :4586    Mean   :4415   Mean   :0.13237   Mean   :0.25427   Mean   :0.2722   Mean   :0.11461
 3rd Qu.:0.023480   3rd Qu.:0.0045580    3rd Qu.:1928   3rd Qu.:2926   3rd Qu.:8411    3rd Qu.:6384   3rd Qu.:0.14600   3rd Qu.:0.33910   3rd Qu.:0.3829   3rd Qu.:0.16140
 Max.   :0.078950   Max.   :0.0298400    Max.   :9981   Max.   :4954   Max.   :9971    Max.   :9936   Max.   :0.22260   Max.   :1.05800   Max.   :1.2520   Max.   :0.29100
 symmetry_worst   fractal_dimension_worst
 Min.   :0.1565   Min.   :0.05504
 1st Qu.:0.2504   1st Qu.:0.07146
 Median :0.2822   Median :0.08004
 Mean   :0.2901   Mean   :0.08395
 3rd Qu.:0.3179   3rd Qu.:0.09208
 Max.   :0.6638   Max.   :0.20750
```

Fuente: elaboración propia

En la tabla 4 se pueden observar medidas de tendencia central para cada una de las variables objeto de análisis. Se destaca que algunas variables ya se encuentran en una escala de 0 a 1, sin embargo hay otras que no, que deberán ser reescaladas para poder aplicar los algoritmos de k-Means y KNN.

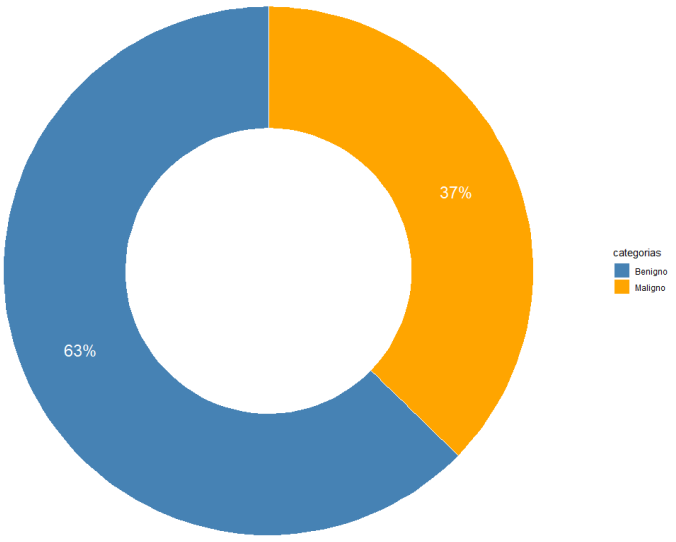
Fase III. Análisis exploratorio de los datos

En esta fase se emplean gráficos y tablas para analizar el comportamiento de los datos, siendo un paso muy importante para poder comprender aspectos como la distribución de los mismos, identificar si existen valores atípicos y los niveles de asociación que se presentan entre el conjunto de variables independientes para los modelos que se aplicarán en la fase IV.

En primer lugar se analiza la variable dependiente, en este caso el tipo de diagnóstico. De 569 registros de información presentes en la base de datos, en el gráfico 1 se puede apreciar que el 63% (357 observaciones) de los registros presentes en el dataset son diagnósticos benignos indicando la no presencia de células cancerígenas y el 37% (212 observaciones) restante son diagnósticos malignos, es decir que se identificó la presencia de células cancerígenas.

De acuerdo con otros estudios, en la vida real no se presenta este tipo de comportamiento en el que la tasa de diagnósticos positivos son superiores a los diagnósticos negativos, generalmente esta relación suele ser al contrario.

