UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Laboratorio N°2

K-Means Clustering

Integrantes:	Felipe González
	Carlos Pérez
Curso:	Análisis de Datos
Sección	0-A-1
Profesor(a):	Dr. Max Chacón Pacheco
Ayudante:	Javier Arredondo Contreras

5 de Diciembre de 2020

Tabla de contenidos

Índice de cuadros

Índice de figuras

1. Introducción

El cáncer es la segunda causa de muerte en el mundo, esto según la Organización Mundial de la Salud (2018). Como se trata de una enfermedad que al pasar el tiempo va evolucionando, su tratamiento también va cambiando. A esto, si se le suma que existe una gran variedad de tipos, resulta crucial tener métodos de detección eficientes.

Se le llama cáncer solo a los tumores que son capaces de realizar metástasis, es decir, aquellos que son capaces de extenderse a los tejidos cercanos o de desprender células cancerosas que viajan a través del sistema circulatorio o linfático para llegar a producir nuevos tumores.

Cuando un cáncer produce metástasis en otra parte del cuerpo, el nuevo tumor posee las características de su origen. Por ejemplo, cuando un cáncer de mama se disemina a los pulmones, el nuevo tumor que se genera tiene las características de un cáncer de mama, e incluso, se le llama cáncer metastásico de mama y no cáncer de pulmón. Esto hace que la clasificación de un tumor recién hallado sea aún mas compleja.

En específico, en este informe se estudia el cáncer de mama. Usualmente, un tumor en el seno puede ser visto a través de un examen de rayos x y se puede sentir. Pero como se explicó anteriormente, no todo cáncer hallado en la zona corresponde a cáncer de mama, pues puede corresponder a un linfoma o sarcoma.

Es crucial estudiar este tipo de cáncer, puesto que "en las mujeres, el cáncer de mama es el segundo tipo de cáncer más común"(NIC, 2020). Además, posee una mortalidad del 10 %.

En esta investigación se realizó un estudio de *Clustering* mediante el uso del algoritmo *K-means Clustering*, usando el dataset *Breast Cancer Wisconsin*. Uno de los objetivos de este laboratorio es variar aspectos importantes, como el preprocesamiento, número de clusters y el método de computación de distancias, para analizar cómo se ve afectado la clasificación de los tumores.

2. Marco Teórico

2.1. Clustering

El *Clustering* corresponde a un amplio conjunto de técnicas para encontrar subgrupos dentro de un determinado conjunto de datos inicial. Cuando se agrupan elementos, se busca que aquellos pertenecientes a un mismo grupo sean lo mas similares posible entre sí, y a su vez, lo mas diferentes a los elementos de los otros subgrupos. Esto es necesario, porque no existe una **variable explicativa** que permita categorizar a cada una de las observaciones. Formalmente, *clustering* pertenece a las técnicas denominadas como de **Aprendizaje No Supervisado**, lo que implica que busca encontrar relaciones entre un rango de n observaciones sin haber sido entrenado por una variable de respuesta. Es utilizado en muchos campos, como *machine learning*, reconocimiento de patrones, análisis de imágenes, recuperación de información, bioinformática, compresión de datos y gráficos por computadora.

2.2. K-Means Clustering

K-Means Clustering corresponde a uno de los algoritmos de *machine learning* No Supervisados mas comunes, donde k es un parámetro de entrada que representa el número de grupos que el analista quiere formar. En este, cada *cluster* o grupo está representado por su *centroide*, el cual corresponde a la media de los puntos asignados al grupo. La idea básica detrás del agrupamiento *k-means* es definir los *cluster* de forma que la variación *intra-cluster* sea mínima. El algoritmo es el siguiente.

1. En primer lugar, es necesario establecer el número k de *clusters* a crear (por el analista)

2. Seleccionar aleatoriamente k elementos u observaciones de el conjunto de datos como los centroides iniciales de cada *cluster* o media.

3. Asignar cada observación a su centroide mas cercano basado en la distancia entre el objeto y el centroide.

4. Para cada uno de los k clusters actualizar su respectivo centroide, calculando los nuevos valores medios para cada una de las variables de cada uno de los puntos. El centroide de

un k_{th} *cluster* es un vector de largo p que contiene las medias para todas las variables de las observaciones en el k_{th} *cluster*.

5. Iterativamente minimizar la variación *intra-cluster*. Esto es iterar sobre los pasos 3 y 4 hasta que las asignaciones de cada cluster se detengan, o se alcance el máximo número de iteraciones. Considerando que se utiliza el entorno y lenguaje R, este define por defecto un numero de 10 iteraciones.

2.3. Distancia

La distancia matricial o también denominada *dissimilarity* en inglés, como se dijo anteriormente, es un paso crucial en el proceso de *clustering*, define cómo es calculada la similaridad entre dos elementos *(x, y)* y puede influenciar en la forma de los clusters. En esta experiencia, se consideran tres distancias utilizadas para el proceso de *clustering*.

2.3.1. Distancia Euclidiana

Es la distancia por defecto y una de las mas comunes. Geométricamente, la Distancia Euclideana entre dos puntos de un plano, corresponde a una linea recta entre estos. La ecuación 1 define la Distancia Eculideana.

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

2.3.2. Distancia de Manhattan

Otro tipo de distancia común. A diferencia de la Distancia Euclideana, si la linea recta que une dos puntos fuera la hipotenusa de un triángulo rectángulo, la Distancia Manhattan correspondería a la suma de sus catetos.

$$d_{man}(x,y) = \sum_{i=1}^{n}|x_i - y_i| \tag{2}$$

Donde, x e y son dos vectores de largo n.

2.3.3. Distancia de Gower

Distancia utilizada cuando las variables son mixtas, es decir, hay variables tanto cualitativas como cuantitativas. La distancia de Gower se calcula como el promedio de las diferencias parciales entre individuos. Cada disparidad parcial (y, por lo tanto, la distancia de Gower) varía en 0 y 1. Se define bajo la siguiente ecuación:

$$d_{gow}(i,j) = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^{(f)} \tag{3}$$

Las diferencias parciales $d_{ij}^{(f)}$ se calculan de distintas formas dependiendo del tipo de dato de la variable. Si es cualitativo, para observaciones que son distintas, se le asigna el valor 1, 0 en otro caso. Si es cuantitativo, la ecuación 4 representa el valor de la diferencia parcial. R_f corresponde al rango máximo observado.

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f} \tag{4}$$

2.3.4. Método del Codo

Se basa principalmente en la suma de los cuadrados dentro de los clusters para un determinado rango de valores para k, dibujando una curva continua desde k con valor 0 hasta el máximo escogido. Cada transición de un k al siguiente significa una variación en la WCSS, al aumentar en uno el valor de k, la idea básica es escoger un k cuya transición signifique una caída significante de la WCSS, lo suficiente para compensar el ingreso de un nuevo cluster.

2.3.5. Método de la Silueta

En resumen, la aproximación del promedio de la silueta mide la calidad de un clustering. Esto significa que determina que tan bien un objeto cae dentro de un determinado cluster, por consiguiente, una gran promedio en el ancho de la silueta indica un buen clustering.

2.3.6. Método de la Brecha Estadística

Este método compara la WCSS para diferentes valores de k con sus valores esperados bajo una distribución con nula referencia de los datos, es decir una distribución sin un clustering tan obvio. Este conjunto es generado utilizando las simulaciones de Monte Carlo en el proceso de muestreo. Y encuentra el óptimo fijandose en el

3. Pre-Procesamiento

Antes de aplicar cualquier algoritmo, es preferente que los datos estén ordenados y/o estructurados. Pero en el mundo real, la mayoría de los datos que resultan de observaciones no lo están. Entonces, para que estén ordenados y para aplicar aún más cualquier algoritmo, los datos deben limpiarse. La razón principal por la que los datos no están ordenados es la presencia de valores perdidos mas conocidos como *missing values* y valores atípicos u *outliers*, a esto se suman también la naturaleza de los datos como el manejo de Variables Categóricas, Normalización de datos y Reducción de Dimensionalidad, por nombrar algunos.

