Universidad del Rosario 26 de febrero de 2022

Características de las células del cáncer de mama

Profesor: *Santiago Alférez* Autor: *Felipe Florián, Luisa Salcedo, Camilo Martínez*

1. Resumen Ejecutivo

Las células cancerígenas del cáncer de seno poseen ciertas características que permiten clasificar un tumor como maligno o benigno, esto lo determinan ya sean medidas geométricas o particularidades en su contorno. Por ello es de principal importancia estudiar los cambios que tienen las células cancerígenas, con el fin de clasificar un tumor como maligno o benigno. Para esto es posible analizar factores como el área, radio, textura, concavidad o simetría del tumor con el fin de determinar una clasificación. Vale tener en cuenta que los tumores malignos suelen tener medidas más grandes, es decir las medias del radio, área, perímetro, concavidad o demás, toman valores que sobresalen, por lo que se puede catalogar como un tumor maligno.

Por otro lado, la clasificación de estos tumores solo se puede realizar en dos categorías maligno o benigno y son determinadas por factores como: Los puntos cóncavos, perímetro, área, dimensión fractal, o el radio. Además, factores como: Textura o suavidad, no son de mayor relevancia para la clasificación del tumor. Finalmente, los métodos de clustering, clasificación y discriminación, confirman la clasificación de los tumores, la cual es única.

2. Introducción y descripción del problema

El cáncer de mama el más común entre las mujeres se origina cuando las células mamarias comienzan a crecer sin control, este tipo de cáncer puede crecer en distintas partes del seno, además, puede propagarse fuera del seno por medio de los vasos sanguíneos. Aunque, la *Sociedad Americana del Cáncer* afirma que la mayoría de bultos en los senos son benignos (los cuales no representan un peligro para la vida) eso no le quita la seriedad o el cuidado que se debe tener al momento de encontrar alguna anomalía de este estilo.

La organización mundial de la salud (OMS), para el 2004 informó que fallecieron alrededor de 519.000 mujeres por cáncer de mama, donde el 69 % de las defunciones ocurrieron en países en vía de desarrollo. Según Globocan en Colombia se estiman 13.380 nuevos casos y 3.702 muertes por año de esta enfermedad.

Por ello en este proyecto se realizó un análisis estadístico de datos acerca de las características entre tumores benignos y malignos del cáncer de seno. Centrándonos en datos acerca de la geometría y el contorno del tumor (consultados en *Kaggle*), con los cuales se utilizaron herramientas de análisis estadístico multivariado, como: análisis de componentes principales, análisis factorial, clustering, entre otros. Con el fin de determinar las características principales de las células del tumor cancerígeno, de los factores mencionados anteriormente (geometría,contorno) ya sean benignos o malignos.

3. Análisis Descriptivo inicial

Para comenzar haremos un análisis breve del dataset trabajado, para esto emplearemos unas gráficas que nos permitirán ver de manera más adecuada algunos datos relevantes con respecto a la comparación entre los tipos de tumor.

