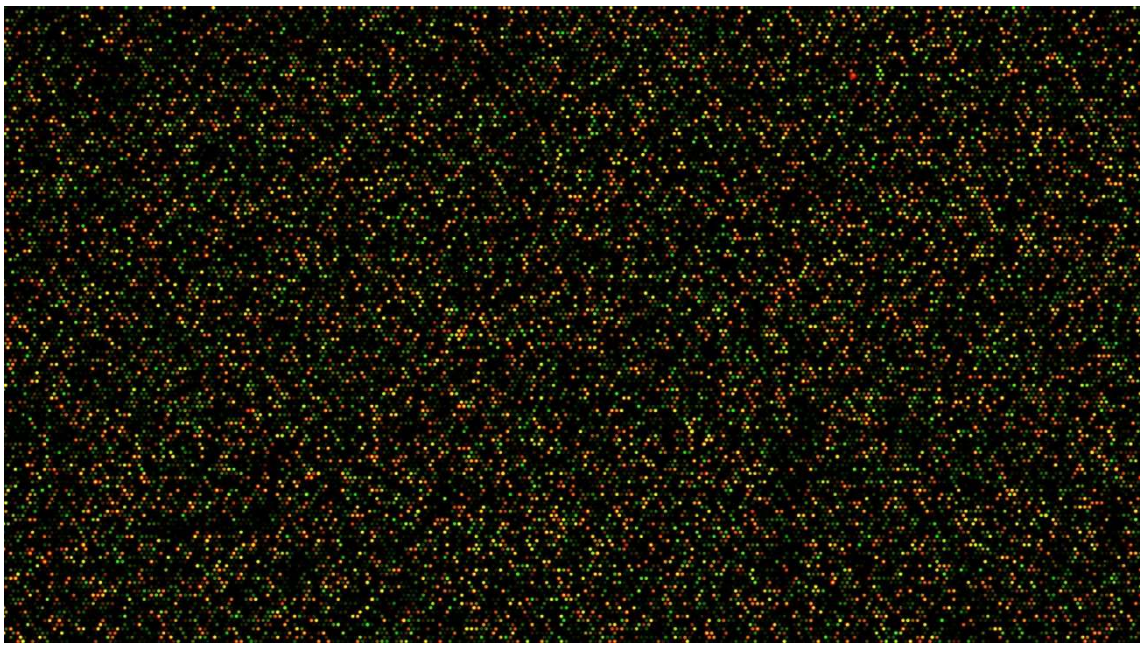
ANALISIS DE DATOS DE MICROARRAY PARA LA PREDICCION Y DIAGNOSTICO DEL TIPO DE CANCER CEREBRAL

Jorge González-Rico Iriarte

Trabajo de Fin de Máster



KSCHOOL

KSCHOOL, Calle de José Picón, 31, 28028 Madrid

05/09/2020

ÍNDICE

1. INTRODUCCION

1.1. ¿**QUÉ ES EL CÁNCER?**

El cáncer es un problema de salud a nivel mundial. En Estados Unidos, es la segunda causa de mortalidad, y en 2020 se estima que alrededor de 1,8 millones de personas padecerán esta enfermedad, de los cuales al menos 600.000 personas se estima que morirán. Está claro que el cáncer es uno de los problemas de salud mas importantes del siglo XXI, pero ¿qué es realmente esta enfermedad?

Se conoce como cáncer a una colección de enfermedades relacionadas entre si, en las cuales, algunas células del cuerpo comienzan a dividirse sin control. La diferencia entre estas células cancerígenas y las células sanas de nuestro cuerpo es que las células cancerígenas pueden dividirse sin control y volverse invasivas. Otras diferencias entre las células cancerígenas y las células sanas del cuerpo se basan en la función de las células, puesto que mientras las células sanas del cuerpo tienen una función específica, las células cancerígenas pierden su especialización.

El cáncer es una enfermedad genética. Esto significa que un tumor surge por un fallo o un cambio en el genoma de una célula. Generalmente, cuando una célula sufre un daño y sus genes se ven afectados, dicha célula se somete al proceso de apoptosis o muerte celular programada. Esto significa, que cualquier célula sana que sufra un daño genético, morirá para no transmitir dichos errores en su genoma a las células descendientes. Sin embargo, es posible que dichos fallos ocurran en alguno de los genes que influyen en el crecimiento, división y mecanismo de apoptosis de una célula. Cuando esto ocurre, es posible que una célula sana se vuelva cancerígena. Estos genes que influyen en el desarrollo del cáncer se denominan **proto-oncogenes** o **genes supresores de tumores**.

1. PROTO-ONCO GENES

Los proto-oncogenes son genes que, en estado normal, contribuyen al crecimiento celular. Cuando un proto-oncogén muta, se vuelve un gen “maligno” que puede estar activado de manera permanente. Cuando esto ocurre, una célula puede crecer sin control. Este proto-oncogén se llama ahora **oncogén.**

2. GENES SUPRESORES DE TUMORES

Los genes supresores de tumores son genes que reducen el ratio de división celular, reparan errores en el ADN, o activan el proceso de apoptosis. Cuando un gen supresor de tumores no funciona de manera correcta, las células pueden crecer sin control, lo que deriva en cáncer.

Una manera sencilla de visualizar estos genes es como los pedales de un coche. Necesitamos un acelerador y un freno para poder avanzar de manera controlada. Los proto-oncogenes actúan como el acelerador, contribuyendo al crecimiento y la división celular, necesarias para la vida, mientras que los genes supresores de tumores actúan como el freno, impidiendo que las células crezcan sin control.