Recordar que el conjunto de datos consta de 699 muestras u observaciones de líquido tumoral de pacientes con protuberancias solidas en su pecho, caracterizadas a través de 10 variables, además del código identificador de cada una de estas. A continuación se detalla a grandes rasgos cada una de estas.

Atributo	Tipo	Dominio
Sample code number	Numérico	id
Clump Thickness	Numérico	1 - 10
Uniformity of Cell Size	Numérico	1 - 10
Uniformity of Cell Shape	Numérico	1 - 10
Marginal Adhesion	Numérico	1 - 10
Single Epithelial Cell Size	Numérico	1 - 10
Bare Nuclei	Numérico	1 - 10
Bland Chromatin	Numérico	1 - 10
Normal Nucleoli	Numérico	1 - 10
Mitoses	Numérico	1 - 10
Class	Nominal	2 , 4

Cuadro 1: Variables de Estudio

3.1. Pruebas de Normalidad

Muchas de las técnicas y métodos, empleados para procesar y clusterizar los datos, requieren determinados supuestos para poder ser utilizados, uno de estos supuestos es la distribución que sigue dicho conjunto de datos, por lo que antes de todo, es necesario realizar pruebas de normalidad para conocer la distribución de los datos. Para verificar esto se realizan pruebas tanto visuales, como pruebas de significancia.

3.1.1. Histogramas

Una de las formas de revisar la distribución de un determinado conjunto de datos, es analizando la curva generada por el histograma de cada una de sus variables, viendo si estas tienen o se acercan a la forma característica de la *Campaba de Gauss*. Aplicando la función **hist()**, para las 9 variables explicativas del conjunto de datos, se obtienen los siguientes expuestos en la figura 1

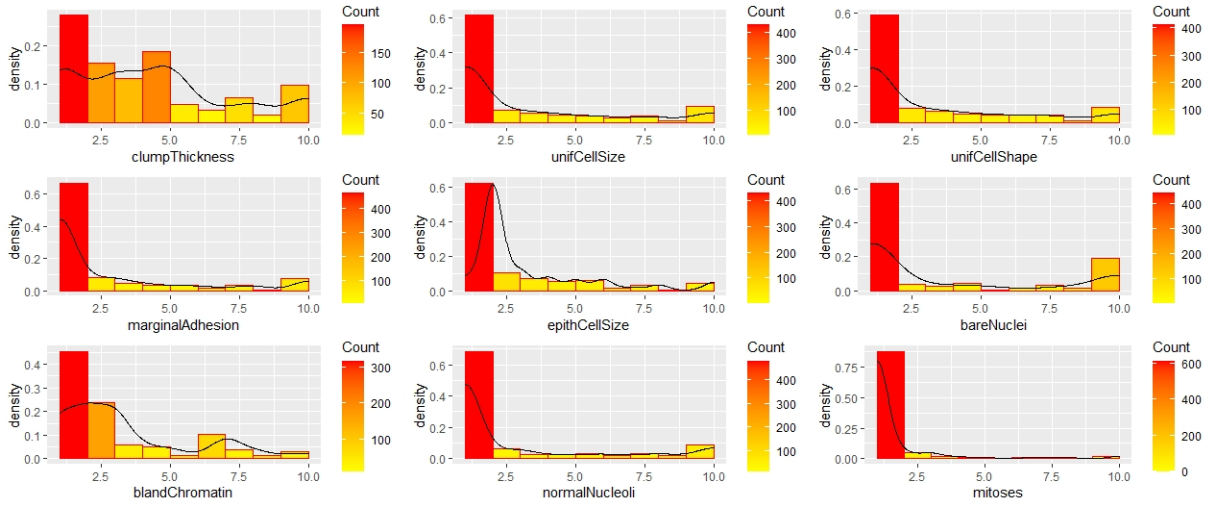

Figura 1: Histograma del Conjunto de Datos

Viendo los gráficos anteriores, resulta difícil poder dilucidar al menos a simple vista que los datos pudiesen seguir una distribución normal, razón por la cual al menos visualmente se puede asegurar que los datos no siguen esta distribución.

3.1.2. Pruebas de Hipótesis

Con el objetivo de tener aun mayor seguridad respecto a los resultados de la prueba anterior, adicionalmente se realiza una serie de pruebas de hipótesis, empleando distintos métodos ofrecidos por distintos paquetes que ofrece el entorno de programación R. Entonces se plantea la siguiente hipótesis:

$$H_0 : \text{La muestra proviene de una distribución normal}$$
$$H_1 : \text{La muestra no proviene de una distribución normal}$$

Además se trabaja con un nivel de significancia $\alpha = 0{,}05$, con el siguiente criterio de decisión:

$$\text{Si } p < \alpha \text{ se rechaza } H_0$$
$$\text{Si } p \geq \alpha \text{ no se rechaza } H_0$$

Los resultados obtenidos para cada método se muestran en la siguiente tabla, en la cual para cada par variable/método se tiene una celda con valor de p obtenido del resultado de aplicar el método de la columna, con la variable de la fila.

V - M	1	2	3	4	5	6
clumpThickness	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$
unifCellSize	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$
unifCellShape	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$
marginalAdhesion	$< \alpha$	$< \alpha$	$< \alpha$	$> \alpha$	$< \alpha$	$< \alpha$
epithCellSize	$< \alpha$	$< \alpha$	$< \alpha$	$> \alpha$	$< \alpha$	$< \alpha$
bareNuclei	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$	$< \alpha$
blandChromatin	$< \alpha$	$< \alpha$	$< \alpha$	$> \alpha$	$< \alpha$	$< \alpha$
normalNucleoli	$< \alpha$	$< \alpha$	$< \alpha$	$> \alpha$	$< \alpha$	$< \alpha$
mitoses	$< \alpha$	$< \alpha$	$< \alpha$	$> \alpha$	$< \alpha$	$< \alpha$

Cuadro 2: Valores de p

Donde,

1. Prueba de Pearson chi-square
2. Prueba de Shapiro-Francia
3. Prueba de Jarque Bera
4. Prueba de Geary
5. Prueba de Agostino
6. Prueba de Shapiro-Wilk

A excepción de unos pocos valores, la gran mayoría de los valores para p resultado de cada uno de los métodos, resultan inferiores a α, por ende se puede definitivamente decir que existe suficiente evidencia estadística para rechazar H_0, en favor de la hipótesis alternativa H_1 lo que significa que la muestra no proviene de una distribución normal.

3.2. Missing Values

Un *missing value* está definido como el valor de un dato que no está almacenado para una variable en una observación de interés, por lo que mantener estos valores mientras se aplica el método puede incurrir en un comportamiento no deseado, o bien en predicciones que no se ajustan a lo que realmente puede suceder. En este caso, afectaría en la asignación de alguna observación a una clase errónea (falsos tumores cancerígenos o falsos tumores benignos).

Particularmente, para este conjunto de datos se sabe que de las 699 observaciones, 16 cuentan con un *missing value* para la variable *Bare Nuclei*, denotados por el caracter "?". Algunos de los métodos mas populares para lidiar con los valores omitidos o perdidos son la **Eliminación** y la **Imputación**. En la practica es posible verificar estos valores gracias a la ayuda de R, con lo que se obtiene el gráfico de la figura 2

Se observa que los *missing values* están distribuidos en una razón de entre el 0.02 y 0.03 de la totalidad de los datos, y que además están distribuidos solamente para la variable *bare-Nuclei*, para ser mas precisos a una razón de 0.02288984, aproximadamente al 2,3 %, lo cual se puede considerar un porcentaje bastante bajo considerando el total de 699 observaciones.