Gráfico 1. Tipo de cáncer



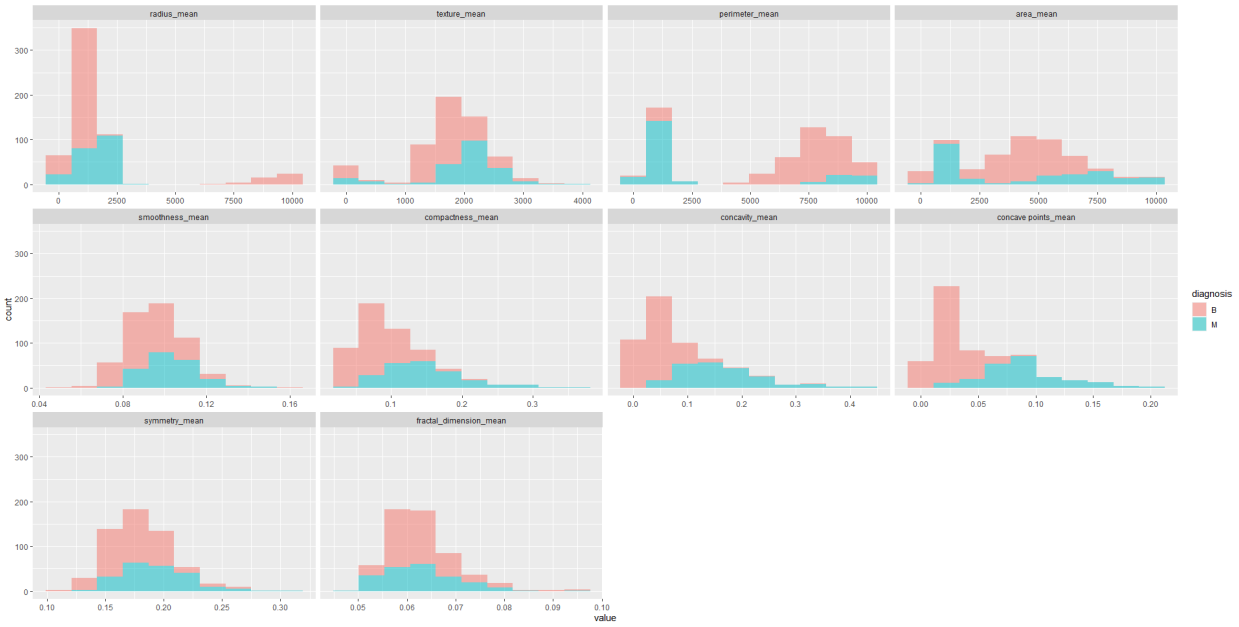
Fuente: elaboración propia

A continuación se procede analizar cada una de las variables predictoras cruzada con el tipo de diagnóstico, esto, con el objetivo de identificar si existen atributos que permitan diferenciar y/o predecir un diagnóstico bueno frente a uno malo.

En el gráfico 2, se presentan las distribuciones para la media de cada una de las 10 características que conforman el grupo de predictoras. Inicialmente no se puede establecer que existan diferencias muy

claras en la distribución de cada variable según el tipo de diagnóstico pues en todas las gráficas las distribuciones se superponen.

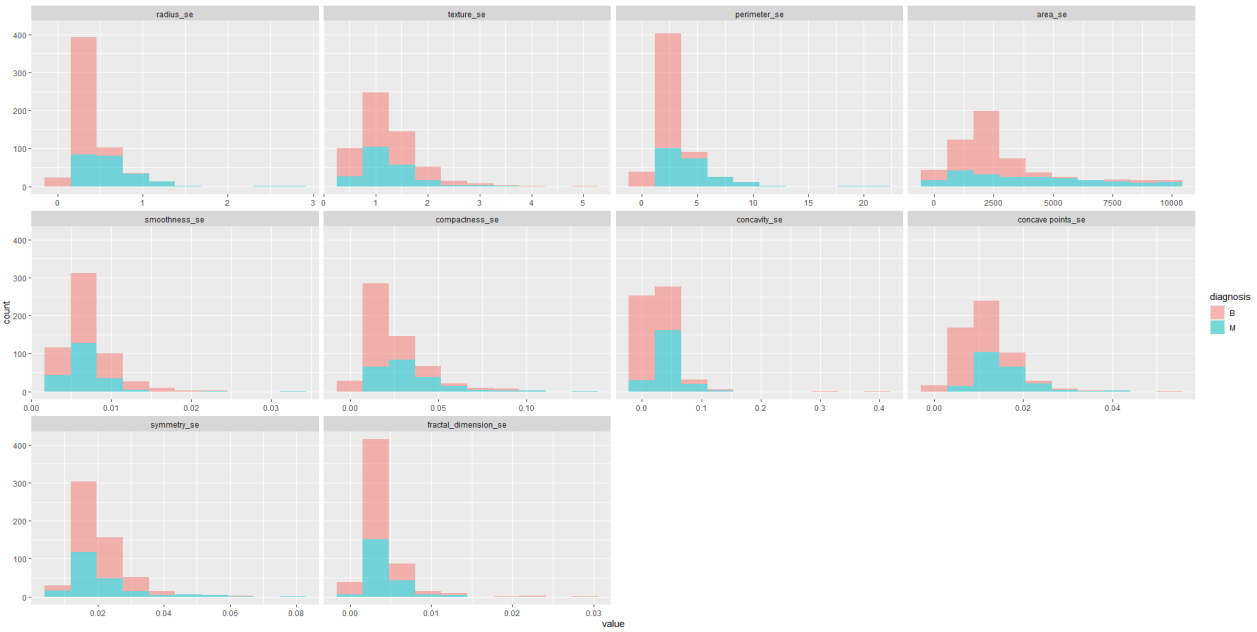
Gráfico 2. Distribución de la media de las 10 variables explicativas por tipo de diagnóstico



Fuente: elaboración propia

Por su parte, el gráfico 3 muestra las distribuciones de cada una de las distribuciones para el error estándar de cada una de las 10 características que conforman el grupo de predictoras

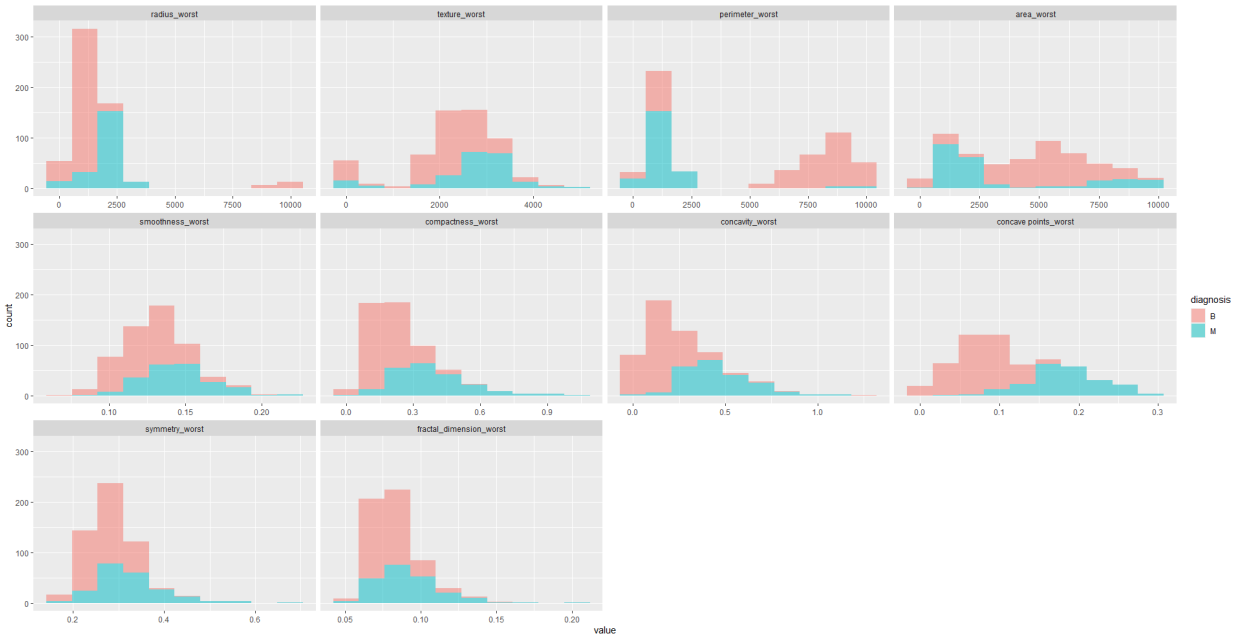
Gráfico 3. Distribución del error estándar de las 10 variables explicativas por tipo de diagnóstico



Fuente: elaboración propia

Finalmente, el gráfico 4 muestra las distribuciones de cada una de las distribuciones para los valores máximos de cada una de las 10 características que conforman el grupo de predictoras

Gráfico 4. Distribución de los valores extremos de las 10 variables explicativas por tipo de diagnóstico



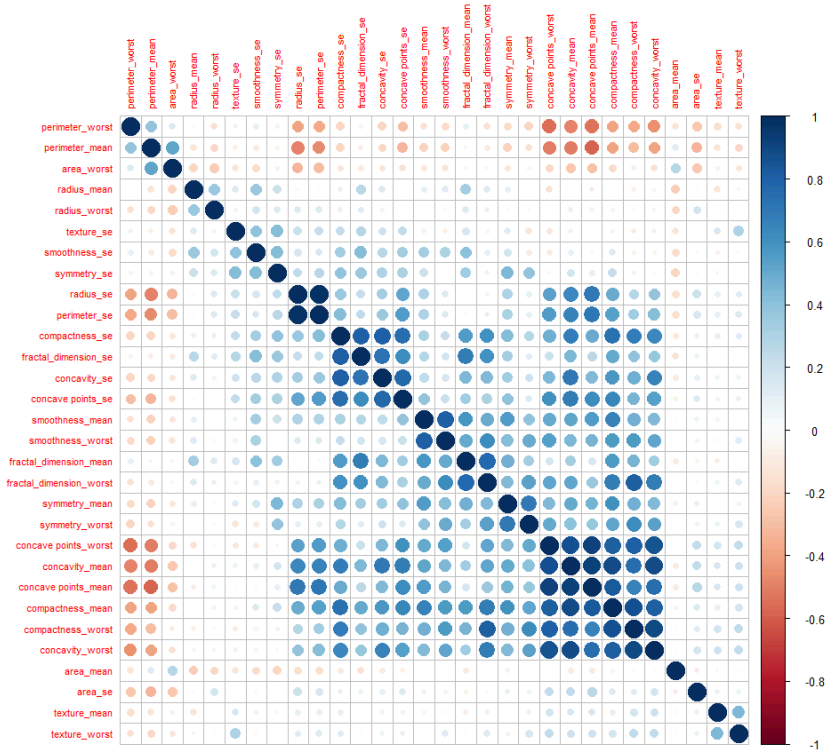
Fuente: elaboración propia

En términos generales no se identifica una separación perfecta entre ninguna de las características; sin embargo algunas variables presentan separaciones bastante buenas, como el caso de concave points_worst, concavity_worst, permiter_worst, y perimeter_mean. También se identifican variables que se distribuyen similar independiente del tipo de diagnóstico ya que su superposición es bastante estrecha como lo es el caso de symetry_se, textture_se, smothness_se y fractal dimension_se. Dicho lo anterior, se podría pensar que aquellas variables cuyas distribuciones difieren entre sí según el tipo de diagnóstico tendrán un mayor poder predictivo frente a aquellas cuya superposición es más estrecha.

