**5. Resultados y discusión**

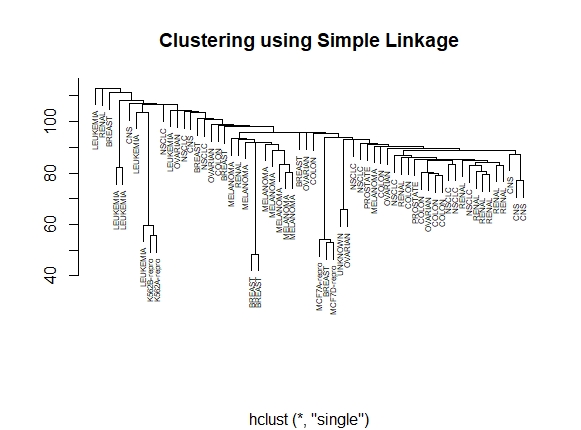
## 5.1 Búsqueda de grupos homogéneos (“class discovery”)

Tal y como se mencionó anteriormente, estudiar el conjunto de datos NCI60 el seleccionado para aplicar los métodos de cluster jerárquico y cluster de k-medias con el objetivo de buscar grupos homogéneos.

Al contrario que en los métodos utilizados en la predicción de clases, los métodos cluster son técnicas del aprendizaje no supervisado y, por tanto, no es necesario dividir la muestra en un conjunto de entrenamiento y un conjunto de test, porque no vamos a poder comparar los resultados obtenidos, debido a que en el aprendizaje no supervisado se trabaja con datos sin etiquetar, es decir, para la elaboración de los clusters trabajaremos con el conjunto de datos completo.

### 5.1.1 Clustering jerárquico

Se comenzó aplicando el método de clustering jerárquico, ya que no requiere determinar el número de clusters que se van a construir. Fue el método aglomerativo el seleccionado para construir los clusters, utilizando los 3 criterios de enlace diferente (Figuras 7, 8 y 9)

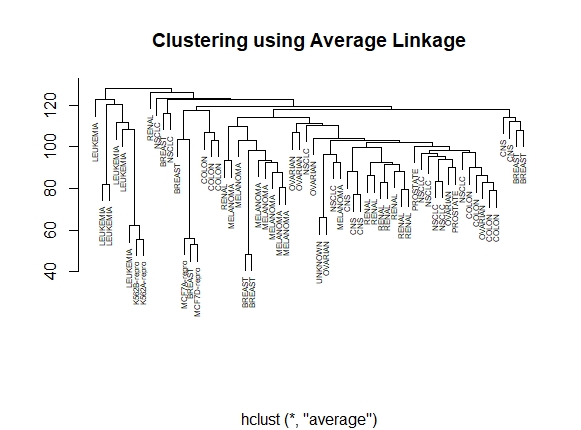


**Figura 7**. Clustering jerárquico utilizando agrupamiento de enlace simple.

**Fuente**: Elaboración propia.

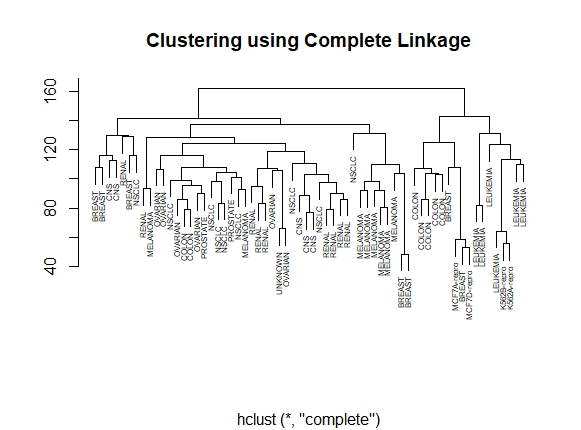
**Figura 8**. Clustering jerárquico utilizando agrupamiento de enlace promedio.