En los últimos años, se han investigado nuevas maneras de detectar y conocer el cáncer de manera mas rápida y cercana. En el campo de la imagen médica, se realizan diversos estudios de reconocimiento de tumores mediante CT, mientras que en el campo de la genética, un interesante avance son los chips de ADN o **Microarrays.**

1.2. MICROARRAYS

En el año 1995, se publica el articulo científico “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray” por Patrick O.Brown , donde se describe el primer microarray. Un microarray consiste en un pequeño chip de cristal que tiene en la superficie unidos fragmentos de ADN. Estos fragmentos se sitúan en parcelas (miles por cada array). Sobre estos chips se deposita una muestra de ADN, y mediante el método de la **hibridación**, los fragmentos de ADN de la muestra se unen a sus fragmentos complementarios en el microchip. Luego, mediante fluorescencia, se mide el nivel de intensidad de cada parcela del array, que es proporcional al nivel de expresión de cada gen.

Esta tecnología resulta crucial para los análisis genético de alto rendimiento (high throughput).

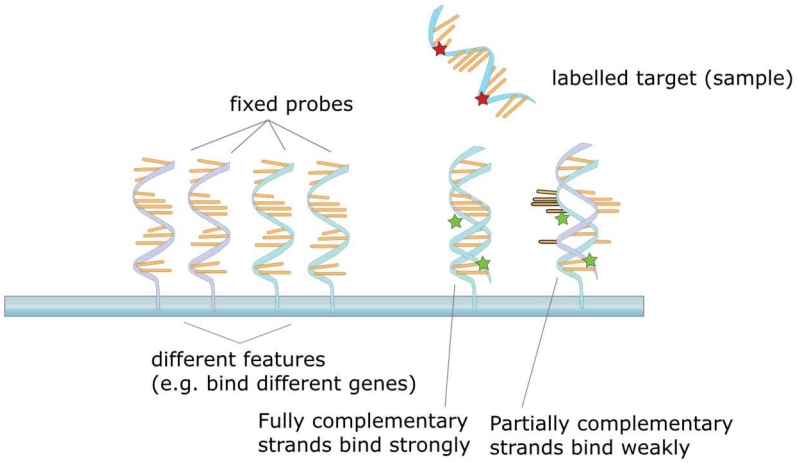


Imagen 1: Principio de funcionamiento de un microarray. La hibridación se produce cuando una hebra de ADN se une por complementariedad a otra.

2. OBJETIVO DEL TRABAJO

El objetivo de este trabajo será el análisis de un set de datos obtenido a partir de 130 microarrays distintos, realizados sobre 130 individuos, algunos de los cuales sufrían un tipo concreto de cáncer cerebral. Se realizará un análisis de conglomerados de los datos, y últimamente, se desarrollará un algoritmo para la detección e identificación de cualquiera de estos tipos de cáncer a partir de los datos de un microarray.

3. DATOS

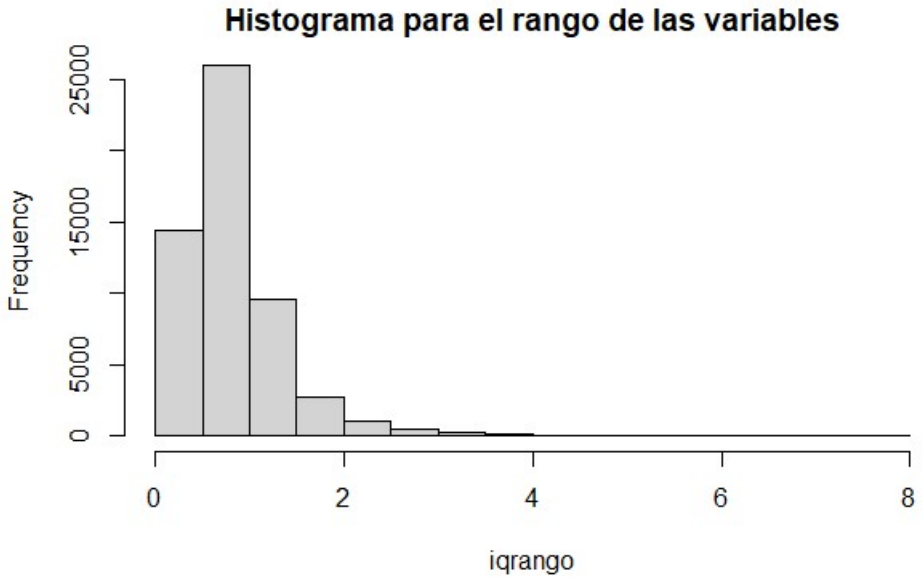
Los datos elegidos para llevar a cabo este análisis provienen de la base de datos CuMiDa. CuMiDa es una base de datos de microarrays obtenidos a partir de la comparación de sets de microarrays. En este caso, emplearemos el set de datos de “Brain GSE50161”, que consta de 130 observaciones de un total de 54675 genes. Estas observaciones están clasificadas en 5 posibles grupos, que resultan del diagnostico de 4 tipos distintos de cancer, así como un quinto grupo de control con individuos sanos. Los 5 grupos son “ependymoma”, “glioblastoma”, “meduloblastoma”, “normal” y “pilocytic_astrocytoma”. Los genes representados son un conjunto de mas de 50000 genes sobre los cuales se cree que pueden tener efecto sobre el cáncer cerebral.

