UNIVERSIDAD NACIONAL DE COLOMBIA

PREGRADO EN ESTADISTICA

Departamento de Estadística
Facultad de Ciencias

— Introducción al Análisis Multivariado —

Analisis de cluster para identificacion de tumores Multivariados

Datos: Cancer Data
tarea número 6

Integrantes:

Ricardo William Salazar Espinal C.C. 1017219472

Medellín, Colombia

Medellin, agosto 23 de 2024

Índice

Índice de figuras

Índice de cuadros

1 Descripcion conjunto de datos

Este conjunto de datos contiene las características de los pacientes diagnosticados con cáncer. el tipo de cáncer (diagnóstico), las características visuales del cáncer y los valores promedio de estas características

Las principales características del conjunto de datos son las siguientes:

Diagnóstico : Indica el tipo de cáncer. Esta propiedad puede tomar los valores “M” (Maligno) o “B” (Benigno).

radio_medio, textura_media, perímetro_medio, área_media, suavidad_media, compacidad_media, concavidad_media, puntos cóncavos_media : Representa los valores medios de las características visuales del cáncer.

Otras características contienen rangos específicos de valores promedio de las características de la imagen del cáncer:

radio_medio, textura_media, perímetro_medio, área_media, suavidad_media, compacidad_media, concavidad_media, puntos cóncavos_media

Cada muestra contiene la identificación única del paciente, el diagnóstico de cáncer y los valores promedio de las características visuales del cáncer.

2 Objetivo del analisis

Como ya describimos en los datos estos estan clasificados en dos tipos Maligno o Benigno, nuestro interes sera realizar un analisis de cluster y ver que tan bien este metodo es capaz de separar los datos en las dos poblaciones de interes, para ellos realizaremos dos modelos el metodo *Ward.D2* que es un metodo de agrupamiento jerarquico mas robusto que el metodo Ward, este lo compararemos con el metodo *K-Means* que tambien es un metodo muy conocido de agrupamiento de datos.

Nuevamente como tenemos caracteristicas que se miden en diferentes unidades, lo primero que debemos realizar es una normalizacion de los datos presentes.

3 Metodo Ward.D2

El método Ward.D2 es una técnica de agrupamiento jerárquico que busca minimizar la suma total de la varianza dentro de los clusters. A continuación se describen los pasos básicos y las fórmulas involucradas:

Para dos clusters (C_i) y (C_j) con (n_i) y (n_j) observaciones respectivamente, y las medias de los clusters $(\bar{\mathbf{x}}_i)$ y $(\bar{\mathbf{x}}_j)$, la disimilitud (D^2) entre los clusters se mide como el aumento en la suma de cuadrados dentro de los clusters si se combinan (C_i) y (C_j).

$$D^2(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \|\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j\|^2$$

donde $\|\cdot\|$ denota la norma Euclidiana.

1. **Inicialización:** Cada observación se considera un cluster individual.
2. **Cálculo de disimilitudes:** Calcular D^2 entre todos los pares de clusters.
3. **Fusión de Clusters:** Identificar y combinar el par de clusters con la menor disimilitud D^2.
4. **Actualización:** Actualizar la disimilitud entre el nuevo cluster combinado y todos los demás clusters.
5. **Repetición:** Repetir los pasos 2 a 4 hasta que todos los clusters se hayan fusionado en un único cluster.

3.1 Indice de silhoutte

Es una medida que se utiliza para evaluar la calidad de un agrupamiento en algoritmos de clustering. Ofrece una forma de determinar cuán bien se ajustan los datos a los clusters en los que han sido agrupados, proporcionando una idea de la separación y cohesión de los clusters.

Tabla 1: Posibles Cantidades de Clústeres con su Índice de Silueta

Cantidad de Clústeres (k)	Índice de Silueta Promedio
2	0.3443195
3	0.3348169
4	0.3031771
5	0.2491277
6	0.1223395

El índice de silueta para un punto i es una medida que combina dos aspectos importantes del clustering:

Cohesión: Qué tan cercano está el punto i a los otros puntos en su propio cluster.

Separación: Qué tan lejos está el punto i de los puntos en el cluster más cercano que no es el suyo.

