Características del tipo de cáncer.

Ashley Dafne Aguilar Salinas
Tecnologías para la Información en Ciencias
UNAM ENES unidad Morelia
ashaguilar06@gmail.com

Mario Alberto Martinez Oliveros
Tecnologías para la Información en Ciencias
UNAM ENES unidad Morelia
mttzoma@gmail.com

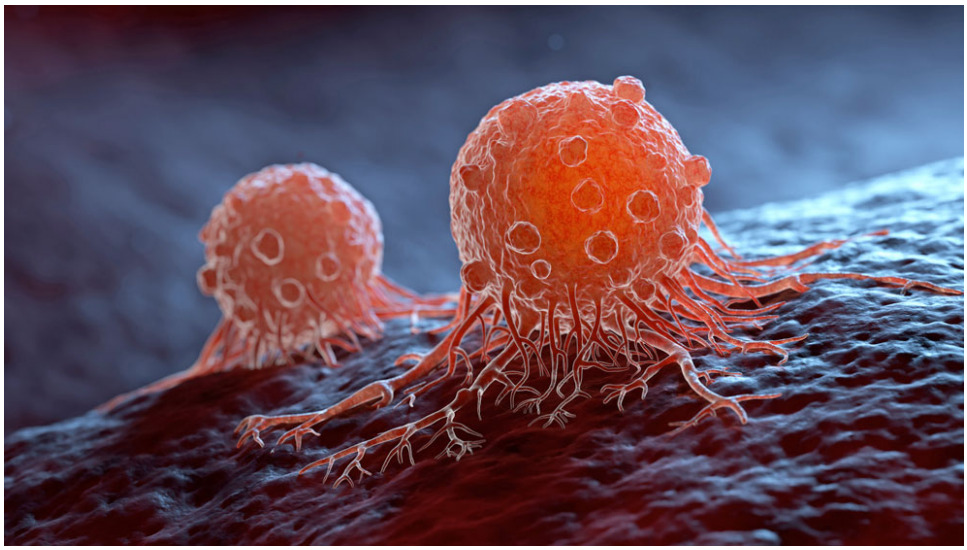

Figure 1: Migración de células cancerosas.

ABSTRACT

The following work aims to perform a statistical analysis of the dataset called *Cancer Data* that is available in the *Kaggle* platform. This with the intention of putting into practice the techniques learned during the multivariate statistics course, some of the techniques planned to be used are: multiple linear regression, principal component analysis (PCA), factor analysis (FA) and cluster analysis.

ACM Reference Format:
Ashley Dafne Aguilar Salinas and Mario Alberto Martinez Oliveros. 2023. Características del tipo de cáncer.. In *Proceedings of Estadística Multivariada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Estadística Multivariada, Diciembre 2023, Morelia, Mich.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCCIÓN

El cáncer es una enfermedad por la que algunas células del cuerpo se multiplican sin control y se diseminan a otras partes del cuerpo. A veces el proceso no sigue este orden y las células anormales o células dañadas se forman y se multiplican cuando no deberían. Estas células tal vez formen tumores, que son bultos de tejido. Los tumores son cancerosos (malignos) o no cancerosos (benignos) [Instituto Nacional de Cáncer [n. d.]].

Las masas malignas suelen diseminarse o invadir los tejidos cercanos, y también es posible que se diseminen a otras partes del cuerpo a través de la sangre y el sistema linfático. También se llama neoplasia y tumoración. Mientras que las masas benignas a veces crecen mucho pero no se diseminan y tampoco invaden los tejidos cercanos ni otras partes del cuerpo [ins [n. d.]].

El siguiente trabajo tiene como objetivo realizar un análisis estadístico del conjunto de datos llamado *Cancer Data* [1] que se encuentra disponible en la plataforma de *Kaggle* [2]. Las técnicas que se aplicaran para realizar el análisis serán la regresión lineal múltiple (ARLM), análisis de componentes principales (ACP), análisis

[1] https://www.kaggle.com/datasets/erdemtaha/cancer-data?rvi=1
[2] https://www.kaggle.com/

factorial (AF) y el análisis de conglomerados. Así como también se realizará el preprocesamiento necesario para limpieza de los datos.

