**Análisis de datos en el dataset de próstata de cáncer**

Para substituir los valores perdidos en el conjunto de datos, se utilizó un KNN-Imputation, con k=10 y los datos fueron escalados antes de calcular los vecinos más cercanos, de esta manera variables con diferentes escalas no tuvieron distintos impactos en el cómputo de distancia.

Para detectar subconjuntos de factores relevantes, es decir aquellos que posibilitan la inducción de modelos predictivos con alta capacidad de generalización, se utilizó el procedimiento descrito en [1], donde primeramente se crea un ranking de factores ordenados de mayor a menor importancia, y luego se realiza una búsqueda heurística sobre dicho ranking para encontrar los mejores subconjuntos de factores.

En este trabajo, en todos los casos se utilizó un *Leave One Group Out cross-validation*, que utiliza en cada *fold-execution* todos los datos como datos de entrenamiento, excepto un par de *samples* pareados que es separado y utilizado como test. Por lo tanto, siguiendo este procedimiento se realizan tantas ejecuciones como pares de muestras pareadas tengamos en el conjunto de datos (83 en nuestro caso), lo cual nos permitió obtener el menor sesgo posible en la estimación de la importancia de los factores y en la construcción de los modelos predictivos. Además, en cada *fold-execution* los datos de entrenamiento son transformados mediante un z-score y un power transformation de YeoJohnson, y la información obtenida es posteriormente utilizada para transformar los datos de test.

A partir de la 1er etapa, obtuvimos el siguiente ranking de factores. Respecto a los tres factores que se esperaban que salieran como más importantes, el ranking es algo contradictorio…pues los factores SRRM1, SNRNP200 y SRSF3 se posicionan al final del ranking, mostrando una menor importancia para separar las muestras en los dos grupos definidos. Para salir de dudas, el ranking también se calculó con la herramienta SAM (Significant Analysis of MicroArrays), la cual computa la importancia de los factores haciendo permutaciones aleatorias. En este último ranking, los tres factores antes citados no terminan al final del ranking, pero tampoco son de los primeros.

A partir de la 2da etapa de la metodología computacional, que propone realizar una búsqueda heurística sobre el ranking de factores, obtenemos los siguientes resultados.

* Utilizando Logistic Regression (LR) se pudieron construir 56 modelos que tenían un AUC mayor de 0.8. De los 56 modelos de LR solo en tres aparece el factor SNRNP200, SRSF3 aparece en 18 modelos, mientras que SRRM1 aparece en 5.
* Utilizando Random Forest (RF) se detectaron 152 modelos que lograban un AUC mayor de 0.8. De los 152 modelos de RF, SNRNP200 solo aparece en 11, SRRM1 aparece en 9 modelos y SRSF3 en 49 modelos.

La combinación (**SRSF3, SNRNP200**, **SRRM1**) solo apareció 1 vez en todos los modelos construidos, y esto corresponde a un modelo RF. Este modelo RF se compone de los siguientes 11 factores y tiene las siguientes características:

Factores: TIA1 SRSF10 U2AF2 SRSF9 SRRM4 SND1 ESRP1 **SRSF3** KHDRBS1 **SNRNP200** **SRRM1**

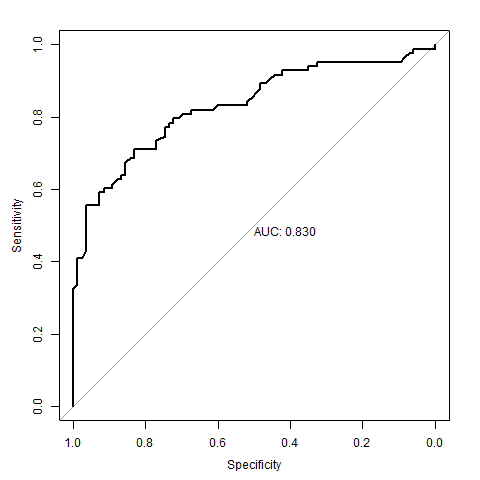
**Nota: Todos estos factores aparecen como relevantes en la imagen que me envió Juan Manuel.**

Para analizar la capacidad de memorización sobre los datos de entrenamiento (análisis que comúnmente se hace en los papers de biología), se probó el modelo sobre el mismo conjunto de datos sobre el cual se construyó, dando los siguientes resultados:

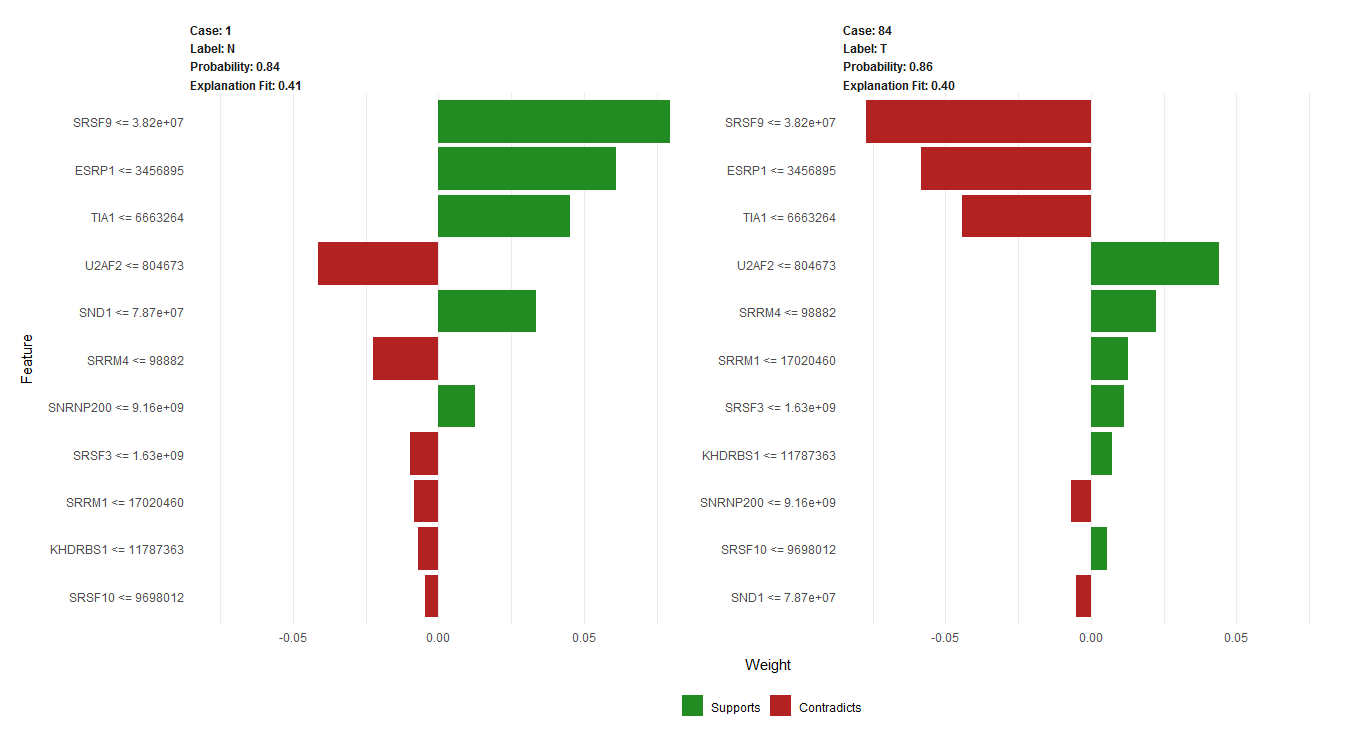
AUC= 1, p-value= 4.824e-29, Sens=1 y Spec= 1.

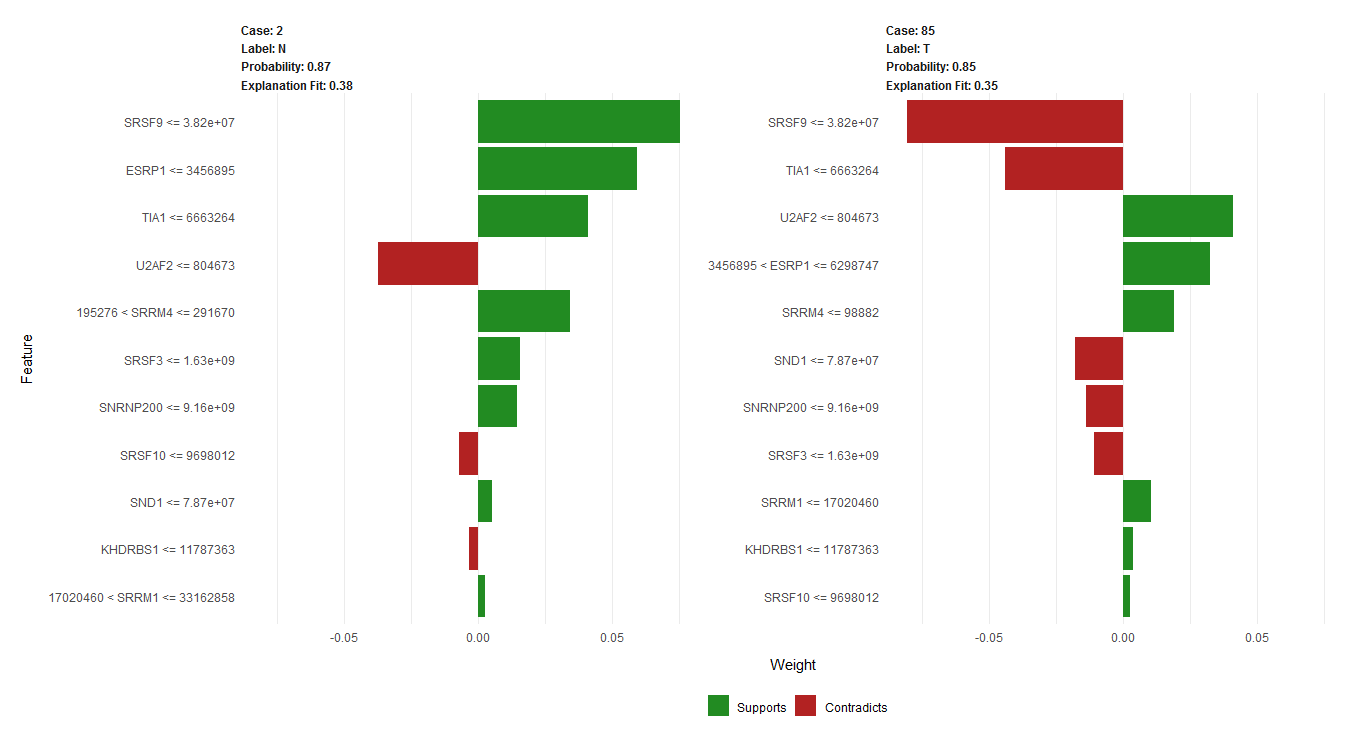
Por otra parte, para analizar la capacidad de generalización del modelo (análisis que realmente muestra qué tan bueno es el modelo para predecir muestras que nunca ha visto), los siguientes resultados se obtuvieron tras ejecutar el proceso de *cross-validation* mencionado al comienzo del documento:

AUC= 0.830 con un 95% CI: 0.7671-0.8932 (DeLong), p-value= 1.034e-13, Sens= 0.831 y Spec= 0.711.

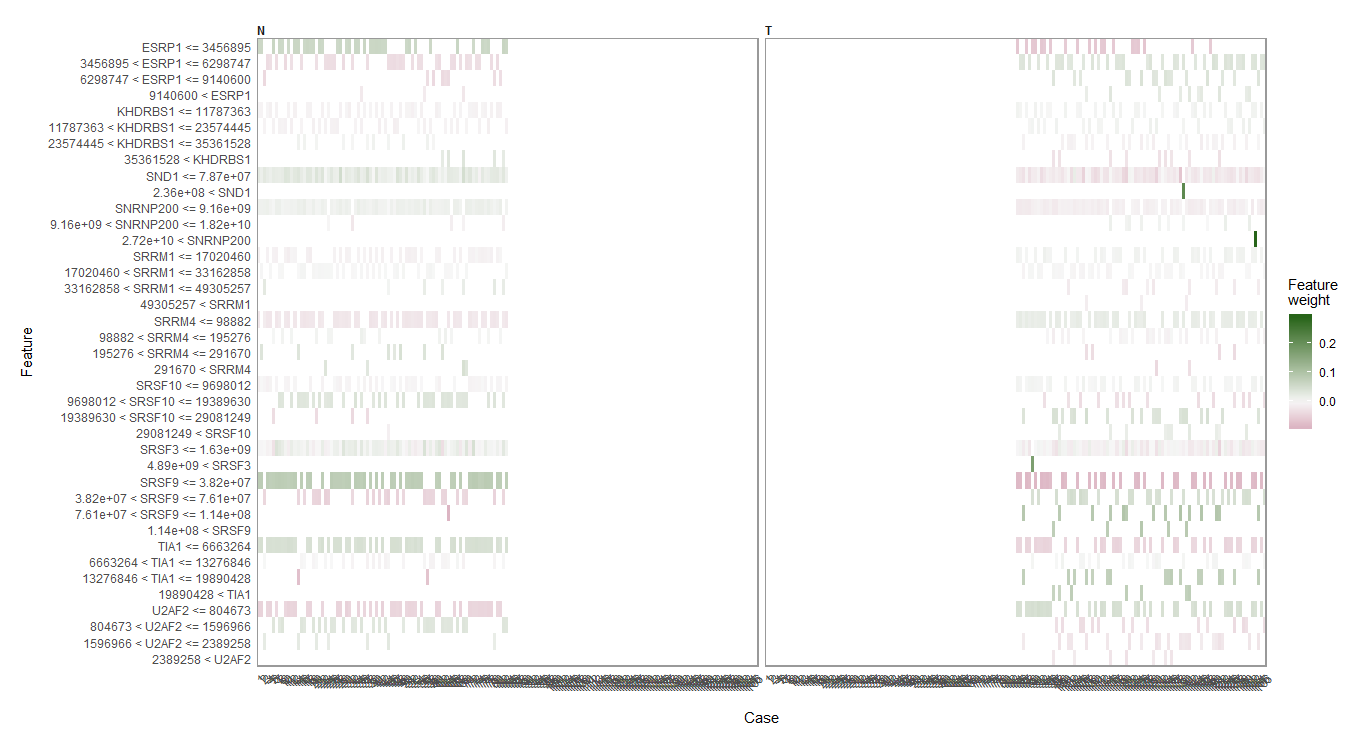


Ahora analicemos que importancia están teniendo realmente estos tres factores de interés en la predicción del modelo.





