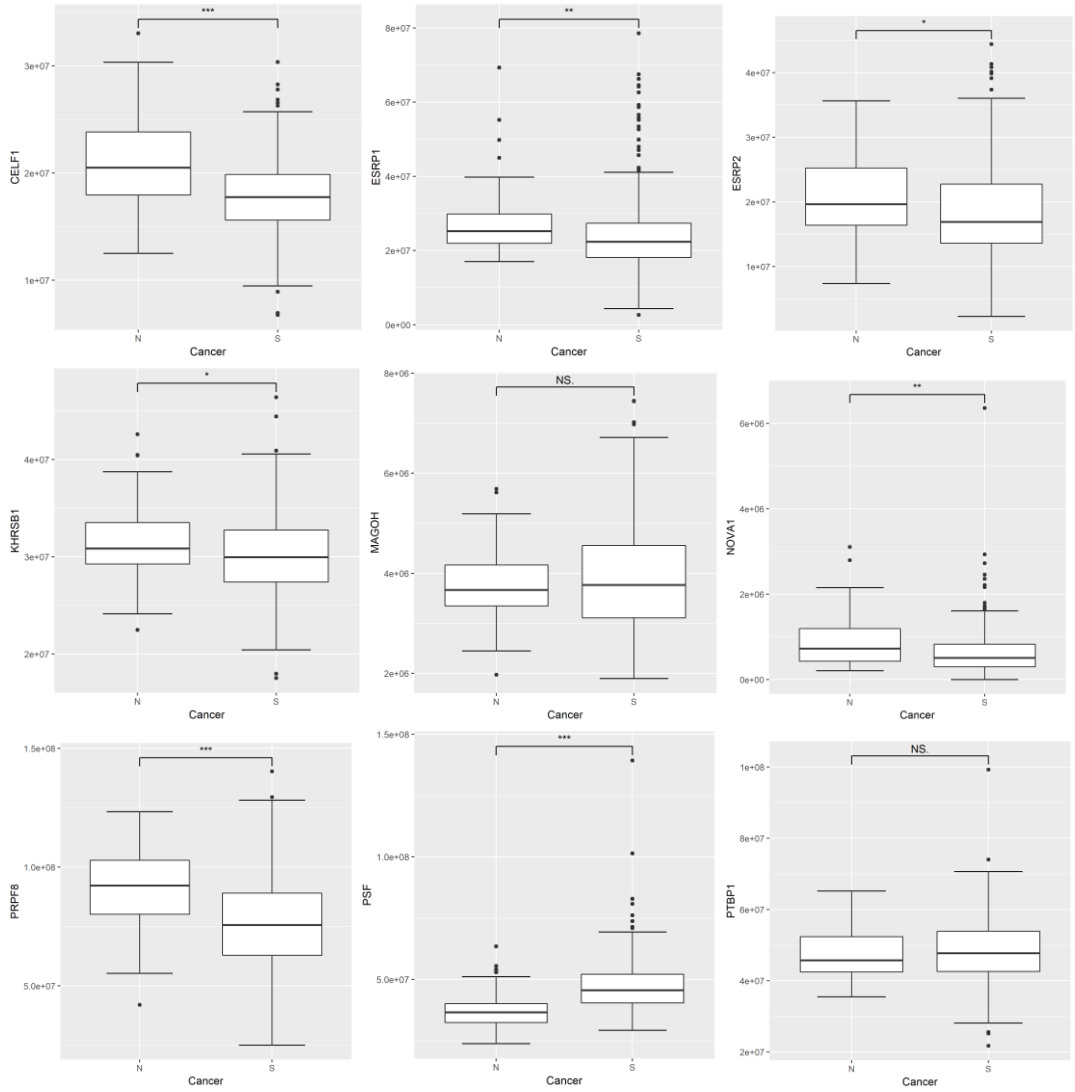
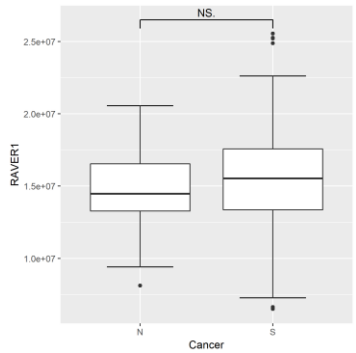
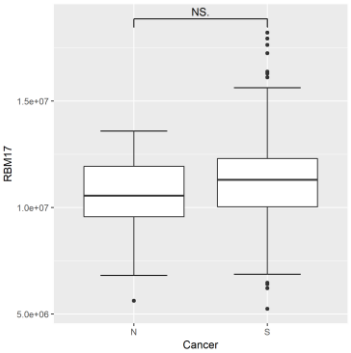
1- Se ejecutó un test de Wilcoxon para detectar diferencias significativas en las expresiones de factores entre los dos grupos. Existen diferencias significativas en los niveles de expresión de los siguientes factores:

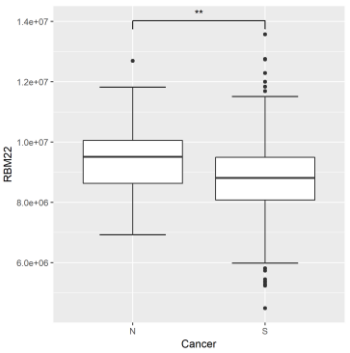
*	**	***
U2AF2	NOVA1	PRPF8
ESRP2	RBM22	PSF
KHRSB1	SF3B1	RBM3
RBM3	ESRP1	RBM45
RBM22		SRSF1
TRA2B		SRSF6
U2AF2		U2AF1
		CELF1
		SRRM4
		SRSF4
		U2AF1

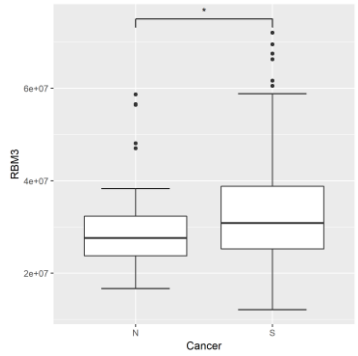
Boxplots para cada factor:

