Análisis de conglomerados en células cancerigenas

Gómez Jiménez Aaron Mauricio

2023-05-20

El archivo NCI.txt contiene 6830 variables de 64 células cancerígenas de distintos tipos de cáncer. Mediante un análisis de conglomerados estudiaremos el posible agrupamiento de las células cancerígenas.

Selección de Medida

Ya que deseamos analizar si las células forman grupos utilizaremos la distancia Euclidiana ya que es la forma más usual y real de medir la distancia entre las células.

Selección del tipo de Modelo

Lo primero que debemos preguntarnos es ¿Queremos clusters grandes o clusters pequeños? Como no tenemos aún la información necesaria para responder esta pregunta, en esta ocasión usaremos dos tipos de modelos, los aglomerativos y los divisivos, en especial el modelo AGNES y DIANA, recordando que los modelos divisivos indentifican de mejor forma los clusters grandes, en contraparte los aglomerativos identifican de mejor forma los clusters pequeños.

Selección del Método de Aglomeración

Existen varios métodos para este modelo, por ejemplo, complete, single, average, de centroide, etc. Por lo cual haremos una prueba de hipotesis con la estadistica de Hopkins para ver si hay tendencia a clusters en los datos

```
get=get_clust_tendency(datos, n=nrow(datos)-1, graph = FALSE)
get$hopkins_stat
```

```
## [1] 0.6333125
```

La prueba de hipotesis es mayor a 0.5, no es tan cercana a 1, por lo cual los datos tienen poca tendencia a agruparse, sabiendo esto haremos varios modelos para ver cual es el que mejor hace los clusters.

Metodo Divisivo DIANA

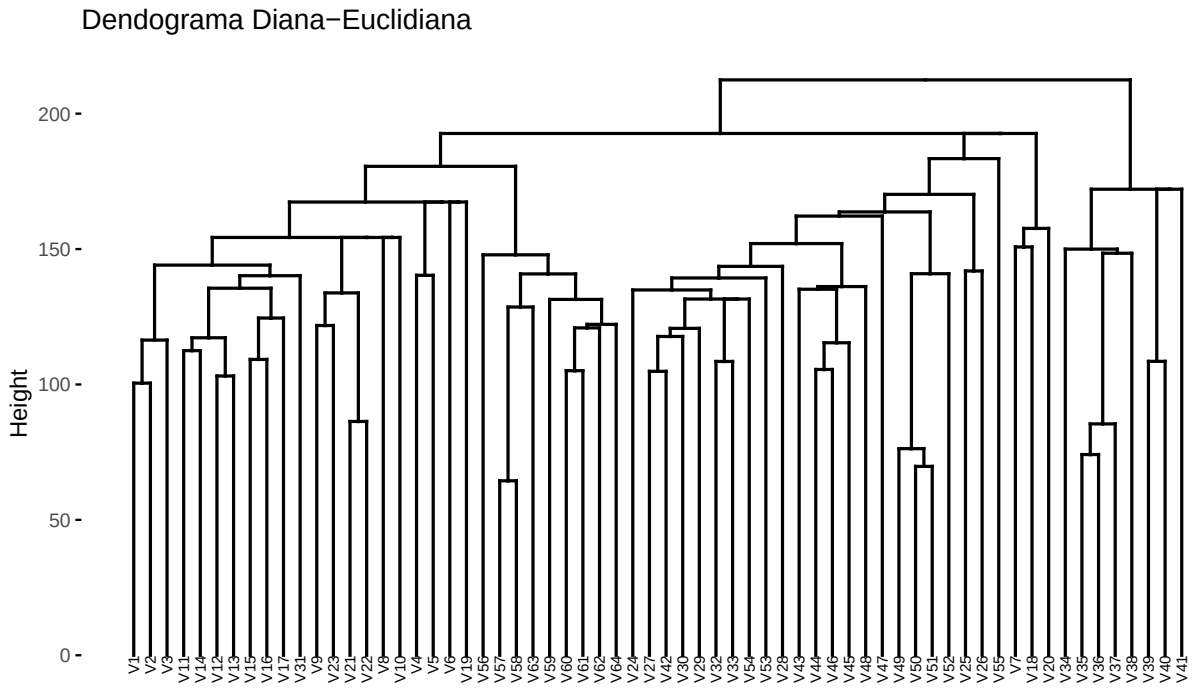
Haremos un modelo Diana con metrica euclideana y analizaremos su nivel de agrupamiento

```
dian=diana(datos, metric ="euclidean", stand = TRUE)
dian$dc
```

```
## [1] 0.4307554
```