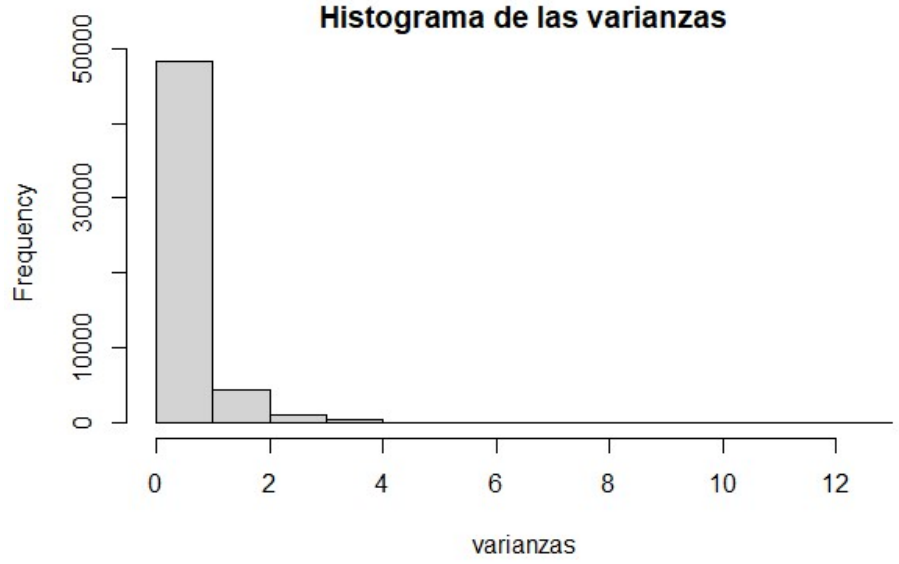
3.1. ANALISIS EXPLORATORIO DE LOS DATOS

El primer paso en nuestro análisis de datos será observar de manera previa el dataset con el que estamos lidiando para ver que pasos previos a la preparación del modelo debemos seguir. Debemos conocer las dimensiones del set de datos, si existen “missing values” y si será necesario realizar una normalización de los datos.

Obtenemos que la dimensión del dataset es de 55677 columnas por 130 observaciones. De estas 55677 columnas, 55675 representan genes.

En cuanto a la necesidad de normalizar los datos, dependerá de los rango en que estos se encuentren, así como la varianza de cada uno de los gene

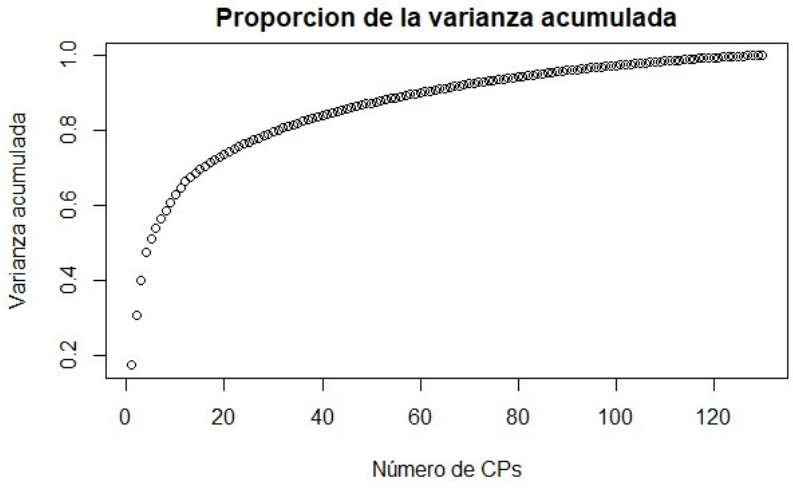




A partir de estas dos graficas vemos que el rango intercuantil de las variables tiene una dispersión pequeña, lo que significa que todos los genes están expresados en los mismos niveles. En cuanto a las varianzas, ocurre lo mismo. La gran mayoría de genes tiene una varianza entre 0 y 1, y los que tienen una varianza superior no supera 5. Además, la normalización en los datos de los microarray debe llevarse a cabo con sumo cuidado, puesto que el nivel absoluto de expresión de un gen juega un papel importante en la célula. Por ello, para el análisis de estos datos, no se llevará a cabo ningún tipo de normalización.

3.2 TRATAMIENTO DE LOS DATOS

Hemos analizado el tamaño del dataset: 130 observaciones de 55677 variables. Existe un claro problema de dimensionalidad en este caso. Para reducir la dimensionalidad del problema, llevaremos a cabo un análisis de componentes principales o **PCA** por sus siglas en ingles. Este método es útil para lidiar con problemas de gran dimensionalidad. Consiste en obtener, a partir de un set de variables colineales entres si, un nuevo set de variables o **Componentes Principales** que son ortogonales entre si. Estas nuevas variables están ordenadas de mayor a menor varianza, siendo la primera componente la que tiene la mayor proporción de varianza de la muestra. Utilizar todas las componentes principales para un análisis proporciona la misma cantidad de información que utilizar el dataset original. Gracias a este método, podemos reducir de manera notable la dimensionalidad del problema sin perder información. Podemos observar en una gráfica la progresión de la proporción de la varianza que representan las *X* primeras componentes principales.



Una vez realizado este análisis, tenemos toda la información que estaba en el dataset completo, pero tan solo en 130 variables. Hemos solucionado el problema de la dimensionalidad.

4. MODELOS

A continuación vamos a realizar una evaluación de distintos modelos de clustering para los datos. Dicha evaluación la vamos a realizar sobre los datos obtenidos del análisis de componentes principales.

La evaluación se realizara en función a la medida "V-measure". Se trata de una medida externalizada para problemas de clustering con etiquetado. Se basa en una combinación de las medidas de homogeneidad y completitud. La homogeneidad se define como asignar a un cluster __solo__ datos de un grupo, mientras que la completitud se defino como asignar __todos__ los datos de un grupo al mismo cluster. La medida "V-measure" se define como la media de estas dos medidas [Rosenberg & Hirschberg, 2007].

