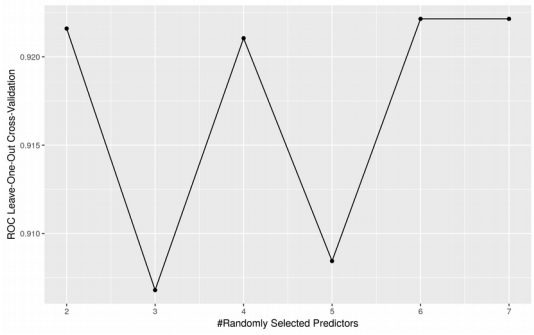
El objetivo de este estudio fue **determinar si la huella molecular (combinación de RBM3, SF3B1, SRRM1, SRSF3, SRRM4, SFPQ, U2AF2) observada anteriormente es capaz de diferenciar también entre pacientes con cáncer de próstata agresivo y pacientes con cáncer de próstata primario y/o con controles en el dataset *Grasso*.**

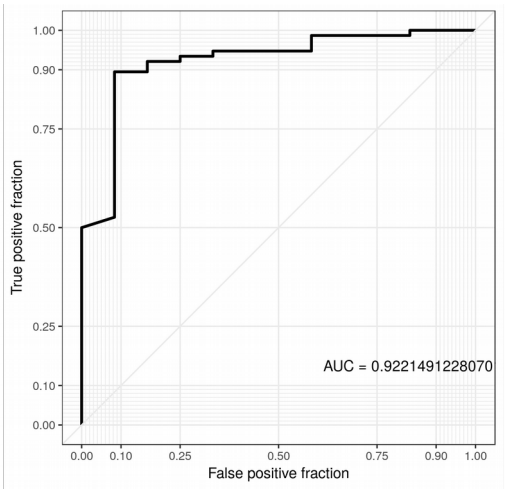
En todos los experimentos se estimó el mejor valor para el párametro *mtry* utilizado por RandomForest. Además, la eficacia del modelo construido se estimó mediante un Leave-One-Out cross-validation. Se utilizó la versión de Random Forest (WSRF) que permite tener en cuenta el desbalance entre clases.

- Cáncer de próstata (tanto primario como metastasis) y controles.

El agoritmo RandomForest logró un AUC de 0.922. La siguiente gráfica representa el proceso de tunning del parámetro *mtry.*



La siguiente figura muestra la curva ROC.



Aunque el AUC reportado por RandomForest es elevado, hay que tener en cuenta que el problema al que nos enfrentamos está desbalanceado, el ratio de desbalance entre clases (grupos de muestras) es de 1:6. A continuación, se muestra la matriz de confusión, donde podemos apreciar que existe un 50% de error de predicción en las muestras pertenecientes a la clase de tejidos benignos.

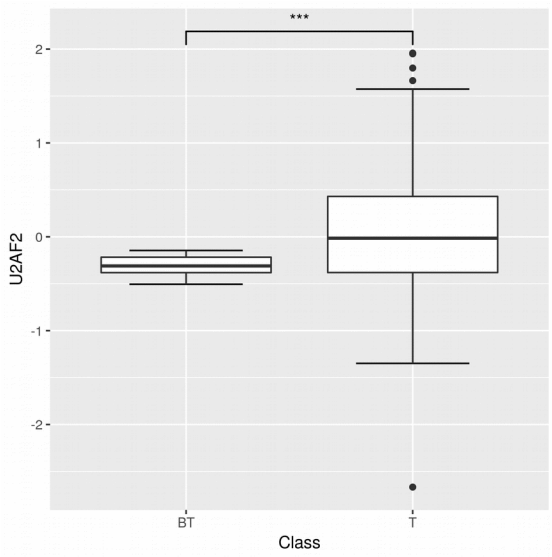
	BT	*T*	*class.error*
BT	6	6	0.50
T	4	72	0.05

Esta situación de desbalance entre clases se verá atenuada cuando hagamos los otros análisis, donde las muestras de tumores se dividirán en dos grupos.

Por otra parte, RandomForest le asignó las siguientes importancias individuales a los factores.