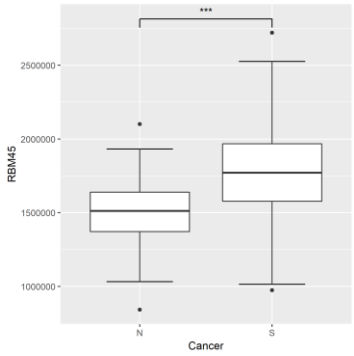


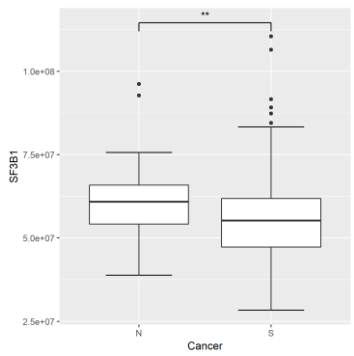

NS.
RAVER1
2.5e+07
2.0e+07
1.5e+07
1.0e+07
N
S
Cancer

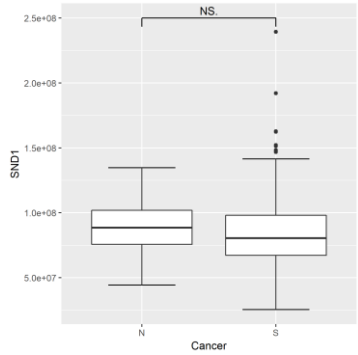

NS.
RBM17
1.5e+07
1.0e+07
5.0e+06
N
S
Cancer

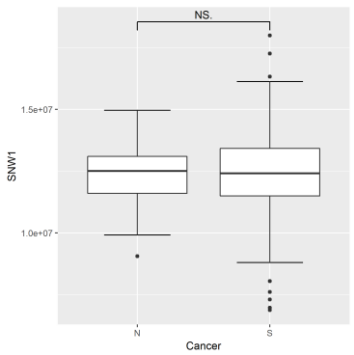

**
RBM22
1.4e+07
1.2e+07
1.0e+07
8.0e+06
6.0e+06
N
S
Cancer

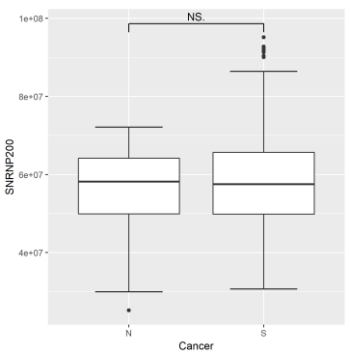

*
RBM3
6e+07
4e+07
2e+07
N
S
Cancer

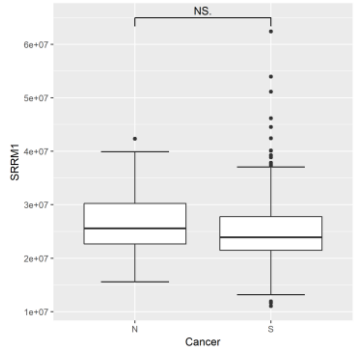

***
RBM45
2500000
2000000
1500000
1000000
N
S
Cancer

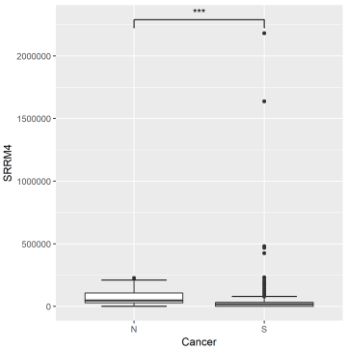

**
SF3B1
1.0e+08
7.5e+07
5.0e+07
2.5e+07
N
S
Cancer

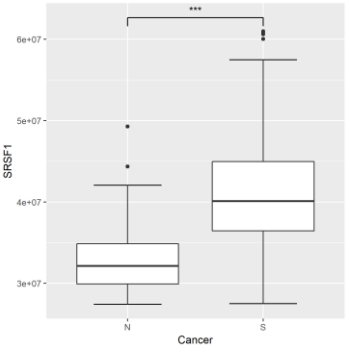

NS.
SND1
2.5e+08
2.0e+08
1.5e+08
1.0e+08
5.0e+07
N
S
Cancer

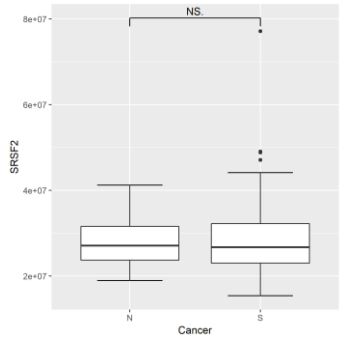

NS.
SNW1
1.5e+07
1.0e+07
N
S
Cancer

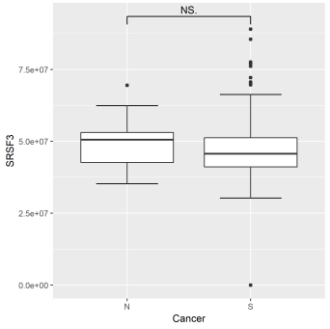

NS.
SNRNP200
1e+08
8e+07
6e+07
4e+07
N
S
Cancer

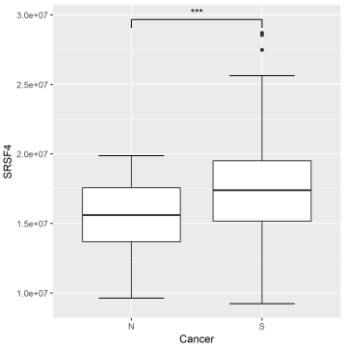

NS.
SRRM1
6e+07
5e+07
4e+07
3e+07
2e+07
1e+07
N
S
Cancer

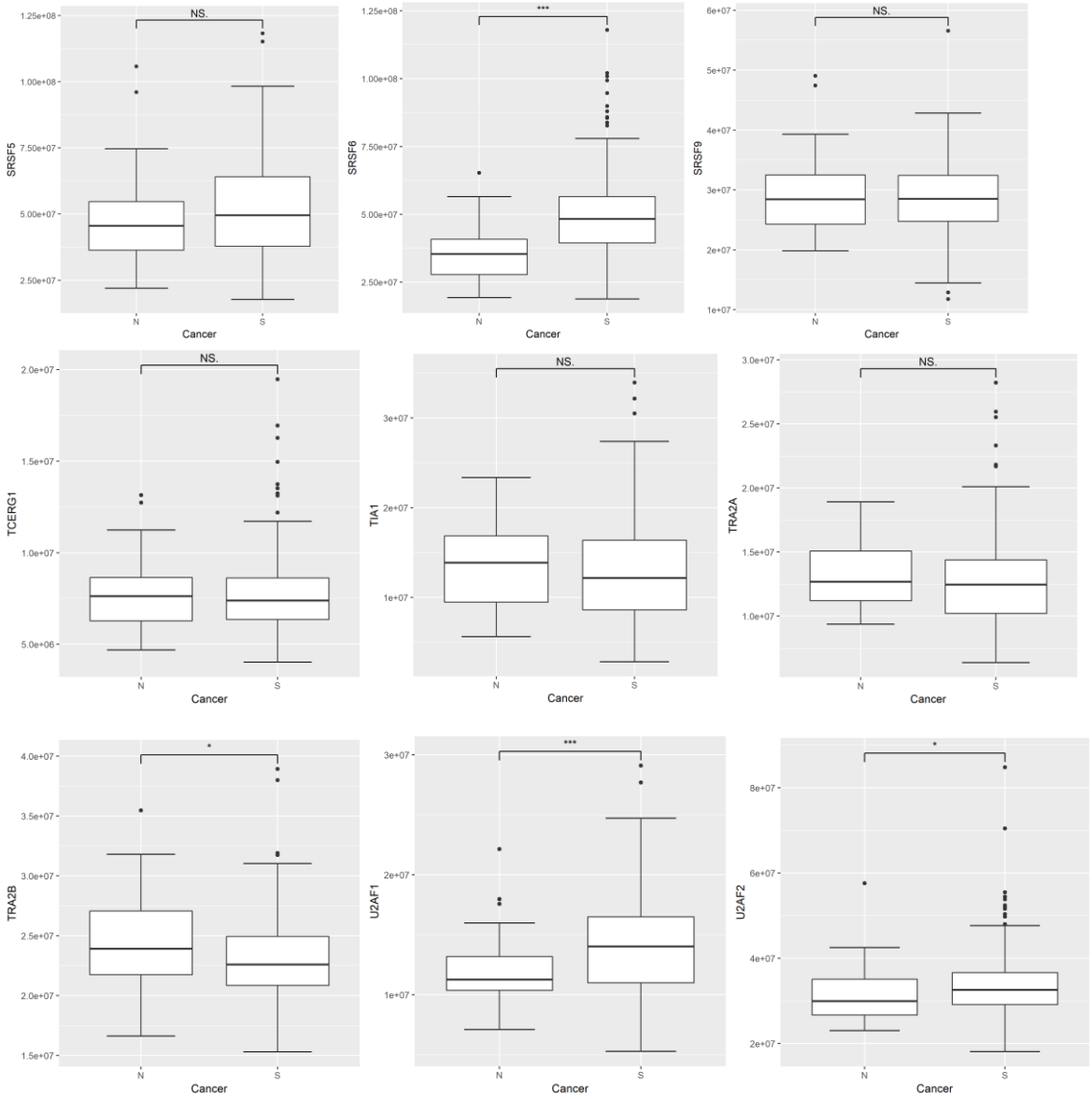

***
SRRM4
2000000
1500000
1000000
500000
0
N
S
Cancer


***
SRSF1
6e+07
5e+07
4e+07
3e+07
N
S
Cancer


NS.
SRSF2
8e+07
6e+07
4e+07
2e+07
N
S
Cancer


NS.
SRSF3
7.5e+07
5.0e+07
2.5e+07
0.0e+00
N
S
Cancer


***
SRSF4
3.0e+07
2.5e+07
2.0e+07
1.5e+07
1.0e+07
N
S
Cancer



2- El dataset no tiene valores perdidos.

3- Teniendo en cuenta los grupos de muestras, se eliminaron todos aquellos valores que estaban por encima de 3er cuartil + 1.5*IQR y por debajo de 1er cuartil-1.5*IQR. Los outliers fueron sustituidos por la mediana de los valores de un mismo grupo.

"Ouliers of var MAGOH in class N: 3" "Ouliers of var MAGOH in class S: 5"
"Ouliers of var RBM22 in class N: 1" "Ouliers of var RBM22 in class S: 15"
"Ouliers of var SRRM4 in class N: 1" "Ouliers of var SRRM4 in class S: 34"
"Ouliers of var PSF in class N: 5" "Ouliers of var PSF in class S: 9"
"Ouliers of var NOVA1 in class N: 2" "Ouliers of var NOVA1 in class S: 11"
"Ouliers of var SRSF6 in class N: 1" "Ouliers of var SRSF6 in class S: 12"
"Ouliers of var RBM3 in class N: 5" "Ouliers of var RBM3 in class S: 6"
"Ouliers of var RBM45 in class N: 2" "Ouliers of var RBM45 in class S: 2"
"Ouliers of var SRSF1 in class N: 3" "Ouliers of var SRSF1 in class S: 4"
"Ouliers of var SNRNP200 in class N: 1" "Ouliers of var SNRNP200 in class S: 8"
"Ouliers of var SRRM1 in class N: 1" "Ouliers of var SRRM1 in class S: 18"
"Ouliers of var SF3B1 in class N: 2" "Ouliers of var SF3B1 in class S: 6"
"Ouliers of var U2AF1 in class N: 4" "Ouliers of var U2AF1 in class S: 2"
"Ouliers of var U2AF2 in class N: 1" "Ouliers of var U2AF2 in class S: 11"
"Ouliers of var TCERG1 in class N: 2" "Ouliers of var TCERG1 in class S: 9"
"Ouliers of var PRPF8 in class N: 1" "Ouliers of var PRPF8 in class S: 2"
"Ouliers of var CELF1 in class N: 1" "Ouliers of var CELF1 in class S: 9"

"Ouliers of var ESRP1 in class N: 4" "Ouliers of var ESRP1 in class S: 25"
"Ouliers of var ESRP2 in class N: 0" "Ouliers of var ESRP2 in class S: 7"
"Ouliers of var RBM17 in class N: 1" "Ouliers of var RBM17 in class S: 11"
"Ouliers of var SNW1 in class N: 1" "Ouliers of var SNW1 in class S: 8"
"Ouliers of var SND1 in class N: 0" "Ouliers of var SND1 in class S: 8"
"Ouliers of var TIA1 in class N: 0" "Ouliers of var TIA1 in class S: 3"
"Ouliers of var SRSF2 in class N: 0" "Ouliers of var SRSF2 in class S: 4"
"Ouliers of var SRSF4 in class N: 0" "Ouliers of var SRSF4 in class S: 3"
"Ouliers of var SRSF5 in class N: 2" "Ouliers of var SRSF5 in class S: 2"
"Ouliers of var SRSF9 in class N: 2" "Ouliers of var SRSF9 in class S: 3"
"Ouliers of var TRA2A in class N: 0" "Ouliers of var TRA2A in class S: 6"
“Ouliers of var TRA2B in class N: 1" "Ouliers of var TRA2B in class S: 4"
"Ouliers of var KHRSB1 in class N: 4" "Ouliers of var KHRSB1 in class S: 5"
"Ouliers of var RAVER1 in class N: 1" "Ouliers of var RAVER1 in class S: 6"
"Ouliers of var PTBP1 in class N: 0" "Ouliers of var PTBP1 in class S: 5"
"Ouliers of var SRSF3 in class N: 1" "Ouliers of var SRSF3 in class S: 10"

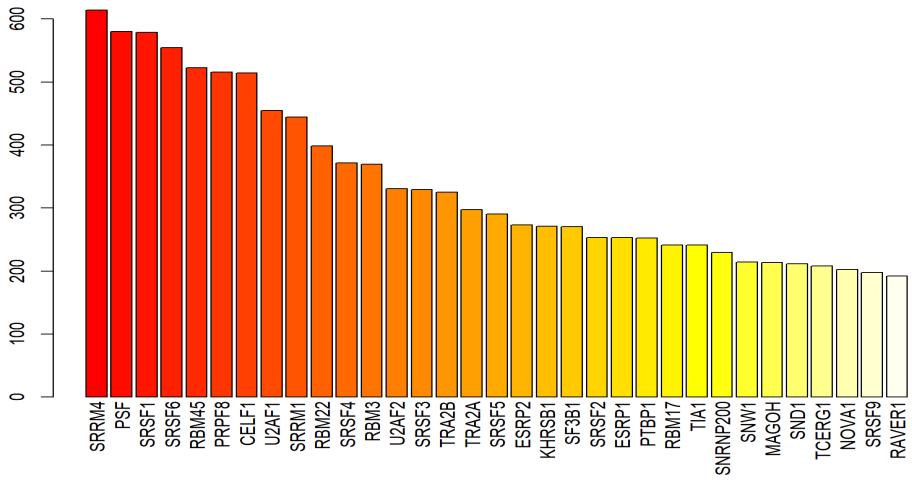
4- Se centraron los datos restando cada valor por la media, de esta manera se enfoca en las diferencias y no en las similaridades entre factores.

5- Se escalaron los datos dividiéndolos por la desviación estándar, de esta manera todos los factores pasan a tener igual importancia.

6- Se utilizó la transformación de Yeo-Johnson para corregir la heteroscedasticity en los datos.

7- No se detectaron correlaciones lineales mayores que 0.8 entre pares de factores.

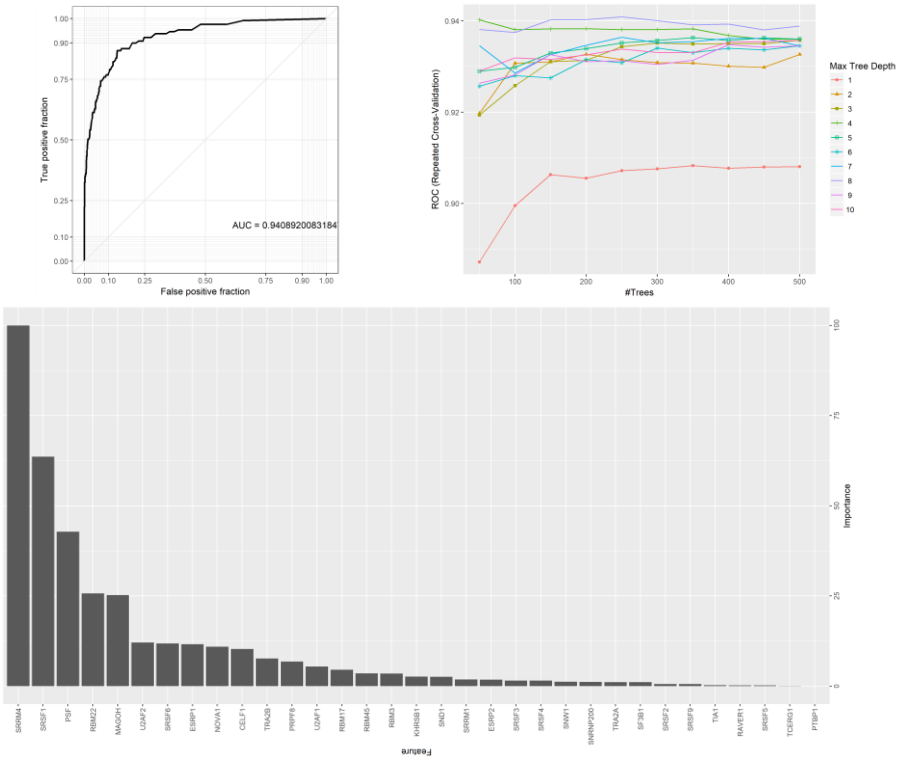
8- Se ejecutaron varios algoritmos de selección de características supervisados de tipo filter. Se genera un ranking de la importancia promedio de los factores. Los algoritmos ejecutados fueron: information.gain, gain.ratio, symmetrical.uncertainty, cfs, consistency y reliefF (para k=5..10). La siguiente figura muestra la relevancia promedio de los factores, los factores más a la izquierda son los más importantes.