La formula para el indice se calcula de la siguiente forma:

$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$

donde:

$a_i$es la distancia media entre el punto i y todos los otros puntos en el mismo cluster (cohesión).

b_i es la distancia media entre el punto i y todos los puntos en el cluster más cercano (separación).

el indice toma valores entre $[-1, 1]$ donde los valores cercanos a 1 indican que los puntos estan bien agrupados, 0 que el punto está en el límite entre dos clusters es decir La distancia entre los clusters es similar a la distancia dentro del cluster.-1 el punto puede estar mal agrupado. Está más cerca de los puntos en un cluster diferente que de los puntos en su propio cluster.

A continuacion calculamos el indice para la cantidad de clusters K, donde tenemos que el indice de Silhouette nos dice que el valor optimo de Clusters es dos, lo cual va en concordancia con lo explicado en la explicacion de los datos donde solo tenemos dos tipos de categorias.

A continuacion podemos ver como el indice de Silhouette va perdiendo valor a medida que la cantidad de cluster aumento, indicando que la cantidad recomendada para el analisis es 2 cluster.

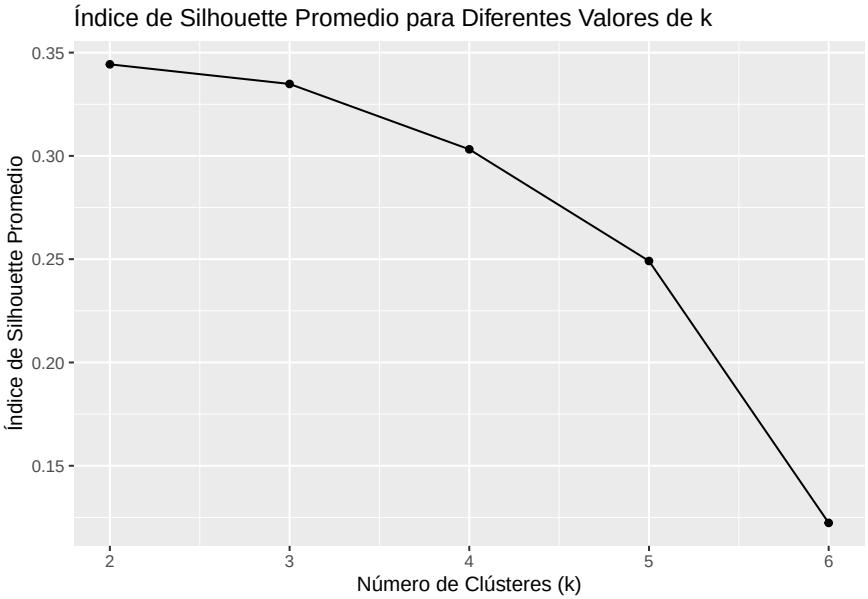

Figura 1: Indice de Silhouette para diferentes valores de K

3.2 Regla del codo para numero de cluster

Adicionalmente usamos la regla del codo donde podemos ver que el numero optimo de klusters es 2 por lo tanto tenemos dos indicadores de que este es el numero optimo con el cual desarrollaremos la investigacion.

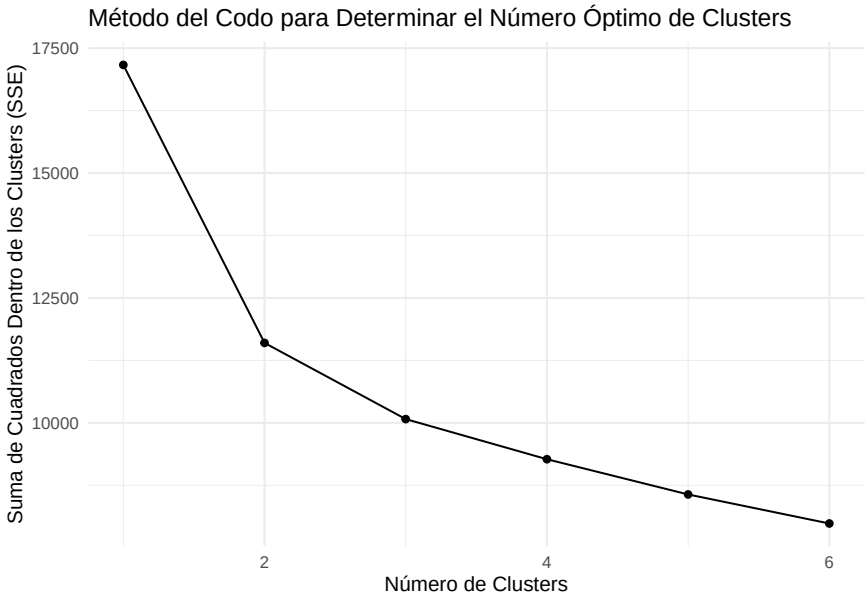

Figura 2: Regla del codo para numero de kluster

Cuando desarrollamos el metodo de Silhouette para k = 2, obtenemos que el cluster tipo 2