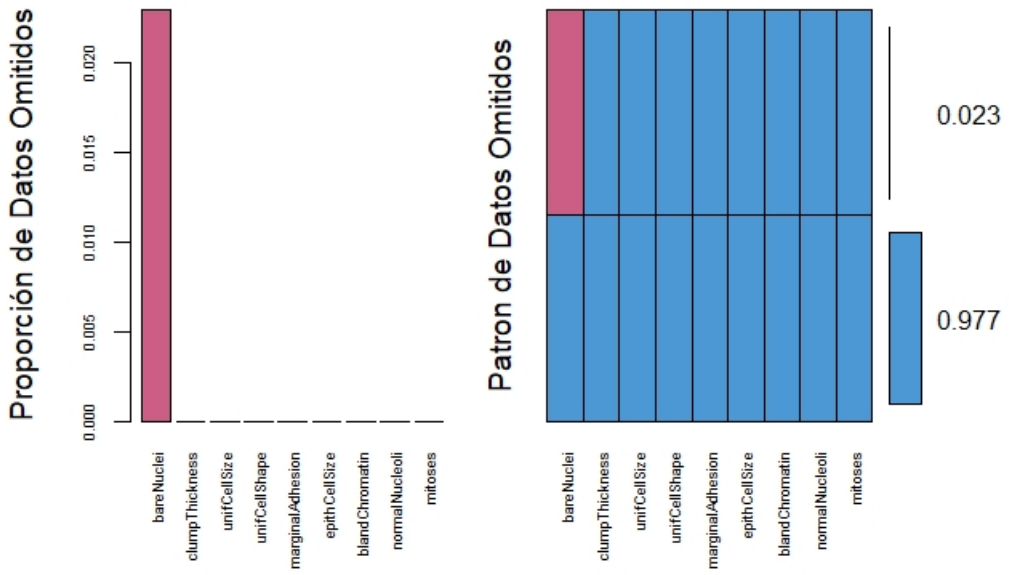

Figura 2: Grafico de los *Missing Values*

- **Eliminación:** Existe mas de un tipo de eliminación, más en esta ocasión se emplea la mas simple y denominada como *Listwise*, donde simplemente se sacan aquellas observaciones que presentan *missing values*. La simplicidad de este método es una de sus principales ventajas, por otro lado reduce el poder del modelo al reducir el tamaño de la muestra, por lo que dependiendo del número de observaciones que presenten *missing values* podría resultar mas desventajoso emplearlo.

  Es importante comprender que en la gran mayoría de los casos, una suposición importante para usar cualquiera de estas técnicas es que sus datos faltan completamente al azar (MCAR), lo que quiere decir que no existe una relación de causalidad que relacione un valor omitido, con alguna variable u otra observación, en otras palabras, el valor omitido depende únicamente de la observación a la que pertenece.

- **Imputación:** Si lo que se busca es manejar estos valores sin incurrir en la perdida de información, puede resultar mas pertinente utilizar los denominados métodos de 'imputación', cuyo funcionamiento se traduce básicamente en tratar de predecir que valores podrían tomar aquellas variables, que para una determinada observación presentan un *missing value*, esto a través de la generación de modelos de regresión lineal,

empleando diferentes estadísticos acordes a la naturaleza del problema y de las mismas características, como lo son por ejemplo la 'media', 'mediana', 'moda', etc. En pocas palabras para una determinada observación, predecir los valores de aquellas variables que presentan *missing values* a partir de aquellas variables que no lo hacen.

Para esto se emplean distintos paquetes de R, con el objetivo de comparar los resultados finales.

1. *mice:* A través de la función "mice"se pueden generar distintos conjuntos de datos a partir de un conjunto de datos inicial, estos conjuntos generados difieren precisamente en los valores que son imputados para cada missing value. Para usarlo es necesario que al igual como sucedía con la .Eliminación", los datos sigan una "MAR". En esta ocasión, se generan 5 conjuntos distintos y los resultados se presentan a continuación.
2. *missForest:* Otra de las formas de realizar imputaciones, consiste en implementar un método basado en el algoritmo de Random Forest", aplicable para datos no perimétricos, osea aquellos que no siguen una distribución normal. Para esto se crea un modelo de Random Forest para cada variable, para predecir los missing values de la variable basados en los valores observados.
3. *Hmisc:* Hmisc es otro de los paquetes que permite manejar missing values, a través de la imputación de sus valores, a través de una de sus funciones principales impute(), esta simplemente imputa el missing value utilizando algún método estadístico a elección, como media, máximo, etc.

A continuación se muestran los valores imputados para la variable *bareNuclei*, para cada una de las 16 observaciones respectivas.

Observación	lw	m1	m2	m3	m4	m5
1	X	3	9	10	10	10
2	X	10	10	10	10	10
3	X	1	1	1	1	1
4	X	1	1	1	1	1
5	X	1	1	1	1	1
6	X	1	1	1	1	1
7	X	1	1	4	1	1
8	X	1	1	1	2	3
9	X	1	1	1	1	1
10	X	10	2	10	3	10
11	X	1	2	1	1	1
12	X	1	1	1	1	1
13	X	1	6	7	6	1
14	X	1	1	1	2	1
15	X	1	1	1	1	1
16	X	1	1	1	1	1

Cuadro 3: Valores Imputados I

- **lw (Listwise):** Referido al método *Listwise Deletion*, como se puede observar el símbolo 'X' denota que dicho valor ha sido eliminado del conjunto de datos, lo que desde luego incluye toda la observación que contiene el *missing value*.
- **m (mice):** Corresponden a los 5 conjuntos generados de la imputación utilizando el paquete *mice*, por lo que cada conjunto es independiente uno del otro, por lo que los valores si bien coinciden en algunos valores, son diferentes.

Observación	mf	H1	H2	H3	H4	H5
1	7	4	5	1	1	10
2	9	4	1	1	1	10
3	1	4	2	1	1	10
4	1	4	1	1	1	10
5	2	4	8	1	1	10
6	1	4	10	1	1	10
7	1	4	5	1	1	10
8	1	4	1	1	1	10
9	1	4	1	1	1	10
10	5	4	5	1	1	10
11	1	4	1	1	1	10
12	5	4	1	1	1	10
13	7	4	1	1	1	10
14	1	4	1	1	1	10
15	1	4	1	1	1	10
16	1	4	1	1	1	10

Cuadro 4: Valores Imputados II

- **mf (missForest):** Relativo al método que emplea el algoritmo de *Random Forest*, dado que la naturaleza de los datos en si es una especie de escala que mide porcentajes de las distintas características y estos son representados por números enteros, los valores imputados fueron todos redondeados para acercarlos a su valor entero.

- **H (Hmisc):** Corresponden a los conjuntos generados por la imputación provista por el paquete *Hmisc*, empleando en cada una de estas como estadístico, la **media (H1)**, un valor **aleatorio (H2)**, la **mediana (H3)**, el **mínimo (H4)** y el **máximo (H5)** de los datos, en ese orden respectivo. En el caso de la media también se aplica redondeo.

Considerando la gran cantidad de alternativas generadas anteriormente, surge la incógnita de cual de estas resulta mas conveniente utilizar de aquí en adelante, para responder esta pregunta se consideran las siguientes acotaciones:

1. En primer, lugar considerando la naturaleza del problema, no resulta muy conveniente emplear un método basado en la generación de valores aleatorios para reemplazar los *missing values* como es el caso del conjunto **H1**, el cual emplea una imputación -*andomizada"*, menos aun si el estudio tiene directa relación con la detección de células cancerosas, y del cual podría depender la salud de una persona, razón por la cual se descarta este conjunto.