2 PREPROCESAMIENTO

2.1 Limpieza de los datos

El conjunto de datos original cuenta con alrededor de 569 registros y 33 columnas. Aunque no se tomaran en cuenta todas para este análisis. Al consultar los datos nulos o vacíos que se encuentran en nuestro conjunto de datos no percatamos que la columna **Unnamed: 32** tenía todos los datos vacíos, por lo que se tomó la decisión de eliminar esa columna, así como también con las columnas **id**, **diagnosis**. Por lo que ahora contamos con 30 columnas que sí serán consideradas para el análisis del conjunto de datos.

2.2 Estandarización de los datos

Los datos de conjunto se encuentran en diferentes escalas, por lo que difieren mucho en los valores, para evitar errores que pueden ocasionar malas interpretaciones de los resultados se decidió estandarizar los datos mediante siguiente fórmula: $z_i = \frac{x_i - \mu}{\sigma}$.

2.3 Muestreo

Por último, para finalizar con el preprocesamiento de los datos y poder aplicar los métodos vistos a lo largo del curso, usaremos una muestra de nuestros datos estandarizados con la cual estaremos trabajando a lo largo del análisis. El tipo de muestreo será aleatorio y estaremos usando alrededor del 30% de los datos.

3 ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE (ARLM)

El análisis de regresión múltiple sirve para predecir y describir si la relación entre las variables explicativas y las variables dependientes es significativa, así como qué variables explicativas son las más importantes. Lo que buscamos es generar un modelo lineal de la forma: $Y = Xb + \epsilon$.

Como análisis de regresión lineal múltiple nos ayuda a predecir el valor de datos desconocidos, lo usamos para predecir el valor de la variable **diagnosis**, nuestra y, la cual suponemos que no conocemos pues la hemos eliminado en el paso de limpieza de datos, esto lo hace mediante el uso de otro valor de datos relacionado y conocido, en este caso todas nuestras X que son las variables. Nuestra ecuación que estima el valor de y es la siguiente:

$$y \approx 0.628 + 0.5289X_1 - 0.1718X_2 - 0.1023X_3 - 0.1932X_4 - 0.04091X_5 + 0.2872X_6 - 0.569X_7 + 0.1115X_8 + 0.03222X_9 + 0.1063X_{10} - 0.07365X_{11} - 0.09035X_{12} - 0.1468X_{13} + 0.2139X_{14} - 0.0224X_{15} + 0.04783X_{16} + 0.09933X_{17} - 0.05222X_{18} + 0.003046X_{19} + 0.1255X_{20} - 1.18X_{21} + 0.1767X_{22} + 0.2358X_{23} + 0.5367X_{24} - 0.0008232X_{25} - 0.06987X_{26} + 0.1949X_{27} - 0.1106X_{28} - 0.05114X_{29} - 0.2716X_{30}$$

Observamos que la ecuación anterior muestra una "receta" para crear y y las b_s, es decir, los coeficientes de nuestra ecuación que son la "cantidad" que hay que poner de cada variable.

Como realmente conocemos los valores de nuestra variable **diagnosis** podemos evaluar nuestro modelo, pensando en la ecuación anterior como la función que nos ayuda a detectar el tipo de cáncer 0 si es Maligno y 1 si es Benigno además hay que considerar que la función arrojaba valores reales entre $(0, 1)$ por lo cual agregamos la restricción de si el valor arrojado tras sustituir los valores de X_i eran menor a 0.5 entonces se ponía un 0 y de ser mayor se ponía 1. Al obtener los valores predichos con las reglas anteriores obtuvimos la siguiente matriz de confusión:

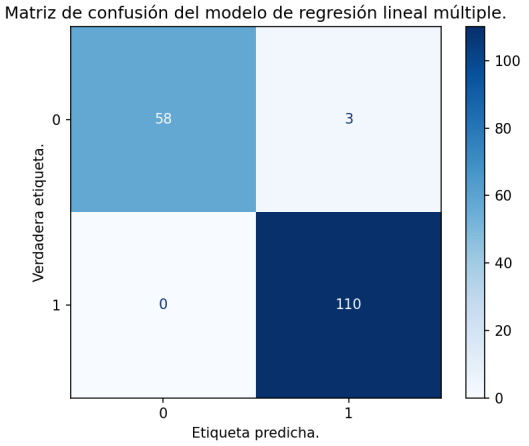

Figure 2: Matriz de confusión del modelo de regresión lineal múltiple.