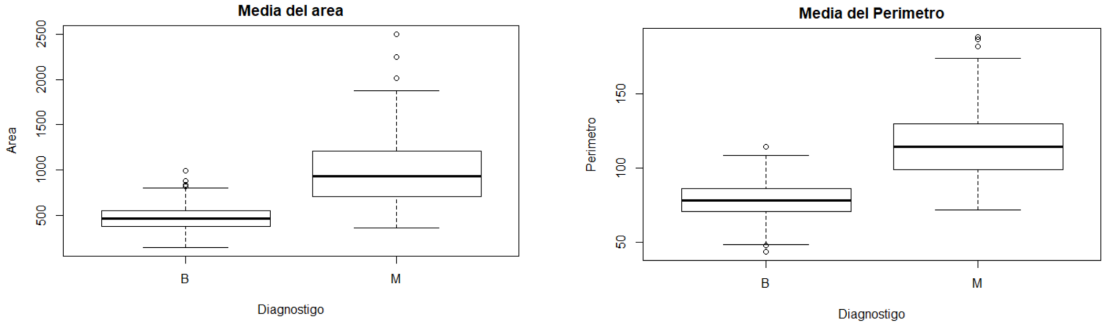

Figura 0.1: Comparación de datos

En la figura 0.1 podemos ver que los tumores benignos tienen rasgos más sutiles que los malignos, cualidad que se ve reflejada en todos los datos que hemos analizado. Lo cual, nos muestra que hay diferencias claras entre algunos de los tumores, sin embargo se ven intervalos en los que se tienen rasgos en común, siendo esto así, con los análisis anteriormente planteados se podrán identificar más eficientemente las diferencias entre ellos y la categorización más adecuada.

4. Análisis estadístico de datos

Análisis Factorial

Para poder analizar los datos con mayor facilidad aplicamos análisis factorial (FA), para esto utilizamos la función `factanal` en R.

Primero buscamos agrupar los datos en 2 factores (uno que se considerará cómo las variables geométricas y el otro las de contorno) sin aplicar una rotación, con lo que obtuvimos la **Unicidad ($\hat{\Psi}$)** (es la proporción de variabilidad, una unicidad alta para una variable, indica que la variable es muy variable), los **Loadings** $\hat{\Lambda}$ con los que producimos la **Comunalidad** (La comunalidad es la proporción de variabilidad).

Unicidad $\hat{\Psi}$: son datos cuyos valores pertenecen al intervalo $[0, 1]$, que muestra la proporción de variabilidad de una variable, cómo bien se mencionó anteriormente, una alta singularidad de una variable significa que no encaja perfectamente en nuestros factores lo que probablemente signifique que toque aumentar el número de factores a trabajar. En los resultados obtenidos ilustrados en la figura 0.2 pudimos notar que las variables **symmetry_mean, symmetry_se, symmetry_worst, texture_worst, texture_mean, texture_se** y **smoothness_se, smoothness_worst, smoothness_mean** no encajan en estos factores, lo cual nos da un indicio de que deberíamos aumentar el número de factores a trabajar.

```
Uniquenesses:
            radius_mean            texture_mean          concavity_se       concave.points_se
                  0.100                   0.874                 0.528                   0.528
          perimeter_mean              area_mean           symmetry_se    fractal_dimension_se
                  0.100                   0.100                 0.879                   0.515
         smoothness_mean         compactness_mean          radius_worst           texture_worst
                  0.558                   0.100                 0.100                   0.874
          concavity_mean      concave.points_mean        perimeter_worst              area_worst
                  0.100                   0.100                 0.100                   0.100
           symmetry_mean  fractal_dimension_mean        smoothness_worst       compactness_worst
                  0.610                   0.201                 0.614                   0.238
             radius_se              texture_se          concavity_worst    concave.points_worst
                  0.440                   0.977                 0.211                   0.139
          perimeter_se                area_se         symmetry_worst fractal_dimension_worst
                  0.433                   0.378                 0.700                   0.288
         smoothness_se          compactness_se
                  0.855                   0.371
```

Figura 0.2: `Unicidad`

Loadings $\hat{\Lambda}$: son las correlaciones entre las variables y los factores, ahora, cómo podemos observar en la figura 0.3, las variables **radius_mean, radius_se, radius_worst, perimeter_mean, perimeter_se, perimeter_worst, area_mean, area_se, area_worst, concavity_mean, concave.points_mean** y **concave.points_worst** tienen una correlación bastante alta con el factor uno, lo que significa que se ajustan bien a ese factor, sin embargo, podemos observar que algunas variables no tienen mayor correlación con ninguna de los 2 factores, por ejemplo **texture_se** tiene una correlación nula con el factor 1, y una correlación de 0.131 con el factor 2. razón por la cual, nuevamente podríamos concluir que requerimos un 3 factor para agrupar de manera más satisfactoria nuestro conjunto de datos.

```
Loadings:
                        Factor1 Factor2
radius_mean              0.980
texture_mean             0.347
perimeter_mean           0.982
area_mean                0.980
smoothness_mean          0.215   0.629
compactness_mean         0.545   0.788
concavity_mean           0.718   0.629
concave.points_mean      0.853   0.442
symmetry_mean            0.189   0.595
fractal_dimension_mean  -0.262   0.855
radius_se                0.737   0.130
texture_se                       0.131
perimeter_se             0.731   0.182
area_se                  0.787
smoothness_se           -0.197   0.326
compactness_se           0.230   0.759
concavity_se             0.221   0.651
concave.points_se        0.399   0.560
symmetry_se                      0.336
fractal_dimension_se             0.696
radius_worst             0.982
texture_worst            0.333   0.121
perimeter_worst          0.981
area_worst               0.969
smoothness_worst         0.182   0.594
compactness_worst        0.456   0.744
concavity_worst          0.572   0.680
concave.points_worst     0.782   0.500
symmetry_worst           0.204   0.508
fractal_dimension_worst          0.842
```