2. Continuando con los conjuntos generados a partir de un estadístico en particular, como es el caso de los conjuntos **H1**, **H2**, **H3** y **H4** utilizando *media*, *mediana*, *mínimo* y *máximo* respectivamente, resultan mucho mas convenientes de utilizar que empleando un valor aleatorio, ya que a diferencia de este toman en cuenta las variables de las observaciones adyacentes, sin embargo pierden fuerza comparándolos con un método que emplea un modelo de regresión lineal, ya que este toma en cuanta el conjunto de datos en su totalidad para predecir un valor en particular, y ya que en este caso hay métodos que emplean esta metodología, al igual que el caso anterior se descartan.

3. Tal como se menciona anteriormente, los conjuntos generados a partir de modelos de regresión son mucho mas fiables, ya que estos tratan de predecir el valor de una variable tomando en cuenta las demás, ahora tomando en cuenta la naturaleza del estudio, tiene sentido que para un cierto tejido, una variable desconocida tenga un valor relativamente parecido al de la misma variable pero en otro tejido, si las variables adyacentes de ambos tejidos también son bastante parecidas, o al menos se distingue cierto patrón. En este caso aquellos conjuntos que cumplen esto, son los 5 conjuntos del paquete mice **m1**, **m2**, **m3**, **m4** y **m5**, junto con el conjunto del paquete *missForest*, **mf**.

4. Finalmente se podría considerar simplemente prescindir de aquellas observaciones que presentan *missing values*, siendo este el caso del conjunto **lw**. Por un lado, si se evalúa este método desde la perspectiva de la perdida de información, se tendería a pensar en

descartarlo inmediatamente, ya que bien como se menciono anteriormente el modelo pierde fuerza, sin embargo esta perdida corresponde solo a 16 observaciones de un total de 699, lo cual no representa mas que un 2.3 % de la totalidad de los datos. Si el volumen de datos hubiera sido mucho mas grande, así como el rango de variables afectadas, probablemente no habría discusión en emplear alguno de los conjuntos del punto anterior, pero como no es el caso, resulta mas pertinente considerar solo aquellas observaciones en las que la totalidad de sus variables son explicadas por mediciones reales, y no hechas en base a predicciones de estas. Por esta razón, el método *Listwise Deletion* y el conjunto de datos generado por este, es el que se considera para el estudio de aquí en adelante y contando ya no con 699 observaciones, sino que con 683.

3.3. Reducción de Dimensionalidad

Otro de los factores que se debe tener bastante en consideración a la hora de realizar cualquier tipo de estudio sobre un conjunto de datos, tiene relación con la utilidad de las variables de estudio, muchas veces puede existir el caso de que una variable no entrega suficiente información, o a lo largo de todas las observaciones posee mucha pérdida de información, o tal vez entrega la misma información que otra variable. Casos como estos puede producir que el análisis del conjunto de datos no se comporte como uno se lo espera, por lo que mantener este tipo de variables puede resultar bastante caro, y mas aun, si el orden de variables es bastante elevado en relación con el numero total de observaciones que se tiene.

La **correlación** es una medida de distancia muy útil para observar como es que se comportan las variables en relación a las demás, por lo que puede aportar información relevante a la hora de determinar si una variable es eliminada del conjunto de datos o no.

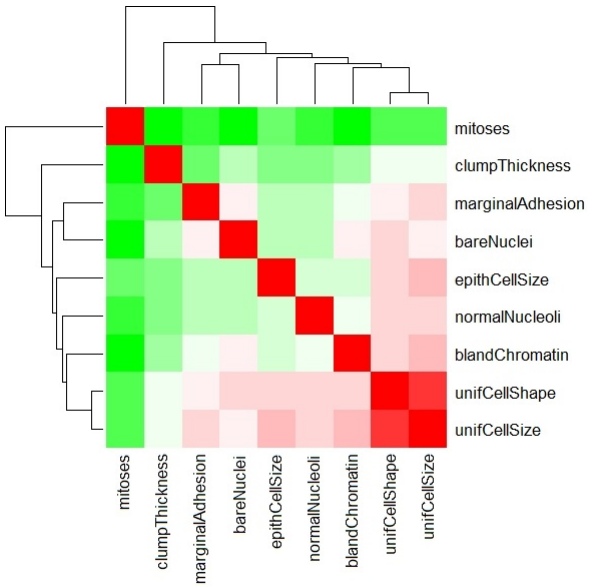

Figura 3: Mapa de Calor de Correlaciones

A simple vista se puede observar que los patrones para las variables *unifCellSize* y *unifCellShape* son iguales para algunas celdas de la matriz, y bastante similares en otros. Para observar mejor esta similitud, se muestran los valores numéricos en la siguiente tabla.

Efectivamente la relación de estas dos variables con las demás es bastante parecida, lo cual se evidencia en los valores de la matriz, por lo que se podría considerar eliminar alguna

Variable	unifCellSize	unifCellShape
clumpThickness	0.64	0.65
unifCellSize	1.00	1.91
unifCellShape	0.91	1.00
marginalAdhesion	0.71	0.69
epithCellSize	0.75	0.72
bareNuclei	0.69	0.71
blandChromatin	0.76	0.74
normalNucleoli	0.72	0.72
mitoses	0.46	0.44

Cuadro 5: Correlaciones de Tamaño y Forma de Células

de estas dos, dejando solamente una en el conjunto de datos. Sin embargo, hay que considerar lo que tratan de explicar estas variables, ya que por mas que numéricamente y en relación con las demás variables, se parezcan bastante, puede que la característica que magnifican no tenga relación alguna.

Analizando la naturaleza de ambas variables y su procedencia, por un lado se tiene a una variable que magnifica el tamaño de la célula de un tejido, y por otro una que magnifica la forma de esta. A primera impresión se podría decir que tanto tamaño como forma tienen una relación, sin embargo en el estudio del cáncer explican cosas que si bien tienen relación, no son lo mismo, una se utiliza de manera mas directa para detectar que la célula sufrió una mala especialización y que puede corresponder a una célula cancerígena, mientras que otra se utiliza mas para determinar ante que tipo de cáncer se esta presente, por lo que prescindir de alguna de estas, cualquiera que sea, significa perder información muy importante, razón por la cual ambas se conservan.

3.4. Tratamiento de valor atípicos

Dado que los valores de todas las variables son discretizaciones de la realidad acotadas del 1 al 10, no tiene sentido tratar los outliers, porque cada posible valor que pueda tener una variable representa un conjunto de posibles magnitudes que fueron clasificadas en dicho valor. Por ejemplo, si un dato tiene 8 en mitosis, que corresponde a un valor atípico, significa que el 80 % de la actividad mitótica presenta una actividad anormal. Además, si 8 fuese un valor atípico, anularía la capacidad que tienen los datos de alcanzar esa cifra. Entonces, eliminar o modificar ese dato significaría perder información que perjudicaría al modelo.

4. Clustering

Ya preparado el conjunto de datos gracias a la etapa de pre-procesamiento, se puede proceder a realizar el proceso de clustering, entonces a modo de retrospectiva el conjunto de datos consta de 683 observaciones de las cuales 16 fueron eliminadas, y explicadas a través de 9 variables.