**Fuente**: Elaboración propia.



**Figura 9**. Clustering jerárquico utilizando agrupamiento de enlace completo.

**Fuente**: Elaboración propia.

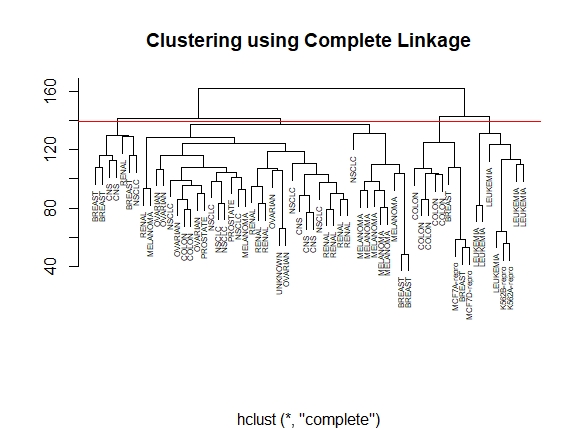


Se puede observar como en función del criterio de enlace seleccionado, los resultados cambian considerablemente. Tal y como se explicó en la metodología, es el enlace completo el que tiende a crear grupos más compactos, y es por ello, que es el que utilizaremos para continuar con los análisis.

Si extendemos una línea horizontal a la altura 139 (Figura 10), observamos como el dendrograma se divide en 4 clusters, y este es el valor de clusters que vamos a utilizar. Hay que considerar, antes de continuar, que se utilizan las etiquetas de los datos única y exclusivamente para poder alcanzar unas conclusiones más fácilmente interpretables, sin embargo, no habría que tenerlas en cuenta ya que los métodos clusters forman parte del aprendizaje no supervisado, en el que se trabaja con datos sin etiquetar.

**Figura 10**. Clustering jerárquico con dividido en 4 clusters. La línea horizontal roja presente en la altura 139, marca el punto en el que el dendrograma se divide en 4 clusters.

**Fuente**: Elaboración propia.



Es lógico que si seleccionaran más de 4 clusters los resultados obtenidos serían más precisos, pero no por ello más adecuados. El clustering jerárquico aglomerativo comienza con tantos clusters como individuos, lo cual significaría una precisión en la agrupación del 100%, sin embargo, claramente no tendría ningún sentido utilizar un método de clustering para crear tantos clusters como datos hubiera en nuestro estudio. Es por ello, que hay que seleccionar el número de clusters más bajo que genere resultados lógicos, para facilitar la interpretación.

A continuación, se muestra la tabla (Tabla 1) en la que se puede observar como de eficaz ha sido la predicción en la elaboración de los clusters, comparando los resultados de estos, con la clasificación original de los datos.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | **BREAST** | **CNS** | **COLON** | **K562A-repro** | **K562B-repro** | **LEUKEMIA** | **MCF7A-repro** |
| **1** | 2 | 3 | 2 | 0 | 0 | 0 | 0 |
| **2** | 3 | 2 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 1 | 1 | 6 | 0 |
| **4** | 2 | 0 | 5 | 0 | 0 | 0 | 1 |
| Cluster | **MCF7D-repro** | **MELANOMA** | **NSCLC** | **OVARIAN** | **PROSTATE** | **RENAL** | **UNKNOWN** |
| **1** | 0 | 8 | 8 | 6 | 2 | 8 | 1 |
| **2** | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| **3** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

Se puede observar como todas las líneas celulares pertenecientes a la leucemia, al cáncer de ovario, al melanoma y al cáncer de próstata están bien agrupadas en los cluster 3 y 1 respectivamente. Las líneas celulalres pertenecientes al cáncer de pulmón y al de riñón, han quedado prácticamente bien agrupadas, ya que tan solo una de las líneas celulares se encuentra clasificada en un cluster diferente. Por último, las líneas celulares propias del cáncer de mama, del cáncer cerebral y de colon no se han agrupado de manera correcta, ya que las líneas celulares se encuentran repartidas entre los diferentes clusters.