Figura 0.3: `Loadings`

Comunalidad $\hat{\lambda}^2$: la comunalidad es la proporción de variabilidad, son los loadings $\hat{\Lambda}$ al cuadrado. con ellos

podemos observar de manera más clara si las variables se adaptan a alguno de los factures, por ejemplo, si observamos la variable **texture_se** en la figura 0.4 podemos ver que es muy baja, por lo que no encaja bien en ningún factor.

```
      radius_mean      texture_mean    perimeter_mean         area_mean   smoothness_mean
        0.96183817        0.12656231        0.96404178        0.96137905        0.44167421
  compactness_mean    concavity_mean concave.points_mean    symmetry_mean fractal_dimension_mean
        0.91807209        0.91072880        0.92225557        0.38993293        0.79881339
         radius_se       texture_se       perimeter_se            area_se      smoothness_se
        0.55989466        0.02345172        0.56687493        0.62226320        0.14486439
   compactness_se     concavity_se  concave.points_se     symmetry_se  fractal_dimension_se
        0.62908355        0.47233615        0.47241205        0.12141043        0.48464417
     radius_worst      texture_worst     perimeter_worst        area_worst     smoothness_worst
        0.96418932        0.12576535        0.96744067        0.93862965        0.38618346
compactness_worst concavity_worst concave.points_worst    symmetry_worst fractal_dimension_worst
        0.76227262        0.78888565        0.86137030        0.29944895        0.71174264
```

Figura 0.4: `Comunalidad`

Luego de esto decidimos volver analizar los datos nuevamente con dos factores pero aplicando una rotación `varimax` en donde obtuvimos resultados muy similares.