A continuación se procede a realizar un análisis de correlaciones entre el conjunto de variables independientes con el objetivo de analizar cómo son las relaciones entre ellas, además, de ver si existen variables tan altamente correlacionadas que se podrían estar explicando entre ellas mismas al ser combinaciones lineales, lo que generaría distorsión en los resultados cuando se proceda a estimar los diferentes modelos.

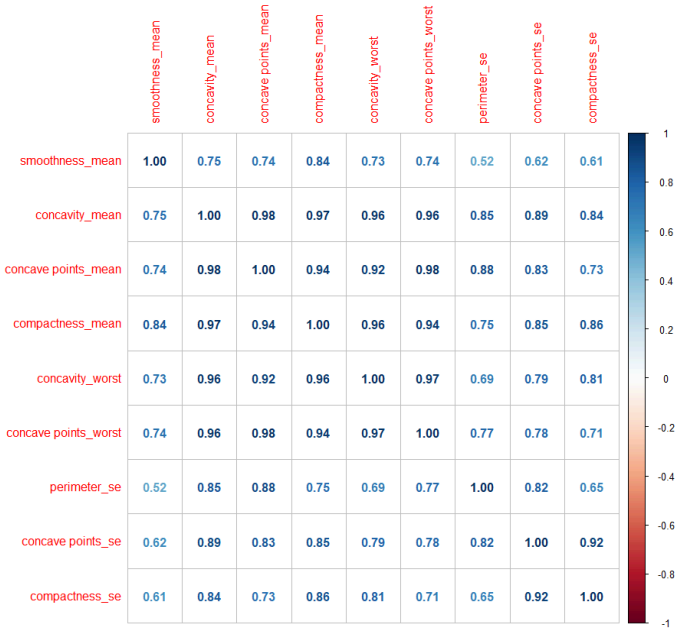
En el gráfico 5, la intensidad del color indica un mayor grado de asociación entre las variables. Entre más azul la correlación es más cercana a 1 y positiva, en contraste, entre más rojo la correlación también es más fuerte pero negativa. En línea con lo anterior, las correlaciones más grandes se identifican entre concavity_mean con concavity_worst y concavity poinst_worst, entre otras más. Para poder saber cuáles son exactamente las variables más correlacionadas el gráfico 6 presenta el valor de las correlaciones más altas.

Gráfico 5. Análisis de correlaciones entre las variables independientes



Fuente: elaboración propia

Gráfico 6. Variables que presentan las correlaciones más fuertes

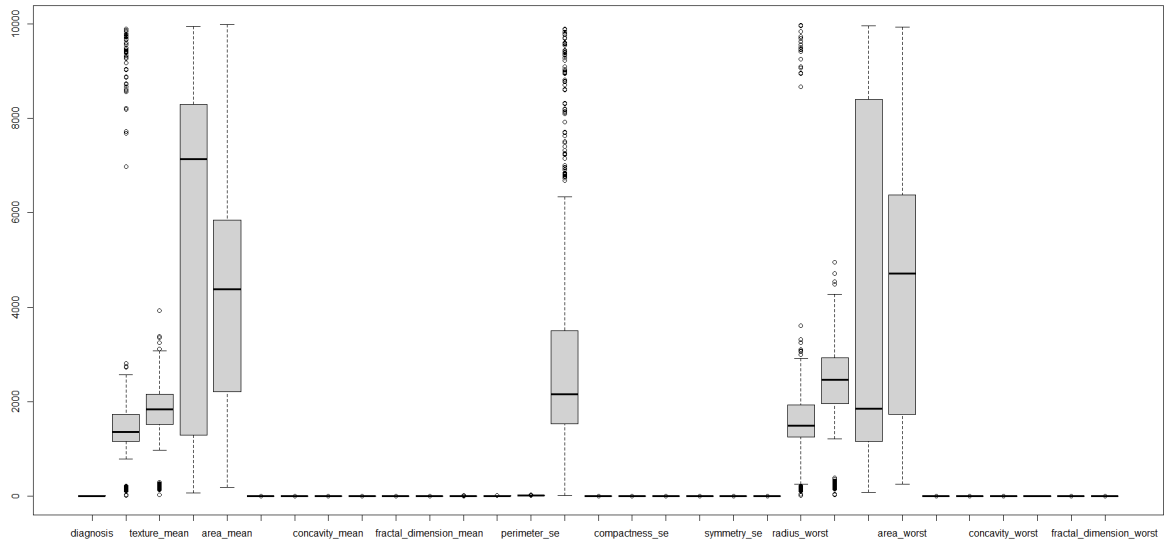


Fuente: elaboración propia

Tal como lo muestra el gráfico 6 las correlaciones más grandes se presentan entre las siguientes variables:

- Smoothness_mean
- Concavity_mean
- Concave points_mean
- Compactness_mean
- Concavity_worst
- Concave points_worst
- Perimeter_se
- Concave points_se
- Compactness_se

Análisis de Outliers



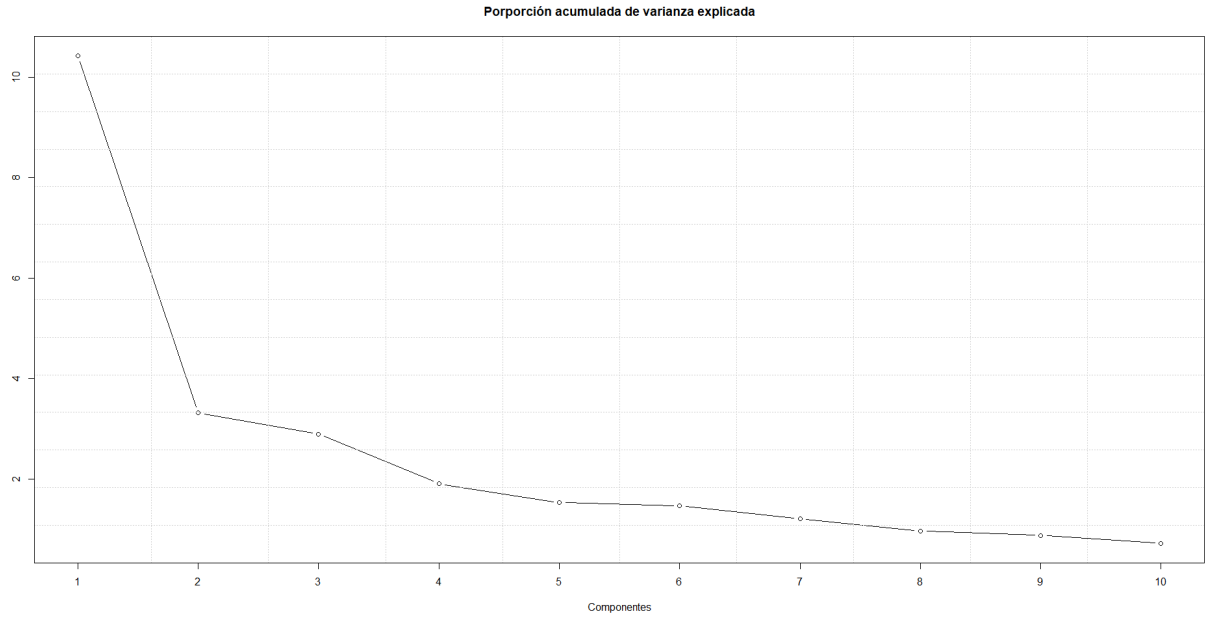
Se observa que algunas variables como radius_mean y area_se, presentan una gran cantidad de valores anómalos, los cuales deben ser tratados en la aplicación de algoritmos como la regresión logística.

Análisis de componentes principales