La cual nos dice que que se predijo 65 valores tipo 0 (Cáncer Maligno = M) y de esos los 65 sí son tipo 0. Por el caso contrario predijo que había 106 valores tipo 1 (Cáncer Benigno = B), de los cuales 103 sí son del tipo 1 y el resto realmente son del tipo 0 Con estos resultados vemos que nuestra predicción es bastante buena, así mismo lo confirma el **accuracy**, obteniendo un valor del 98.24%.

Finalmente tenemos un valor de varianza no explicada no tan grande, con un valor de 3, esto es entendible debido a la evaluación anterior de nuestro modelo, pues la correlación no es perfecta, de ser así no habría varianza no explicada, sin embargo si podemos decir que la correlación no es mala.

4 ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

El análisis de componentes principales (ACP) es un método estadístico para la reducción de la dimensionalidad. Esta técnica se utiliza cuando queremos simplificar la base de datos, ya sea para elegir un menor número de predictores para pronosticar una variable objetivo, o para comprender una base de datos de una forma más simple. El ACP optiene nuevas variables llamadas componentes principales Y_i que son combinaciones lineales de nuestras variables X_i.

Para comenzar con nuestro análisis de componentes principales necesitamos calcular los vectores y valores propios de nuestros

datos, para ello emplearemos la matriz de correlación de nuestros datos.

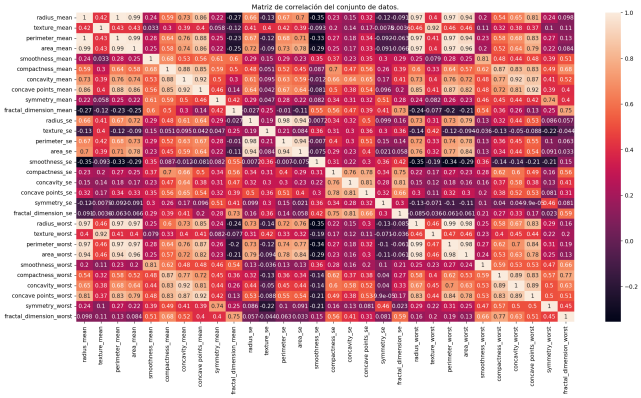

Figure 3: Matriz de correlación del conjunto de datos.

Como la cantidad de variables de nuestros datos que estamos considerando son 30 y estamos usando la matriz de correlación, la varianza total simplemente será la suma de la diagonal de la matriz de correlación que esta es igual a 30. Los componentes principales que seleccionaremos para la reducción de dimensión serán los primeros 3. El porcentaje de varianza que explican las primeras tres componentes principales es de alrededor del 76%, por lo que no estaríamos explicando un 24% de la varianza.

Posterior a obtener los coeficientes de los componentes principales, convertiremos cada punto (x_1, x_2, x_3) que representa un dato del conjunto, en las nuevas coordenadas (y_1, y_2, y_3) donde y_i es un componente principal.

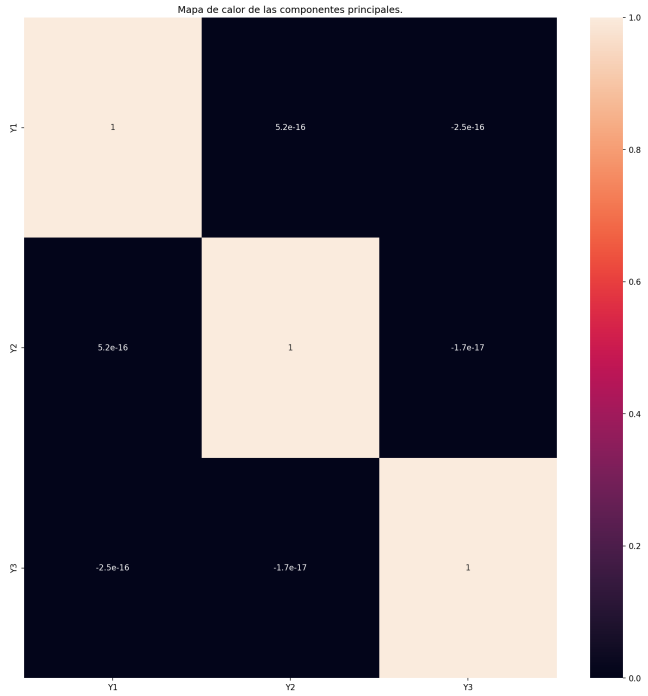

Figure 4: Mapa de calor de las componentes principales.