4.1. Determinar distancia

De acuerdo al algoritmo descrito en el marco teórico, el primer paso del clustering *k-means* consiste en escoger un k óptimo, sin embargo varios de los métodos utilizados para lograr esto requieren de antemano una distancia especifica para funcionar correctamente. Y es que la distancia tanto para los métodos para determinar el k-óptimo, como para el clustering en si, es muy importante y tiene una fuerte influencia en el resultado de este, razón por la cual en primera instancia es necesario analizar las distancias disponibles para su uso, cuales son efectivamente aplicables dada la naturaleza de los datos, y luego cual de estas resulta mas conveniente utilizar.

En particular para este estudio se consideran tres distancias ya definidas anteriormente, y que son la distancia *Euclidiana*, *Manhattan* y *Gower*, y son las que se analizan a continuación.

Ventajas		
Euclidiana	Manhattan	Gower
(1) Fácil de calcular. (2) Trabaja bien con conjuntos de datos con cluster compactos. (3) Trabaja bien con variables continuas y escaladas.	(1) Trabaja bien con conjuntos de datos con cluster compactos.	(1) Util para para conjuntos de datos con variables categóricas y ordinales. (2) Util para cualquier tipo de distribución. (3) Más robusto

Cuadro 6: Principales Ventajas de Distancias

Desventajas		
Euclidiana	Manhattan	Gower
(1) Sensible a *outliers*. (2) Si no se tiene una distribución normal, tiende a agrupar mal.	(1) Sensible a *outliers*.	(1) Al estar basado en correlaciones, es mas algo mas complejo de calcular

Cuadro 7: Principales Desventajas de Distancias

Si se consideran las características del conjunto de datos, se pueden realizar una serie de conclusiones relativas a las distancias. Por un lado considerando que el conjunto de datos si cuenta con *outliers*, y que estos no fueron eliminados dada la naturaleza de los datos, emplear las distancias tanto *Euclidiana* como *Manhattan* tendrían una cierta desventaja respecto a la distancia de *Gower*. Por otro lado, utilizar la distancia *Euclidiana* considerando que la distribución de los datos se aleja bastante de lo que es una distribución normal, no resultaría muy pertinente y se incurre en el riesgo de realizar una mala agrupación de los datos, lo cual no sucede con las distancias de *Manhattan* y de *Gower*. En cuanto al tipo de variables, al contar con un conjunto de datos, en su totalidad de caracter discreto, tampoco seria conveniente emplear la distancia *Euclidiana*.

Finalmente, contrastando las características de los datos, y las ventajas y desventajas para cada una de las distancias, la distancia que se considera como mas adecuada para proseguir con el proceso de clustering es la distancia de *Gower*.

4.2. Determinar k óptimo

Existe una relación entre la cantidad de clusters y el valor de la WCSS (Within cluster sum of squares) del modelo, y es que generalmente un mayor número de grupos implica que el WCSS sea menor. Ahora bien, el fin mismo del clustering consta en agrupar un conjunto de datos inicial, pero manteniendo un número adecuado de grupos, ya que emplear el algoritmo con un k muy elevado provocaría que este pierda el sentido. Entonces, resulta necesario mantener un cierto equilibrio entre el número de grupos formados, y la WCSS.

Para resolver esta incógnita de cual es el k mas conveniente para agrupar los datos, se analiza el comportamiento del conjunto de datos a partir de tres métodos ya mencionados anteriormente, y que son el *Método del Codo*, el *Método de la Silueta* y el *Método de la Brecha Estadística*. Los resultados obtenidos se muestran a continuación.

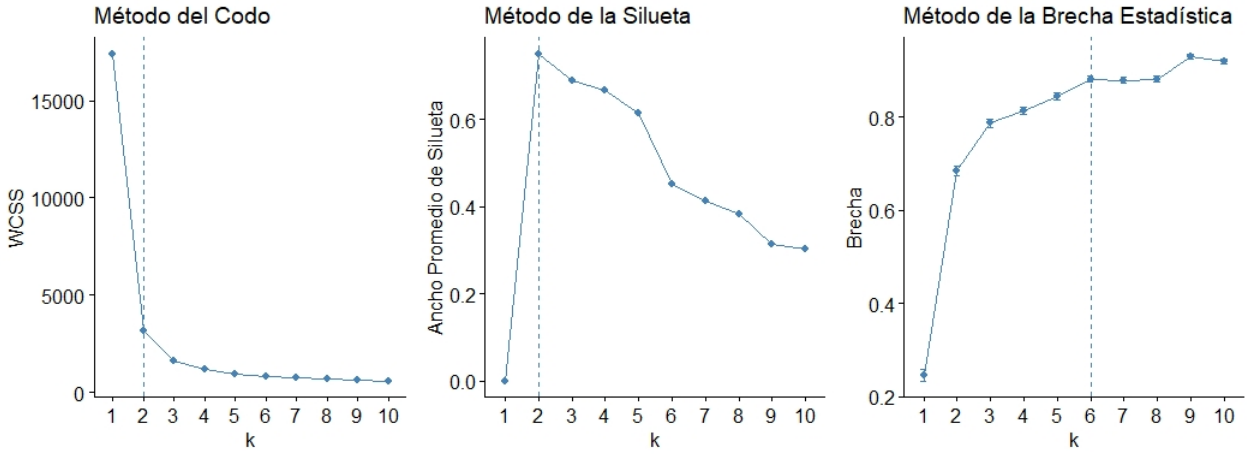

Figura 4: Resultados de los métodos para encontrar el k óptimo

- Gráfico Codo: Como es posible observar, en la medida que el numero de centroides se incrementa, la WCSS cae de manera brusca entre los k 1 y 2, mientras que para los demás lo hace de una forma mucho mas pausada, dándole a la curva la forma de un çodo". Para escoger el valor óptimo de k, la idea consiste en encontrar el punto para el cual la WCSS ya no sufre variaciones significantes al aumentar el k, lo que significa que la WCSS ganada no es suficientemente significante en comparación con aumentar el k una unidad. Finalmente, para este método el valor óptimo para k es 2, tomando en consideración el valor 3 igualmente.

- Gráfico Silueta: Para este caso basta con encontrar el valor de k, para el cual el ancho de la silueta promedio se maximiza, lo cual fácilmente se puede observar para k igual a 2. Finalmente, para este método, al igual que el anterior el valor para k resulta ser óptimo es 2.

- Gráfico Brecha: Para este ultimo método, simplemente basta con fijarse en el valor de k, para el cual el valor de la brecha se maximiza, en este caso particular se especifica un máximo local, razón por la cual apenas la curva tiende al descenso inmediatamente selecciona el valor k anterior como el óptimo. Finalmente el valor óptimo para k según este método es el valor 6.

En retrospectiva, recabando lo arrojado por los gráficos anteriores, se tiene lo siguiente.

Metodo	k
Codo	2 y 3
Silueta	2
Brecha Estadística	6

Cuadro 8: Resultados para k

De acuerdo a la tabla anterior, el valor para k que convendría mas utilizar es el valor 2, debido a un consenso entre los métodos del *Codo* y de la *Silueta*, ya que el ultimo método se aleja bastante de estos valores, aun así, para efectos de prueba y comparación, se toman en cuenta un rango de valores para k desde el 2 hasta el 6, para tener mayo seguridad de que el numero final de clusters es el óptimo, y tal vez descubrir algún comportamiento oculto dentro de estos valores.

4.3. Aplicación algoritmo k-means

Para la aplicación del algoritmo *k-means*, se consideran los siguiente parámetros, expresados en la notación del entorno R.

- k: El numero de centroides o grupos a formar al final del algoritmo. Se emplea un rango de 5 valores para k, que van **desde 2 hasta 6**.
- nstart: El numero de veces que se re-muestrea la configuración inicial de asignaciones de puntos, con los respectivos centroides. Se emplea un único valor **25**, en la practica no se logra verificar una diferencia notable en la asignación de grupos, por eso no se comparan en esta ocasión.
- algorithm: El algoritmo por defecto para las *k-means* es el basado en **Hartigan-Wong**
- iter.max: El número máximo de iteraciones que emplea el algoritmo, en la asignación de clusters y centroides. Por defecto se usa el valor **10**.