Tabla 2: Tabla de Contingencia: Comparación entre Clústeres y Diagnóstico

	Benigno	Maligno
Cluster 1	20	164
Cluster 2	337	48

tiene un mejor indice y agrupamiento de los datos donde 385 de los datos conforman a este cluster, por otro lado el cluster tipo 1 tiene un indice mucho peor el cual cuenta con 184 de los datos, al promediar el indice de ambos se obtiene un indice promedio de 0.34431 el cual es un valor mayor a 0.25, por lo cual podemos decir que es un valor moderadamente positivo.

Esto sugiere que hay una separación razonable entre los clústeres, pero no es ideal. Los puntos están razonablemente bien agrupados en su clúster, pero podría haber margen de mejora en la separación entre clústeres. A un asi dado el analisis anterior y la informacion conocida ante mano de que los datos se separan en dos categorias podemos decir que el valor 0.34431 es un valor aceptable y continuar con el analisis.

3.3 Tabla de contingencia ward.D2 vs Datos reales

Ahora realizaremos una tabla de contingencia y realizaremos un analisis donde concluiremos si hay una relacion entre los datos obtenidos mediante la separacion con el metodo *ward.D2* y los datos reales donde sabemos que tipo de tumor tiene cada persona

Ahora realizaremos una prueba chi-cuadrado para ver si hay una relacion entre las variables en la tabla de contingencia para lo cual proponemos las siguientes hipotesis:

Hipótesis nula (H0): No hay asociación entre las dos variables,esto significa que el diagnóstico no está relacionado con el clúster al que pertenece una muestra. Es decir, la distribución del diagnóstico es independiente del clúster.

Hipótesis alternativa (H1): Hay una asociación entre las dos variables. Esto significa que el diagnóstico y el clúster están relacionados y la distribución del diagnóstico varía según el clúster.

Esta prueba de hipotesis funciona con el siguiente estadistico:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

donde: O_i es la frecuencia observada en la celda i.

E_i es la frecuencia esperada en la celda i bajo la hipótesis nula.

como obtenemos un valor extremadamente pequeño podemos concluir que hay una relacion directa entre los dos cluster y la variable diagnostico.

Tabla 3: Prueba de Hipotesis chi cuadrado

Estadisticos	valores
Chi-Square	309.74
Degrees of Freedom	1.00
P-Value	0.00

3.4 Grafico particion de cluster metodo ward.D2

en el siguiente grafico podemos observar como se comportan la nube de puntos para el metodo *ward.D2* donde podemos observar que en su gran mayoria los datos estan separados y por lo tanto es posible poder diferenciar en su gran mayoria cuando un dato pertenece a un grupo en especifico, a expecion de algunos datos que estan en la interseccion de ambos conuntos donde parece que el metodo no es capaz de captar a cual cluster deben pertenecer exactamente estos puntos, este grafico esta en concordancia con el grafico de silueta para el metodo ward.D2 donde se observo que los datos para el cluster 2 (color rojo) tenian una metrica peor que los del cluster 1 (color azul).