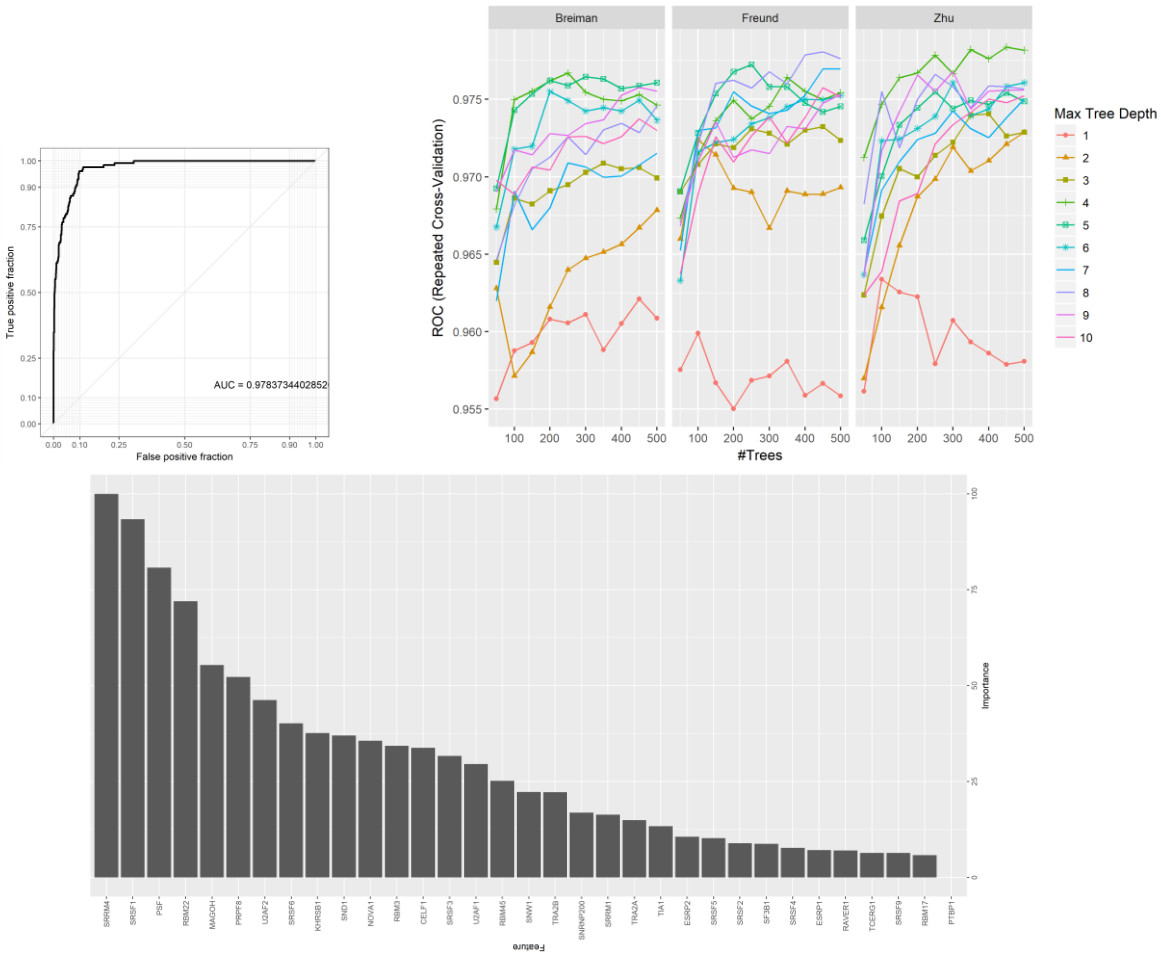
9- Se ejecutaron 26 modelos para clasificar los grupos de muestras, todos los modelos tienen implícito un proceso de selección de características, por lo tanto, a partir de cada modelo se estimó la importancia de los factores. Por cada modelo se realizó un proceso de tunning para encontrar la mejor combinación de parámetros que intervienen en la construcción del clasificador. La precisión de los modelos se estimó mediante un 10-fold cross validation repetido tres veces.

A continuación, se muestra por cada modelo el proceso de tunning, la curva ROC del mejor modelo y la importancia de los factores.