4.3.1. Analisis Gráfico

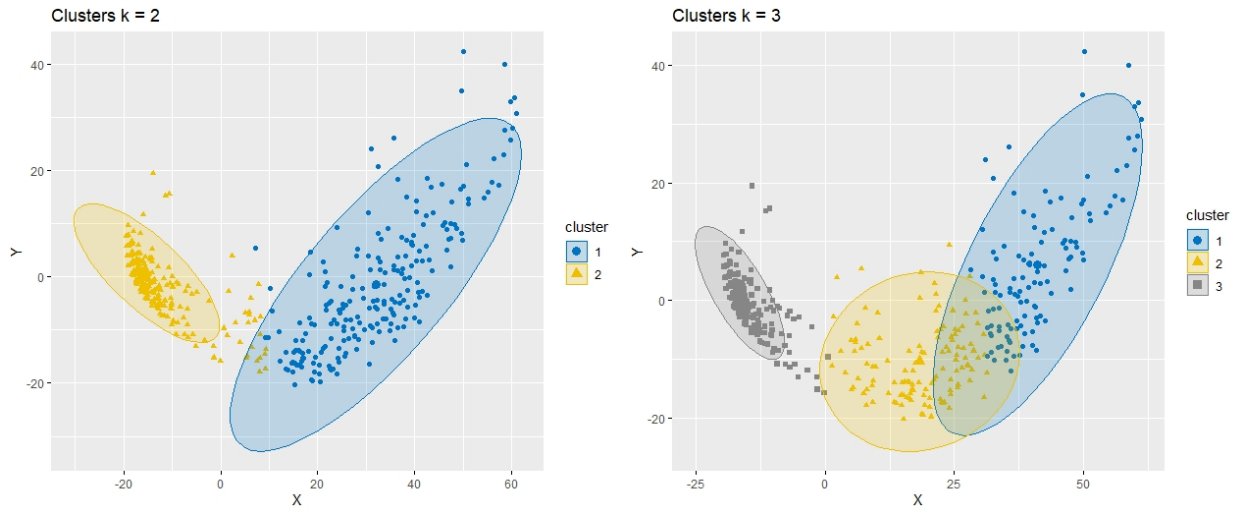

Figura 5: Clustering 2 y 3 grupos

Comenzando por los resultados del clustering de 2 grupos, se puede fácilmente identificar cada uno de estos, y a pesar de que la distancia no es muy grande, no existe una superposición

entre estos permitiendo identificar a que cluster pertenece cada observación. Para 3 grupos, pareciese que el cluster de la izquierda permanece intacto, mientras que el cluster de la derecha es el que se divido en dos nuevos clusters, y a diferencia del caso para 2 grupos, estos si se sobreponen dificultando un poco el trabajo de identificar a que cluster pertenece cada observación, al menos visualmente.

Para ambos casos se puede observar la existencia de una reducida cantidad de outliers, nuevamente debido a que se decidio no eliminar estos.

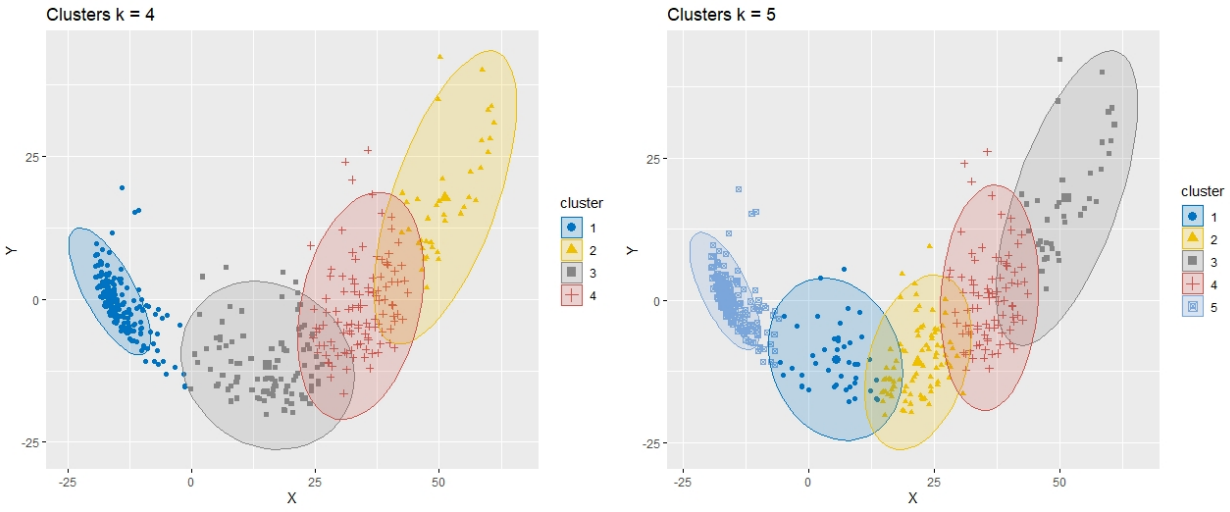

Figura 6: Clustering 4 y 5 grupos

Continuando con el clustering de 4 grupos, nuevamente pareciese como si el cluster de la izquierda permaneciese intacto y el de la derecha es el que se vuelve a reagrupar formando 3 grupos nuevos en este caso, donde al igual que el caso anterior también presenta una superposición entre las observaciones. Para el caso de 5 grupos, pareciese como si el cluster central con forma mas esférica es el que se divide para formar dos nuevos cluster y llegar a un total de 5 grupos, y da la impresión de que comienzan a acercarse mas al cluster de la izquierda.

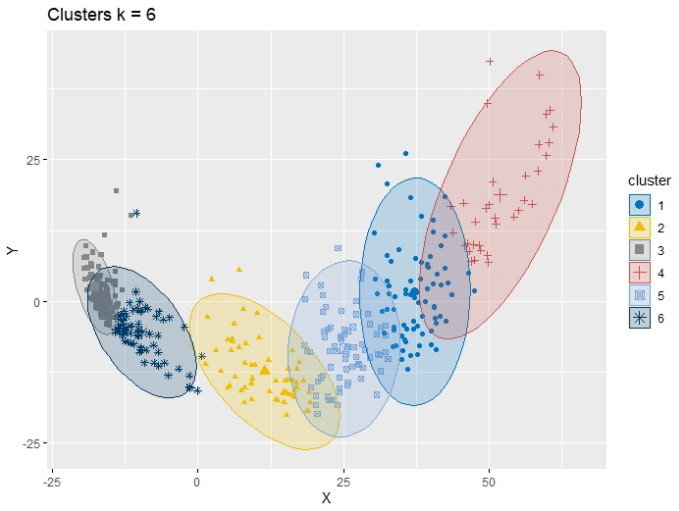

Figura 7: Clustering 6 grupos

Para 6 grupos finalmente se puede observar una superposición, de al menos una porción de áreas, entre todos los grupos formados.