El objetivo de realizar el análisis de componentes principales es para identificar cuales son las variables más importantes al momento de aplicar los algoritmos de clasificación, pues en los análisis exploratorios no ha quedado muy claro cuáles podrían ser aquellas variables que tienen mayor poder predictivo de forma clara, pues como se pudo observar en los histogramas las distribuciones de cada variable según el tipo de diagnóstico no se separan completamente y en el análisis de correlaciones se identificó que existen niveles altos entre algunas variables explicativas.

Como resultado del análisis de componentes principales se va obtener como resultado un conjunto de nuevas variables llamadas componentes principales que son ortogonales, no colineales, ya que cada componente es una combinación lineal de las variables que lo conforman.

Gráfico 7. Varianza aculumada explicada por componentes principales



Fuente: elaboración propia

Aplicando el algoritmo de componentes principales se obtuvo como resultado un total de 30 componentes, de las cuales 13 de ellas explican el 90% de la varianza acumulada como se muestra en la tabla 1, destacando que la primer componente principal explica el 34% de la variabilidad total, la segunda componente principal el 11%, la tercera el 9% y la cuarta el 6%, que en conjunto solo estas cuatro estarían explicando más del 60% de la variabilidad total.

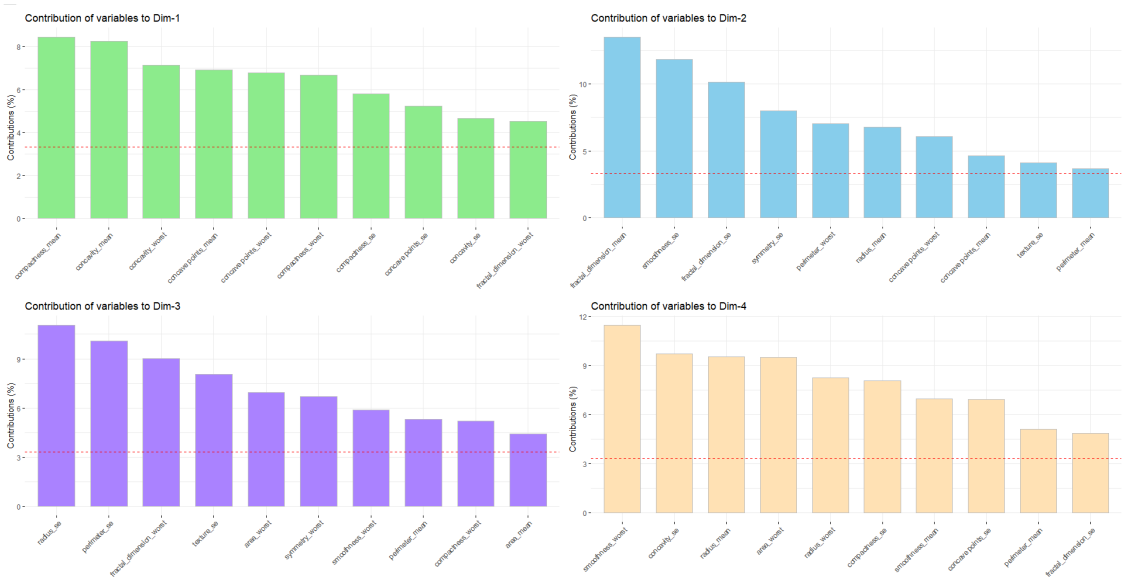
Tabla 1. Componentes principales que explican el 90% de la varianza acumulada.

Componentes	**Desviación estándar**	**proporción de varianza**	**proporción de varianza acumulada**
Componente 1	1,22	0,34	34%
Componente 2	1,82	0,11	45%
Componente 3	1,7	0,09	55%
Componente 4	1,38	0,06	61%
Componente 5	1,23	0,05	66%
Componente 6	1,21	0,048	71%
Componente 7	1,09	0,0443	75%
Componente 8	0,98	0,032	79%
Componente 9	0,93	0,029	81%
Componente 10	0,84	0,024	84%
Componente 11	0,78	0,0205	86%
Componente 12	0,77	0,020005	88%
Componente 13	0,73	0,0181	90%

Fuente: elaboración propia

Ahora se procede a observar cuáles variables conforman las primeras 4 componentes principales que estarían explicando más del 60% de la variabilidad total

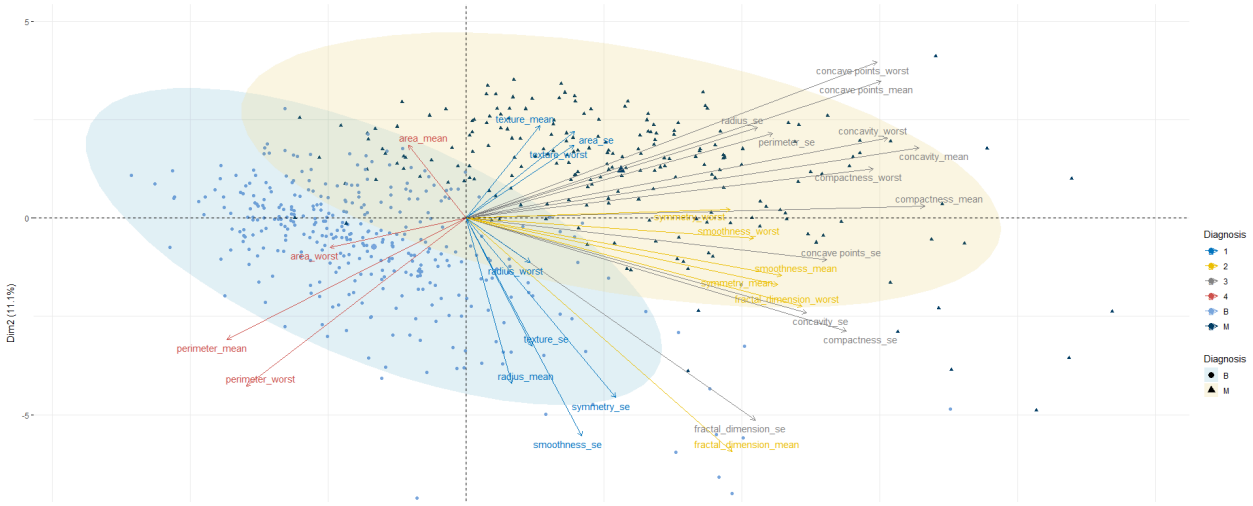
Gráfico 8. Variables que componen las primeras 4 componentes principales