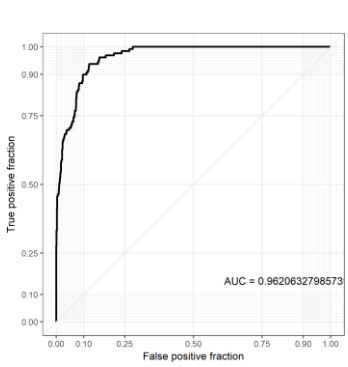
Bagged AdaBoost

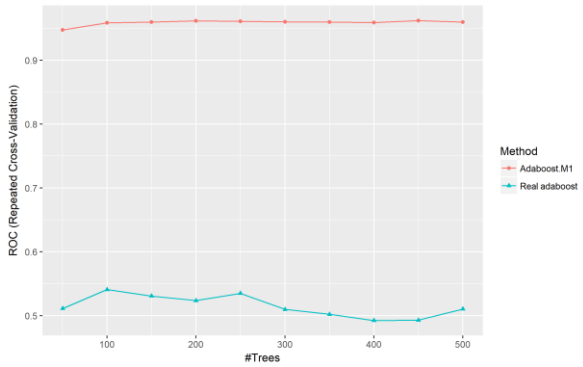


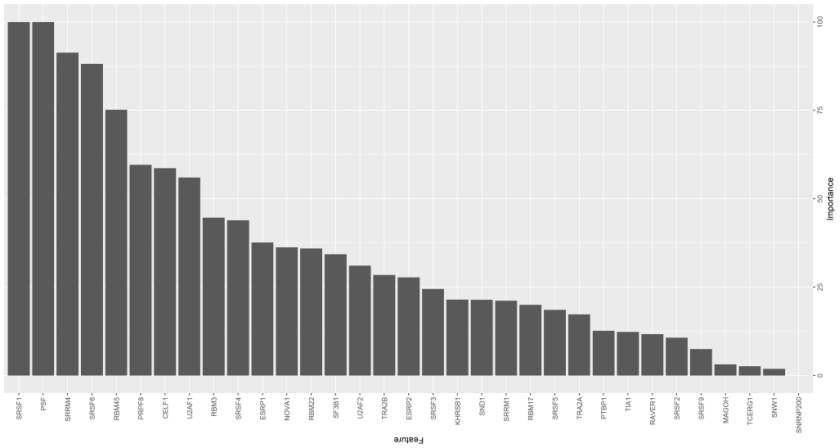
AdaBoost.M1



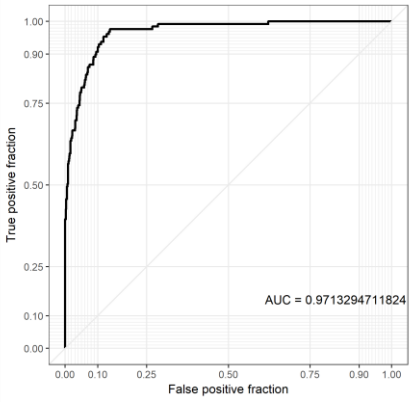
AdaBoost Classification Trees

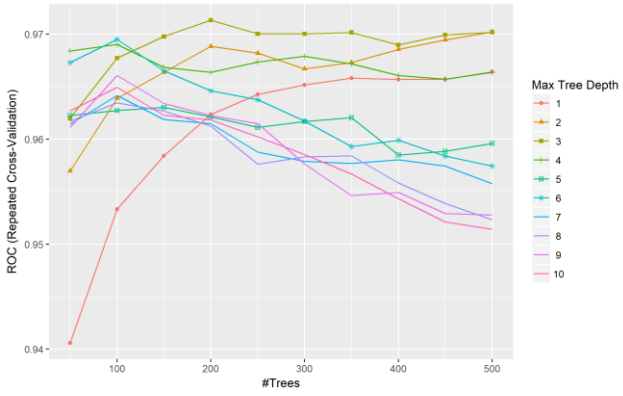


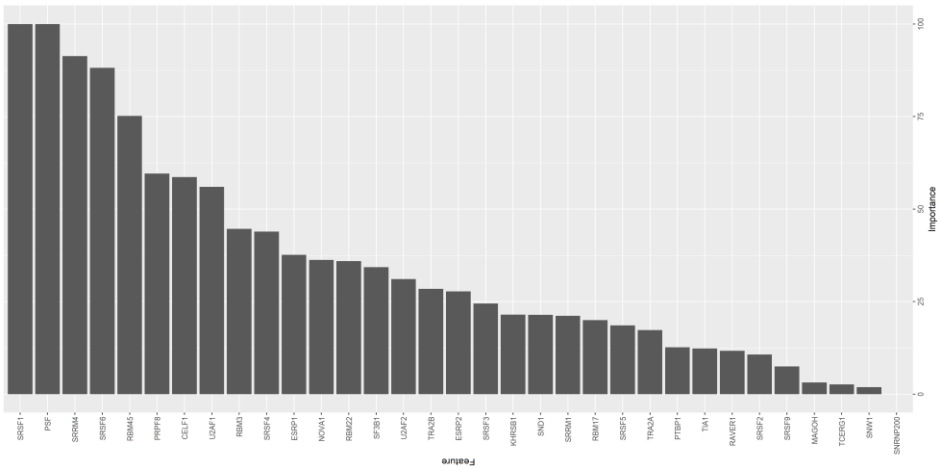




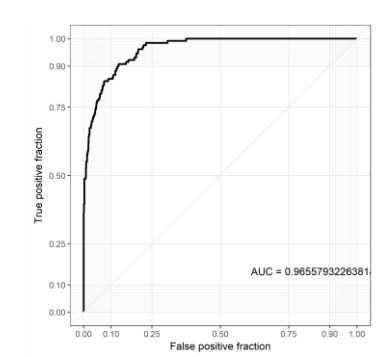
Boosted Classification Trees

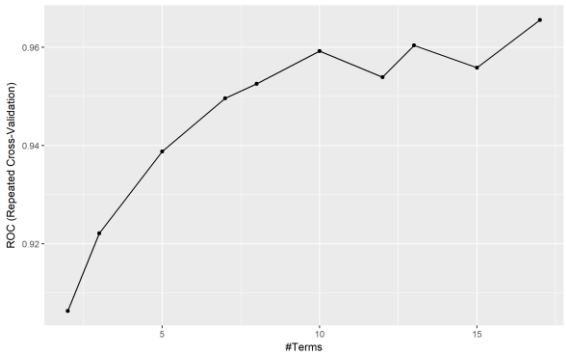


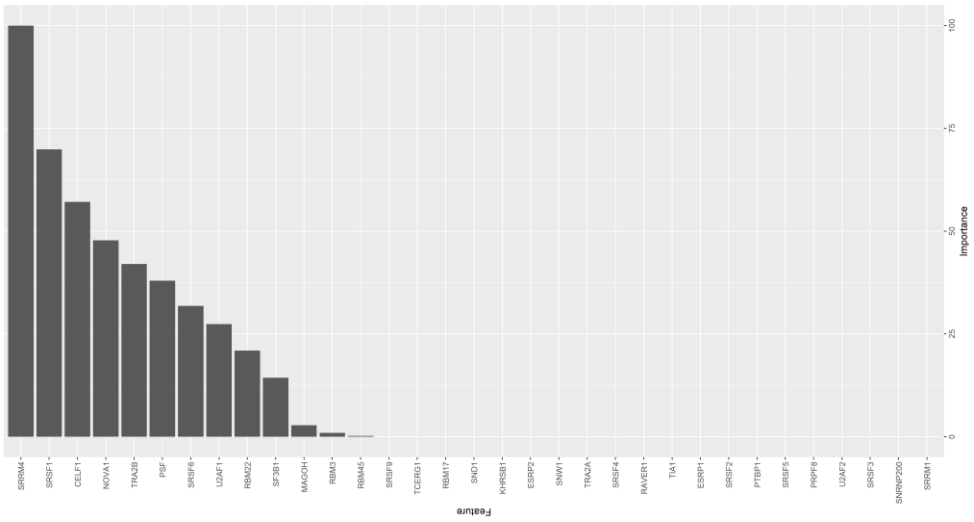




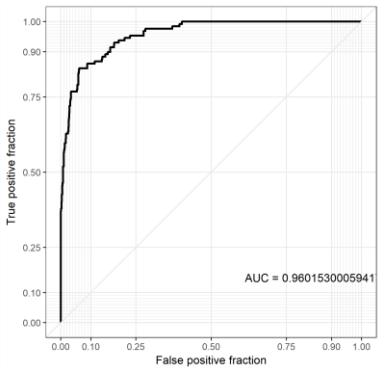
Bagged MARS

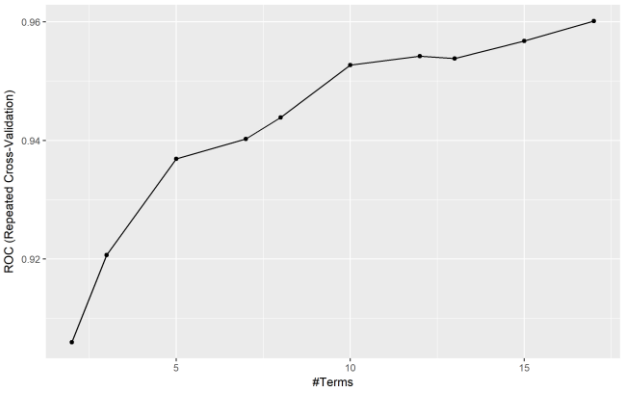


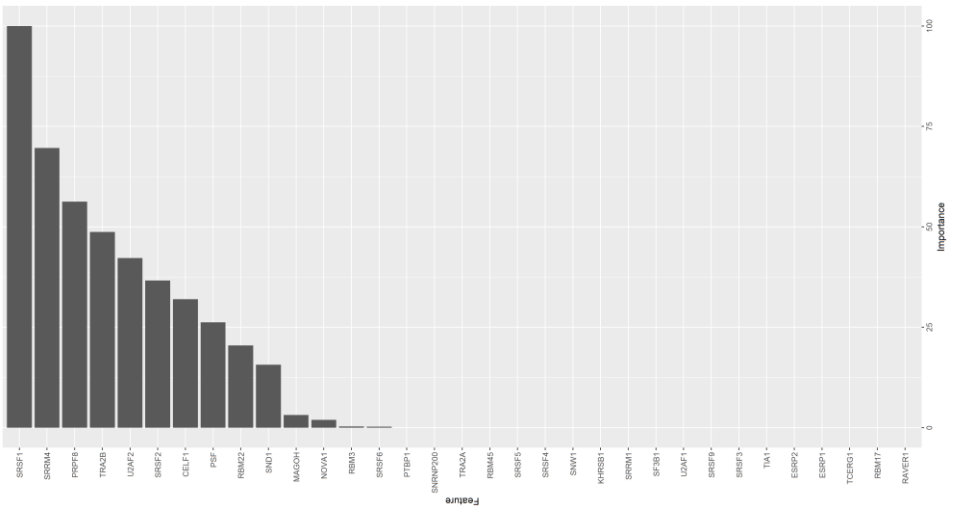