A continuación se muestra una tabla de porcentajes de varianzas explicadas por cada componente principal. La información que se muestra es la varianza por cada componente principal (CP), y la varianza acumulada de los primeros k componentes principales.

	Varianza	Varianza acumulada
CP_1	46.73%	46.73%
CP_2	18.69%	65.42%
CP_3	9.14%	74.57%

Table 1: Porcentajes de varianzas explicadas por cada componente principal.

Por último graficamos nuestros datos transformados por nuestras componentes principales, las coordenadas y_1, y_2, y_3.

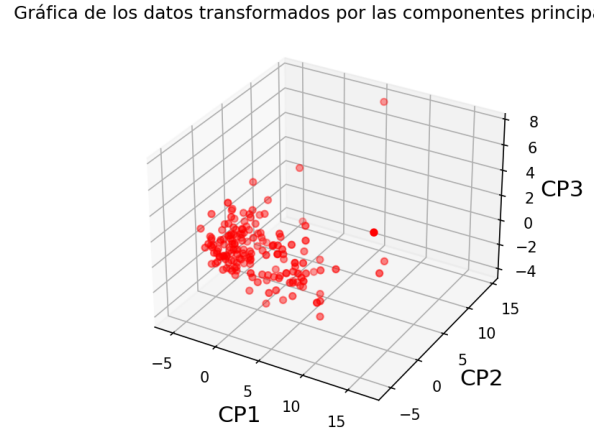

Figure 5: Gráfica de los datos transformados por las componentes principales.

5 ANÁLISIS FACTORIAL (AF)

Como en los demás análisis estadísticos, el objetivo del análisis factorial, es tratar muchas variables mediante una cantidad menor de estas mismas (en este caso llamados factores), sin perder mucha información en el proceso. El análisis factorial escribe cada variable inicial X_i como combinación lineal de las nuevas variables f_i llamados factores.

Como en la sección anterior se mencionó, empezamos calculando los valores y vectores propios que en este caso se utilizara nuevamente la matriz de correlación de nuestros datos, para poder obtener las componentes principales. La siguiente gráfica muestra la varianza explicada por cada componente de nuestros datos.

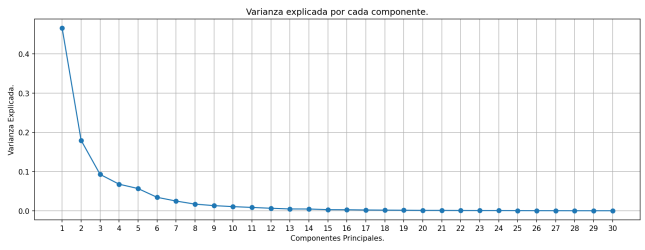

Figure 6: Gráfica de la varianza explicada por cada componente.

En la gráfica 6 podemos observar como a partir de la tercera componente se empieza a notar una caída en la varianza, por lo que se tomó la decisión de realizar el análisis factorial con las primeras tres componentes principales. Puesto que la varianza que explican las tres componentes es de alrededor del 74%.

El primer análisis factorial lo decidimos realizar sin rotación alguna, esto con la intención de analizar los resultados obtenidos y observar si mejoran con una posible rotación usando el método de varimax. A continuación se muestra una tabla con los resultados obtenidos.