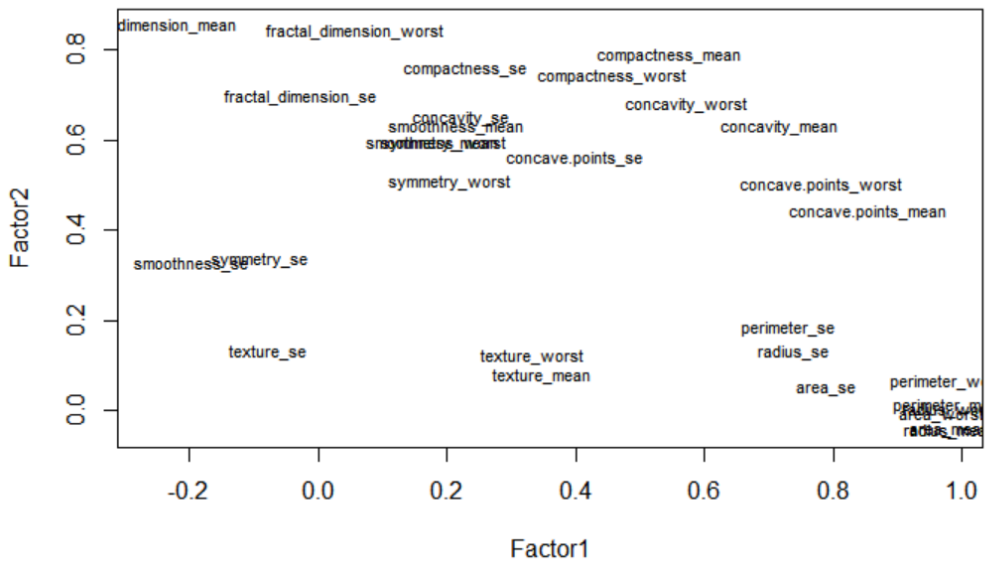

Figura 0.5: `Gráfica con rotación Varimax`

Ahora, por las gráficas podemos observar cómo sería la mejor forma de agrupar los datos, como bien hemos visto anteriormente, lo ideal sería tener más de dos factores, en la figura **??** mostramos la Gráfica con rotación Varimax (0.5) resaltando las mejor manera de agrupar los datos.

Por lo que podríamos concluir que los mejores grupos para maneras los grupos serian:

- Texturas:
  - texture_mean
  - texture_se

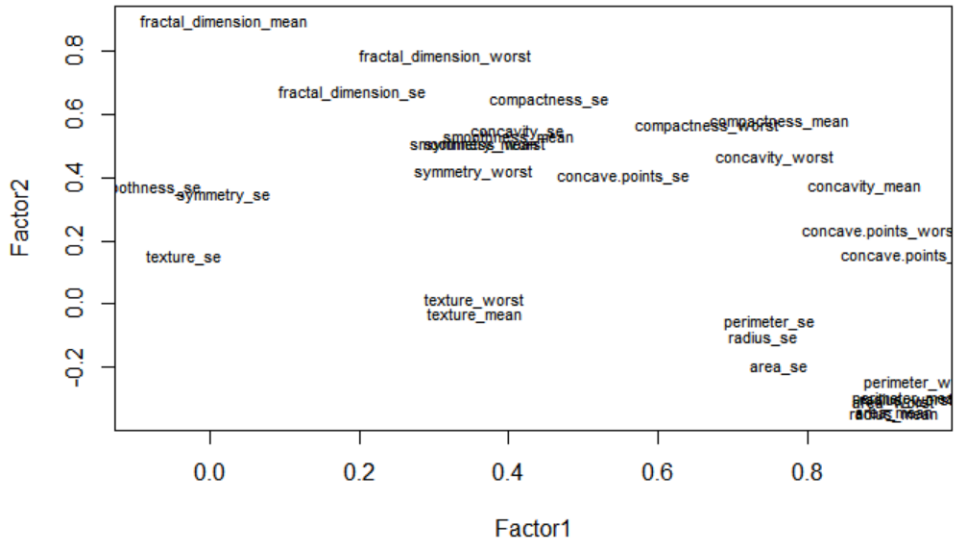

Figura 0.6: `Gráfica sin rotación`

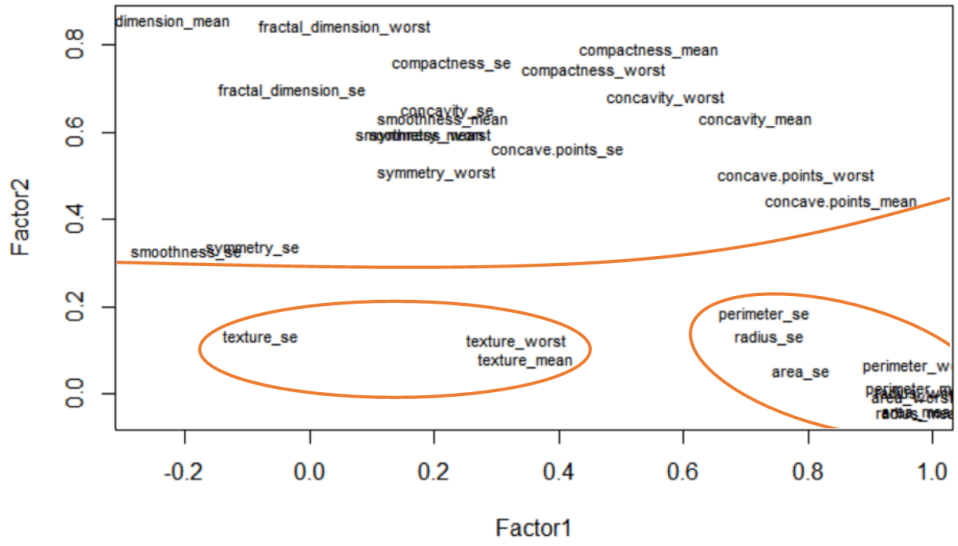

Figura 0.7: `Gráfica con rotación Varimax agrupada`

- texture_worst

- Geométricas:
  - area_mean
  - area_se
  - area_worst
  - radius_mean

  - radius_se
  - radius_worst
  - perimeter_mean
  - perimeter_se
  - perimeter_worst

- Contorno:

  - smoothness_mean
  - smoothness_se
  - smoothness_worst
  - compactness_mean
  - compactness_se
  - compactness_worst
  - concavity_mean
  - concavity_se
  - concavity_worst
  - concave.points_mean
  - concave.points_se
  - concave.points_worst
  - symmetry_mean
  - symmetry_se
  - symmetry_worst
  - fractal_dimension_mean
  - fractal_dimension_se
  - fractal_dimension_worst

Discriminación y Clasificación (LDA y QDA)

Para realizar **LDA** y **QDA** se tomó un entrenamiento del 70 % y un entrenamiento del 30 %, en donde para ambos se obtuvieron las mismas probabilidades a priori, 0. 6306533 y 0. 3693467 para los tipos de tumor benigno y maligno respectivamente. Por otro lado, se hallaron las matrices de confusión para cada uno de los métodos, se usarán las etiquetas "B" para tumores benignos y "M" para los malignos.

Tipo	B	M
B	106	0
M	7	58

Cuadro 0.1: Matriz de confusión para **LDA**.

Para la discriminación lineal se obtuvo que solo 7 tumores malignos fueron clasificados como benignos y la totalidad de los tumores benignos fueron clasificados correctamente, dando un error del 0.04093567, con el cual podemos decir que esta clasificación es efectiva pues su error es mínimo debido a que la gran mayoría de los tumores fueron clasificados correctamente.

Tipo	B	M
B	100	6
M	3	62

Cuadro 0.2: Matriz de confusión para **QDA**.

Se pude observar en la tabla dada anteriormente que la discriminación cuadrática clasifico correctamente la gran mayoría de los tumores 3 tumores de la clase "M" fueron clasificados en la clase "B" y 6 tumores de la clase "B" en la clase "M". Dando un incremento en su error con respecto a **LDA** aumentando alrededor de 0,01. Su resultado fue: 0.05263158. Por lo que podemos decir que el método más efectivo para la muestra tomada es **LDA**.

Asimismo, se realizó una muestra aleatoria de tamaño 100 con un entrenamiento del 70 % y una validación de 30 % para cada una de las observaciones y se realizó un análisis estadístico para los errores en donde se obtuvieron los siguientes resultados:

Método	Máximo	Mínimo	Media	Intervalo de Confianza
LDA	0.09357	0.0117	0.04538	$[0.042135, 0.048625]$
QDA	0.07018	0.01754	0.04713	$[0.04478, 0.04948]$

Cuadro 0.3: Análisis estadístico para la muestra.

En consecuencia, podemos decir que los métodos llegan a presentar errores similares pues se presentan medias similares 0.0453 para **LDA** y 0.0471 para **QDA**. No obstante, al hablar de máximos y mínimos de la muestra notamos que el método de **LDA** difiere más que el de **QDA** esto debido a que el máximo de la discriminación lineal es mayor, mientras que sus mínimos son similares. Por otro lado, los limites del intervalo de confianza del método de discriminación cuadrática son mayores que los del lineal. Sin embargo, la distancia entre el limite superior e inferior de **QDA** es menor que la **LDA** entonces se puede concluir que los errores de la muestra para el método cuadrático difieren menos que los del método lineal. Además, al realizar una prueba de hipótesis de dos colas con hipótesis nula de que las medias difieren y alternativa de que no difieren se obtuvo un valor p de 0.386 entonces se puede concluir que las muestras difieren en su media.