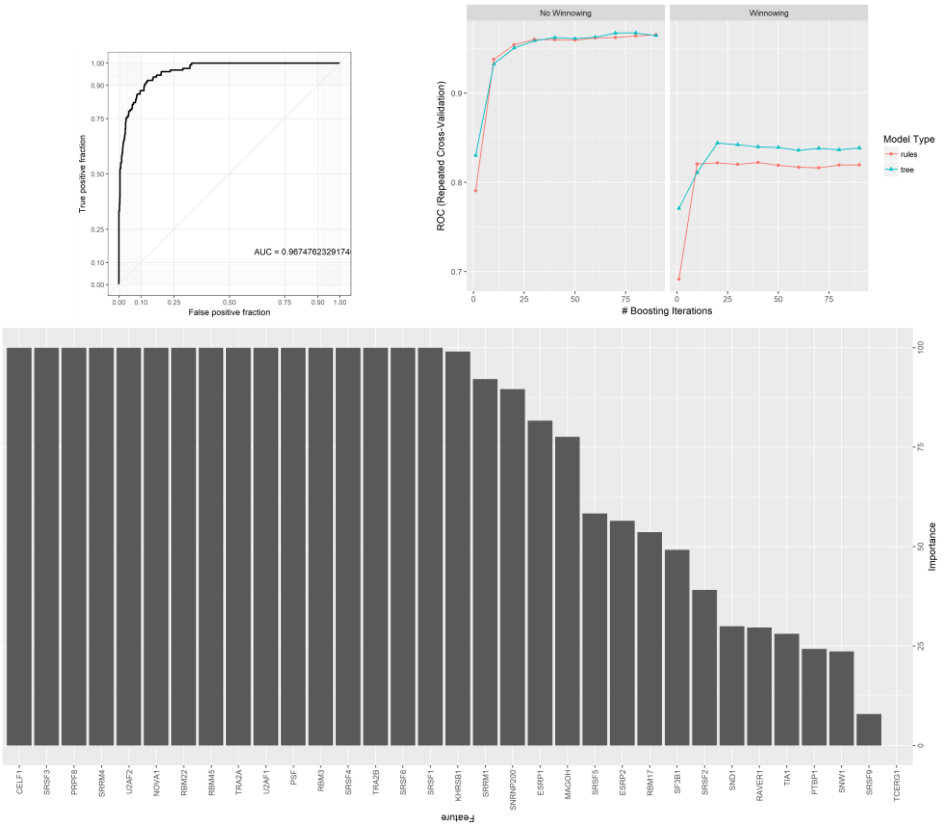
Bagged Flexible Discriminant Analysis



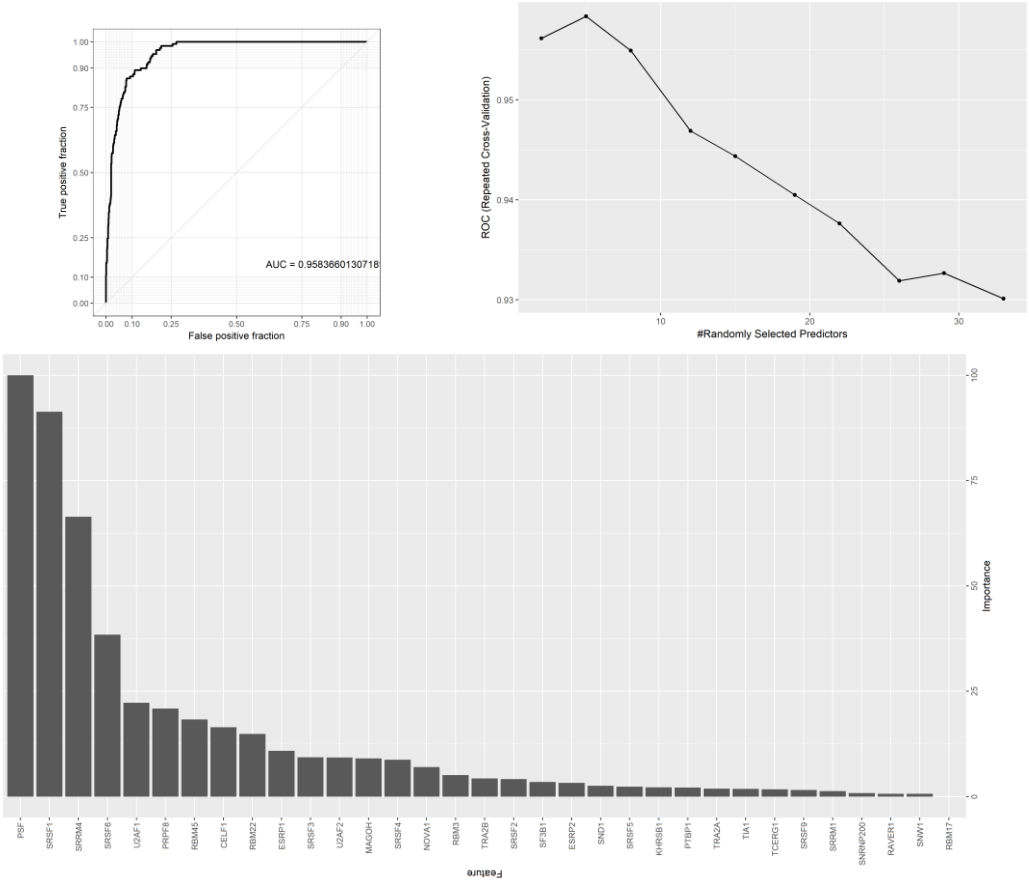




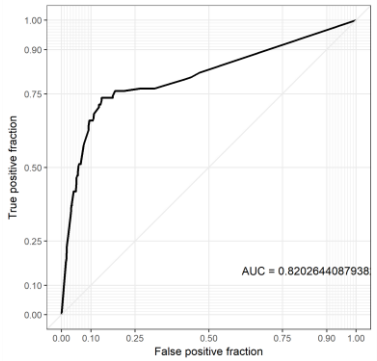
C5.0

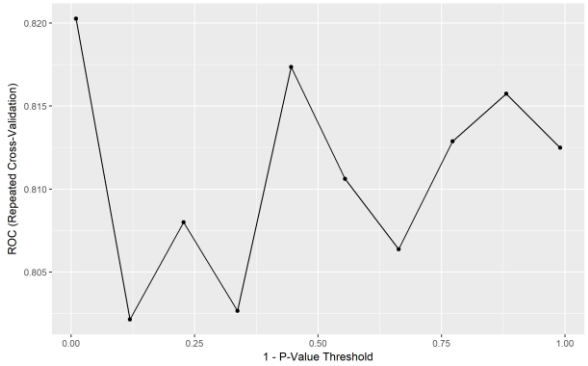


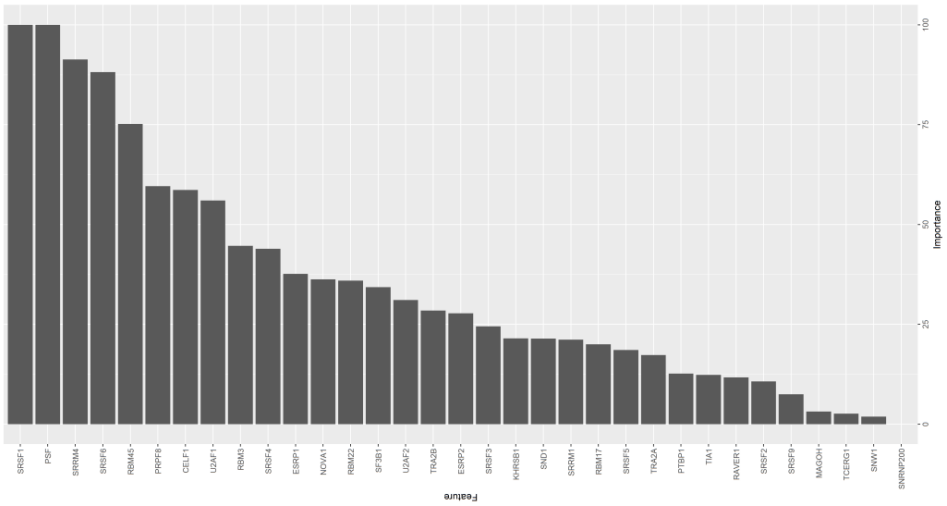
Conditional Inference Random Forest



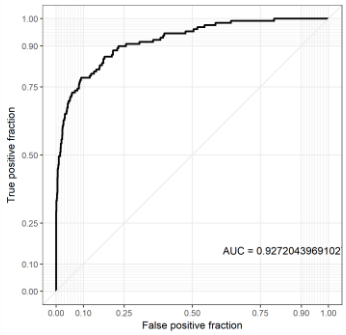
Conditional Inference Tree

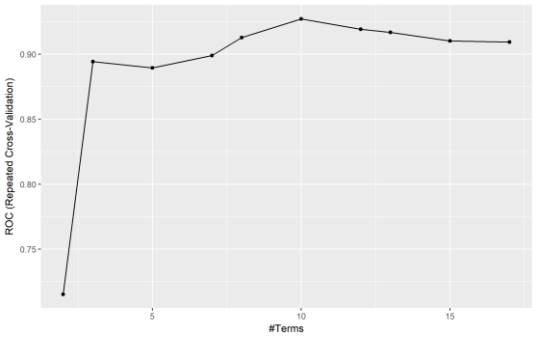


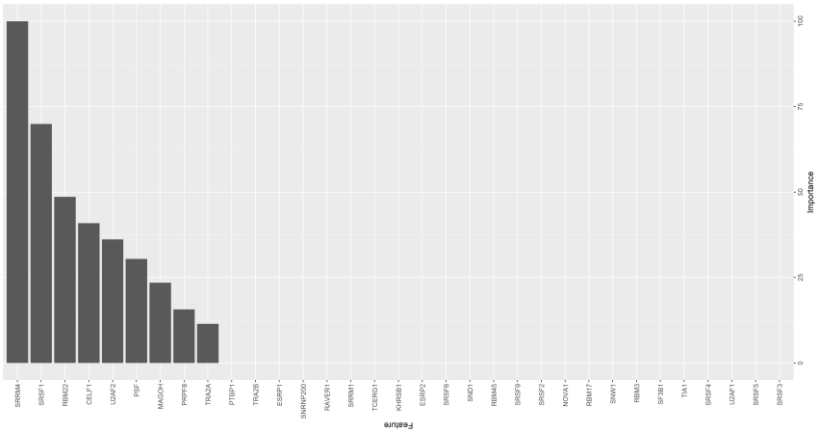




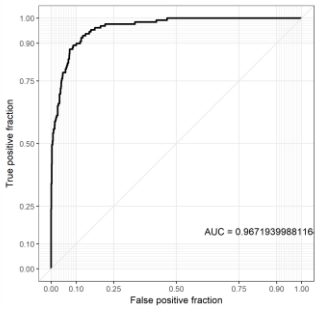
Multivariate Adaptive Regression Spline

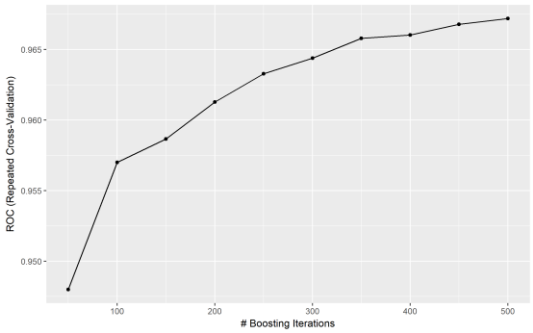


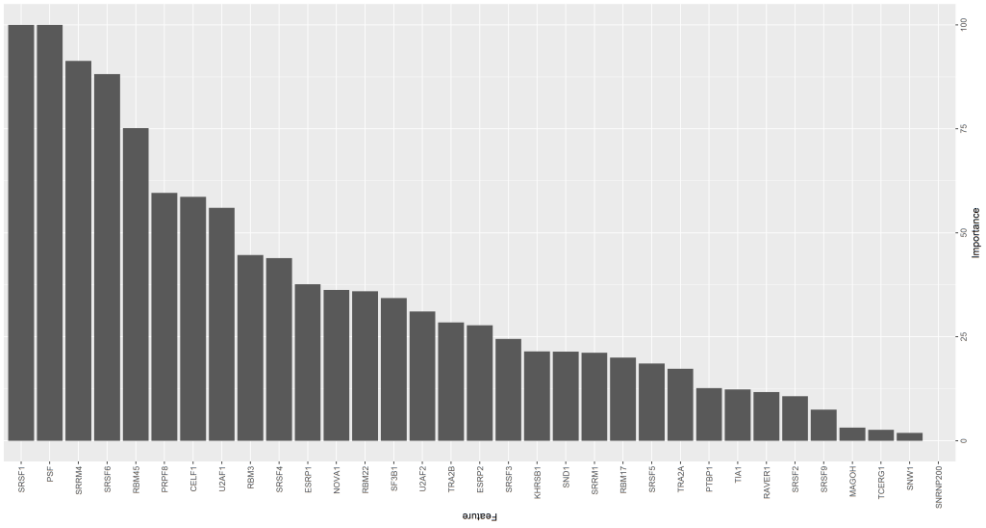