**Tabla 1**. Agrupación de los datos en clusters según clustering jerárquico.

**Fuente**: Elaboración propia.

### 5.1.2 Clustering de k-medias

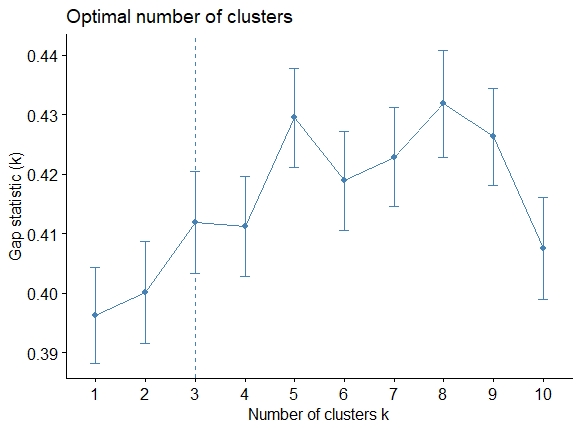
El principal inconveniente de este método es que es necesario determinar el número de clusters que se van a generar. Existen numerosos métodos para estimar el número de clusters que sería más adecuado para nuestros datos, y en este caso concreto se utilizó “gap statistic” (Mohajer et al., n.d.) (Figura 11). El resultado obtenido fue que 3 clusters sería lo óptimo para agrupar todos los datos mediante el método de k-medias.

“Gap statistic” únicamente nos ayuda a determinar cuál es el número óptimo de clusters, sin embargo, eso no quiere decir que obligatoriamente haya que utilizar de manera estricta el número de clusters determinada por el algoritmo. Como se puede observar en la gráfica (Figura 11), la diferencia que existe con respecto a agrupar los datos en 4 clusters es mínima, y esto nos permitiría comparar los resultados obtenidos con el clustering jerárquico, ya que en este también se agruparon los datos en 4 clusters. Es por esta razón por la que se finalmente se decidió agrupar los datos en 4 clusters, a pesar de que 3 clusters sería el número óptimo según la técnica de “gap statistic”.

Los resultados obtenidos tras la agrupación de los datos en clusters son los siguientes (Figura 12) (Tabla 2).

**Figura 11**. Gráfico resultante de aplicar el método “gap statistic”. Se puede observar como 3, es el número óptimo de clusters para nuestros datos.

**Fuente**: Elaboración propia.



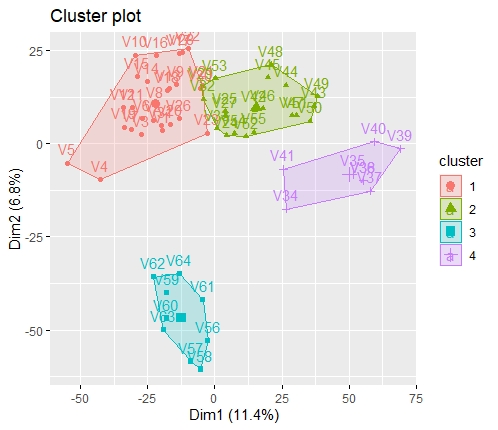
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | **BREAST** | **CNS** | **COLON** | **K562A-repro** | **K562B-repro** | **LEUKEMIA** | **MCF7A-repro** |
| **1** | 3 | 5 | 0 | 0 | 0 | 0 | 0 |
| **2** | 2 | 0 | 7 | 0 | 0 | 0 | 1 |
| **3** | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 1 | 1 | 6 | 0 |
| Cluster | **MCF7D-repro** | **MELANOMA** | **NSCLC** | **OVARIAN** | **PROSTATE** | **RENAL** | **UNKNOWN** |
| **1** | 0 | 1 | 4 | 3 | 1 | 9 | 1 |
| **2** | 1 | 0 | 5 | 3 | 1 | 0 | 0 |
| **3** | 0 | 7 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Tabla 2**. Agrupación de los datos en clusters según algoritmo de k-medias.

**Fuente**: Elaboración propia.

**Figura 12**. Agrupación de los datos en clusters según el algoritmo de k-medias.

**Fuente**: Elaboración propia.



Tal y como cabría esperar, las agrupaciones obtenidas cambian con respecto a lo obtenido en el clustering jerárquico. Líneas celulares que quedaban bien agrupadas en el caso del clustering jerárquico, como las del melanoma, cáncer de ovario y cáncer de próstata, no lo están con el k-medias, sin embargo, sí que hay algunas líneas celulalres que pasan estar bien agrupada mediante este método, y que no lo estaban con el clustering jerárquico, como son el cáncer de cerebro, el de colon y el de riñón.