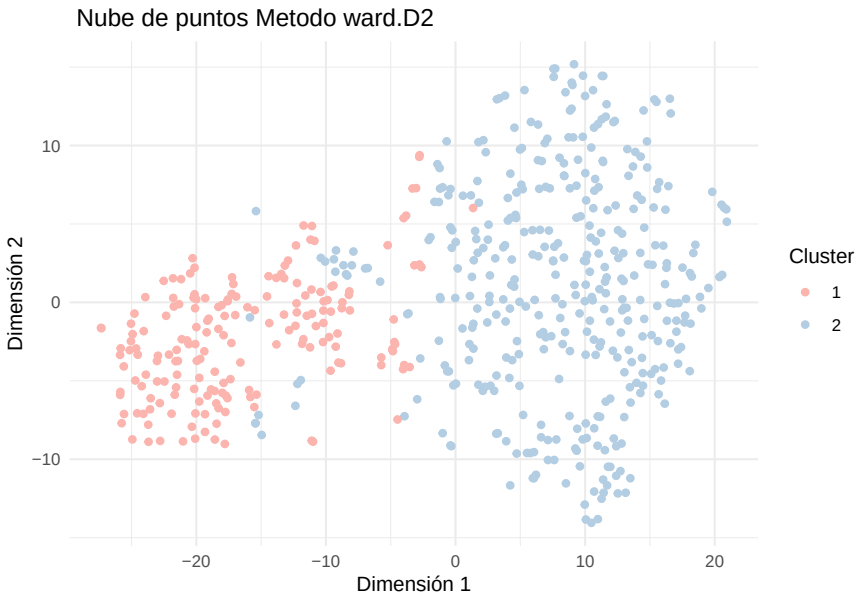

Figura 3: Nube de puntos metodo ward.D2

3.5 Matriz de confusion caso ward.D2

partiendo de la tabla de contingencia y considerando el caso positivo donde la persona tiene un tumor *Maligno* y el caso negativo donde la persona tiene un tumor *Benigno* podemos construir la matriz de confusion como sigue:

Tabla 5: Métricas de Evaluación del Modelo ward.D2

Metricas	Porcentajes
Exactitud	88.049%
Precision	87.53%
Sensibilidad	94.39%
Especificidad	77.35%

Tabla 4: Matriz de Confusion metodo ward.D2

	Maligno	benigno
Cluster M	337	48
Cluster B	20	164

bajo esta matriz tenemos entonces que la exactitud seria:

$$\frac{VP + VN}{VP + FP + FN + VN} = \frac{337 + 164}{337 + 48 + 20 + 164} = 0.88049$$

es decir el metodo tiene una exactitud del 88.049.

tenemos una precision:

$$\frac{VP}{VP + FP} = \frac{337}{337 + 20} = 0.8753$$

por lo tanto el metodo tiene una precision del 87.53

Ahora calculamos la sensibilidad como:

$$\frac{VP}{VP + FN} = \frac{337}{337 + 20} = 0.9439$$

Por lo tanto el modelo tiene una sensibilidad de 94.39.

Calculamos le especificidad como:

$$\frac{VN}{VN + FP} = \frac{164}{164 + 48} = 0.7735$$

tenemos una especificidad de 77.35.

Estos valores podemos resumirlos en la siguiente tabla:

Podemos concluir entonces que el alto valor de exactitud y sensibilidad sugiere que el método de clúster con *ward.D2* está funcionando bien en general, especialmente en la identificación de casos positivos. La precisión también es alta, lo que indica que los casos identificados como positivos son correctos la mayor parte del tiempo.

Aunque la sensibilidad es alta, la especificidad es relativamente más baja. Dependiendo del contexto, se podria considerar ajustar el método para mejorar la especificidad si es importante minimizar los falsos positivos.

4 Metodo K-Means

El metodo de *K-Means* es uno de los algoritmos más populares para la agrupación o clustering en aprendizaje automático.

Este tiene un funcionamiento de la siguiente manera:

Definición de K: K es el número de clústeres que deseas identificar en los datos. Antes de ejecutar el algoritmo, debes especificar el valor de K.

Inicialización: El algoritmo comienza seleccionando aleatoriamente K puntos en el espacio de características como los centros iniciales (o centroides) de los clústeres.

Asignación de Clústeres: Cada punto de datos en el conjunto se asigna al clúster cuyo centroide está más cercano. Esta cercanía se mide típicamente usando la distancia Euclidiana, aunque pueden usarse otras métricas de distancia.

Actualización de Centroides: Una vez que todos los puntos se han asignado a clústeres, el centroide de cada clúster se recalcula como el promedio de todas las observaciones asignadas a ese clúster.

Iteración: Los pasos de asignación de clústeres y actualización de centroides se repiten iterativamente hasta que los centroides ya no cambian significativamente o hasta que se alcance un número máximo de iteraciones.

Convergencia: El algoritmo termina cuando los centroides de los clústeres se estabilizan y no cambian mucho entre iteraciones, indicando que se ha encontrado una solución estable.

El K-Means se basa principalmente en dos métricas para evaluar su rendimiento y para la optimización del algoritmo:

Suma de Cuadrados de Distancias Dentro del Clúster (Within-Cluster Sum of Squares, WCSS):

$$WCSS = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2$$

donde x es un punto en el cluster C_i y μ_i es el centroide del cluster. El objetivo del algoritmo K-Means es minimizar esta métrica. Un menor valor de WCSS indica que los puntos están más cerca de sus centroides, lo que generalmente significa que los clústeres están más compactos y bien definidos.