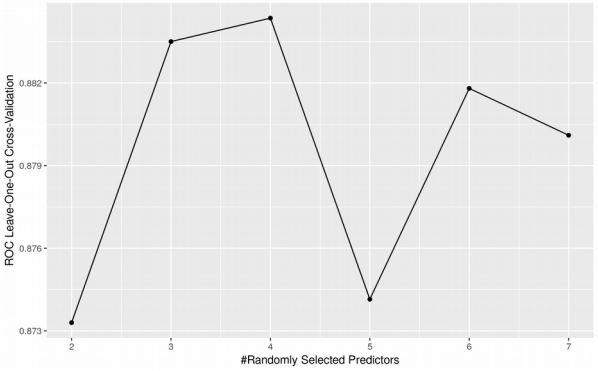
U2AF2, 100
SF3B1, 77.297
SRRM1, 75.675
SRSF3, 41.621
SRRM4, 18.378
SFPQ, 2.702
RBM3, 0

Según este ranking, el factor más importante para discriminar entre clases de muestras fue U2AF2. Precisamente este factor presenta una diferencia significativa entre los dos grupos de muestras. Los tejidos tumorales presentan en promedio una expreción significativamente mayor de este factor que los tejidos benignos.

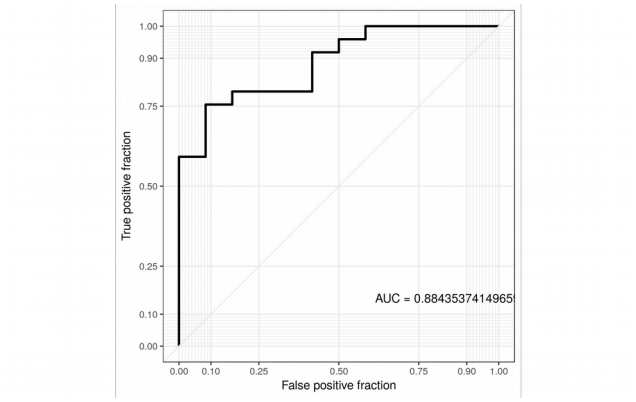


- Cáncer de próstata primario y controles

Para este caso, el agoritmo RandomForest logró un AUC de 0.883. La siguiente gráfica representa el proceso de tunning del parámetro *mtry.*



La siguiente figura muestra la curva ROC.



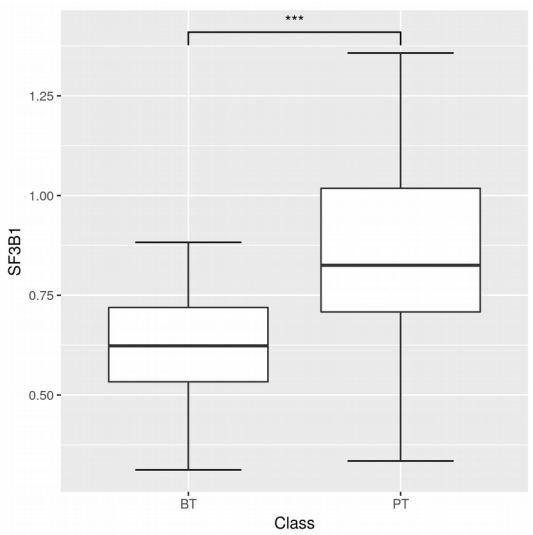
Esta es la matriz de confusión obtenida por RandomForest, aun se puede ver que existen muestras de la clase tejidos benignos que se clasifican mal.

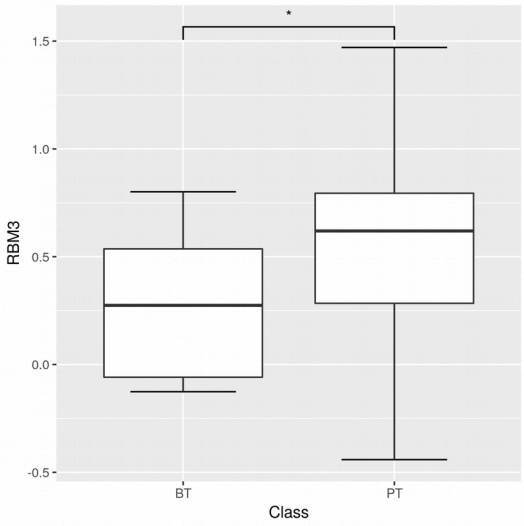
	BT	PT	class.error
BT	6	6	0.50
PT	1	48	0.02

RandomForest le asignó las siguientes importancias individuales a los factores considerados.