En cuanto a la distribución del numero de observaciones para cada clusterización, se muestra a continuación.

k	C1	C2	C3	C4	C5	C6
2	221	462	X	X	X	X
3	120	125	438	X	X	X
4	436	38	99	110	X	X
5	44	85	37	89	428	X
6	79	59	371	34	73	67

Cuadro 9: Distribución de los *Cluster*

4.3.2. Análisis de Calidad

La calidad de una partición se puede determinar calculando el *Porcentaje de TTS* explicado por la partición, a través de la siguiente formula.

$$\frac{BSS}{TSS} x 100\,\% \tag{5}$$

Donde *BSS* y *TSS* representan *Between Sum of Squares* y *Total Sum of Squares* rspectivamente. Un porcentaje mas alto, significa un mejor *score*, y por ende una mejor calidad del *clsutering*, lo cual significa que la *BSS* es mucho mas grande, o bien que la *TSS* es mucho mas pequeña que la otra.

k	BBS	TSS	BBS/TSS x 100 %
2	14210	17393	81.7 %
3	15758	17393	90.6 %
4	16215	17393	93.2 %
5	16445	17393	94.5 %
6	16569	17393	95.3 %

Cuadro 10: BBS y TSS

Como se puede observar, en la medida en que el numero de clusters se incrementa, el valor de *BBS* lo hace también, mientras que la *TSS* se mantiene invariante, produciendo que el porcentaje de calidad aumenta. Partiendo del hecho de que un 80.7 % de calidad es una buena medida, pasando desde k igual a 2 hasta k igual a 3 existe un importante aumento de un 8.9 %, lo que justifica en cierta medida lo expuesto por el *Método del Codo*. Por otro lado, pasando de los 3 clusters en adelante el aumento ya no es tan significante.

A continuación se pueden ver también, la suma de los cuadrados para cada uno de los clusters formados para cada valor de k.

k	C1	C2	C3	C4	C5	C6
2	2262	922	X	X	X	X
3	740	499	396	X	X	X
4	370	141	329	338	X	X
5	101	192	135	223	297	X
6	191	129	146	119	162	77

Cuadro 11: WCSS

Finalmente, considerando el análisis gráfico en primera instancia y el análisis de calidad en segunda, se puede llegar a una mejor conclusión respecto a cual de las configuraciones anteriores es mejor para un determinado valor para k. Para el análisis gráfico, el clustering para k igual a 2 arroja mejores resultados, ya que se puede ver una mejor separación de los grupos generados, lo que denota una mayor distancia entre estos. En segundo lugar, respecto a la calidad de los clusters, tanto la primera como la segunda configuración son buenas opciones.

5. Analisis de Resultados

Ahora se procede a interpretar los clusterings realizados para poder generar conclusiones a partir de la información que podamos extraer. Para hacer esto, se estudia la relación que tienen los resultados con la variables class y se estudian las medias y medianas de cada varaible para cada cluster

5.1. Comparación con la variable class

Dado que en el preprocesamiento se decidió descartar la variable class, que define si un tumor corresponde a uno benigno o canceroso, se procede a utilizar esa información para evaluar si las agrupaciones ayudan a detectar la naturaleza de los tumores.

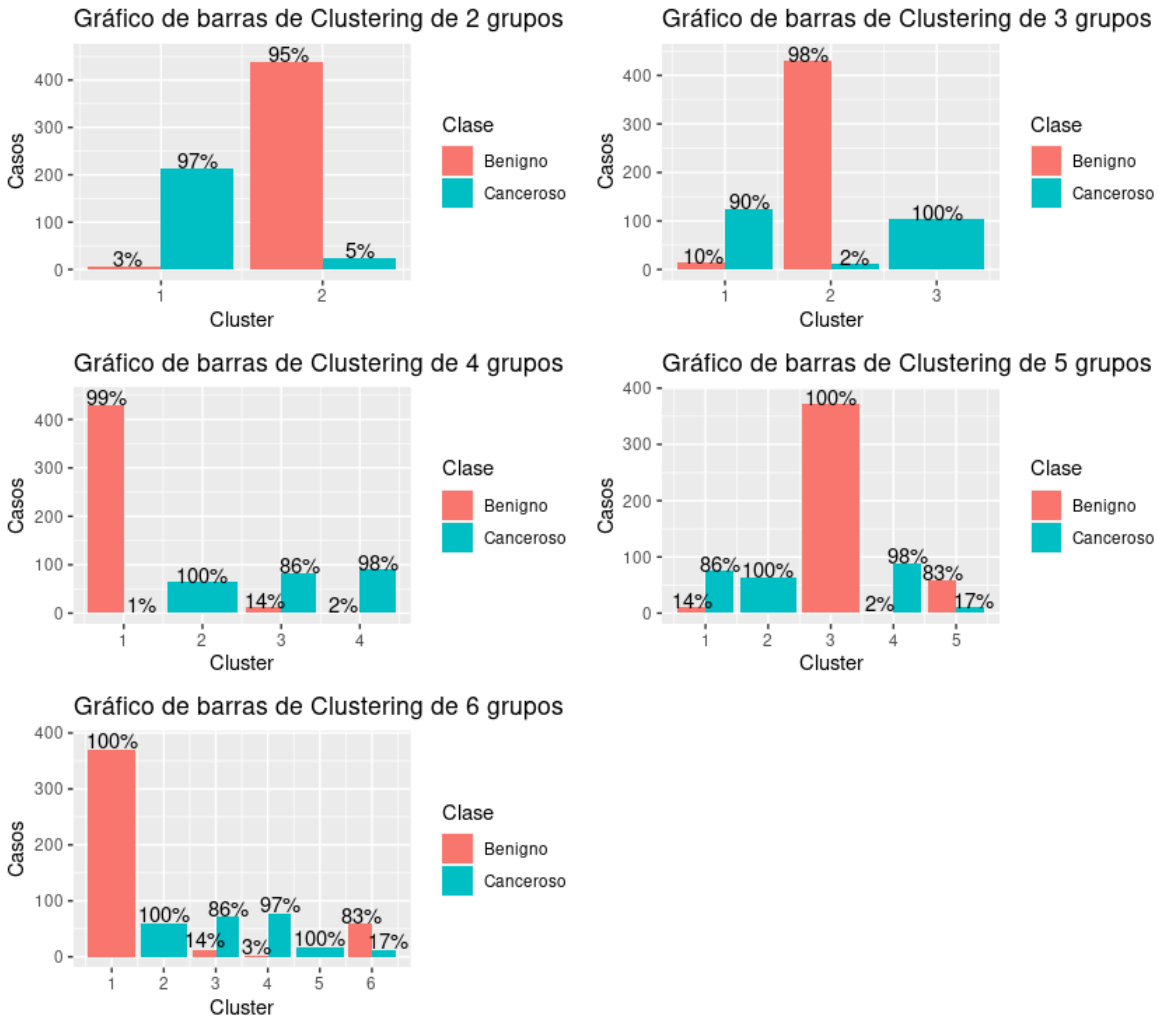

Figura 8: Gráficos de barra para contrastar con la variable class

Como se observa en la figura 8, cada agrupación tiende a englobar un tipo de tumor (benigno o canceroso). Esto ocurre para los seis casos de estudio. Sin embargo, se nota que para los grupos de 4, 5 y 6 clusters, existen agrupaciones que no tienen esta tendencia tan marcada como en los demás casos. Además para los clusterings con un k inicial a partir de 3, existen grupos que solo contienen un tipo de tumor.

Si se considera que cada grupo formado debería contener tumores de un solo tipo y si pasara lo contrario, serían errores del modelo, los errores presentados en cada uno serían 32, 25, 17, 26 y 26 respectivamente. Los cluster que realizaron mas subconjuntos fuueron los que erraron menos, en específico, el cluster con k = 4 fue el que erró menos. Sin embargo, esta diferencia no de errores no es significante en comparación con la cantidad de datos que se

estudia. Además, a partir del cluster con k = 4, está el subgrupo que presenta mayor error, lo que hace pensar que en esa área se encuentran los tumores con valores atípicos.

5.2. Medias y medianas de los clusters

Dado que ninguna variable sigue una distribución normal, también se consideró la mediana para realizar este análisis.