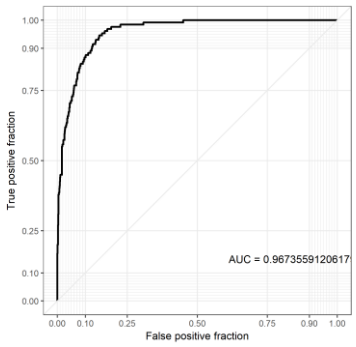
Boosted Generalized Additive Model

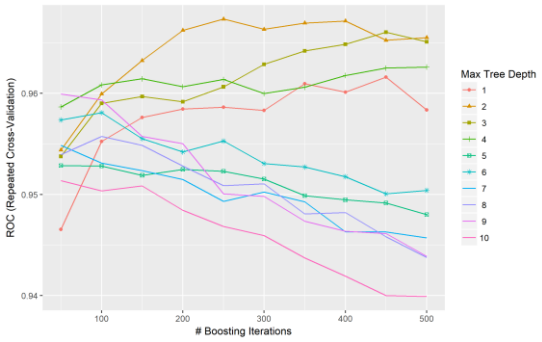


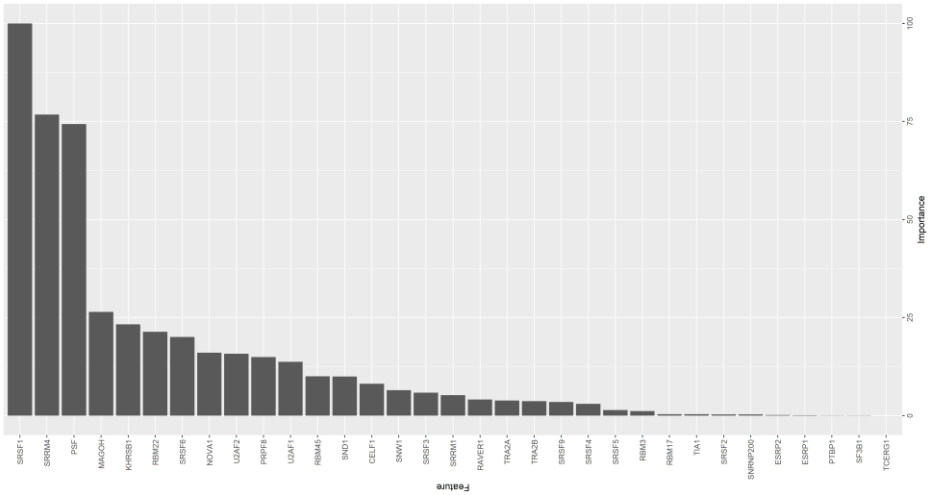




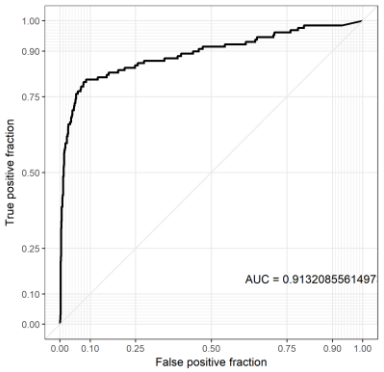
Stochastic Gradient Boosting

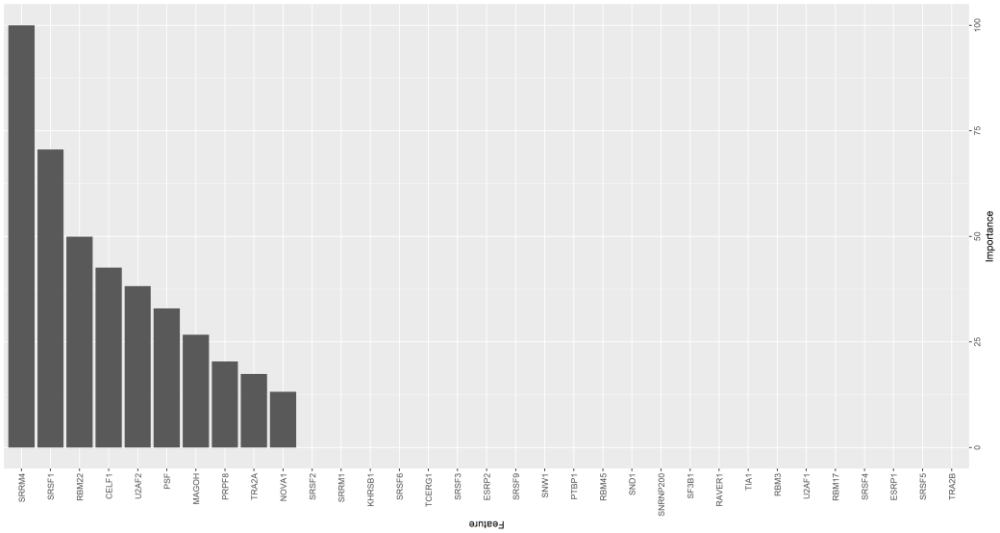




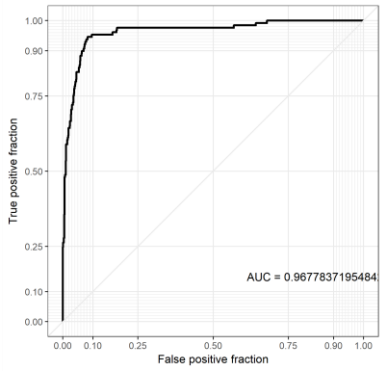


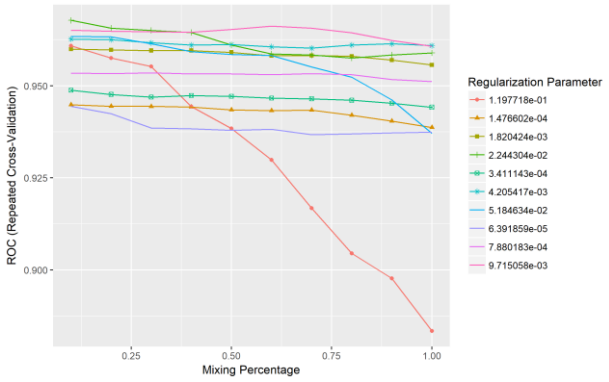
Gcv Earth

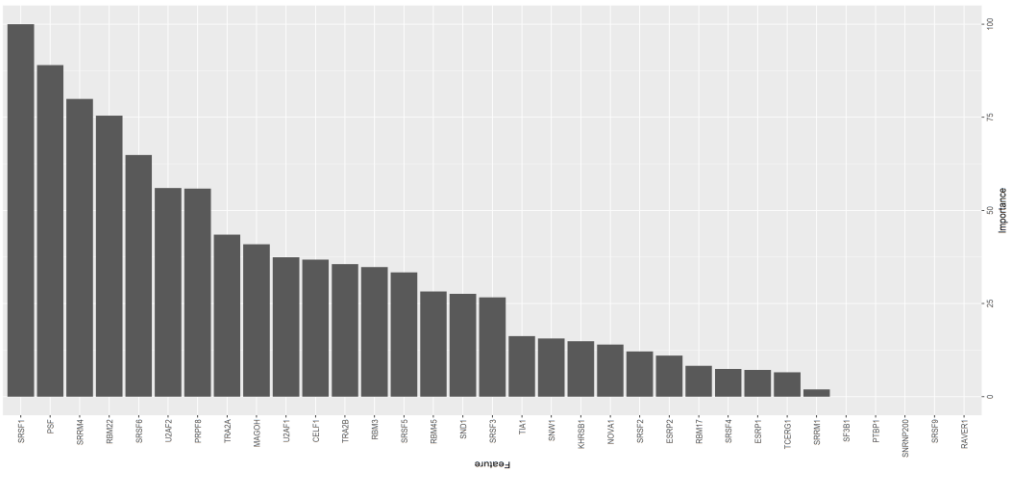




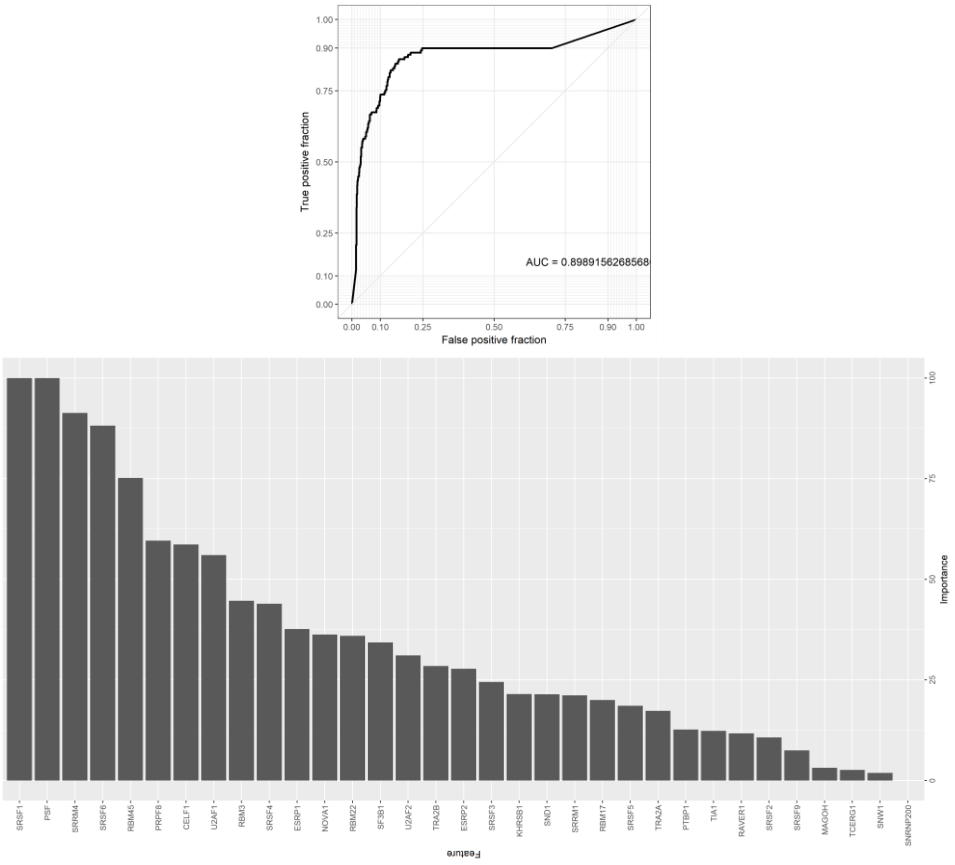
GLMNET



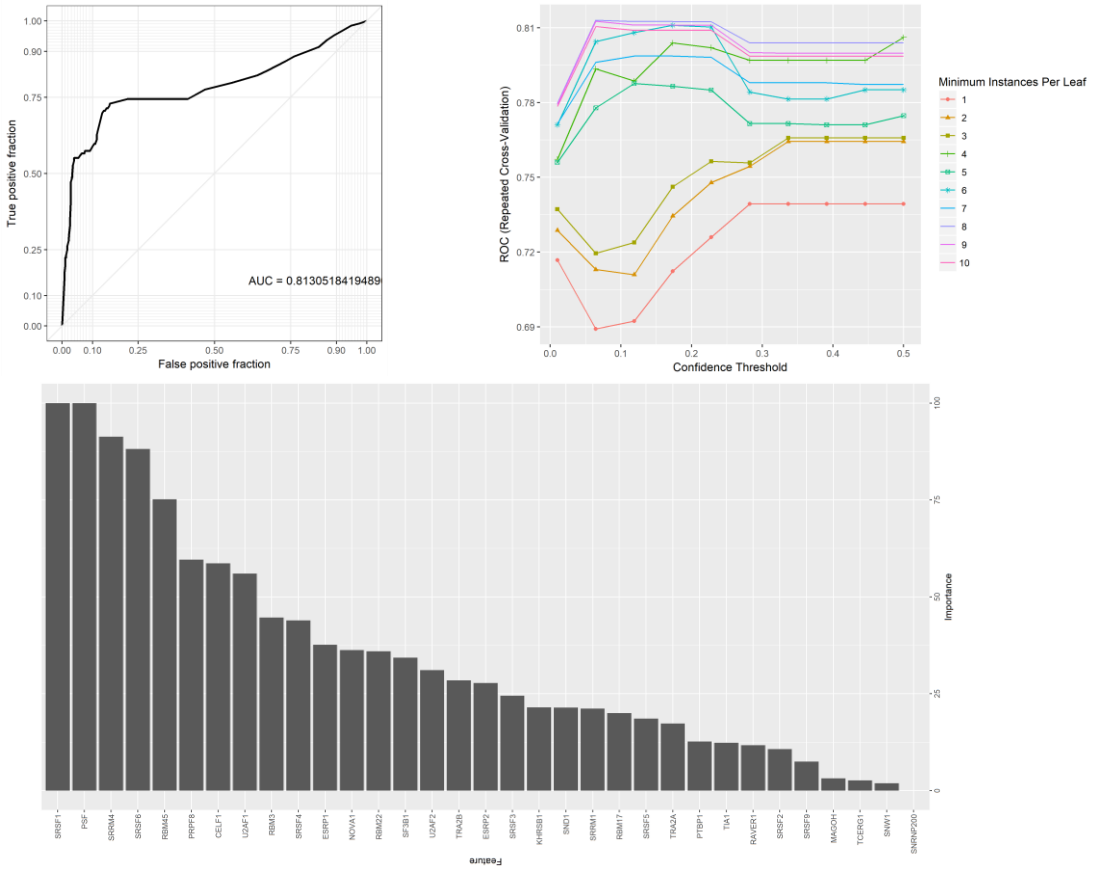




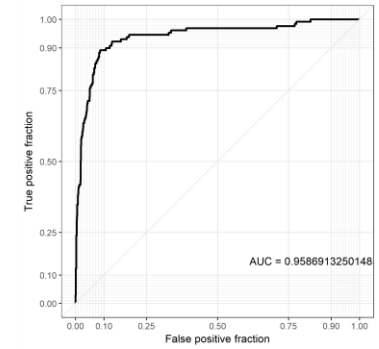
Generalized Linear Model with Stepwise Feature Selection