4.1. CLUSTERING JERÁRQUICO

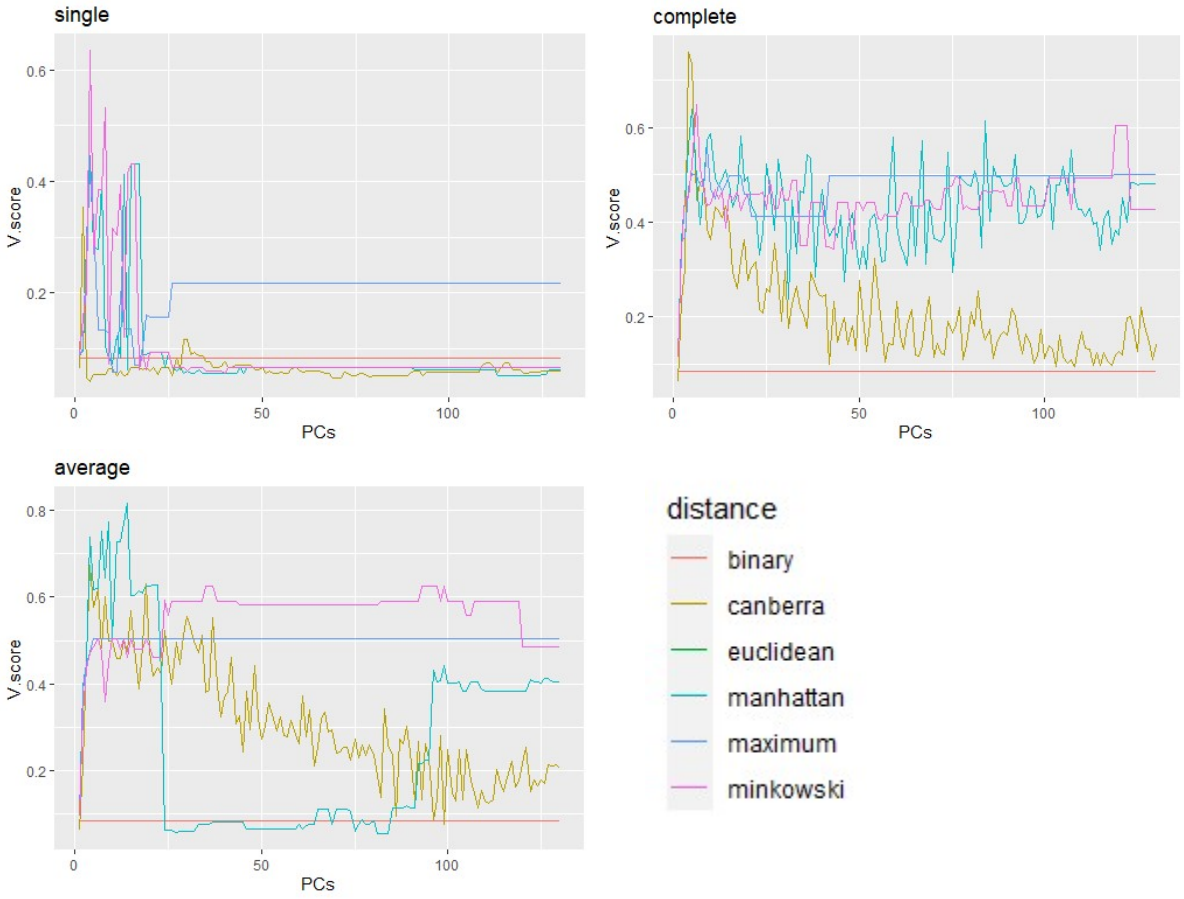
Realizamos la evaluacion del clustering jerárquico teniendo en cuenta todas las posibles combinaciones de factores:

-Utilizamos las 7 distancias posibles

-Utilizamos las 3 maneras posibles de dividir el arbol en 5

Tambien realizaremos un barrido por todas las componentes principales, utilizando las X primeras componentes principales para cada caso.

En las siguientes gráficas podemos ver como progresa la puntuación (V-measure) para cada una de las maneras de calcular la distancia en cada una de las maneras de dividir el árbol jerárquico en 5 clusters.

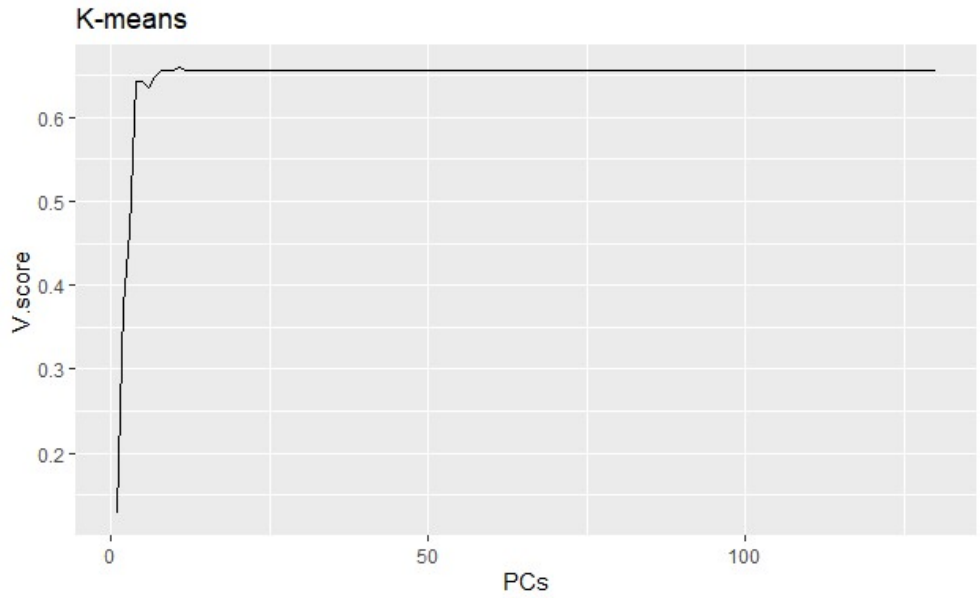


Observamos como la mejor puntuación nos la reporta el método “average” para dividir el árbol, y por lo general, las distancias manhattan, minkowski y máximum resultan las mejores. El mejor de los resultados es el siguiente.

```
     distance clust_method PCs Homogeneity Completeness   V.score
822 manhattan      average  14    0.760486     0.879317 0.8155958
```