SF3B1, 100
RBM3, 70.857
SFPQ, 42.285
SRRM1, 39.428
U2AF2, 22.857
SRSF3, 17.714
SRRM4, 0

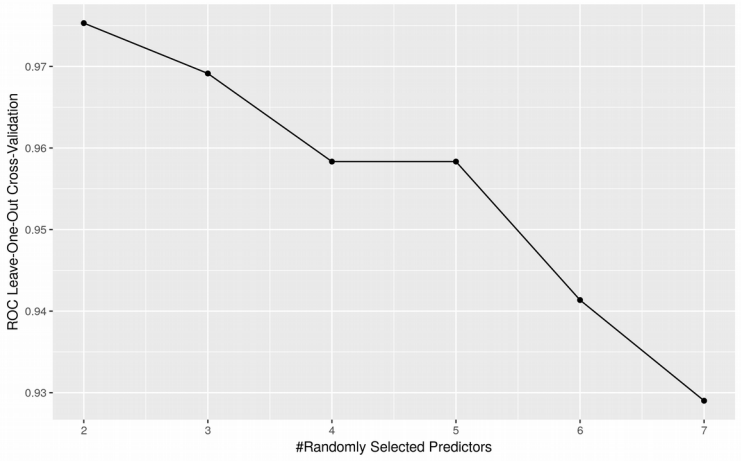
Según este ranking, los dos factores más importante para discriminar entre clases de muestras fueron SF3B1 y RBM3. Precisamente estos dos factores son los que muestran diferencias significativas entre los dos grupos. Las muestras de tumores primarios tienen en promedio mayores valores de SF3B1 y RBM3.



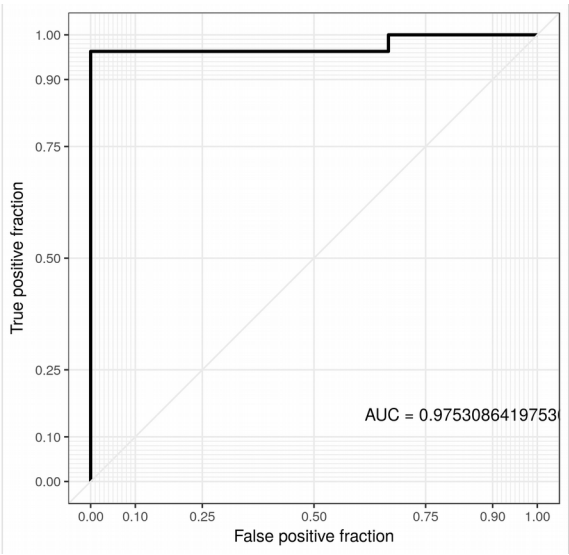


- Cáncer de próstata metastásico y controles.

Para este caso, el agoritmo RandomForest logró un AUC de 0.975. La siguiente gráfica representa el proceso de tunning del parámetro *mtry.*



La siguiente figura muestra la curva ROC.



Esta es la matriz de confusión obtenida por RandomForest, como se puede observar para este caso la clasificación es casi perfecta.

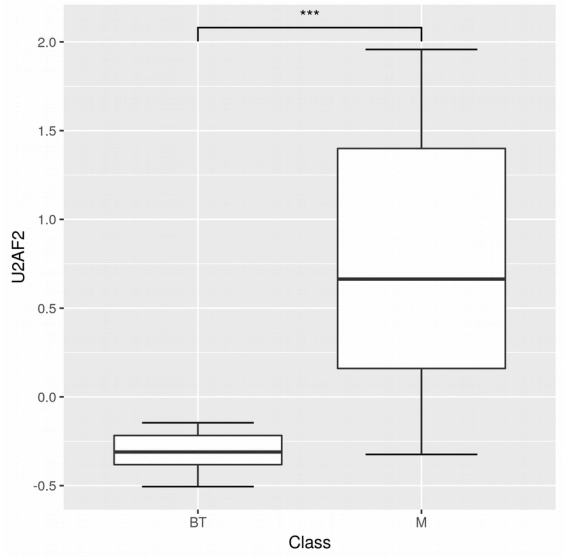
	BT	M	class.error
BT	12	0	0.00000000
M	1	26	0.03703704

RandomForest le asignó las siguientes importancias individuales a los factores considerados.