Como ya vimos anteriormente el indice de silhoutte nos dio que el numero de cluster adecuados es de 2

al aplicar el metodo de *K-Means* tenemos que el cluster numero 1 tiene un tamaño de 194 mientras que el cluster 2 tiene un tamaño de 375, tenemos un indice de silhoutte de 0.3433822 que es un valor casi identico al obtenido en el metodo *ward.D2* por lo tanto este valor no sera un valor para definir cual de los dos metodos funciona mejor.

4.1 Grafico particion de cluster metodo K-Means

este grafico nuevamente muestra lo indicado en el grafico de particion del metodo *ward.D2*, pero en este caso parece que la cantidad de puntos en la interseccion de ambos conjuntos no estan tan mezclados como en el metodo anterior, esto puede indicar que el metodo de *K-Means* es un metodo mas eficiente para este tipo de datos que estamos investigando, a un que es importante tener mas evidencias para llegar a una conclusion de este tipo.

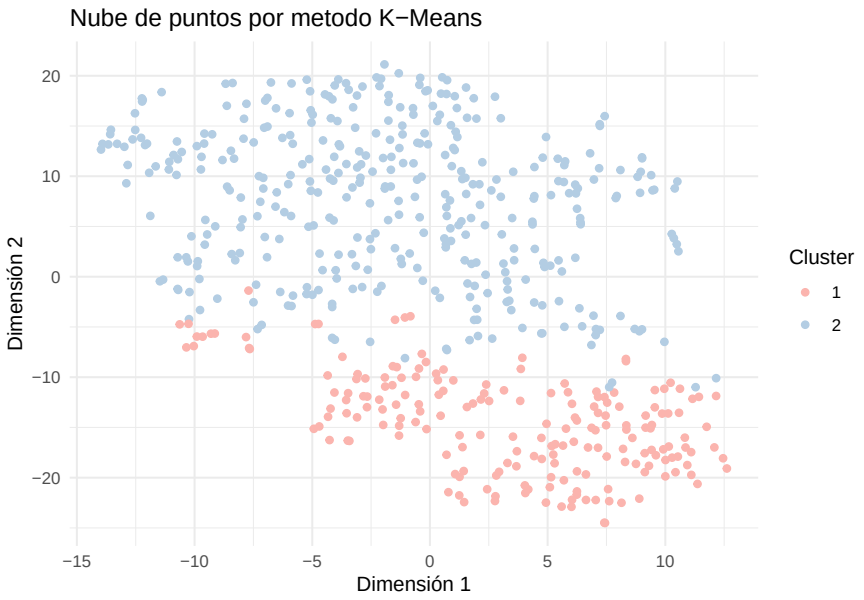

Figura 4: Nube de puntos por metodo K-Means

4.2 Tabla de contingencia Metodo K-Means vs Datos Reales

Ahora tenemos la tabla de contingencia para el metodo *K-Means* donde se contrastara los datos reales donde sabemos que tipo de tumor tiene cada persona con los cluster obtenidos por el metodo

realizamos nuevamente la prueba χ^2 como se indico anteriormente y obtenemos la siguiente salida:

donde nuevamente notamos que el valor p es muy pequeño y por lo tanto podemos concluir que hay una asociación estadísticamente significativa entre las dos variablesen la tabla de contingencia. Es decir, hay una relacion entre los cluster obtenidos por el metodo *K-Means* con los datos reales de diagnostico.

4.3 Matriz de confusion metodo K-Means

Ahora definimos la matriz de confusion para el metodo de *K-Means*, procederemos a calcular las metricas como: Exactitud, Precisión, Sensibilidad y Especificidad

Tabla 6: Matriz de Confusion metodo K-Mens

	Maligno	benigno
Cluster M	339	18
Cluster B	36	176

Tabla 7: Métricas de Evaluación del Modelo K-Means

Metricas	Porcentajes
Exactitud	90.05%
Precisión	94.95%
Sensibilidad	90.4%
Especificidad	90.72%

bajo esta matriz tenemos entonces que la exactitud seria:

$$\frac{VP + VN}{VP + FP + FN + VN} = \frac{339 + 176}{339 + 18 + 36 + 176} = 0.905$$

es decir el metodo tiene una exactitud del 90.05.

tenemos una precision:

$$\frac{VP}{VP + FP} = \frac{339}{339 + 18} = 0.9495$$

por lo tanto el metodo tiene una precision del 94.95

Ahora calculamos la sensibilidad como:

$$\frac{VP}{VP + FN} = \frac{337}{337 + 36} = 0.904$$

Por lo tanto el modelo tiene una sensibilidad de 90.4.