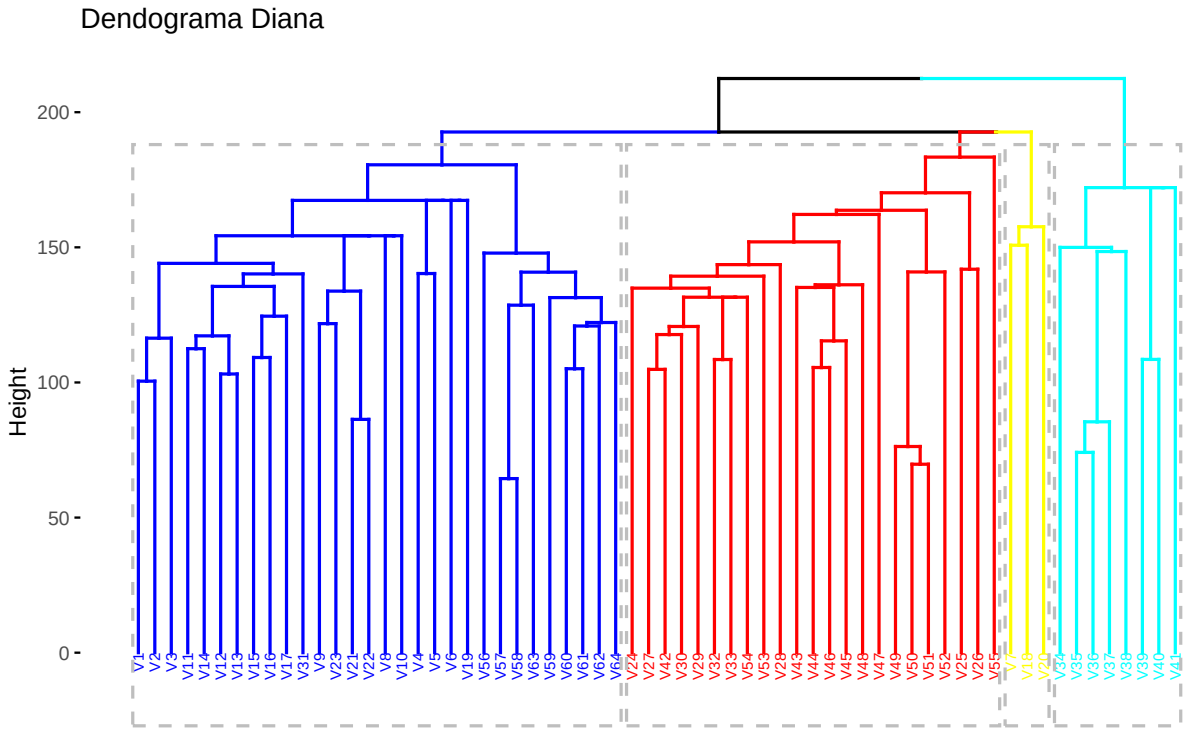
Nuevamente obtenemos un nivel de agrupamiento muy pobre, por lo cual haremos un dendograma para visualiazar los clusters más grandes.

```
fviz_dend(dian,
          cex = 0.5,
          k_colors = c("blue", "red"),
          color_labels_by_k = TRUE,
          rect = TRUE,
          main="Dendograma Diana-Euclidiana"
)
```

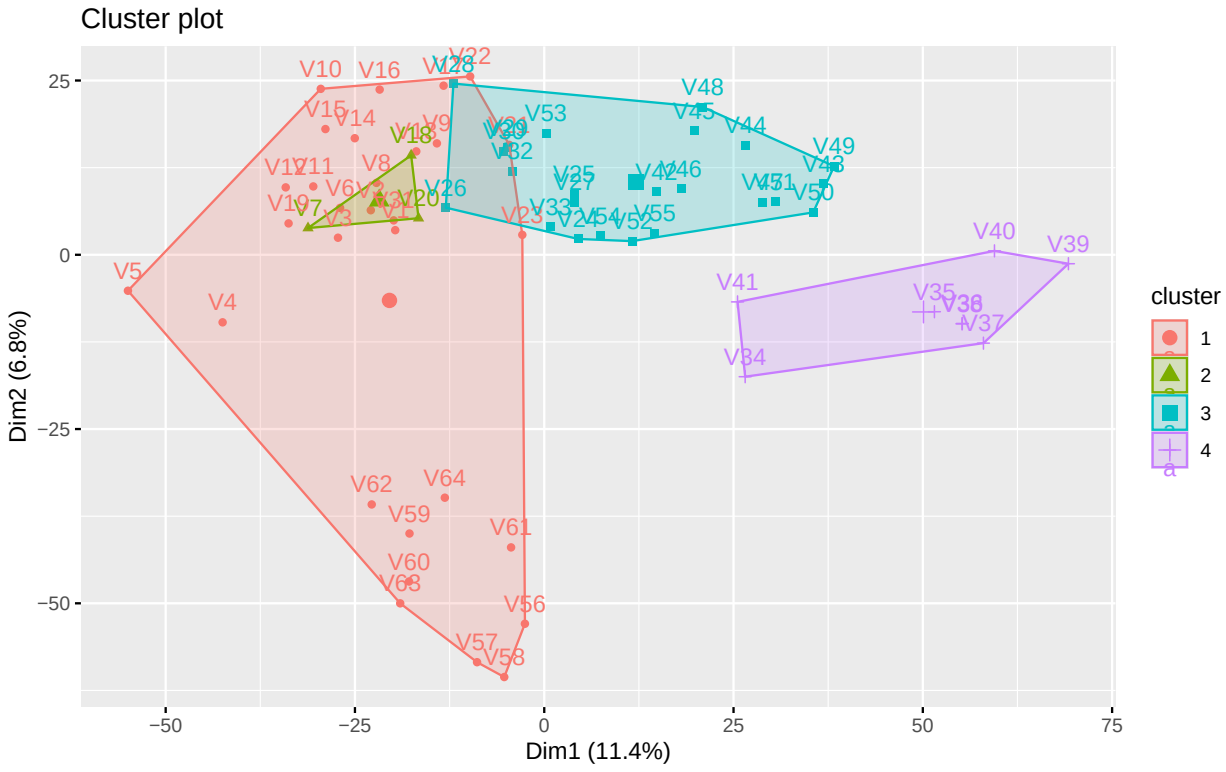


Podemos observar 4 cluster más significativos, sin embargo, estos se dividen en muchos clusters más pequeños lo cual nos indica que no existe una tendencia importante a agruparse.

```
fviz_dend(dian,k=4,
 cex=0.5,
 k_colors=c("blue","red", "yellow", "cyan" ),
 color_labels_by_k=TRUE,
 rect=TRUE,
 main="Dendograma Diana"
 )
```




```
grp = cutree(dian, k = 4)
fviz_cluster(list(data = datos, cluster = grp))
```