Finalmente, en la Figura 0.8 se muestran los `boxplots` respectivos para cada uno de los errores de las muestras y en un punto rojo las medias. Así pues, por lo dicho anteriormente y por la distribución de las gráficas se concluye que el método de discriminación cuadrática difiere menos que el de la discriminación lineal. Empero, nos es concreto afirmar que **QDA** es más efectivo que **LDA** puesto que sus errores no varían mucho y tanto sus medias como intervalos de confianza son similares.

PCA y Clustering(k-means)

La finalidad de emplear clustering en el dataset empleado radica en encontrar posibles agrupaciones distintas a la presentadas en la variable diagnostico, la cual me clasifica los datos en dos categorías: maligno y benigno;

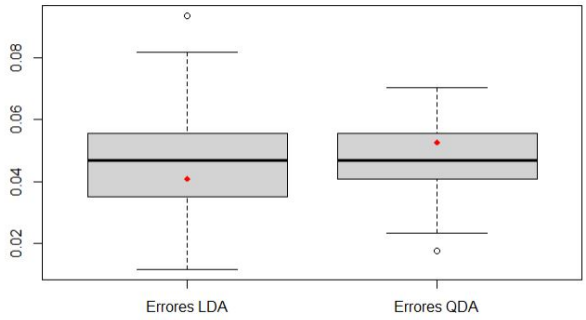

Figura 0.8: `boxplots` errores

o de ratificar mediante un método alternativo que estas agrupaciones son "naturales". Ya que contamos con 30 variables predictoras, emplearemos la técnica de PCA para reducir la dimensionalidad y así poder emplear el algoritmo de k-means sobre unas pocas componentes principales.

Una vez aplicado PCA, y con base en los datos de la desviación estándar para cada componte principal calculamos las proporciones de varianza explicada (PVE) de cada componente principal y generamos la gráfica que se muestra en la Figura 0.9.

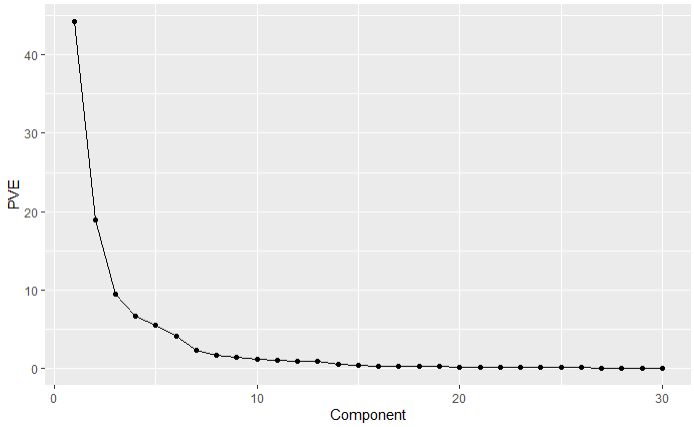

Figura 0.9: Gráfica de la proporción de varianza explicada (PVE) de cada componente principal.

A partir de la Figura 0.9 y empleando el criterio del *elbow* podemos tomar las primeras 7 componentes principales y descartar el resto. Además, al calcular las proporciones de varianza acumulada encontramos que desde la segunda componente principal ya se tiene mas de la mitad con valor de 63 % y que para la componte principal 7 se obtiene un valor aproximado de 91 %.

La Figura 0.10 presenta una gráfica de la dos primeras componentes principales donde los datos o puntos han sido etiquetados dependiendo del valor que toma la variable *diagnosis* para cada uno. En esta gráfica es fácil ver que los datos pueden ser agrupados en dos, y que estos grupos pueden coincidir con la clasificación previa a partir del diagnostico. De esta manera esto nos podría indicar que bastaría con solo emplear las dos primeras componentes principales para emplear algún algoritmo de clustering sobre ellas.