Otra diferencia con respecto a los resultados obtenidos entre los dos métodos de clustering estudiados, es que las líneas celulalres de la leucemia y de K562A-repro y K562B-repro, a pesar de encontrarse bien agrupadas en los dos casos, lo hacen en clústeres diferentes, es decir, en el caso del clustering jerárquico se agrupan en el cluster 3 y en el caso del clustering de k-medias lo hacen en el cluster 4. Esto podría significar que lo que en el clustering jerárquico era el cluster 3, pasa a ser el cluster 4 mediante el método de clustering de k-medias.

En definitiva, se puede observar como ninguno de los dos métodos de clustering es tan preciso como cabría esperar, ya que lo que se encontraba bien agrupado en uno, pasa a estar mal agrupado en el otro y viceversa. Obviando la parte de que trabajar con conjuntos de datos reales dificulta la obtención de resultados muy precisos, el hecho de que los conjuntos de datos de expresión génica sean tan grandes, hace más difícil aún el poder obtener una muy buena clasificación/agrupación de los datos, mediante una técnica tan sencilla como son los métodos clusters. En este caso solo se busca ejemplificar las técnicas de clustering aplicándolo a un conjunto de datos reales, por ello los resultados no son del importantes. No cabe duda de que, si se tratara de un estudio en el que los resultados son fundamentales, lo obtenido mediante estos métodos de clustering no serviría para dar una conclusión definitiva, y debería de ser complementado con resultados obtenidos mediante otras técnicas de agrupación/clasificación más complejas y precisas.

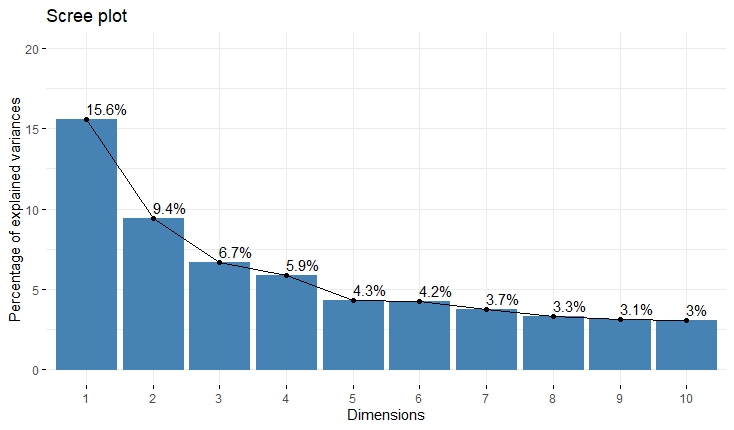
## 5.2 Predicción de clases (“class prediction”)

### 5.2.1 Regresión logística

Realizar una regresión logística con las 3051 variables que formaban parte del conjunto de datos no es apropiado, por lo que se realizó un ACP, con el objetivo de utilizar estas componentes como variables para la elaboración del modelo, más concretamente, las coordenadas de los individuos sobre las componentes.

El primer paso fue realizar un “scree plot” (Cattell, 1966) (Figura 13) para determinar el número de componentes principales que se iban a extraer. Un scree plot no es más que una representación de los valores propios de las componentes principales. Los valores propios de estas se ordenan de manera descendente, de modo que el número óptimo de componentes principales a extraer será el punto en el que se produce un “codo” en la gráfica, es decir, el punto a partir del cual, los valores propios de las componentes se mantienen más o menos constante, o descienden muy ligeramente. En este caso se observa como a partir de la componente principal 5, los valores comienzan a descender de manera muy ligera, es decir, 5 es el número óptimo de componentes principales a extraer de nuestro conjunto de datos.

Al observar la significancia de las 5 componentes extraídas, solo las dos primeras resultaron ser estadísticamente significativas, por ello, a pesar de que 5 fuera el número de componentes óptimas según el “scree plot”, se utilizaron solo las dos primeras para construir el modelo de regresión logística.



**Figura 13**. Scree plot. Esta gráfica nos permite determinar que 5 es el número óptimo de componentes principales a extraer para nuestro estudio.

**Fuente**: Elaboración propia.

En un principio, se recurrió a una regresión logística binaria estándar, como la explicada en la metodología, sin embargo, el modelo no convergía como consecuencia de la distribución de los datos, ya que había una separación perfecta de los mismos, por lo que hubo que optar por una regresión logística binaria penalizada, en concreto, se utilizó la penalización de Ridge (James et al., 2013). El valor de la penalización utilizado para la elaboración del modelo fue 0.2, que es el que usa por defecto en la función.