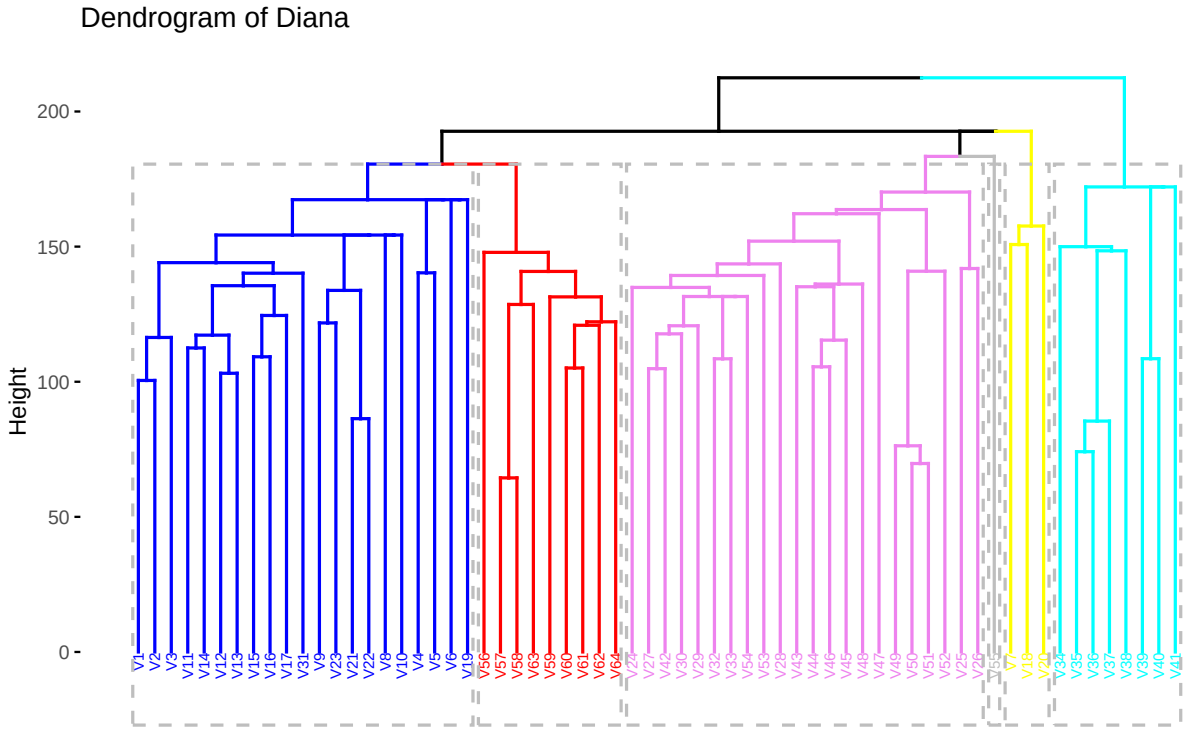
Como podemos observar los clusters se traslapan, es decir, tienen datos en común lo cual nos sugiere que los grupos no se estan haciendo de buena forma, veamos la distribución por grupo

```
table(grp)
```

```
## grp
##  1  2  3  4
## 30  3 23  8
```

Notamos que hay 2 clusters que tiene pocas variables, por esta razón y que los clusters se traslapan probaremos haciendo 6 clusters

```
fviz_dend(dian, k=6,
          cex = 0.5,
          k_colors = c("blue","red", "violet", "grey", "yellow", "cyan" ),
          color_labels_by_k = TRUE,
          rect = TRUE,
          main="Dendrogram of Diana"
)
```




```
grp_1 = cutree(dian, k = 6)
fviz_cluster(list(data = datos, cluster = grp_1))
```


```
table(grp_1)
```

```
## grp_1
##  1  2  3  4  5  6
## 21  3 22  8  1  9
```

El cambio mas notorio es que el cluster 1 en el modelo con corte en k=4, se dividio en 2, además se crea un nuevo cluster dentro del cluster 3 del modelo 1, se mantiene el cluster de una sola variable.

Como mencionamos anteriormente, los métodos divisivos son buenos para crear clusters grandes.