4.2. CLUSTERING POR K-MEANS

A continuación realizamos el análisis de conglomerados por el método de k-vecinos. Al igual que previamente, realizaremos un análisis de la puntuación “v-measure” con respecto a las *X* primeras componentes principales.



En este caso, como podemos observar, el resultado es mucho mas estable a partir de 11 componentes principales. La máxima puntuación que conseguimos es de 0.65 sobre 1.

En este punto, hemos analizado y comparado dos métodos distintos de aglomeración de los datos y hemos obtenido unos resultados de fidelidad de alrededor del 65%. Sin embargo, podemos seguir indagando en los datos y en otros métodos que nos permitan obtener mejores resultados a la hora de aglomerar los datos, y, en última instancia, crear un protocolo de clasificación para la detección y diagnóstico del cáncer cerebral.

5. OPTIMIZACIÓN

Una vez hecho un análisis exploratorio de los datos y de los principales métodos de clustering a utilizar, vamos a intentar optimizar tanto los datos como la metodología para obtener unos resultados más favorables. Nos serviremos del hecho de que estamos lidiando con un problema de clustering, pero los datos están etiquetados. Vamos a cuestionar dos suposiciones que hemos hecho desde el principio:

1. Debemos usar todos los genes que tenemos a nuestra disposicion para realizar el clustering
2. Existen 5 grupos en los que realizar el clustering

1. ¿Debemos usar todos los genes que tenemos a nuestra disposicion para realizar el clustering?

A continuación, sirviéndonos de la información que nos brinda el dataset, elegiremos los genes que son mas distintivos de cada tipo de cáncer, que son aquellos con el menor “within class variance”, pero con el mayor “between class variance”. El test que se realiza para este problema se denomina "análisis de la varianza" o **ANOVA** por sus siglas en ingles.

Queremos comprobar si existe alguno de los genes que no varíe entre distintos tipos de cáncer, es decir, que la media de la expresión de todos los genes sea la misma para el tipo de cáncer que sea. Definimos el análisis de la siguiente manera:

$$H_0 : \mu_{ependymoma} = \mu_{glioblastoma} = \mu_{medulloblastoma} = \mu_{normal} = \mu_{pilocytic}$$

$$H_1 : no\ todas\ las\ medias\ son\ iguales$$

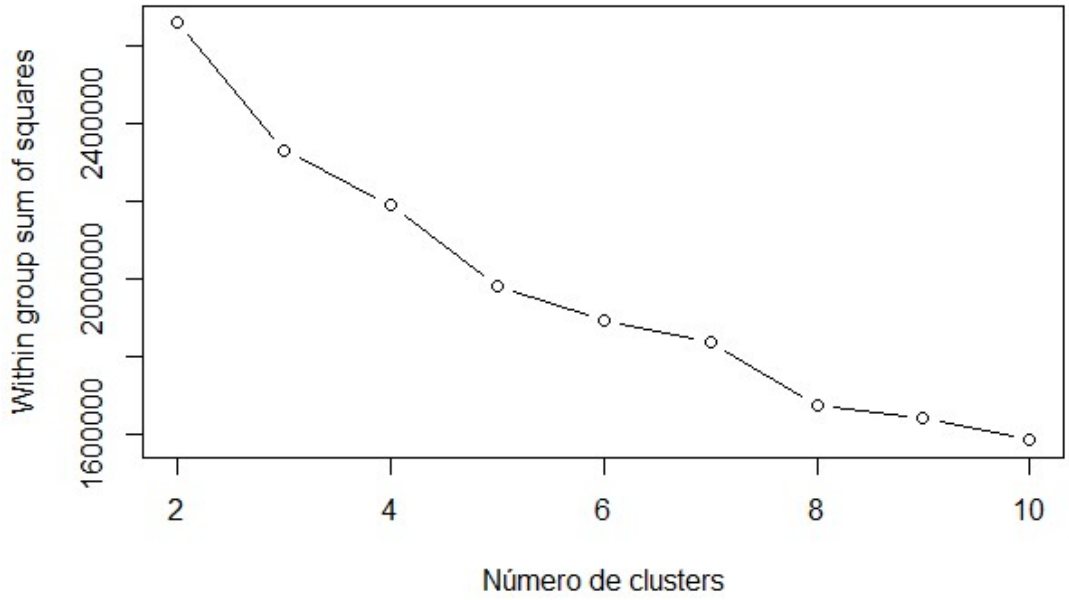
El test nos devolverá un valor p (p-value) para el estadístico F que se calcula para cada una de las variables. Si el valor p es menor que 0.05 (valor de significancia elegido), podemos afirmar que para dicho gen, existe una diferencia estadística en su expresión para al menos uno de los grupos.

El resultado final nos dice que solo 37367 genes de los 54677 genes son representativos desde un punto de vista estadístico. Siguen siendo muchos, pero hemos eliminado casi la mitad de ellos.

2. ¿Existen 5 grupos en los que realizar el clustering?