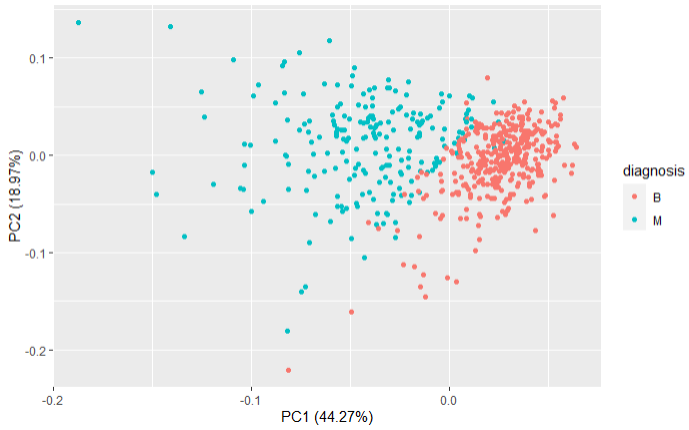

Figura 0.10: Gráfica de las dos primeras componentes principales.

Ahora bien, empleamos k-means sobre distintos valores de k usando las primeras 7 componentes principales y calculamos la suma de cuadrados dentro de cada cluster, un componente por cluster; luego sumamos cada una de estas sumas para obtener una métrica del desempeño de la agrupación para cada k. De esta manera obtenemos lo que llamaremos TWCSS (Total Within-Cluster Sum of Squares) para cada valor de k. La figura 0.11 muestra el comportamiento de los valores de TWCSS con respecto al número de clusters k.

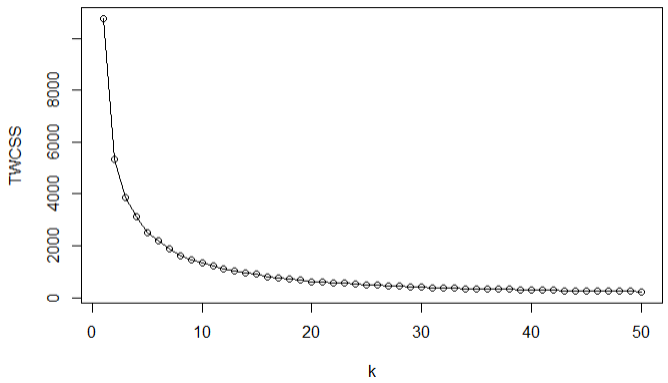

Figura 0.11: Gráfica del número de clusters k contra la suma total de cuadrados dentro de cada cluster(TWCSS).

A partir de la figura 0.11 y empleando el criterio del *elbow* podemos decir que el k adecuado corresponde a $k = 2$. Esta suposición nos afirma que a partir de los datos una agrupación en dos cluster es adecuada, esto concuerda con la clasificación hecha y establecida por la variable *diagnosis* en la que se dividen los datos en los grupos maligno y benigno.

Una vez establecido $k = 2$ implementamos k-means para las 7 primeras componentes principales y obtenemos la Figura 0.12. Una vez mas se puede observar que existe una división clara en dos grupos para las dos primeras componentes principales, y que la primera componente en combinación con una de las otras componentes también nos lleva a establecer una división similar.

Resulta conveniente elaborar una tabla de contingencia, Tabla 0.4, para comparar la agrupación que logramos

Figura 0.12: Gráfica de resultados empleando K-means, con k = 2, para las primeras 7 componentes principales.

con k-means respecto de la clasificación original entre maligno y benigno.

Cluster	Benigno	Maligno
1	14	175
2	343	37

Cuadro 0.4: Tabla de contingencia de la agrupación empleando k-means para las 7 primeras componentes principales respecto a la clasificación original entre maligno y benigno.

Con respecto a la Tabla 0.4 podemos decir que la agrupación mediante k-means agrupa la mayor parte de datos con etiqueta "benigno" en el grupo 2 (343 datos) y los de etiqueta "maligno" en el grupo 1 (175 datos); además se agrupan de forma incorrecta 14 datos en el grupo 1 y 37 datos en el grupo 2. Es decir, se agrupó correctamente el 96 % de los datos etiquetados como "benigno" y el 82,5 % de los etiquetados como "maligno". Cabe mencionar que para diversas iteraciones del algoritmo k-means pueden cambiar los etiquetas de los grupos.