Método Aglomerativo AGNES

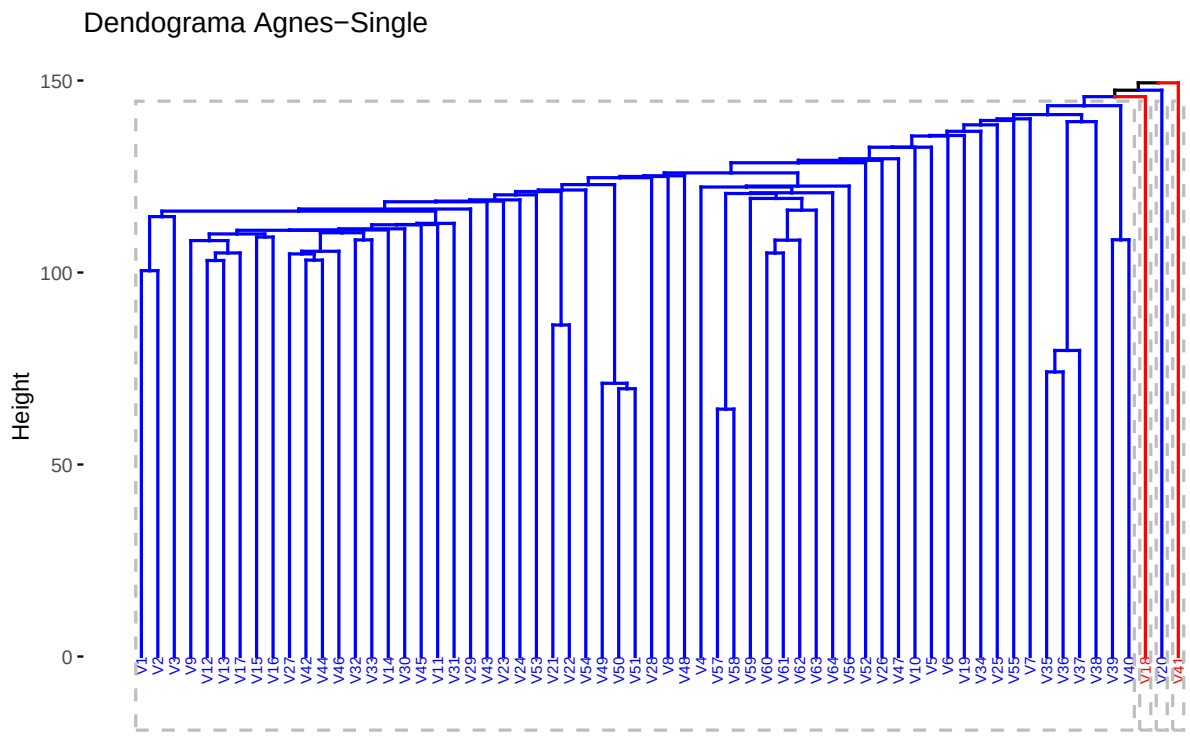
Primero haremos un modelo AGNES con la metrica euclidiana y metodo single (teoricamente clusters mas grandes), otra con metodo complete (en teoria clusters mas compactos), y por ultimo uno con Ward (fusiona clusters muy cercanos).

```
agnes_single=agnes(datos, metric="euclidean", stand=TRUE, method="single")
agnes_complete=agnes(datos, metric="euclidean", stand=TRUE, method="complete")
agnes_ward=agnes(datos, metric="euclidean", stand=TRUE, method="ward")
```

AGNES - Single

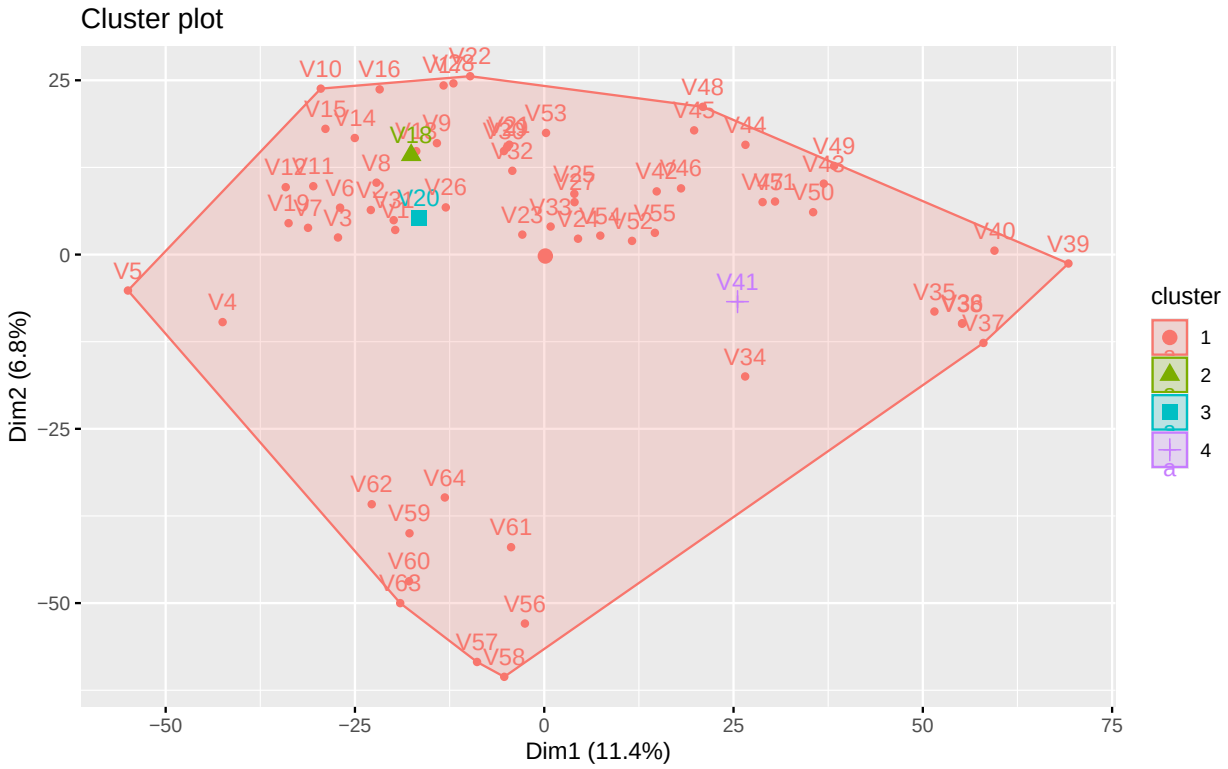
```
fviz_dend(agnes_single, k = 4,
          cex = 0.5,
          k_colors = c("blue", "red"),
          color_labels_by_k = TRUE,
```

```
          rect = TRUE,
          main="Dendograma Agnes-Single"
)
```