También debemos cuestionar esta suposición, puesto que es posible que dos canceres sean genéticamente muy similares, o bien que dentro del diagnóstico de uno de los canceres, existan dos variantes genéticas o más distintas.

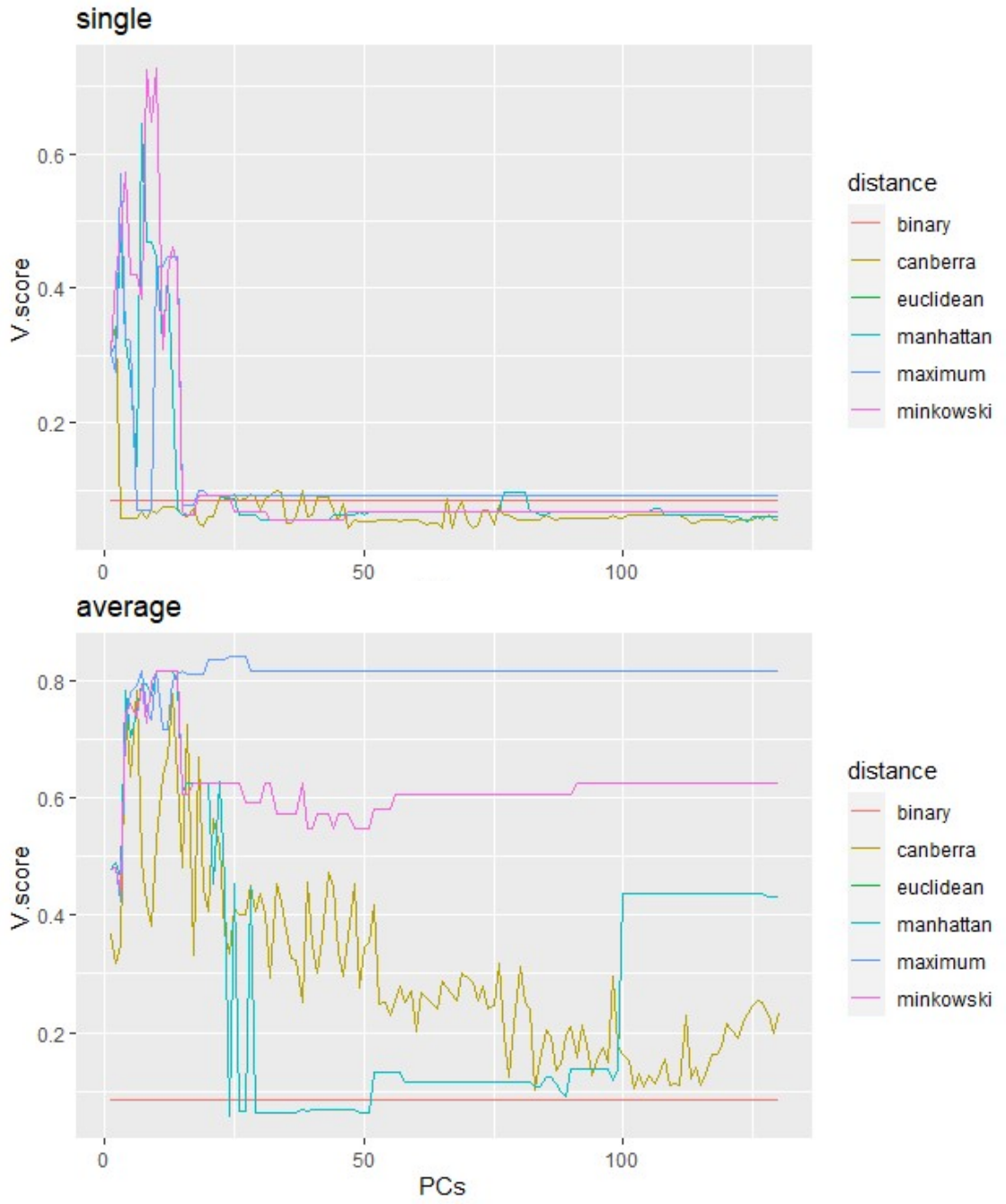
Por ello vamos a realizar un análisis del número de clústeres que minimizará el “within-groups sum of squares” contra el número de clusters.