Las dos figuras nos muestran la importancia que tuvieron cada uno de los 11 factores que componen el modelo RF en la predicción de los pares de muestras (1, 84) y (2,85). Podemos apreciar que los factores SNRNP200, SRRM1 y SRSF3 contribuyen en menor grado a la predicción, mientras que los factores SRSF9, ESRP1, TIA1, etc, son los que más importancia tienen.



En la figura anterior podemos ver un resumen de la importancia de cada factor en todas las muestras. Aunque podemos observar que en algunos casos específicos si que SRSF3 y SNRN200 tuvieron una importancia mayor, de manera general los factores con más importancia fueron ESPR1 y SRSF9.

**Dejando a un lado los tres factores de interés preestablecidos, es de destacar que existen muchos modelos RF que obtienen mejores valores de AUC. Por ejemplo, el modelo RF compuesto por “ESRP2 SRSF5 SRSF4 RNU6 SRSF10 PRPF40A” obtiene un AUC = 0.952, Sens= 0.867 y Spec= 0.759. Tendrían que decirme si hay algún modelo de los que aparecen en los ficheros rf.csv y glm.csv que os interese analizar con más profundidad, hasta ahora lo que hecho es concentrarme en aquellos modelos que tuvieran los tres factores que me indicó Juanma.**

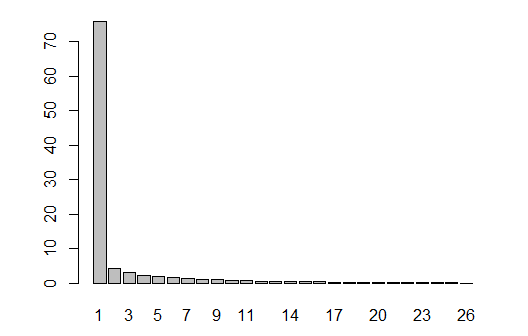
2- Como segundo paso de este análisis, siguiendo la misma lógica utilizada para construir los modelos predictivos, tratamos de construir modelos de clustering. En este caso, utilizando un *Hierarchical Clustering* y todas las funciones de distancias posibles, no se encontró ningún subconjunto de factores que produzca un agrupamiento con una *purity* mayor de 0.75.

3- Por último, se realizó un análisis PCA sobre los datos. Este fue el resultado:

The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied.

The first two dimensions of PCA express **80.39%** of the total dataset inertia; that means that 80.39% of the individuals (or variables) cloud total variability is explained by the plane. This percentage is high and thus the first plane represents an important part of the data variability. This value is strongly greater than the reference value that equals **14.44%**, the variability explained by this plane is thus highly significant (the reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 1007 data tables of equivalent size on the basis of a normal distribution).

An estimation of the right number of axis to interpret suggests to restrict the analysis to the description of the first 1 axis. These axis present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (75.97% against 7.63%). This observation suggests that only this axis is carrying a real information.



**Figure. Decomposition of the total inertia on the components of the PCA** *The first factor is major: it expresses itself 75.97% of the data variability.* *Note that in such a case, the variability related to the other components might be meaningless, despite of a high percentage.*

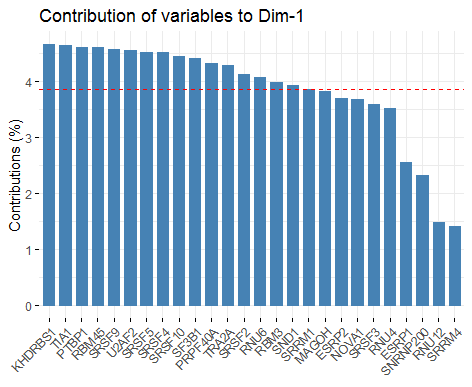


Figure. Contribution of splicing factors to dimension 1.

As shown in above Figure, the factors SNRNP200, SRSF3 y SRRM1 did not significantly contribute to explain the variability of individuals.

[1] Gahete, M. D., del Rio-Moreno, M., Camargo, A., Alcala-Diaz, J. F., Alors-Perez, E., Delgado-Lista, J., ... & Lopez-Miranda, J. (2018). Changes in Splicing Machinery Components Influence, Precede, and Early Predict the Development of Type 2 Diabetes: From the CORDIOPREV Study. *EBioMedicine*, *37*, 356-365.