Graficando el modelo

```
grp_3 = cutree(agnes_single, k = 4)
fviz_cluster(list(data = datos, cluster = grp_3))
```




```
table(grp_3)
```

```
## grp_3
##  1  2  3  4
## 61  1  1  1
```

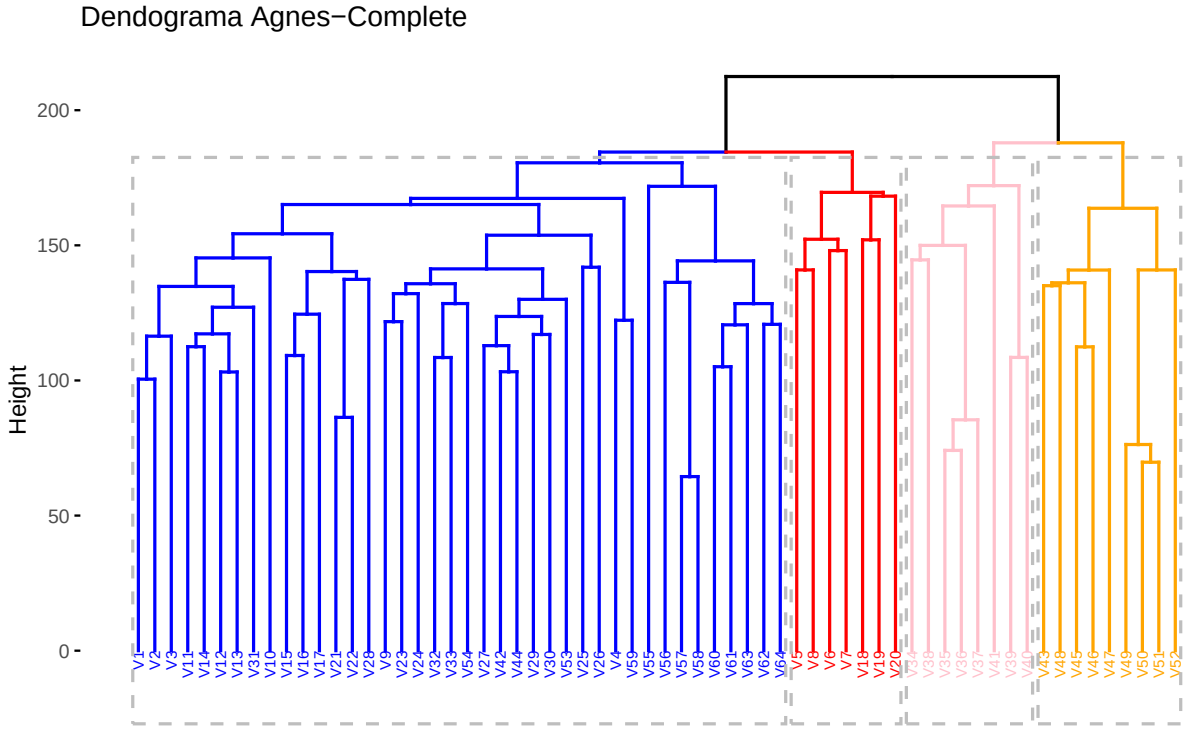
```
agnes_single$ac
```

```
## [1] 0.2496627
```

Podemos ver que con 4 clusters 3 solo tiene una variable, este modelo obviamente no es bueno ya que no agrupa de buena forma los datos ya que se sobreponen los clusters y hay un cluster muy grande, ademas que el coeficiente de cohesion es muy bajo.

AGNES- Complete

```
fviz_dend(agnes_complete, k = 4,
          cex = 0.5,
          k_colors = c("blue", "red", "pink", "orange"),
          color_labels_by_k = TRUE,
          rect = TRUE,
          main="Dendograma Agnes-Complete"
)
```



En este modelo vemos que existen 2 grandes ramas y de cada rama observamos que se divide en 2 y despues en muchas otras ramas, es decir tendriamos 4 clusters relevantes.

Graficando el modelo

```
grp_4 = cutree(agnes_complete, k = 4)
fviz_cluster(list(data = datos, cluster = grp_4))
```


```
table(grp_4)
```

```
## grp_4
##  1  2  3  4
## 40  7  8  9
```