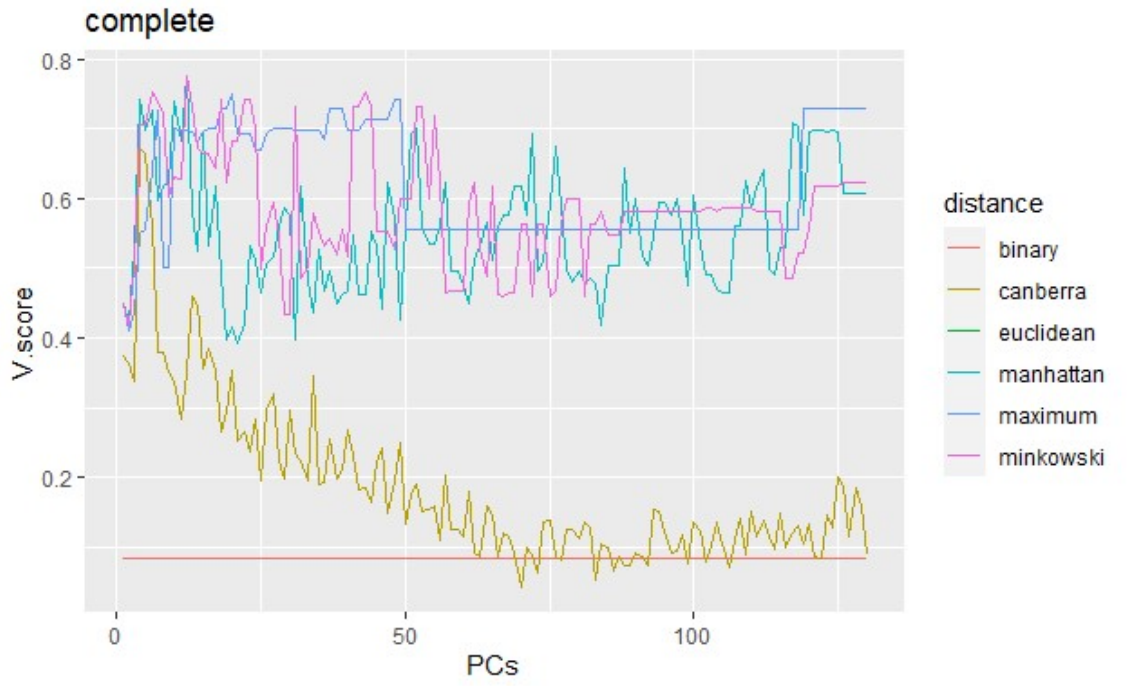


No encontramos en esta gráfica evidencia suficiente para afirmar que el número de clusters con el que debemos trabajar no es 5.

6. MODELOS CON EL DATASET REDUCIDO

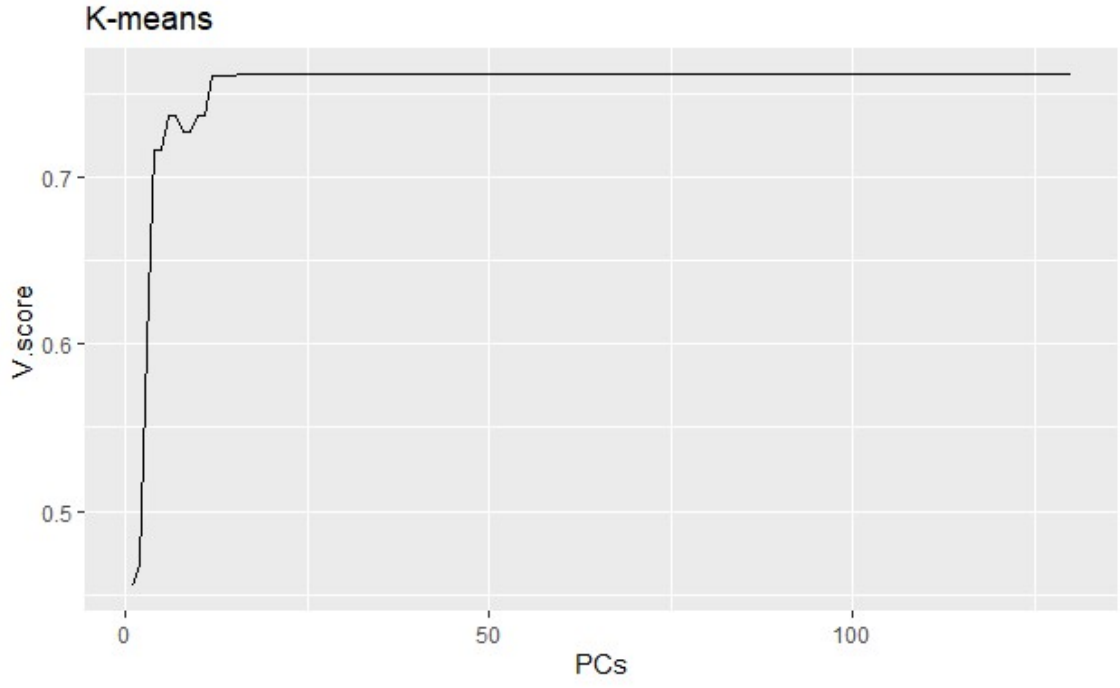
6.1. CLUSTERING JERÁRQUICO





Observamos como en este caso, la puntuación del método jerárquico ha aumentado considerablemente, además de que parece más estable.

6.2. K- MEANS CLUSTERING

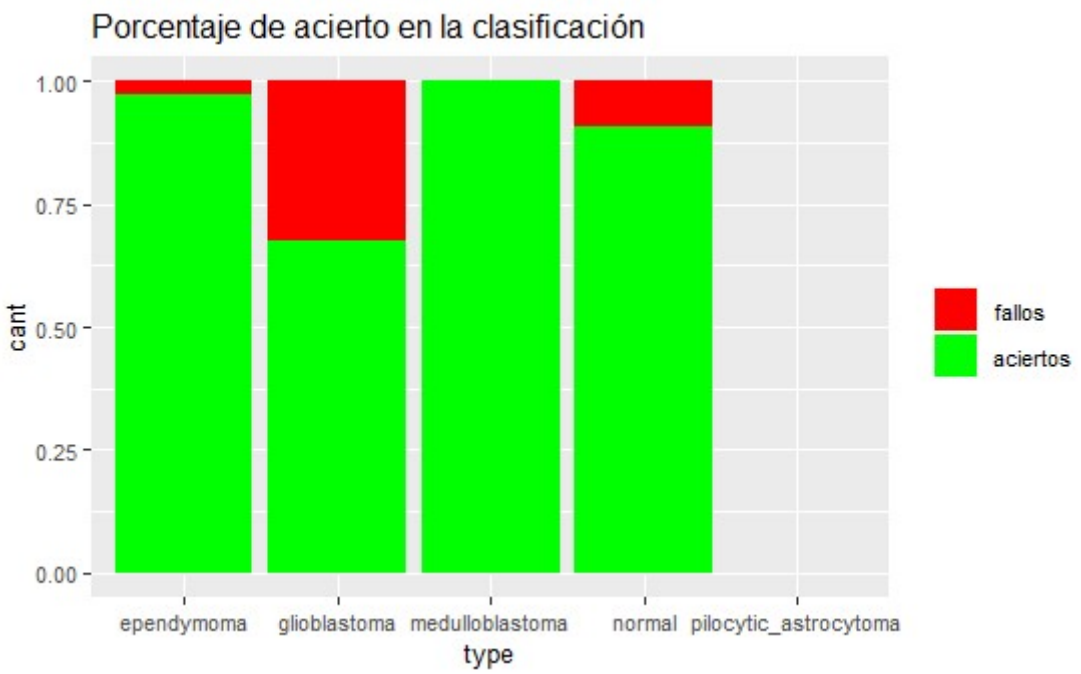


En este caso, el resultado del clustering por k-vecinos también mejor substancialmente.