Los coeficientes obtenidos tras la realización del modelo fueron -1,597 para la ordenada en el origen y 3,024 y -2,245 para las variables 1 y 2 respectivamente (que corresponden con las componentes principales 1 y 2).

Uno de los valores más importantes a analizar, resultantes de una regresión logística son los OR. Los valores de estos OR fueron de 20,58 para la primera componente principal y de tan solo 0,11 para la segunda componente principal. En condiciones normales, en las que se trabaja con las variables originales, los OR nos permiten observar o analizar la relación que existe entre las variables, sin embargo, trabajando con componentes principales, que son combinación lineal de las variables originales, la interpretación no es la misma. La interpretación que se puede dar a los resultados obtenidos es que la combinación lineal de variables que constituye la primera componente principal, tiene una mayor capacidad para separar los grupos (LMA y LLA) que la segunda componente principal.

El último paso es comprobar el poder predictivo del modelo de regresión logística sobre nuestro conjunto de datos. En la siguiente tabla (Tabla 3), como quedan clasificados los datos en los grupos LMA y LLA utilizando el modelo de regresión logística

Predicción

Clasificación original

|  |  |  |
| --- | --- | --- |
|  | 0 (LMA) | 1(LLA) |
| 0 (LMA) | 26 | 1 |
| 1 (LLA) | 1 | 10 |

**Tabla 3**. Clasificación de los individuos obtenida tras regresión logística.

**Fuente**: Elaboración propia.

Tal y como se puede observar en la tabla, tan solo 1 individuo ha quedado mal clasificado tras la aplicación del modelo. A partir de la tabla hay dos valores que nos pueden aportar información importante sobre como de buena ha sido la predicción: la sensibilidad y la especificidad (Shalev-Shwartz & Ben-David, 2013). La sensibilidad, es la proporción de casos positivos que fueron bien clasificados por el algoritmo, mientras que la especificidad hace referencia a la proporción de verdaderos negativos bien clasificados por el algoritmo. Considerando verdaderos positivos a los casos de LMA clasificados correctamente y verdaderos negativos a los casos de LLA bien clasificados, los resultados obtenidos son:

Sensibilidad = 26/27 = 0,96

Especificidad = 10/11 = 0,91

Los resultados para la sensibilidad y la especificidad son de 0,96 y 0,91 respectivamente, es decir, se trata de valores altos (cercanos a 1), lo cual nos permite concluir que la predicción que ha realizado el modelo es realmente buena.