Calculamos le especificidad como:

$$\frac{VN}{VN + FP} = \frac{176}{176 + 18} = 0.9072$$

tenemos una especificidad de 90.72.

Estos valores podemos resumirlos en la siguiente tabla:

Las métricas obtenidas para la evaluación del modelo muestran un rendimiento sólido y equilibrado en su capacidad de clasificación. La exactitud del 90.05 indica que el modelo ha realizado una gran cantidad de predicciones correctas en general, reflejando una buena capacidad para clasificar los datos en la mayoría de los casos. La precisión de 94.95 es particularmente alta, lo que sugiere que cuando el modelo clasifica un elemento como positivo, es muy probable que sea realmente positivo, minimizando las falsas alarmas. La sensibilidad

del 90.4 muestra que el modelo es eficiente en la identificación de positivos verdaderos, capturando la mayoría de los casos positivos disponibles. Finalmente, una especificidad de 90.72 demuestra que el modelo también es competente en identificar correctamente los negativos verdaderos, evitando errores en la clasificación de casos negativos. En conjunto, estas métricas revelan que el modelo ofrece un buen equilibrio entre precisión y capacidad de detección, proporcionando una clasificación confiable en ambas categorías de interés.

5 Comparacion de metodos

Compararemos algunas metricas obtenidas durante el analisis y compararemos que modelo termina siendo mas eficiente

Tabla 8: Comparacion metricas ward.D2 y K-Means

	wardD2	KMeans
Exactitud	88.049%	90.05%
Precision	87.53%	94.95%
Sensibilidad	94.39%	90.4%
Especificidad	77.35%	90.72%
Silhouette	0.34431	0.34338

Exactitud y Precisión: K-Means supera a Word.D2 en ambas métricas, sugiriendo que es más confiable en general y que es menos probable que cometa errores cuando hace predicciones positivas.

Sensibilidad: Word.D2 es superior en términos de sensibilidad, lo que implica que es mejor para detectar positivos verdaderos, una característica crucial si se requiere alta cobertura.

Especificidad: K-Means es notablemente mejor en especificidad, lo que es beneficioso si se busca minimizar el número de falsos positivos.

Índice de Silhouette: Dado que el índice de Silhouette es prácticamente el mismo para ambos métodos, no hay una ventaja clara en términos de calidad del clustering entre Word.D2 y K-Means.

6 Recomendaciónes Generales

K-Means parece ser superior en términos de exactitud, precisión y especificidad.Word.D2 podría ser preferido si la detección de todos los positivos verdaderos es crítica, debido a su mayor sensibilidad.

La elección del mejor método depende de cuál métrica se considere más importante para la aplicación específica. Si es crucial minimizar errores en la predicción de positivos, se podria optar por Word.D2. Si se prefiere un modelo que proporcione menos falsos positivos y tenga una mayor exactitud general, K-Means sería la mejor opción.

7 Conclusiones

K-Means muestra una ventaja en términos de exactitud, precisión y especificidad. Esto sugiere que es generalmente más confiable para clasificar correctamente los casos y minimizar los errores, especialmente en aplicaciones donde es importante evitar falsos positivos.

Word.D2 tiene una mayor sensibilidad, lo que puede ser preferible si es crucial no perder ningún caso positivo verdadero, como en aplicaciones de detección temprana.

Las matrices de confusión y los gráficos de puntos en 2 dimensiones son herramientas complementarias. Mientras que las matrices de confusión ofrecen una evaluación cuantitativa detallada del rendimiento del modelo, los gráficos de puntos proporcionan una visualización intuitiva de la estructura y distribución de los datos.

Juntas, estas herramientas ayudan a tomar decisiones informadas sobre la eficacia del modelo y la calidad de la segmentación o clasificación, facilitando tanto el análisis cuantitativo como la interpretación visual de los resultados.

Referencias

Luque-Calvo, P.L. (2017). *Escribir un Trabajo Fin de Estudios con R Markdown*. Disponible en http://destio.us.es/calvo.

Porras, J.C. (2016). Comparacion de pruebas de normalidad multivariada. *Anales Cientificos*, pp. 141-146. Universidad Nacional Agraria La Molina.

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and computing*, **2**, 117-119.

Taha, E. (2021). Cancer Data. URL https://www.kaggle.com/datasets/erdemtaha/cancer-data