7. PREDICTOR DE CANCER

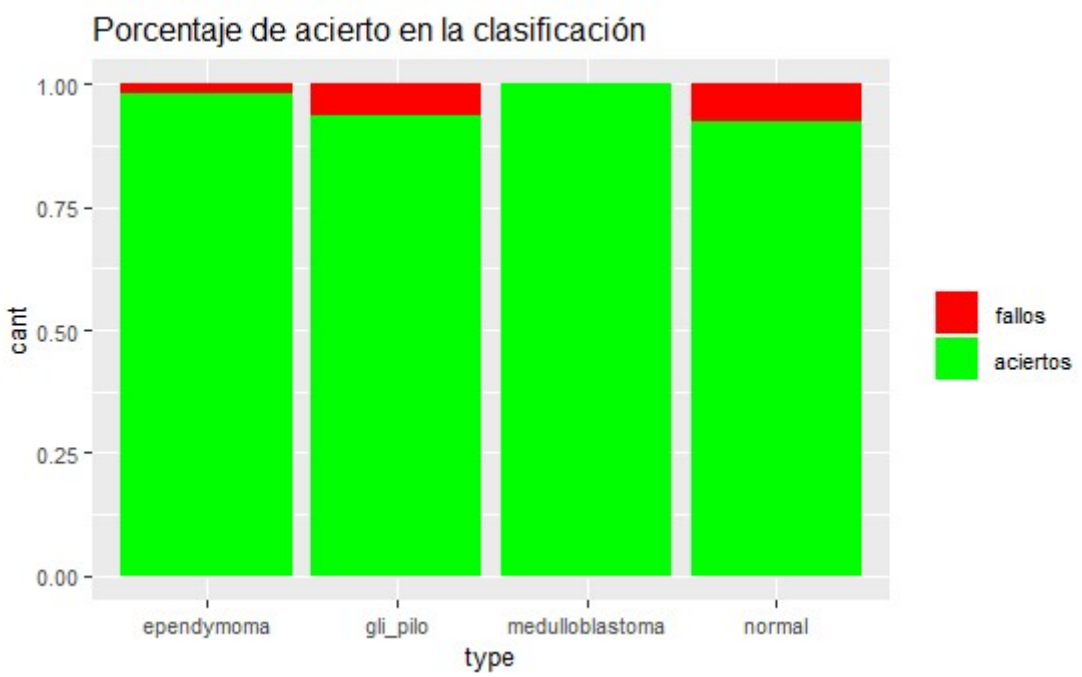
Una vez llevado a cabo la optimización de los métodos de clustering, realizaremos un programa para clasificar, en base a un dataset preexistente, un nuevo individuo en uno de los grupos de cáncer. De esta manera, podremos diagnosticar de que tipo de cáncer se trata.

Descubrimos que la eficiencia de la calificación del cancer es de alrededor del 85%. Esto significa, que mediante un screening genético del paciente, podemos predecir con un 85% de precisión si tiene cancer o no, y que tipo de cancer tiene.



En la gráfica podemos observar el porcentaje de acierto para cada tipo de cancer. Observamos que es muy alto para los grupos 1, 3 y 4, mientras que el segundo grupo es mucho menos preciso, y el ultimo grupo no tiene resultados. Esto puede ser debido al hecho de que los grupos 2 y 5 estén extremadamente cerca en estas dimensiones, lo que significa que el grupo 2 engloba a los grupos 2 y 5, con lo que una gran porción de sujetos está erróneamente etiquetados.

Para comprobar nuestra suposición, uniremos los grupos 2 y 5 en un único grupo que englobe a ambos, y realizaremos la misma clasificación pero esta vez con estos 4 grupos. Como podemos observar, parece que nuestra suposición era cierta. Sacrificando algo de precisión en el diagnóstico, ganamos seguridad en la clasificación.



En este caso, los porcentajes de acierto son altos para todos los grupos.

8. CONCLUSIONES

En este trabajo hemos lidiado con uno de los principales problemas del análisis de microarrays, el problema de la dimensionalidad, puesto que cada vez que se realiza un estudio, obtenemos información de miles de genes (variables). Hemos empleado el método de análisis de componentes principales para lidiar con este problema. Además del problema de la dimensionalidad, otro problema con la gran cantidad de datos que obtenemos es que muchos de ellos no son relevantes. Esto puede eclipsar los datos que sí lo son. Por ello, hemos hecho uso de la ventaja de ese dataset, que es que las entradas están etiquetadas, para optimizar el método de clustering. Hemos descubierto que mediante el análisis de la varianza hemos sido capaces de eliminar una gran cantidad de datos que no proporcionaban ninguna información relevante, de tal manera que hemos sido capaces de mejorar la precisión en la clasificación de los individuos. Por último, descubrimos que para la clasificación efectiva de los tipos de cáncer, debemos tomar la opción de juntar dos clusters en uno debido a su proximidad. Esta perdida de información nos devuelve un gran beneficio en cuanto al porcentaje de acierto.