Si implementamos k-means para las dos primeras componentes principales obtenemos la Figura 0.13.

Elaborando una tabla de contingencia para comparar la agrupación que logramos con k-means, usando las dos primeras componentes principales, respecto de la clasificación original en maligno y benigno obtenemos la Tabla 0.5

Con respecto a la Tabla 0.5 podemos decir que se agrupó correctamente el 95,5 % de los datos etiquetados

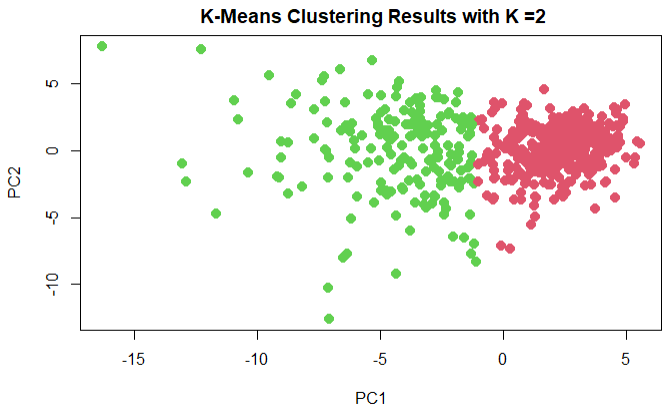

Figura 0.13: Gráfica de resultados empleando K-means, con k = 2, para las primeras 2 componentes principales.

Cluster	Benigno	Maligno
1	341	37
2	16	175

Cuadro 0.5: Tabla de contingencia de la agrupación empleando k-means para las 2 primeras componentes principales respecto a la clasificación original entre maligno y benigno.

como "benigno" y el 82,5 % de los etiquetados como "maligno". De esta manera se ve que se presenta una diferencia del 0,5 % en la agrupación de datos etiquetados como "benigno" con respecto al porcentaje hallado para k-means usando las 7 primeras componentes principales, mientras que se mantiene el porcentaje con respecto a los etiquetados como "maligno".

5. Conclusiones

- Con el fin de poder analizar mejor los datos y reducir la dimensión de los datos aplicando Análisis Factorial (FA) pudimos observar que lo mejor es agrupar los datos en 3 factores, **Texturas, Geométricas** y **Contorno**.
- Los métodos de discriminación lineal y cuadrática presentan errores similares al realizar un entrenamiento del 70 % y una validación del 30 % puesto que sus matrices de confusión varían alrededor de los mismos valores. Sin embargo, la prueba de hipótesis de la muestra de los errores nos dice que sus medias difieren en consecuencia se puede decir que uno de los métodos erra menos que el otro. No obstante, no es posible concluir cual de los dos presenta un menor error pues datos como las medias o intervalos de confianza son similares.
- El uso de PCA nos permitió conocer que para las primeras dos componentes principales se tiene mas de la mitad de la proporción de varianza acumulada con valor de 63 % y que para la componente principal 7 se obtiene un valor acumulado aproximado de 91 %. Además, se observó que puede resultar sencillo agrupar en dos clusters empleando las primeras dos componentes principales.

- Al emplear una técnica de clustering como k-means sobre el dataset trabajado se encontró que hay un claro agrupamiento en dos grupos. Esto quiere decir que las variables que contiene el dataset son adecuadas para agrupar y posteriormente categorizar lo datos como benigno y maligno.

Referencias

[1] Johnson, Richard A; Wichern, Dean W. *Applied Multivariate Statistical Analysis, 6th Ed.*

[2] Kaggle. *Breast Cancer Wisconsin data.*

[3] American Cancer Society. *¿Qué es el cáncer de seno?.*

[4] OMS. *Cáncer de mama: prevención y control.*

[5] Globocan. *Colombia.*