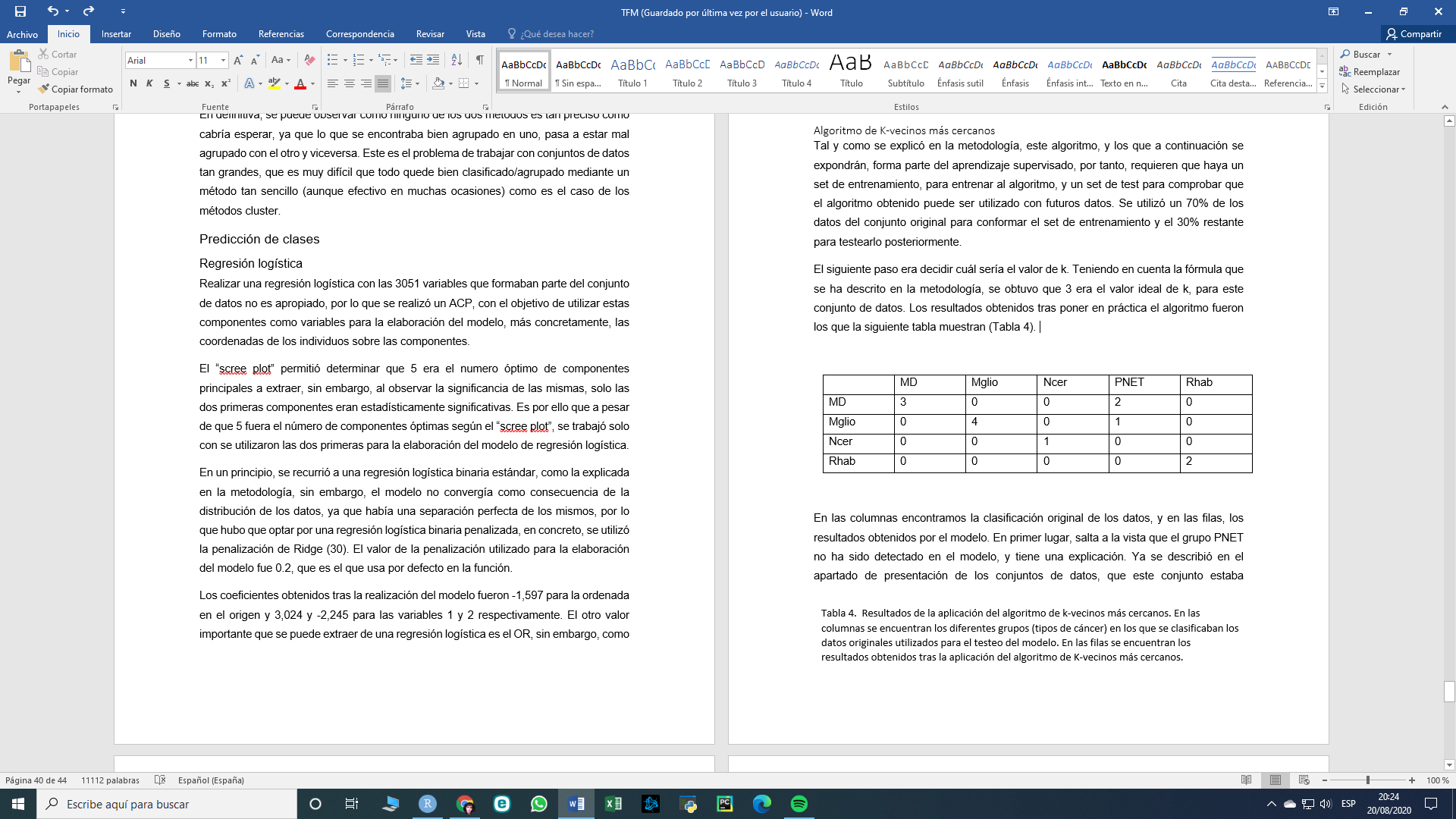
### 5.2.2 Algoritmo de k-vecinos más cercanos

El único paso necesario para aplicar el algoritmo de los k-vecinos más cercanos, es determinar cuál es el valor de k que se va a utilizar. Como se explicó en la metodología existen varias formas de estimar el valor de k, pero en este caso se utilizó la fórmula , lo cual permitió estimar que 3 sería el valor óptimo para el conjunto de datos con el que se estaba trabajando.

En la siguiente tabla (Tabla 4), se muestra la clasificación obtenida con el modelo (basándose en la clasificación original de los datos).

**Tabla 4**. Resultados de aplicar el algoritmo de k-vecinos más cercanos. Esta tabla muestra la predicción realizada por el algoritmo de k-vecinos más cercanos, para la clasificación de los datos.

**Fuente**: Elaboración propia.



El primer punto a resaltar, que salta a la vista al observar la tabla, es que el grupo PNET no ha sido detectado en la predicción que ha proporcionado el modelo, y tiene una explicación. Ya se describió en el apartado de presentación de los conjuntos de datos, que este conjunto estaba constituido por tan solo 42 individuos, que es un número bastante reducido, para además tener que dividir la muestra en un grupo de entrenamiento y uno de prueba. El valor seleccionado de k=3 nos proporciona la mejor clasificación posible (de todos los valores de k que se podrían utilizar) y aun así el grupo PNET no aparece en la predicción del modelo. También se intentó modificar los porcentajes de la muestra que constituirían el conjunto de entrenamiento y el conjunto de prueba, pero en ninguno de los casos aparecía el grupo PNET en la predicción (finalmente se utilizó un 70% de la muestra para construir el conjunto de entrenamiento). Este tipo de problemas se suele dar con mucha frecuencia en los estudios de datos de expresión génica, porque siempre se trabaja con miles de variables, pero un número de individuos considerablemente pequeño. Si observamos la tabla (Tabla 4), podemos comprobar que tanto los individuos pertenecientes a los grupos MD, Mglio, Ncer y Rhab han sido bien clasificados, sin embargo, los individuos que formaban parte del grupo PNET en los datos originales, han sido redistribuidos entre los grupos MD (2) y Mglio (1), como consecuencia de que el grupo PNET no ha sido detectado por el modelo.