	F1	F2	F3
radius_mean	0.76	-0.62	-0.12
texture_mean	0.45	-0.01	0.30
perimeter_mean	0.79	-0.57	-0.13
area_mean	0.76	-0.62	-0.04
smoothness_mean	0.46	0.48	0.03
compactness_mean	0.80	0.52	-0.15
concavity_mean	0.91	0.14	-0.11
concave points_mean	0.91	-0.08	-0.05
symmetry_mean	0.44	0.50	-0.20
fractal_dimension_mean	0.24	0.83	0.17
radius_se	0.66	-0.35	0.57
texture_se	0.25	0.42	0.31
perimeter_se	0.70	-0.32	0.52
area_se	0.75	-0.44	0.41
smoothness_se	0.04	0.61	0.62
compactness_se	0.60	0.59	-0.11
concavity_se	0.61	0.48	-0.04
concave points_se	0.61	0.24	0.12
symmetry_se	0.13	0.55	0.15
fractal_dimension_se	0.52	0.76	0.17
radius_worst	0.79	-0.60	-0.02
texture_worst	0.53	0.02	0.16
perimeter_worst	0.82	-0.57	-0.05
area_worst	0.78	-0.59	0.05
smoothness_worst	0.53	0.40	0.17
compactness_worst	0.80	0.35	-0.36
concavity_worst	0.87	0.21	-0.33
concave points_worst	0.89	-0.15	-0.24
symmetry_worst	0.43	0.31	-0.36
fractal_dimension_worst	0.61	0.60	-0.10

Table 2: Resultados obtenidos de realizar AF sin rotación.

En la tabla anterior 2 nos podemos dar cuenta que el AF sin rotación llega a ser bueno, ya que si revisamos los valores de los factores en cada variable, la mayoría de variables sí se llegan a distinguir unas de otras con algún factor; sin embargo, existen casos en los que incluso dos factores llegan a distinguir la misma variable.

Una de las cosas que podemos tener en cuenta también es la especifidad ψ_i y la comunalidad h_i^2 de las variables. Las gráficas que se muestran a continuación nos deja ver esto precisamente.

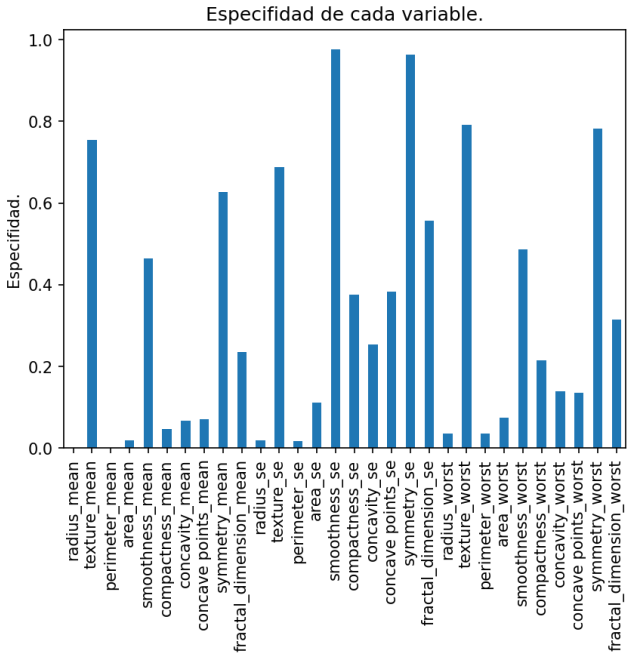

Figure 7: Gráfica de la especifidad de cada variable.

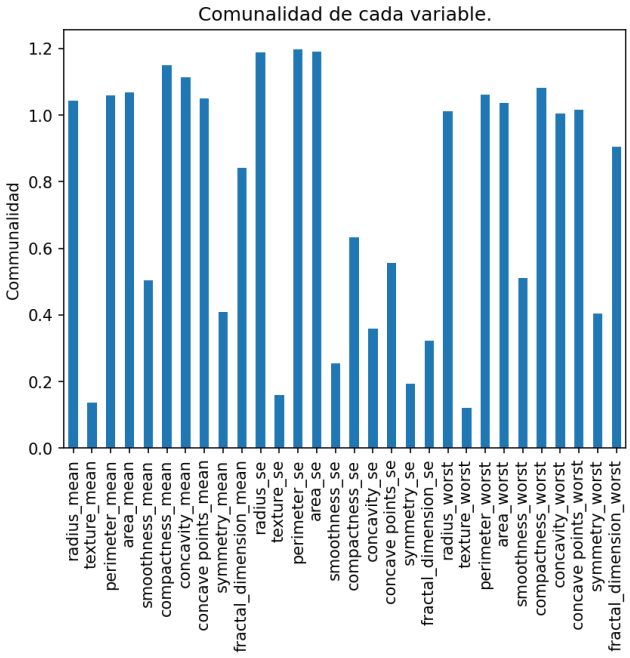

Figure 8: Gráfica de la comunalidad de cada variable.