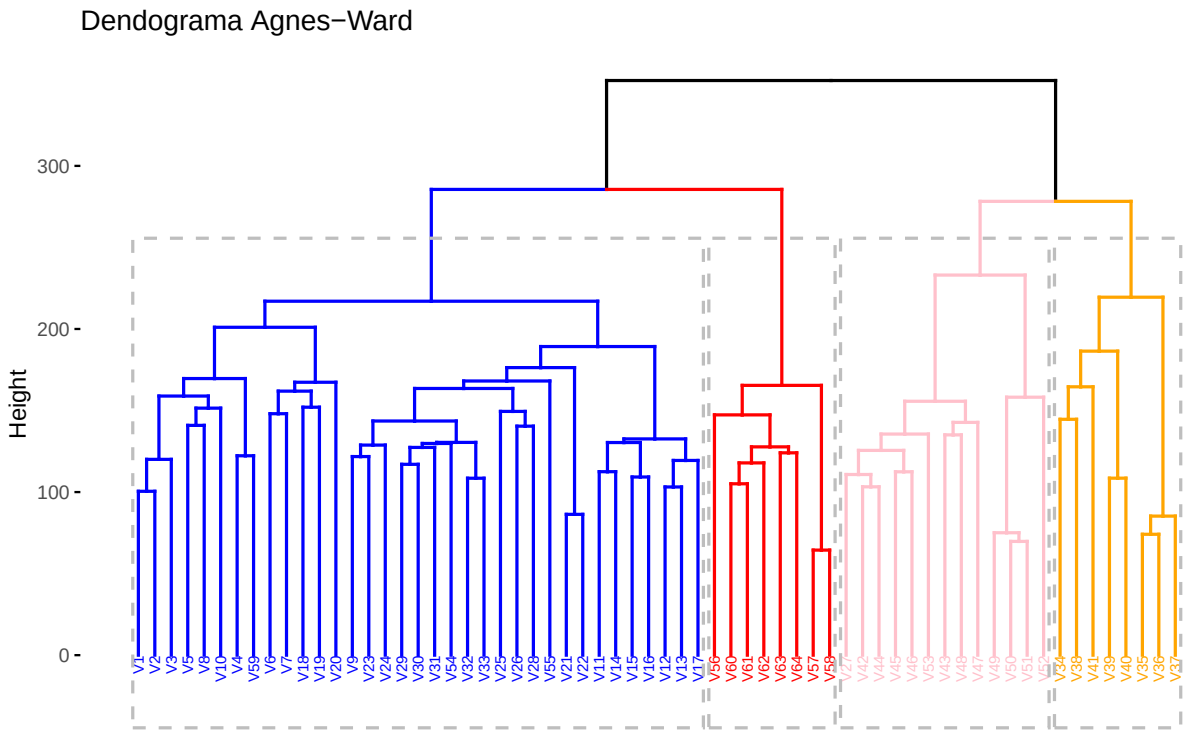
```
agnes_complete$ac
```

```
## [1] 0.4437809
```

Al analizar este modelo podemos observar una mejoría con un número de clusters igual a 4, donde hay un grupo mas significativo con 40 tipos de cancer, y los otros 3 son casi del mismo tamaño, el coeficiente de cohesión mejora.

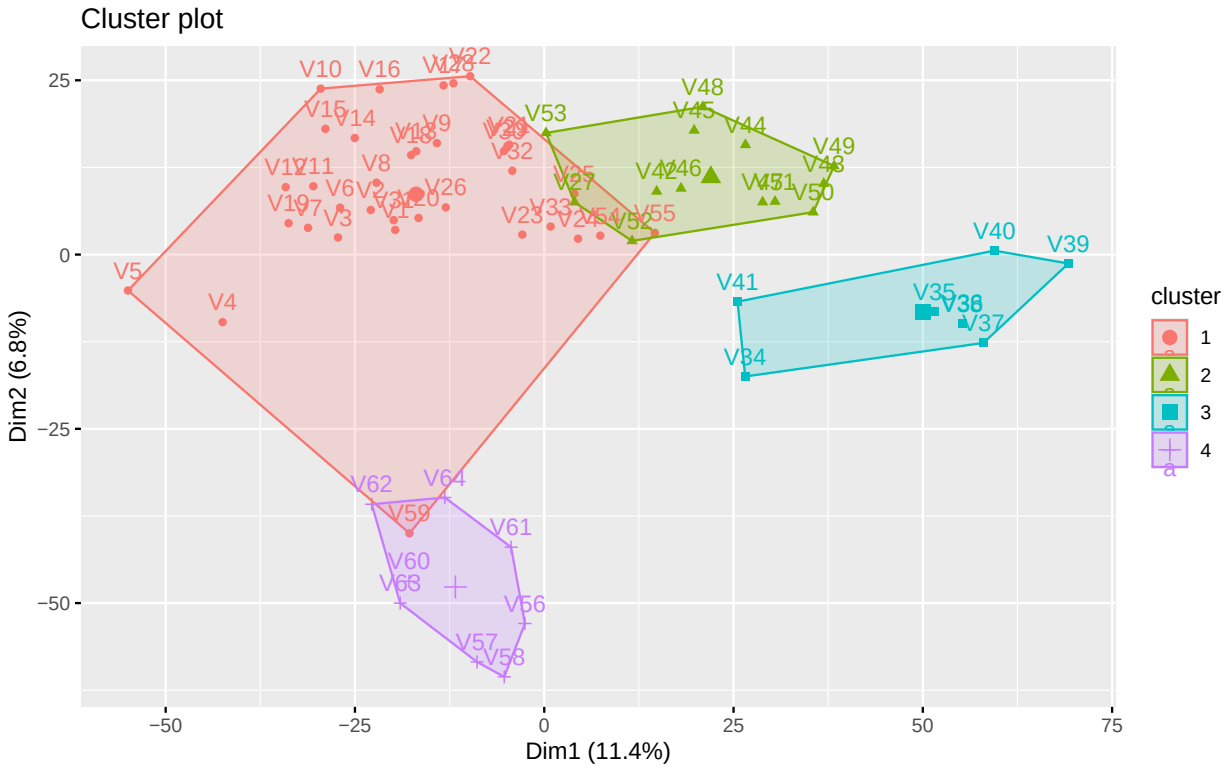
AGNES- Ward

```
fviz_dend(agnes_ward, k = 4,
          cex = 0.5,
          k_colors = c("blue", "red", "pink", "orange"),
          color_labels_by_k = TRUE,
          rect = TRUE,
          main="Dendograma Agnes-Ward"
)
```



De igual forma que con el metodo complete, en este metodo Ward obtenemos 4 clusters relevantes pero con una mejor organización y mayor distancia de esos 4 clusters principales con las demas ramas.

```
grp_5 = cutree(agnes_ward, k = 4)
fviz_cluster(list(data = datos, cluster = grp_5))
```




```
table(grp_5)
```

```
## grp_5
##  1  2  3  4
## 35 13  8  8
```

```
agnes_ward$ac
```

```
## [1] 0.6626535
```