La componente principal 1, al ser la que mayor variabilidad aporta es clave para comprender cuáles son las variables que la componen, pues es una señal de cuáles dentro de las 30 que se tienen son las que podrían aportar más al momento de clasificar. De acuerdo con los resultados obtenidos, estas variables serían:

- Compactness_mean
- Concavity_mean
- Concavity Worst
- Concave_points_worst
- Concave_points_mean
- Compactness_worst
- Compactness_se
- Concavity_points_se
- Concavity_se
- Fractal_dimension_worst

Gráfico 9. Variables que conforman las primeras 4 componentes principales, segmentadas según tipo de diagnóstico



Fuente: elaboración propia

Fase IV. Modelado - aplicación de las técnicas de machine learning

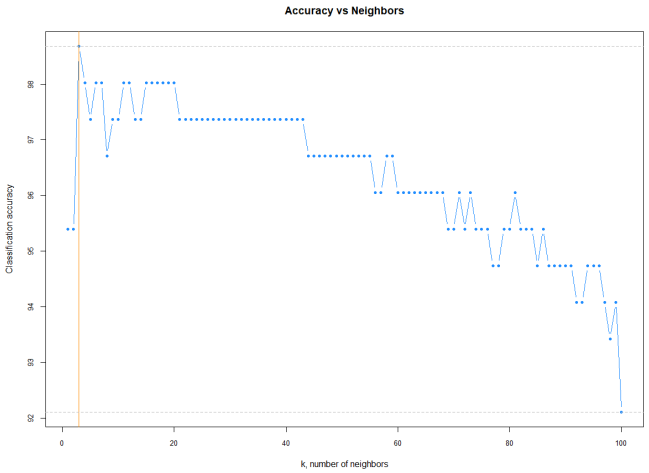
KNN

Se comienza ejecutando el algoritmo de K vecinos más cercanos, el cual pertenece a la familia de métodos supervisados y está basado en distancias. Este algoritmo puede predecir o clasificar, en este caso como la variable objetivo de análisis es cualitativa el ejercicio se enfoca en la clasificación. A continuación se describen los pasos a seguir en la estimación del modelo:

1) El primer paso consta de la normalización de todas la variables que serán introducidas en el modelos
2) Partimos el dataset en el conjunto de entrenamiento y de prueba, en este caso se designó el 70% para entrenar y 30% para testear.
3) Mediante un ciclo analizamos diferentes valores de K entre 1 y 100, al mismo tiempo que se van estimando los modelos de clasificación para obtener la medida de exactitud y poder obtener cuáles serían los valores óptimos de K.

Resultados:

Gráfico 9. Comapración de los valores de K frente al porcentaje de exactitud.



a pesar de que es evidente que el mejor es 3 se escoge k = 2 porque salgado es gay

el sobre ajuste conlleva a problemas futuros así que k =2

matriz de confusión (hacerla bonita)

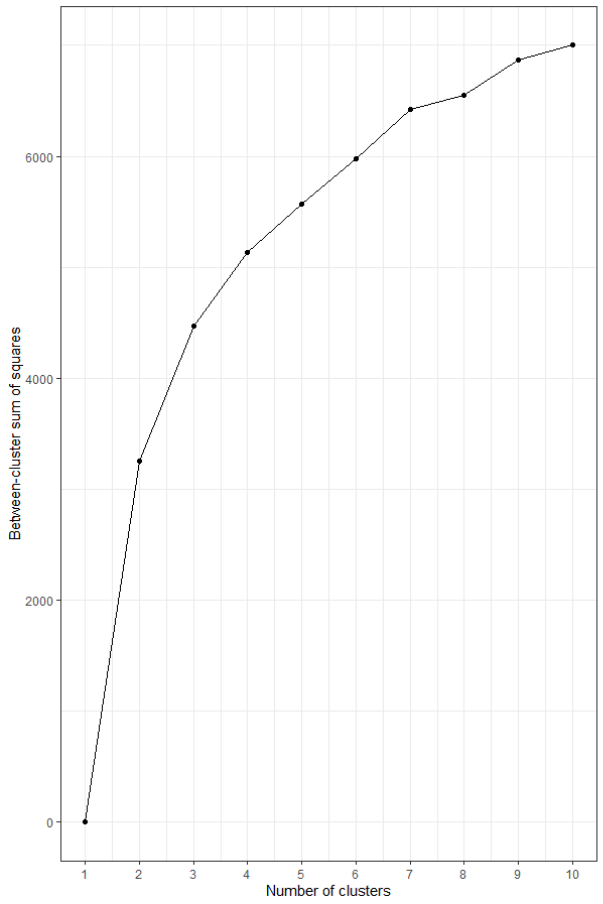
	Benigno	**Maligno**
Benigno	**90**	**4**
Maligno	**1**	**57**

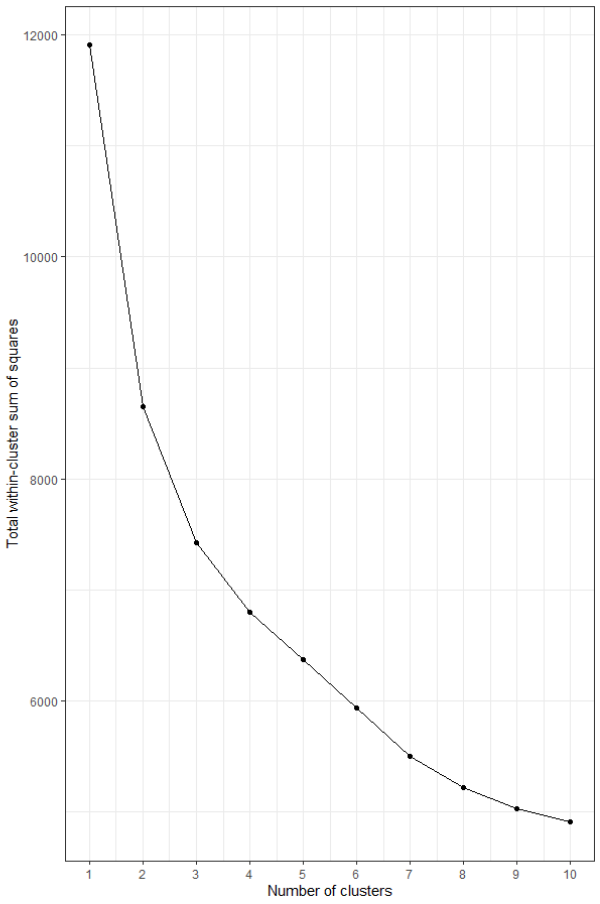
accuracy

- 96.71053

Kmeans

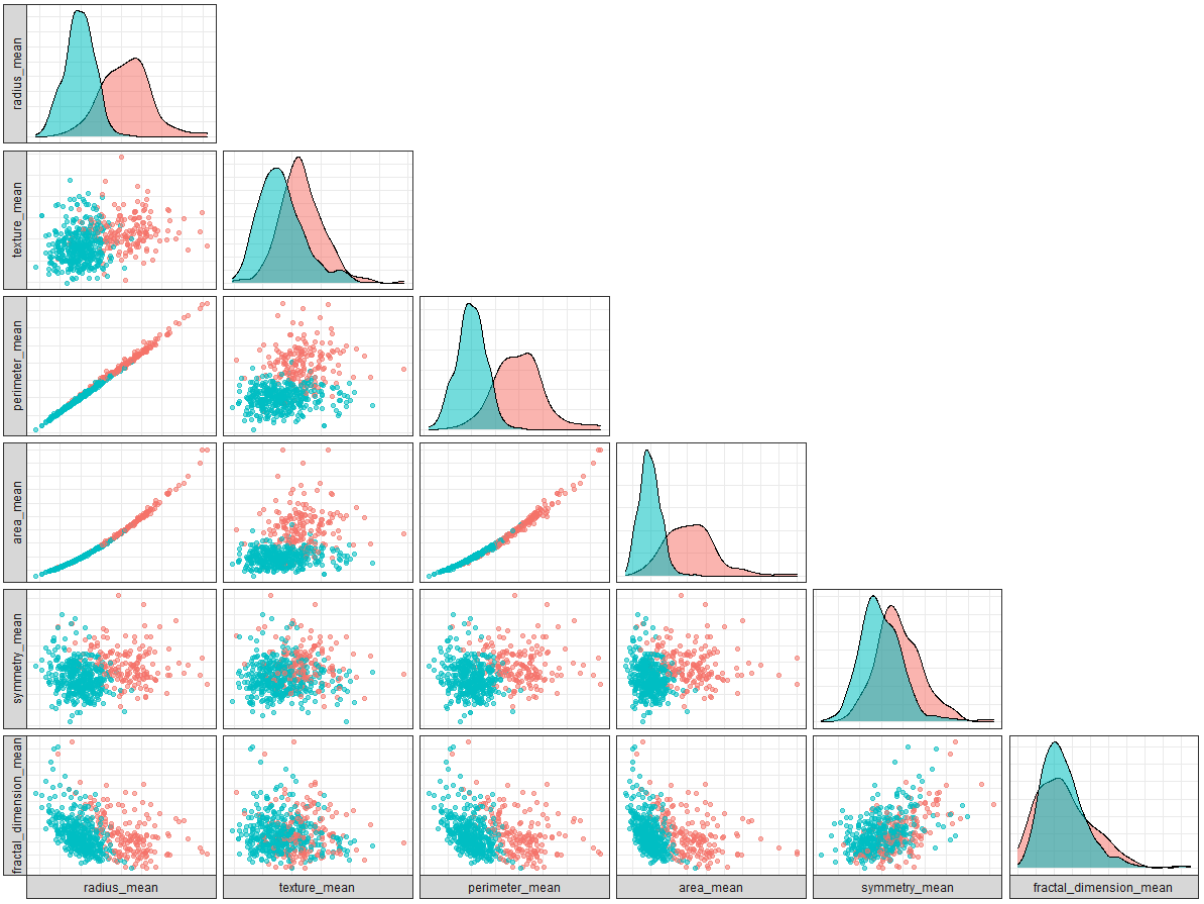
Diagrama de codo para ver el numero de k





lo mismo que en el modelo anterior k = 2

clusters para los 6 primeros(muchos se ve feo)



MATRIZ DE CONFUSION