Lo que buscamos en la especifidad ψ_i es que los valores de las variables sean pequeños y como nos podemos dar cuenta así lo son, aunque algunas variables tiene valores altos no sobrepasa el valor de 0.7, en cambio, en la comunalidad h_i^2 de las variables, lo que buscamos son valores altos como se observa en la gráfica de la comunalidad 8 la gran mayoría si tienen valores altos incluso muy cercanos a 1.

Para finalizar el AF vamos a realizar de nuevo este análisis, pero ahora aplicando una rotación de varimax y analizaremos los resultados obtenidos. A continuación se muestra una tabla con los factores rotados.

	F1	F2	F3
radius_mean	0.92	0.15	-0.30
texture_mean	0.41	0.14	0.32
perimeter_mean	0.92	0.20	-0.29
area_mean	0.95	0.11	-0.24
smoothness_mean	0.07	0.58	0.33
compactness_mean	0.26	0.90	0.24
concavity_mean	0.58	0.71	0.10
concave points_mean	0.73	0.55	0.05
symmetry_mean	-0.01	0.68	0.14
fractal_dimension_mean	-0.28	0.59	0.59
radius_se	0.85	-0.06	0.41
texture_se	0.01	0.28	0.51
perimeter_se	0.85	0.01	0.38
area_se	0.93	0.01	0.24
smoothness_se	-0.20	0.11	0.84
compactness_se	0.07	0.80	0.28
concavity_se	0.16	0.70	0.29
concave points_se	0.35	0.48	0.30
symmetry_se	-0.19	0.35	0.42
fractal_dimension_se	-0.02	0.72	0.59
radius_worst	0.96	0.13	-0.21
texture_worst	0.42	0.27	0.22
perimeter_worst	0.95	0.19	-0.22
area_worst	0.96	0.10	-0.14
smoothness_worst	0.20	0.51	0.42
compactness_worst	0.31	0.89	-0.03
concavity_worst	0.46	0.84	-0.06
concave points_worst	0.71	0.59	-0.16
symmetry_worst	0.06	0.63	-0.09
fractal_dimension_worst	0.08	0.81	0.30

Table 3: Resultados obtenidos de realizar AF con rotación varimax.

Con referencia en la tabla anterior podemos observar que aplicado el análisis factorial con rotación varimax si mejora la interpretación de los factores con los datos, ya que podemos ver como cada factor describe cierto tipo de variables a comparación de los otros, así como también ya no es tan común encontrar que dos o tres factores expliquen bien una variable.

6 ANÁLISIS DE CONGLOMERADOS (AC)

El análisis de conglomerados consiste en buscar grupos (comúnmente llamados conglomerados) en un conjunto de observaciones de forma tal que aquellas que pertenecen a un mismo grupo se

parezcan, mientras que aquellas que pertenecen a grupos distintos sean disímiles, según algún criterio de distancia o de similitud.

En esta sección tiene sentidos hacer algoritmos de partición, por lo cual hicimos uso del método KMeans. Para empezar debemos proporcionar el número k de conglomerados que deseamos tener. Como recordarás nuestros datos están clasificados en dos grupos o conglomerados de acuerdo al tipo de cáncer, estos son en Maligno 0 y Benigno 1, si bien es evidente que elegiríamos $k = 2$ corroboramos con la siguiente gráfica que el "mejor" valor para k es 2.

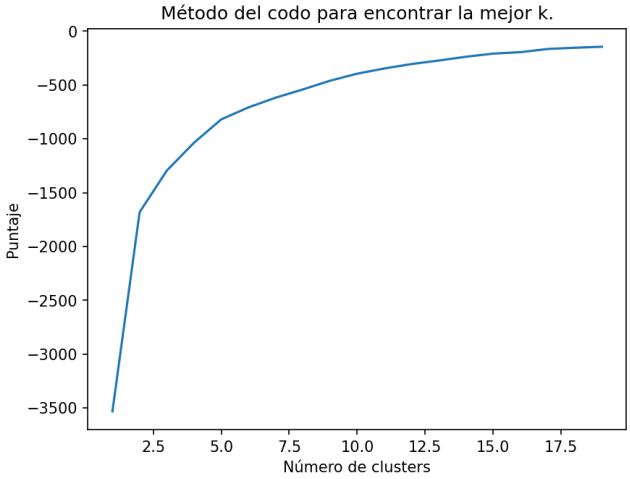

Figure 9: Gráfica del método del codo para encontrar la mejor k.