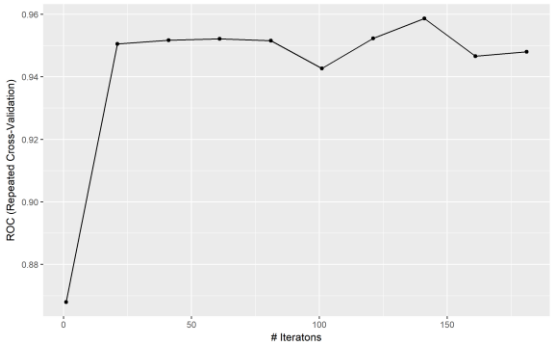


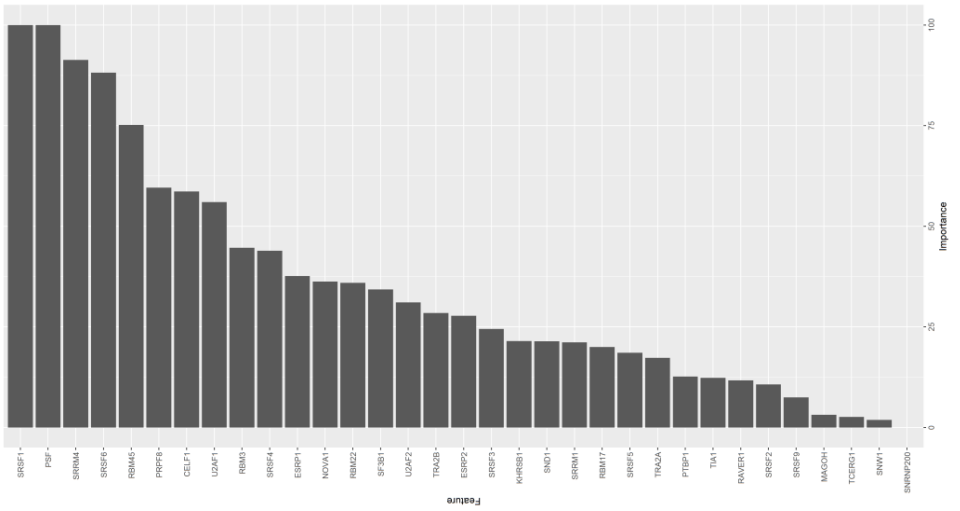
C4.5



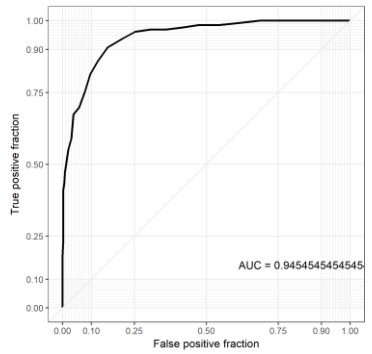
Logistic Model Trees

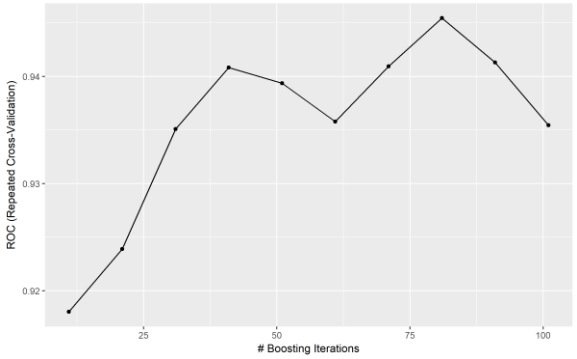


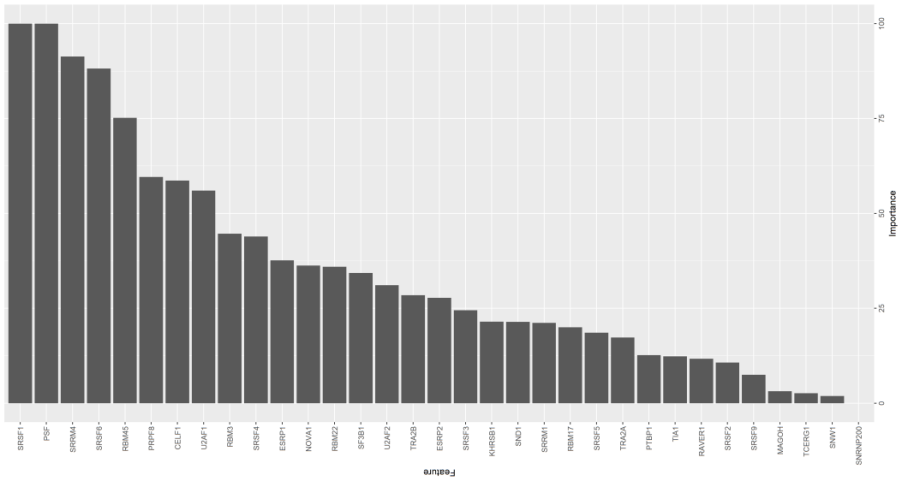




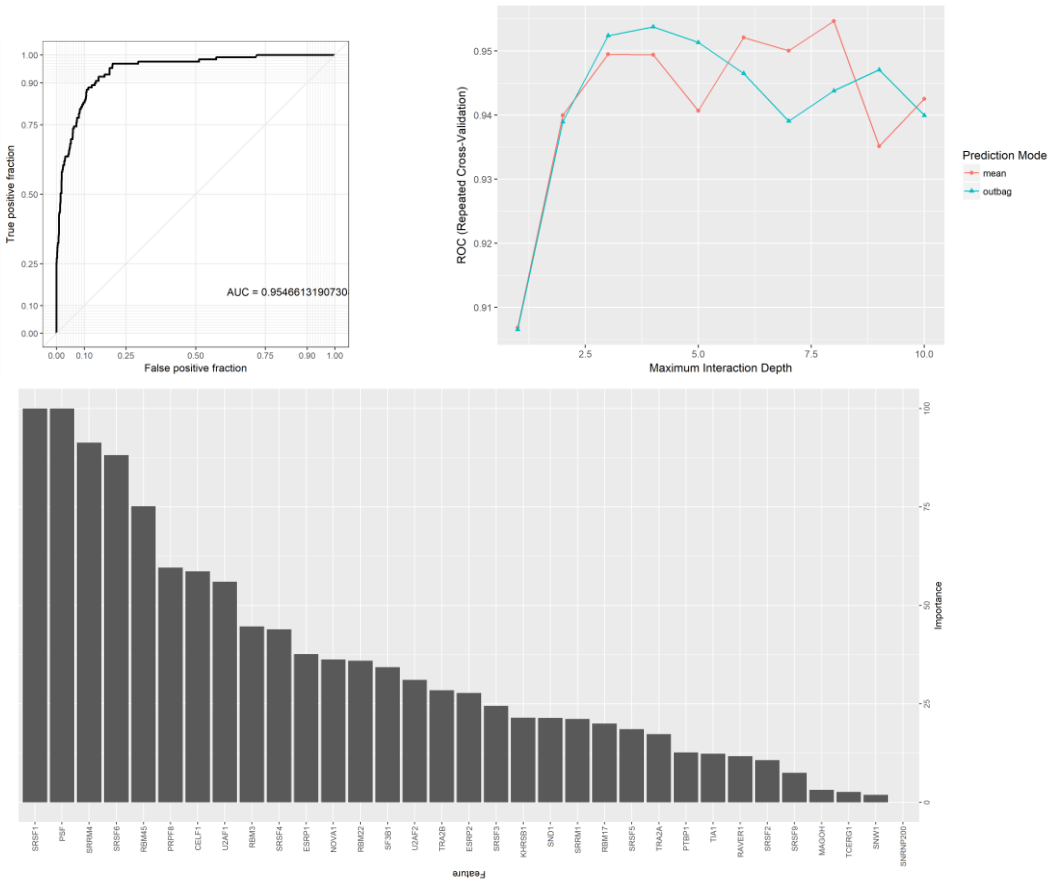
Boosted Logistic Regression



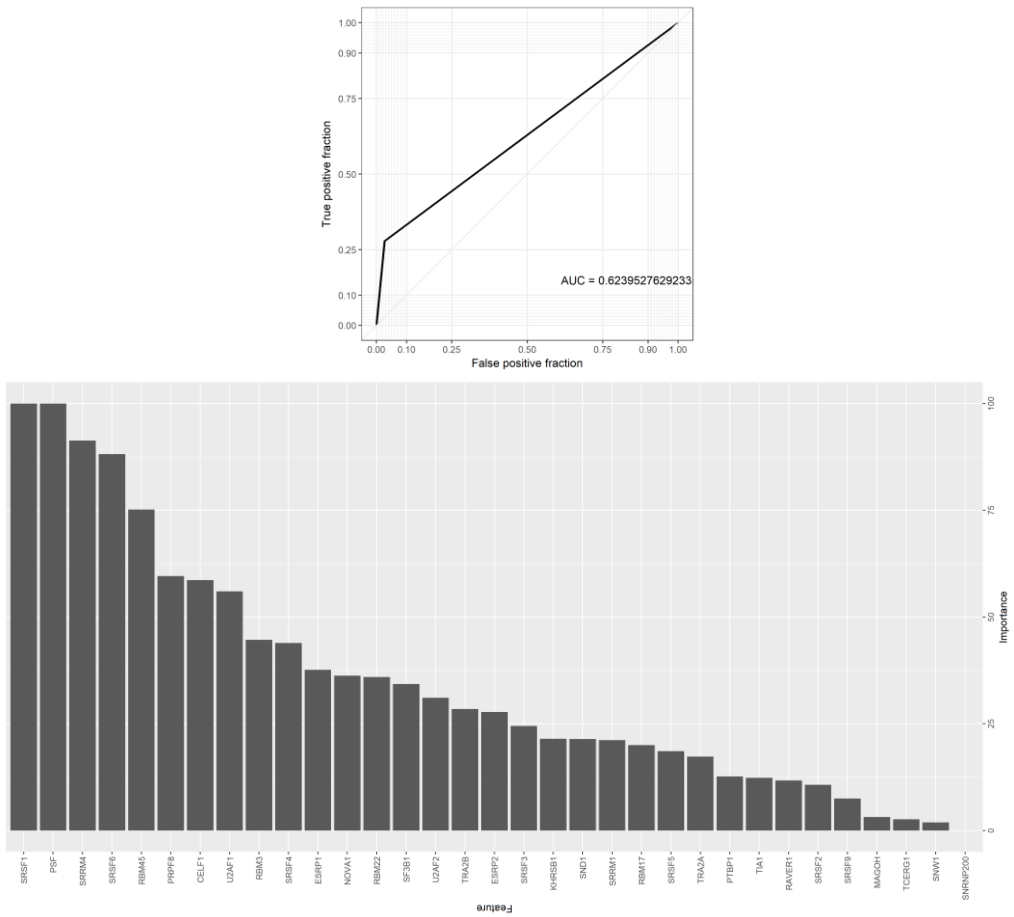




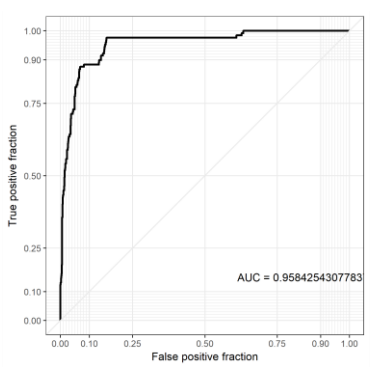
Tree-Based Ensembles

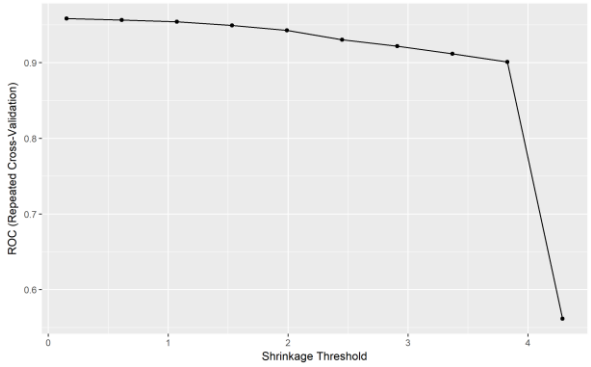


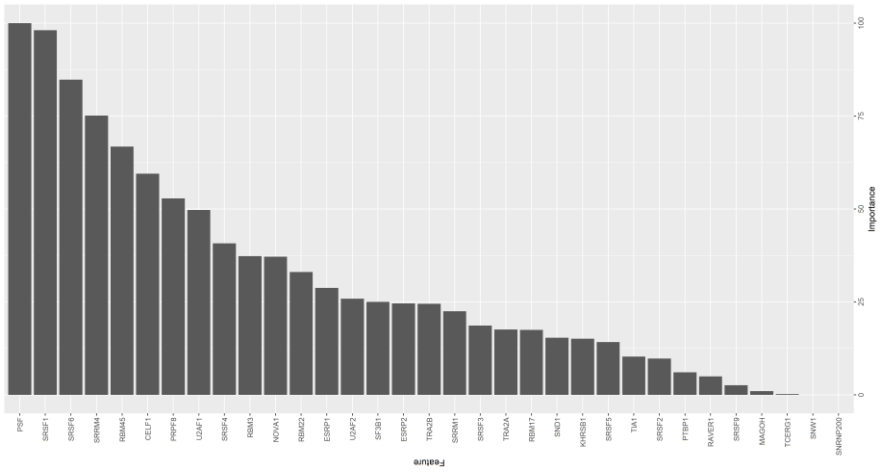
Single Rule Classification



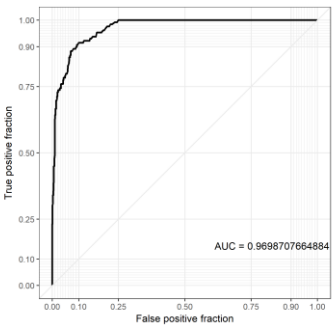
Nearest Shrunken Centroids

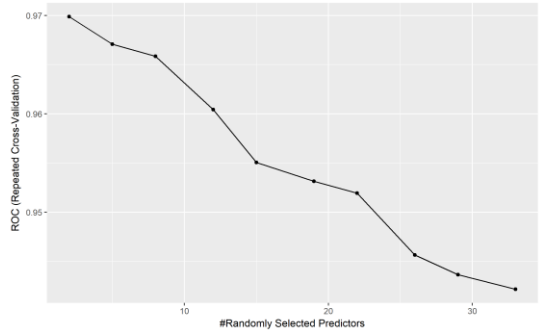


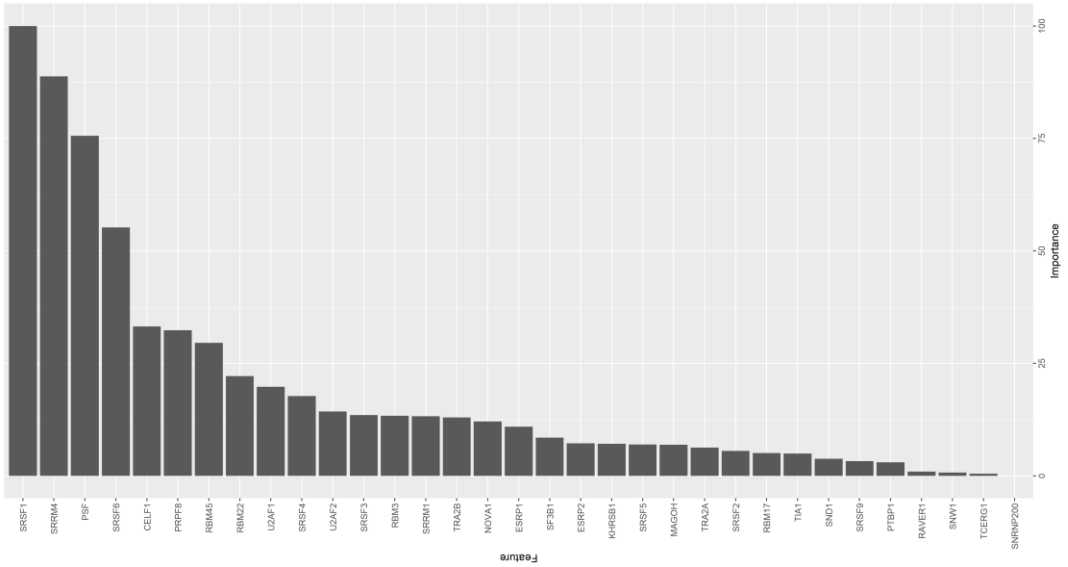




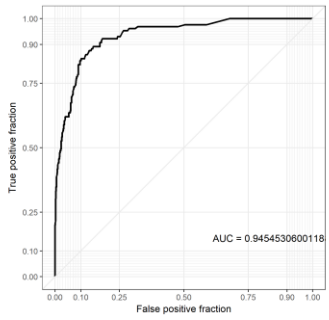
Random Forest

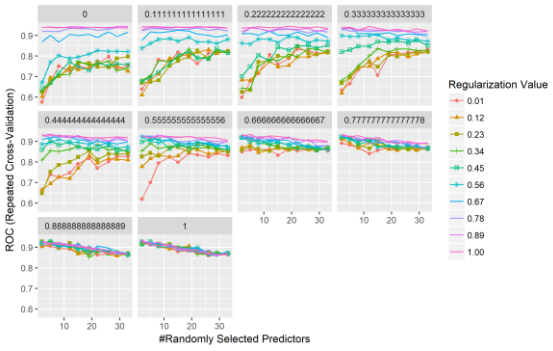


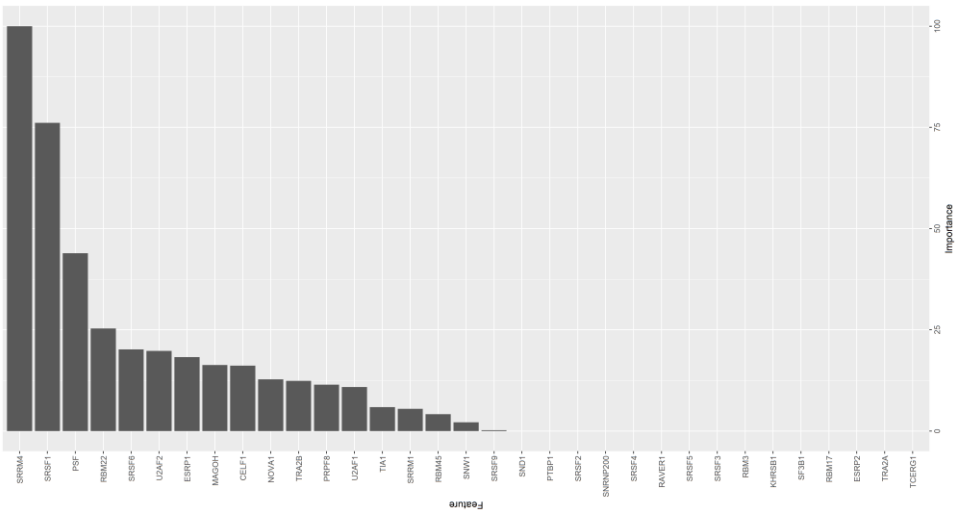




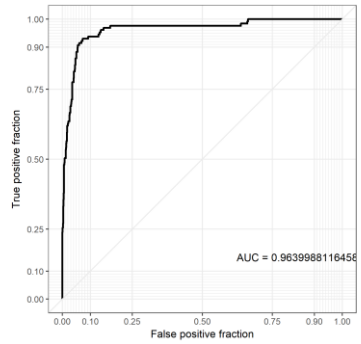
Regularized Random Forest

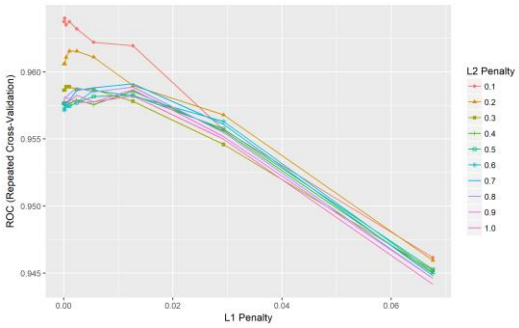


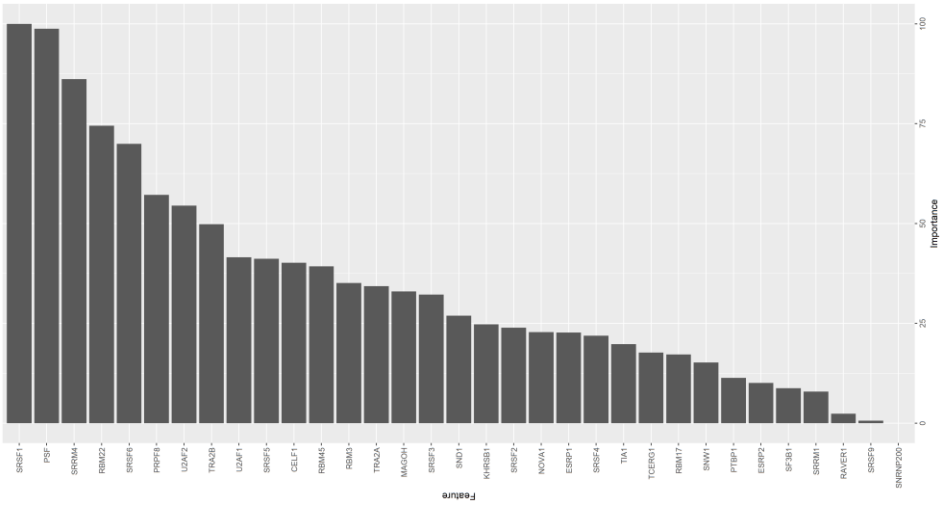




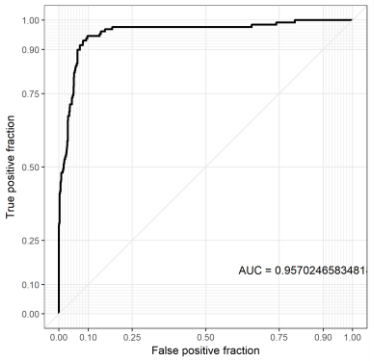
Sparse Distance Weighted Discrimination

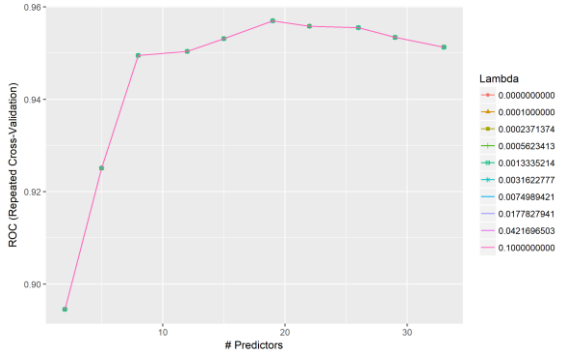


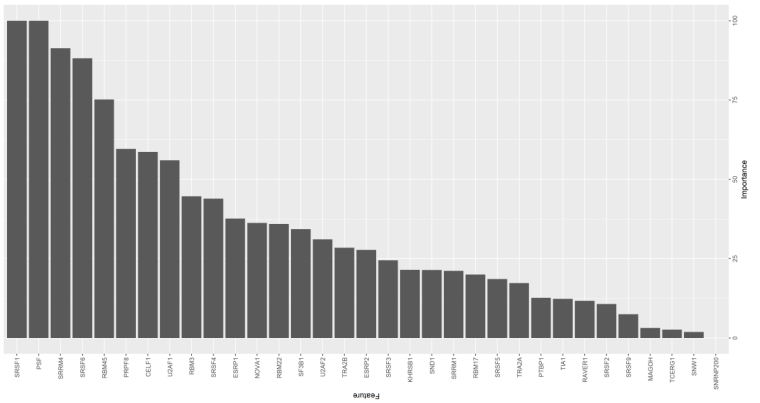




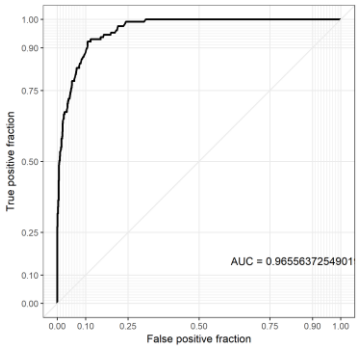
Sparse Linear Discriminant Analysis

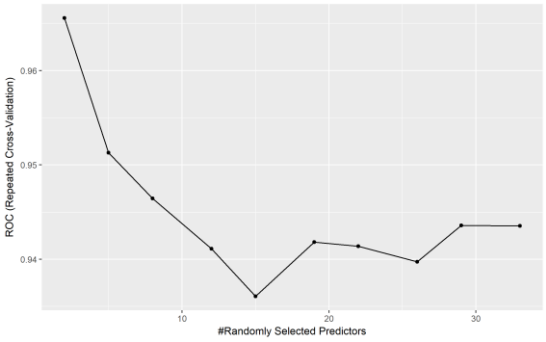


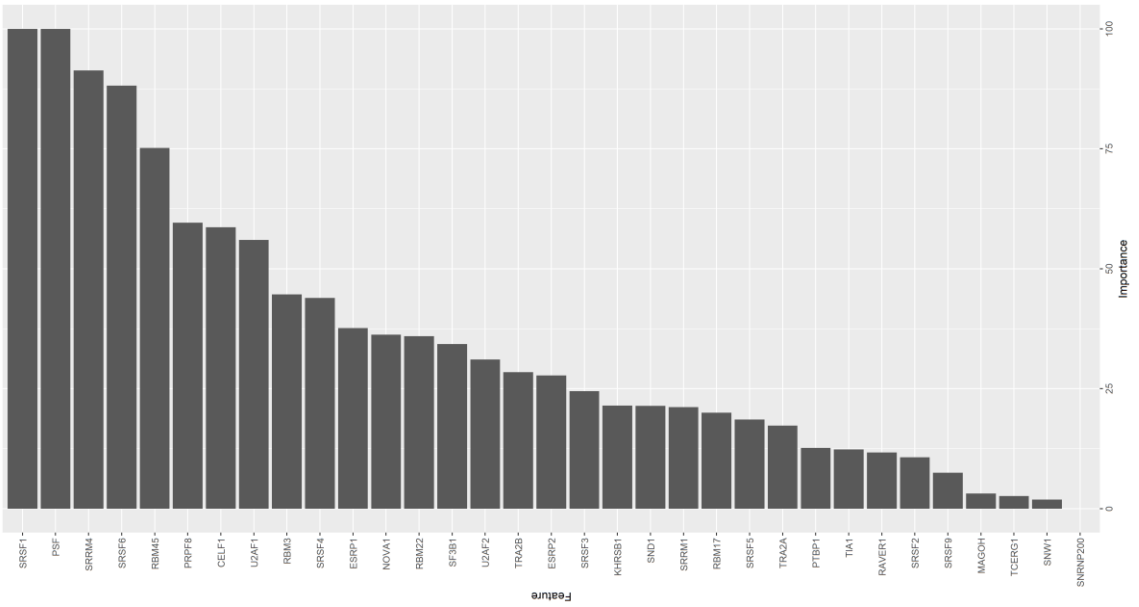




Weighted Subspace Random Forest

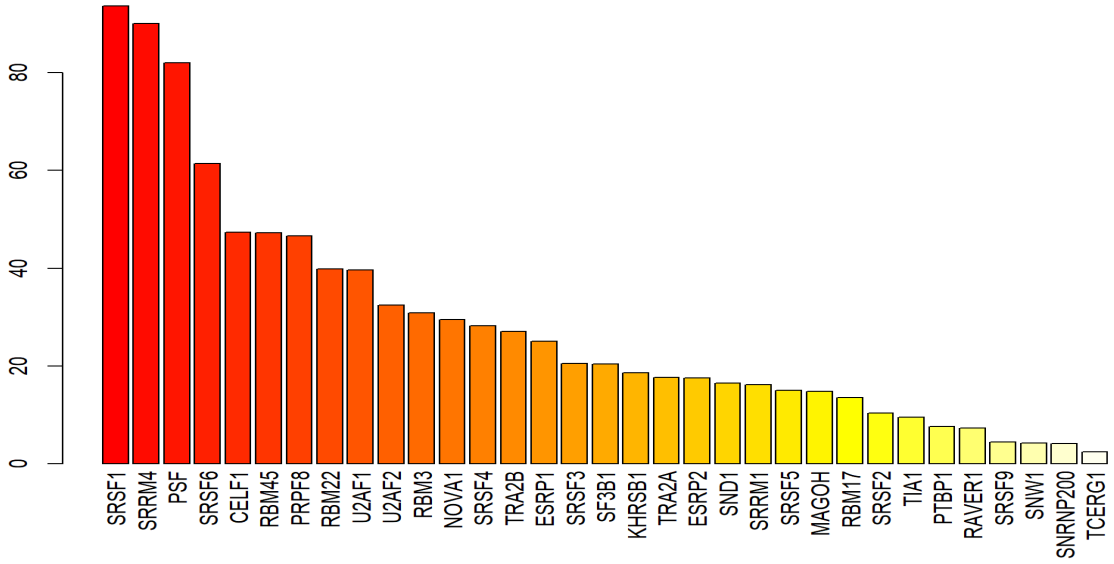






A partir de los resultados obtenidos se puede apreciar que 17 modelos obtuvieron curvas ROC con un área mayor de 0.95, 5 modelos obtuvieron un AUC entre 0.90 y 0.95, 1 modelo con un AUC entre 0.85 y 0.90, 2 modelos con AUC entre 0.80 y 0.85, y 1 modelo con un AUC igual a 0.62. Los resultados mostraron que varios algoritmos de clasificación de diferentes naturalezas pueden clasificar los dos grupos de muestras con niveles de precisión bastante altos.