Para continuar, hay que recordar que las variables están discretizadas de tal forma que un menor valor significaría que la célula presenta características normales, es decir, no es cancerosa.

A continuación se estudia el cluster seleccionado, el que realizó dos subgrupos. Las métricas a evaluar se encuentras en la siguiente tabla.

Variable	$\overline{x}_{C1}$	$\overline{x}_{C2}$	ME_{C1}	ME_{C2}
clumpThickness	7.21	3.12	8	3
unifCellSize	6.96	1.33	7	1
unifCellShape	6.86	1.47	7	1
marginalAdhesion	5.87	1.37	6	1
epithCellSize	5.53	2.14	5	2
bareNuclei	4.84	2.44	3	2
blandChromatin	6.14	2.15	7	2
normalNucleoli	6.22	1.27	7	1
mitoses	2.69	1.08	1	1

Cuadro 12: Medias y medianas de los clusters para k = 2

En general, los valores de la mitosis suelen ser bastante bajos, esto se explica, porque los datos no presentan anomalías en su actividad mitótica, esto se puede comprobar revisando los histogramas presentados anteriormente.

La proporción de núcleos rodeados con citoplasma y el tamaño de las células epiteliales no alcanza valores tan altos, lo que da a entender que las células cancerígenas no necesariamente tienen estos indicadores tan altos.

En cuanto al cluster C_1, las métricas de tendencia central son notoriamente mas altas, lo que permite pensar que las células cancerosas fueron agrupadas en este conjunto. Por el contrario, para el cluster C_2, la mayoría de las medianas tienen el valor mínimo, lo que indica que engloba los datos de los tumores no cancerosos.

Las variables que presentaron mayor diferencia en las medidas de tendencia central fueron *normalNucleoi* o la tendencia de los nucleolos a presentar citoplasma, *unifCellShape* o la tendencia de la células a presentar deformidades y *unifCellSize* o la tendencia de las células a presentar anomalías con su tamaño. Corresponden a variables discretas del 1 al 10, donde 1 significa que la característica no representa anomalías y 10 que el 100 % de las células de una muestra de tumor presentan anomalías en esa característica.

Una gran variación en la mediana en una variable significa que la característica que representa corresponde a un factor a considerar en el estudio de células cancerígenas. En específico, para estas tres variables, la mitad de las muestras de tumores que resultaron ser cancerosos presentaron por lo menos un 70 % de células con anomalías en su forma, tamaño y/o nucleolo.

Estos resultados tienen sentido, porque las células cancerosas, desde su fase temprana, presentan mutaciones en su material genético, lo que explicaría alteraciones en el nucleolo. Además, lo que diferencia un tumor canceroso de uno que no lo es, es su capacidad de generar metástasis. Esto es importante, porque la forma y tamaño de una célula es usada para identificar la capacidad que tiene una célula de desprenderse y viajar por el sistema circulatorio o linfático. Incluso existen tratamientos para manejar estos parámetros para evitar la propagación del cáncer.

6. Conclusión

A modo de retrospectiva y considerando el objetivo general de esta experiencia, se logra realizar una implementación del algoritmo de clustering *k-means* aplicado al conjunto de datos *Breast Cancer Wisconsin*, analizando distintos factores que influyen en la calidad de la agrupacion final, llegabdo a las siguientes conclusiones.

- La razón de ser de este estudio, tiene principal relación con el caracter **no supervisado** del algoritmo *k-means*, pues se busca determinar si las observaciones a través de sus variables pueden de alguna forma entregar información, que permita explicar algún fenómeno, en este caso considerando la naturaleza del problema, determinar que aquellas observaciones que comparten un cierto rango de valores para sus variables, tienden a relacionarse bastante y tienen una menor distancia, respecto a aquellas observaciones cuyas variables están fuera de esos rangos.

- De la mano de lo anterior, y considerando los resultados obtenidos, se logra aplicar el método para un determinado rango de valores k, que corresponden al numero de grupos finales a formar, y comparar los resultados de cada uno, identificando en base a determinadas métricas, tales como la calidad en base a la BSS y TSS, cual de estas configuraciones representan un mejor agrupamiento de datos. Lo anterior apoyado también por el análisis de los gráficos generados, comparando la distancia entre los cluster para cada configuración, así como la superposición de estos.

- Las configuraciones generadas además de depender del valor de k, dependen de todo un pre-procesamiento de los datos, considerando el tratamiento de los *missing values*, donde dependiendo de la naturaleza del problema se pueden tratar de una forma u otra, como el conjunto de datos se basa en un estudio de diagnostico de células cancerosas, prescindir o tratar de predecir el valor de estos valores puede no ser algo que se tome tan a la ligera, por lo que dependiendo de la magnitud de la perdida de información, se puede tomar una opción o la otra.

- La distribución de los datos también juega un papel fundamental, e influye directamente sobre la métrica de distancia mas conveniente de utilizar, ya que utilizar una métrica con

una distribución para la cual no esta hecha, puede resultar en una mala agrupación de los datos. Llevando esto al campo de investigación del problema, una mala *clusterización* puede resultar en un falso diagnostico para un paciente que se realiza un examen para determinar la existencia de células cancerosas en su organismo.

- Conocer de antemano la variable explicativa de los datos, como es el caso de la variable 'class', puede influenciar en cierta forma la interpretación de los resultados finales, ya que en el proceso se puede esperar que los resultados arrojen el numero de clases provistos por el conjunto de datos. Sin embargo, este proceso puede ayudar también a encontrar comportamientos desconocidos en los datos, como una nueva clase.

- Finalmente, contrastando los clusters asignados a cada una de las observaciones, y la clase a la cual pertenecen, es posible evidenciar que gran parte de estas están asignadas a la clase correspondiente y solo un bajo porcentaje falla en la asignación, por lo que se puede concluir que al menos para el caso que es comprobable, que es con 2 cluster, el algoritmo *k-means* explica lo que sucede en la realidad.

En cuanto a lo que queda pendiente se puede considerar tal vez realizar una comparativa aun mas exhaustiva en cuanto a las distancias utilizadas para el proceso de clustering, con el objetivo de visualizar de mejor manera la diferencia entre utilizar una métrica y otra, así también como una mejor comparación entre los métodos para tratar los *missing values*.z

Bibliografía

[enu]

[2] Borges, L. (2015). Analysis of the wisconsin breast cancer dataset and machine learning for breast cancer detection.

[3] Institute, N. C. (2020a). Breast cancer—patient version. [Online] https://www.cancer.gov/espanol/tipos/seno.

[4] Institute, N. C. (2020b). Cancer stat facts: Female breast cancer. [Online] https://seer.cancer.gov/statfacts/html/breast.html.

[5] Institute, N. C. (2020c). ¿qué es el cáncer? [Online] https://www.cancer.gov/espanol/cancer/naturaleza/que-es.

[6] MERZOUKI, R. (2017). User manualbreast cancer diagnosis web user interface. [Online] https://www.rai-light.com/docs/BCD_User_Manual_v01.pdf.

[7] Salom, E. V. (2017). ¿qué son las células epiteliales? [Online] https://cienciatoday.com/que-son-celulas-epiteliales/.

[8] Society, A. C. (2020a). Tipos de cáncer de seno. [Online] https://www.cancer.org/es/cancer/cancer-de-seno/comprension-de-un-diagnostico-de-cancer-de-seno/tipos-de-cancer-de-seno.html.

[9] Society, A. C. (2020b). What is cancer? [Online] https://www.cancer.org/cancer/cancer-basics/what-is-cancer.html.

[10] Society, A. C. (2020c). What is cancer? [Online] https://medlineplus.gov/spanish/cancer.html.

[11] UCI (2020). Breast cancer wisconsin (original) data set. [Online] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original).