Usamos el predict de KMeans para obtener la etiqueta que tendrá cada una de nuestras instancias $(0o1)$, posteriormente evaluamos la predicción con los valores reales de la variable, **diagnosis**. Obtenemos la siguiente matriz de confusión:

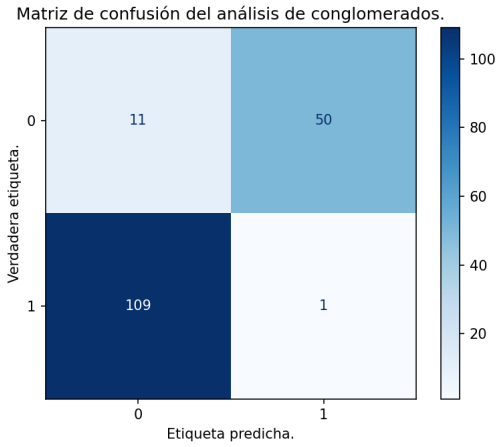

Figure 10: Matriz de confusión del análisis de conglomerados.

Con la matriz de confusión vemos que se predijo 52 valores tipo 0 (Cáncer Maligno = M) y de esos los 50 sí son tipo 0 y 2 son del tipo 1. Por el caso contrario predijo que había 119 valores tipo 1 (Cáncer Benigno = B), de los cuales 101 sí son del tipo 1 y el resto realmente son del tipo 0 Con estos resultados vemos que nuestra predicción es bastante buena, así mismo lo confirma el **accuracy**, obteniendo un valor del 88.30%.

Finalmente mostramos una visualización en $2D$ de los clusters predichos con KMeans, donde se pueden distinguir cada cluster por el color y el triangulo es el centroide de cada cluster respectivamente.

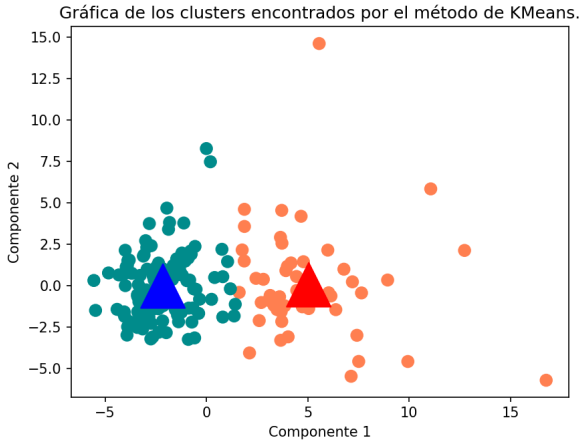

Figure 11: Visualización de los clusters encontrados con KMeans.

7 CONCLUSIONES

Es interesante ver como hay métodos que nos ayudan a estimar las etiquetas de los datos a partir de cierta información con la que contamos o podemos obtener a partir de ella. Vimos como la reducción de dimensionalidad reduce los costes del aprendizaje automático en KMeans y permite la resolución de problemas complejos con modelos simples, esto visualizando los datos en gráficas de dimensión 2D. Esta técnica permite reducir el tiempo de entrenamiento del modelo, pero también presenta ciertos inconveniente como la pérdida de algunos datos.

En cuanto a nuestros resultados vemos que obtuvimos buenos modelos, esto debido a que algunos de ellos les aplicamos algunas métrica de evaluación de modelos, y en algunos otros vimos cuanta varianza explicaban los nuevos componentes principales o factores.

REFERENCES

[n. d.]. *Diccionario de cáncer del NCI.* https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/tumor

Instituto Nacional de Cáncer. [n. d.]. *¿Qué es el cáncer?* https://www.cancer.gov/espanol/cancer/naturaleza/que-es