### 5.2.3 Máquinas de vector soporte

El conjunto de datos utilizado para aplicar esté método fue el conjunto de datos del artículo publicado por Golub (Golub et al., 1999), debido a que los datos están clasificados en dos grupos (LMA y LLA), lo cual permitirá utilizar el algoritmo para generar una clasificación binaria. El primer paso, y probablemente el más problemático a la hora de aplicar esta técnica, en el caso de trabajar con un conjunto de datos similar al utilizado aquí, es el de dividir el conjunto de datos original en conjunto de entrenamiento y de prueba, debido a la reducida cantidad de individuos que presenta. En condiciones normales, en las que la muestra original tiene un número suficiente de individuos, se suelen utilizar valores que oscilan entre un 70 y un 80% del total de individuos del conjunto original para construir el conjunto de entrenamiento, y en este caso hubo que optar por el valor mínimo recomendado, que es un 70%, con el objetivo de obtener la mejor predicción posible.

El último paso para poder aplicar el algoritmo, era determinar el tipo de kernel que iba a ser utilizado. Se probaron los tres tipos de kernel que con más frecuencia se usan en este tipo de estudios, que son el kernel lineal, el kernel radial o gaussiano y el kernel polinomial. Las predicciones obtenidas por el modelo se muestran en las siguientes tablas (Tabla 5, 6 y 7).

Predicción

Clasificación original

|  |  |  |
| --- | --- | --- |
|  | 0 (LMA) | 1(LLA) |
| 0 (LMA) | 9 | 0 |
| 1 (LLA) | 0 | 2 |

# 

**Tabla 5**. Clasificación de los individuos utilizando un kernel lineal.

**Fuente**: Elaboración propia.

Predicción

Clasificación original

|  |  |  |
| --- | --- | --- |
|  | 0 (LMA) | 1(LLA) |
| 0 (LMA) | 9 | 1 |
| 1 (LLA) | 0 | 1 |

**Tabla 6**. Clasificación de los individuos utilizando un kernel radial.

**Fuente**: Elaboración propia.

Predicción

Clasificación original

|  |  |  |
| --- | --- | --- |
|  | 0 (LMA) | 1(LLA) |
| 0 (LMA) | 9 | 2 |
| 1 (LLA) | 0 | 0 |

**Tabla 7**. Clasificación de los individuos utilizando un kernel polinomial.

**Fuente**: Elaboración propia.

Si calculamos la sensibilidad y la especificidad, tal y como se hizo en las predicciones obtenidas mediante el modelo de regresión logística, para las tres tablas de clasificación obtenidas con cada uno de los kernels, poderemos hacernos una idea más precisa de como de buenas han sido cada una de las predicciones obtenidas mediante SVM.

Los valores de sensibilidad y especificidad para la clasificación obtenida mediante el kernel lineal son de 1 y 1 respectivamente, es decir, la predicción ha sido completamente perfecta.

Para las predicciones obtenidas mediante el uso de kernel radial y kernel polinomial, los valores de sensibilidad y especificidad han sido, como cabía esperar al observar las tablas, inferiores. En el caso del kernel radial, los resultados de sensibilidad y especificidad han sido de 1 y 0,5 respectivamente. En el caso del kernel polinomial, la sensibilidad ha sido de 1, mientras que la especificidad ha sido de 0, ya que el algoritmo a pesar de haber clasificado de manera adecuada todos los verdaderos positivos, ha sido incapaz de clasificar correctamente ningún verdadero negativo, probablemente debido a la reducida cantidad de individuos que constituían el conjunto de datos de prueba.

Analizando los resultados obtenidos mediante el cálculo de la sensibilidad y la especificidad de las predicciones, se puede concluir que el kernel lineal es el que mejor clasificación de los datos ha conseguido.