A partir de las estimaciones de importancia de factores realizada por cada modelo, se calculó la importancia promedio de cada factor. En la siguiente figura se muestra la importancia global.



A partir del cálculo de la importancia de los factores, podemos apreciar que los factores más importantes son SRSM4, SRSF1, PSF, SRSF6, RBM45 y PRPF8. Este grupo de factores coincide con el cálculo que se hizo mediante los algoritmos de estimación de características de tipo filter (ver punto 8).

10- Se ejecutó un algoritmo de clustering jerárquico determinando el mejor subconjunto de factores que produce un clustering con AUC mayor de 0.85. Se encontraron 41 subconjuntos de factores que producen clusterings con AUC mayores de 0.85, ninguno de ellos supera los 0.9. Los heatmaps generados pueden verse en el compactado “clustering.zip”.

El factor SRSF1 aparece en los 41 modelos, PSF aparece en 40 modelos, SRSF6 aparece solo en 9, CELF1 aparece en 19, RBM45 aparece en 36 modelos, PRPF8 aparece en 16 modelos. Curiosamente, el factor SRRM4 no apareció en ningún modelo, aun cuando este factor fue estimado como uno de los más importante por todos los modelos de clasificación y algoritmos de selección de tipo filter.

Conclusiones

A partir de los resultados obtenidos, podemos decir que los factores más relevantes para este estudio son SRSF1 y PSF, aunque también pudiera considerarse a RBM45. Estos factores fueron seleccionados como relevantes por los algoritmos de selección de características, los algoritmos de clasificación, y además aparecieron en la mayoría de los subconjuntos de factores que logran un clustering de la población con un AUC entre 0.85 y 0.90.