```
a      1   2
  1 174  38
  2   1 355
```

naives bayes

probabilidades a priori (investigar comojueque con eso)

```
A-priori probabilities:
Y
        B         M
0.6298077 0.3701923
```

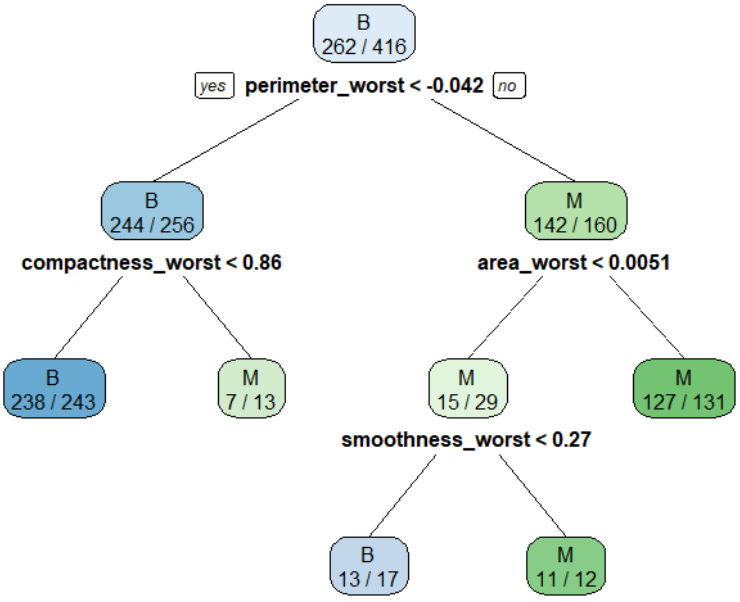
probabilidades condicionales (no se cuán necesario sea)