A pesar de que, según los resultados obtenidos, el kernel lineal parece ser el que mejore predicciones ha proporcionado, es necesario aclarar una cosa con respecto a la interpretación de estos. El conjunto original de datos presentaba un número relativamente bajo de individuos (38), y del mismo modo los conjuntos de entrenamiento y de prueba construidos a partir de este (27 y 11 respectivamente). Trabajar con muestras tan pequeñas, en las que el número de individuos es tan reducido, puede y suele dar lugar a errores de clasificación, y es lo que se denomina “overfitting” o sobreajuste (Zhang & Ma, 2012). El “overfitting” hace referencia a la incapacidad de un modelo para ajustarse a nuevos datos como consecuencia de que el algoritmo ha quedado ajustado a datos muy particulares (los del conjunto de entrenamiento) y es incapaz de generalizar para clasificar nuevos datos. Como consecuencia del “overfitting” se pueden obtener clasificaciones extremadamente erróneas o perfectas (como es este caso).

Para lo que acomete este trabajo, que es una revisión teórica de las técnicas, los resultados son perfectamente válidos, porque nos permiten observar diferencias de clasificación según los diferentes tipos de kernel, pero si se tratara de un estudio en el que las predicciones obtenidas son el propósito del trabajo, sería arriesgado concluir que el kernel lineal ha sido capaz de proporcionar una predicción perfecta. Es prácticamente imposible conseguir que una predicción proporcione una clasificación perfecta de los datos si se trabajó con datos reales, por ello, si el conjunto de datos con el que se realizó este análisis hubiera presentado un mayor número de individuos, evitando así este “overfitting”, es muy probable que los valores de sensibilidad y especificidad de la misma hubieran sido inferiores.

**6. Conclusiones**

1. El machine learning es fundamental si se pretenden conseguir avances en cualquier campo de estudio que requiera de análisis de datos de expresión génica, ya que la cantidad de información que generan este tipo de estudios es inmensa, y desarollar modelos que aprendan de manera automática, facilita su estudio.

2. Los datos de expresión génica son relativamente complejos de analizar como consecuencia de la estructura que presentan. Por lo general están formados por un número muy alto de variables (desde unos pocos cientos a incluso decenas de miles) y por un número de individuos, de manera habitual, bastante reducido.

3. De manera habitual, se tiende a asociar el machine learning con procedimientos complejos, que nada tienen que ver con las técnicas de la estadística clásica, pero esto no es en absoluto cierto. Existen muchas técnicas propias de la estadística clásica que funcionan como un complemento perfecto para algunos procedimientos propios del machine learning. Por ejemplo, algunas técnicas dedicadas a la reducción de la dimensionalidad, como el ACP o el análisis factorial (AF), pueden ser de gran utilidad para trabajar con conjuntos de datos de expresión génica, debido a la enorme cantidad de variables que presentan.

4. Las técnicas presentadas en este trabajo son algunas de las más frecuentemente observadas en la literatura, sin embargo, cuando se trata de analizar datos reales, se observa que su eficacia no es tan alta como cabría esperar. Obviamente existen técnicas más complejas, que podrían dar lugar a mejores resultados, pero la mejor solución es crear conjuntos de datos más completos. Lógicamente hay una explicación para que, por regla general, los conjuntos de datos de este tipo estén constituidos por un número tan reducido de individuos, y es el coste de extracción de los datos. Extraer información de la expresión génica (del grupo de genes que se quiera estudiar) de un número grande de individuos es un procedimiento bastante costoso, sin embargo, si se quieren obtener resultados reveladores de un estudio, evitando en la medida de lo posible el “overfitting”, no se puede trabajar con conjuntos de datos tan reducidos. Es en estos casos cuando se acude al estadístico/analista de datos en busca de una solución milagrosa, pero tanto el investigador como el propio estadístico saben que, en la mayoría de los casos, lo único que se consigue es maquillar los resultados. Sí que es verdad que hay procedimientos para reducir el “overfitting” más allá de aumentar el tamaño de la muestra (tener una muestra balanceada, reducir el número de dimensiones del conjunto de datos...), sin embargo, con unos tamaños de muestra tan reducidos como los que se han utilizado en este trabajo, el riesgo de “overfitting” es prácticamente inevitable.