Este sin duda es el mejor modelo obtenido mediante el modelo AGNES, ya que los clusters se agrupan de mejor forma y el coeficiente de cohesión ya es mayor a 0.66 lo cual ya es una buena señal, por lo tanto se concluye que el metodo ward en el modelo AGNES es el que mejor resultado nos da para estos datos, con un corte en k=4 clusters, sin embargo existen traslapes en algunos clusters

Método K-Means

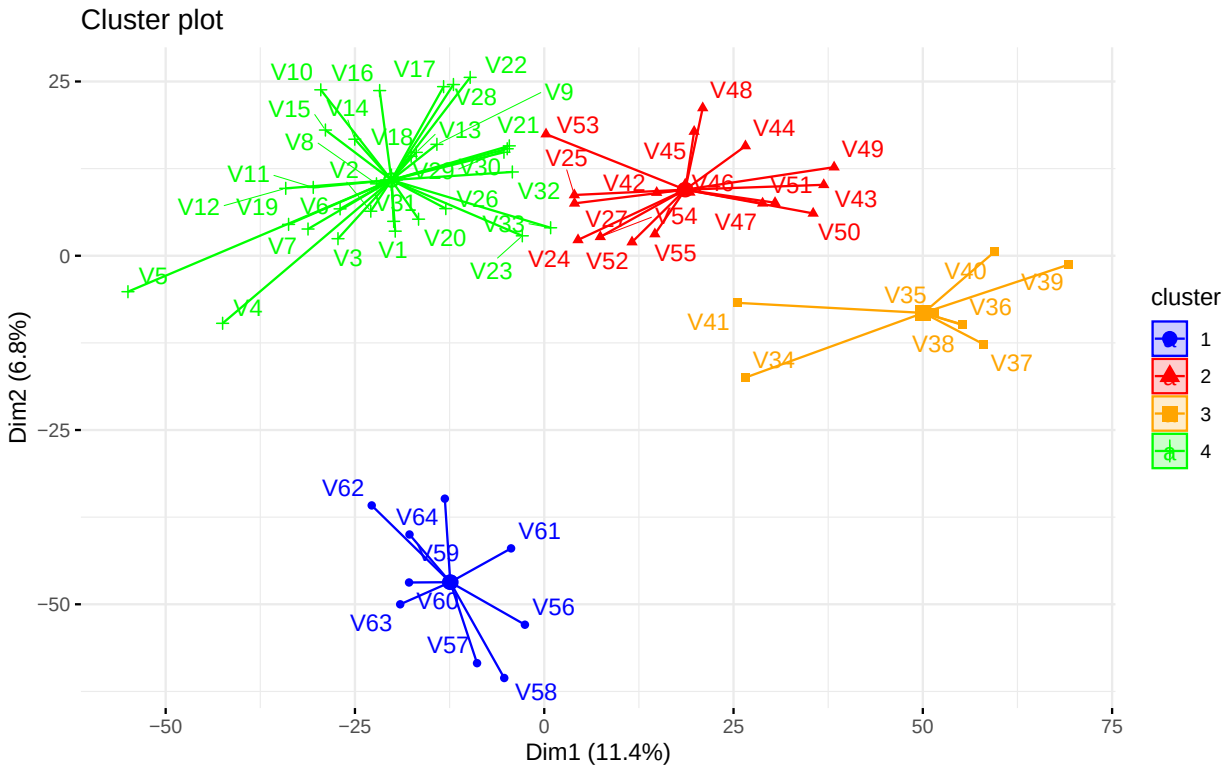
Hacemos el modelo de k-means, empezaremos haciendo 30 asignaciones iniciales y con 4 particiones como lo habiamos inferido con los modelos anteriores

```
k_mean=eclust(datos, "kmeans", k = 4, nstart = 30, graph = FALSE)
```

Graficando

```
fviz_cluster(k_mean, data = datos,
             palette = c("blue", "red", "orange", "green"),
             ellipse.type = "euclid",
```

```
            star.plot = TRUE,
            repel = TRUE,
            ggtheme = theme_minimal()
)
```



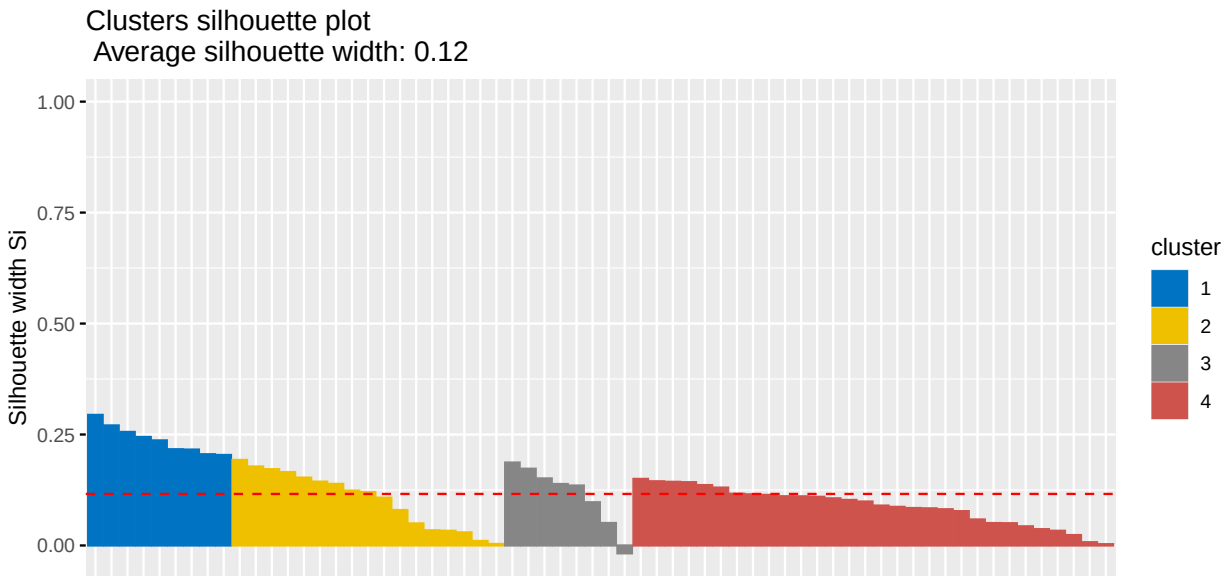

```
k_mean$size
```

```
## [1]  9 17  8 30
```