```
Conditional probabilities:
   radius_mean
Y         [,1]      [,2]
  B -0.5790864 0.4903765
  M  0.9494934 0.9340400

   texture_mean
Y         [,1]      [,2]
  B -0.3087818 0.9247307
  M  0.4574092 0.8970040

   perimeter_mean
Y         [,1]     [,2]
  B -0.5885375 0.472861
  M  0.9623197 0.927534

   area_mean
Y         [,1]      [,2]
  B -0.5594420 0.3695857
  M  0.9238392 1.0857930

   symmetry_mean
Y         [,1]     [,2]
  B -0.2320749 0.942190
  M  0.4191512 1.053558

   fractal_dimension_mean
Y          [,1]      [,2]
  B  0.01034111 0.9383048
  M -0.02580529 1.1016681

   radius_se
Y         [,1]     [,2]
  B -0.4173336 0.421403
  M  0.7511092 1.317940

   texture_se
Y          [,1]      [,2]
  B  0.053417214 1.0696231
  M -0.007900471 0.9199392

   area_se
Y         [,1]      [,2]
  B -0.4143578 0.2028052
  M  0.7258873 1.4710618

   smoothness_se
Y          [,1]      [,2]
  B  0.02643521 0.9700837
  M -0.11994663 0.8508268

   concavity_se
Y         [,1]      [,2]
  B -0.1860962 1.1870727
  M  0.2730367 0.6395022

   symmetry_se
Y         [,1]      [,2]
  B 0.03389420 0.8757353
  M 0.03852649 1.2896832

   fractal_dimension_se
Y          [,1]      [,2]
  B -0.01588152 1.2048973
  M  0.05062868 0.7375195

   radius_worst
Y         [,1]      [,2]
  B -0.6131401 0.3986987
  M  0.9923472 0.8921138

   texture_worst
Y         [,1]      [,2]
  B -0.3411288 0.8814737
  M  0.5087494 0.8845601

   perimeter_worst
Y         [,1]      [,2]
  B -0.6176122 0.3932508
  M  0.9898735 0.8847174

   area_worst
Y         [,1]      [,2]
  B -0.5765759 0.2788249
  M  0.9349537 1.0602818

   smoothness_worst
Y         [,1]      [,2]
  B -0.3770624 0.8857974
  M  0.4984732 0.9390252

   compactness_worst
Y         [,1]      [,2]
  B -0.4635554 0.5878595
  M  0.6633984 0.9712693

   symmetry_worst
Y         [,1]      [,2]
  B -0.3298772 0.6818157
  M  0.5484136 1.2476451

   fractal_dimension_worst
Y         [,1]      [,2]
  B -0.2560732 0.7414603
  M  0.3383998 1.1069903
```

matriz de confusión y estadisticos

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 92  3
         M  2 55

               Accuracy : 0.9671
                 95% CI : (0.9249, 0.9892)
    No Information Rate : 0.6184
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9301

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9787
            Specificity : 0.9483
         Pos Pred Value : 0.9684
         Neg Pred Value : 0.9649
             Prevalence : 0.6184
         Detection Rate : 0.6053
   Detection Prevalence : 0.6250
      Balanced Accuracy : 0.9635

       'Positive' Class : B
```

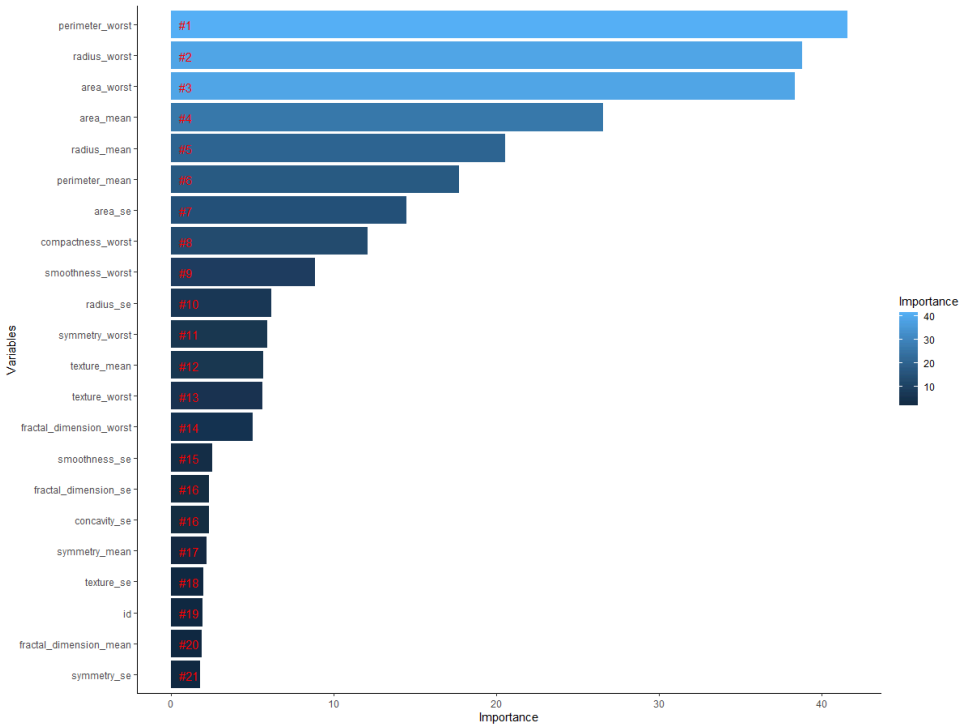
Arbol de desicion



```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 88  2
         M  6 56

               Accuracy : 0.9474
                 95% CI : (0.8989, 0.977)
    No Information Rate : 0.6184
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8899

 Mcnemar's Test P-Value : 0.2888

            Sensitivity : 0.9362
            Specificity : 0.9655
         Pos Pred Value : 0.9778
         Neg Pred Value : 0.9032
             Prevalence : 0.6184
         Detection Rate : 0.5789
   Detection Prevalence : 0.5921
      Balanced Accuracy : 0.9508

       'Positive' Class : B
```

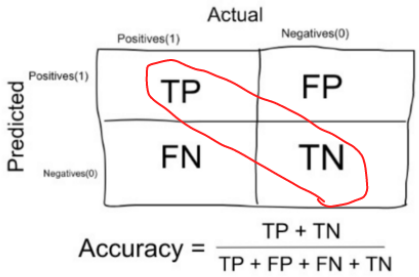
Random Forest

Se observa cuál o cuales son las variables más importantes para la clasificación de los diagnosticos como benignos o malignos



```
                Number of trees: 150
No. of variables tried at each split: 4

        OOB estimate of  error rate: 3.7%
Confusion matrix:
    B   M class.error
B 349   7  0.01966292
M  14 198  0.06603774
```

toca sacar el accuracy con fórmula



Regresión logística, falta porque salgado es un vago

```
Confusion Matrix and Statistics

                test_y
conv_13_logit_t  0  1
              0 96  1
              1  1 72

               Accuracy : 0.9882
                 95% CI : (0.9581, 0.9986)
    No Information Rate : 0.5706
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.976

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9863
            Specificity : 0.9897
         Pos Pred Value : 0.9863
         Neg Pred Value : 0.9897
             Prevalence : 0.4294
         Detection Rate : 0.4235
   Detection Prevalence : 0.4294
      Balanced Accuracy : 0.9880

       'Positive' Class : 1
```

Conclusiones

luego de hacer todo y sacar métricas, el mejor algoritmo es el _____ ya que blablabla,

de igual forma para un trabajo posterior se recomienda hacerlo con Redes neuronales y otros algoritmos buscando una mejora blablabla

5.0

Referencias:

Centro para el control y la prevención de enfermedades (USA). (2021). Cáncer de mama. Consultado el 9/05/2022. Disponible en: https://www.cdc.gov/spanish/cancer/breast/index.htm

Gallegos. A, torres. D, Álvarez. F & Torres. A. (2017). Identificación de características de células de cáncer de mama por medio de testores típicos. Universidad Autónoma de Aguascalientes, Departamento de Ciencias de la Computación, Aguascalientes, Aguascalientes, México. Recuperado

de: https://rcs.cic.ipn.mx/2017_140/Identificacion%20de%20caracteristicas%20de%20celulas%20de%20cancer%20de%20mama%20por%20medio%20de%20testores%20tipicos.pdf

Instituto Nacional del Cáncer (USA). (2012). Enfermedad de Paget de seno**.** Consultado el 9/05/2022. Disponible en: https://www.cancer.gov/espanol/tipos/seno/hoja-informativa-paget-seno#191qu233-es-la-enfermedad-de-paget-de-seno

Millán. J, Robles. B. (2020). Modelo en machine learning para el diagnóstico de cáncer de mama. (tesis de posgrado). Universidad Distrital Francisco José de Caldas. Bogotá, Colombia. Disponible en: https://repository.udistrital.edu.co/handle/11349/25070

Organización Panamericana de la Salud. (2020). Cáncer de mama. Consultado el el 9/05/2022. Disponible en: https://www.paho.org/es/temas/cancer-mama

Sociedad Americana contra el Cáncer. (2020 & 2022). *Cancer Facts & Figures 2022* y *Cancer Facts & Figures 2020.* Disponible en: https://www.cancer.net/es/tipos-de-c%C3%A1ncer/c%C3%A1ncer-de-mama/estad%C3%ADsticas