Podemos observar que al realizar el algoritmo k-means con 4 particiones obtenemos clusters bien definidos, con diferentes tamaños pero mejor distribuidos, donde sí es posible diferenciar los clusters a simple vista, a pesar de esto el promedio de concordancia con el cluster asignado no es alto (0.12), haremos una grafica de silueta para buscar variables que esten mal clasificadas.

```
fviz_silhouette(k_mean, palette = "jco")
```

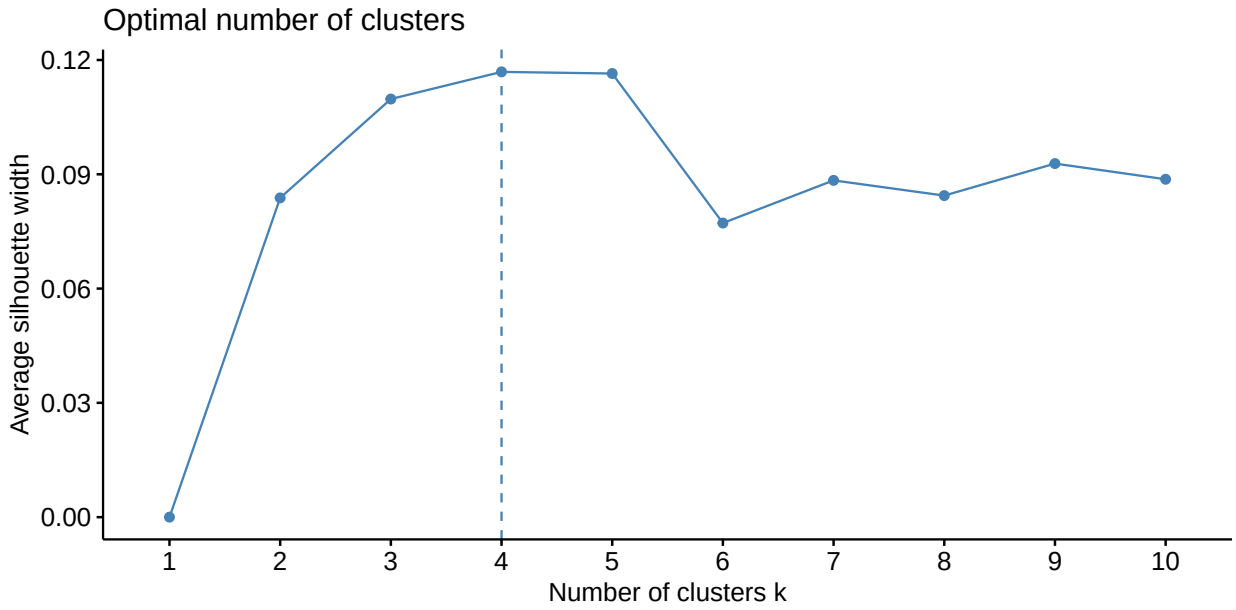
```
##   cluster size ave.sil.width
## 1       1    9          0.24
## 2       2   17          0.10
## 3       3    8          0.11
## 4       4   30          0.09
```



Podemos ver que existe alguna variable que posiblemente este mal calsificada, pero al estar muy cerca de el cluster rojo, seguramente podemos aceptar este modelo aun con esa variable, pensando que podria pertencer a un 5 cluster haremos la regla del codo para ver cual es el número optimo de clusters.

Para ver el numero optimo de clusters en k-means

```
fviz_nbclust(datos, kmeans, method='silhouette')
```



Al parecer tenemos el número correcto de particiones, asi que haremos la comparación entre AGNES con metodo Ward y K-means.

Conclusiones

Al comparar los modelos AGNES_Ward con el modelo K-means, notamos que el corte debe ser en k=4 en ambos modelos, al analizar gráficamente los clusters de cada modelo podemos observar que en el modelo de

K-means los clusters se ven mejor clasificados ya que los clusters no se tocan ni comparten datos, comparado con el modelo AGNES que hay grupos que se llegan a tocar.

```
k_mean$size
```

```
## [1]  9 17  8 30
```

```
table(grp_5)
```

```
## grp_5
##  1  2  3  4
## 35 13  8  8
```

El número de variables en cada cluster es parecido, pero no igual, notamos que en k-means el cluster mas grande tiene 30 variables, mientras en AGNES tiene 35, podemos intuir que estas variables faltantes se fueron al 2 cluster mas grande ya que tiene 17 y 13 respectivamente, los dos ultimos clusters son iguales practicamente. Podemos concluir que el modelo K-Means hace una mejor clasificación de los datos.