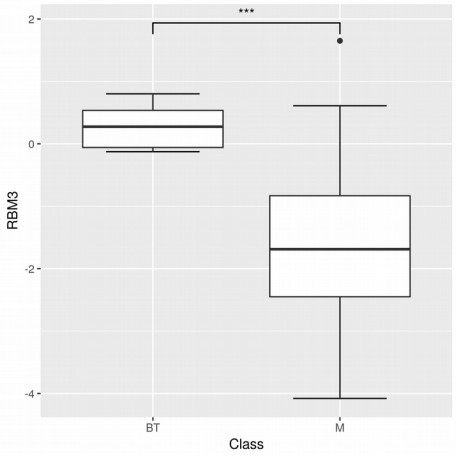
U2AF2, 100
RBM3, 55.849
SRSF3, 26.159
SFPQ, 13.629

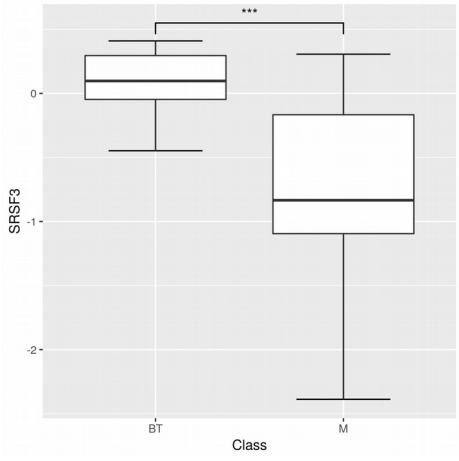
SRRM4, 4.216
SF3B1, 4.969
SRRM1, 0

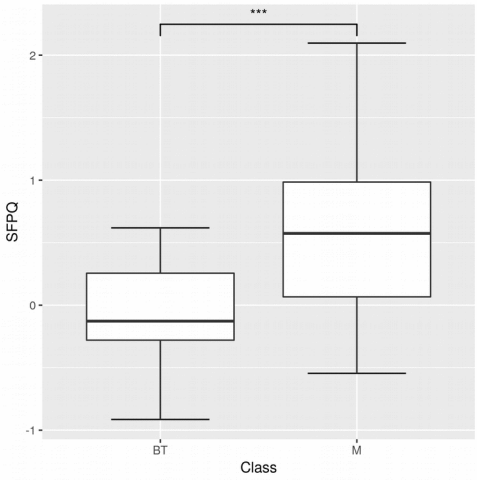
Según este ranking, el factor más importante para discriminar las muestras en los dos grupos es U2AF2, y precisamente este factor tiene valores significativamente mayores en las muestras de cáncer metastásico que en las muestras de tejidos benignos.



RandomForest le dio también importancia (aunque mucho menor) a los factores RBM3, SRSF3 y SFPQ, que también se expresan significativamente diferentes entre clases de muestras.



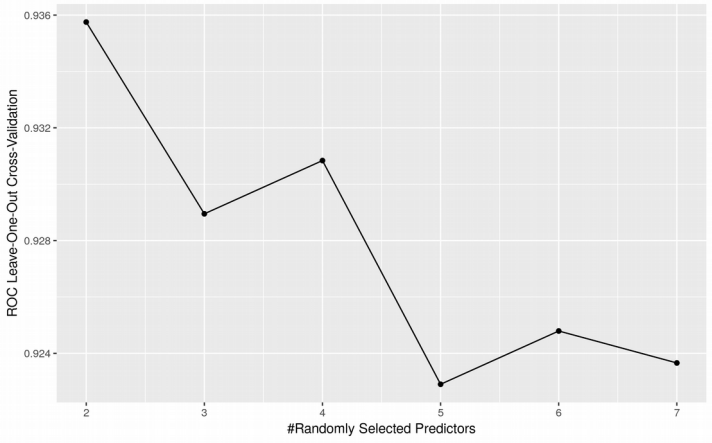




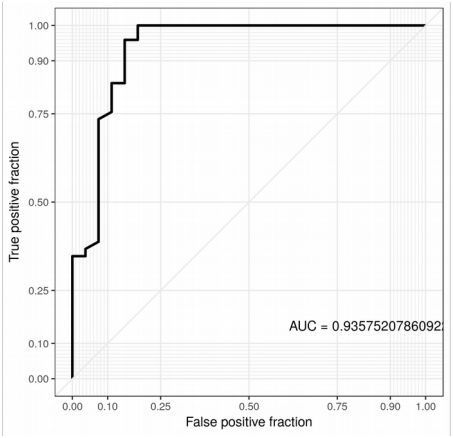
Las muestras de cáncer de prostata metastásico se caracterizaron por tener valores elevados de U2AF2 y SFPQ, y valores pequeños de RBM3 y SRSF3.

- Cáncer de próstata primario y cáncer de próstata metastásico

Para este caso, el agoritmo RandomForest logró un AUC de 0.935. La siguiente gráfica representa el proceso de tunning del parámetro *mtry.*



La siguiente figura muestra la curva ROC.



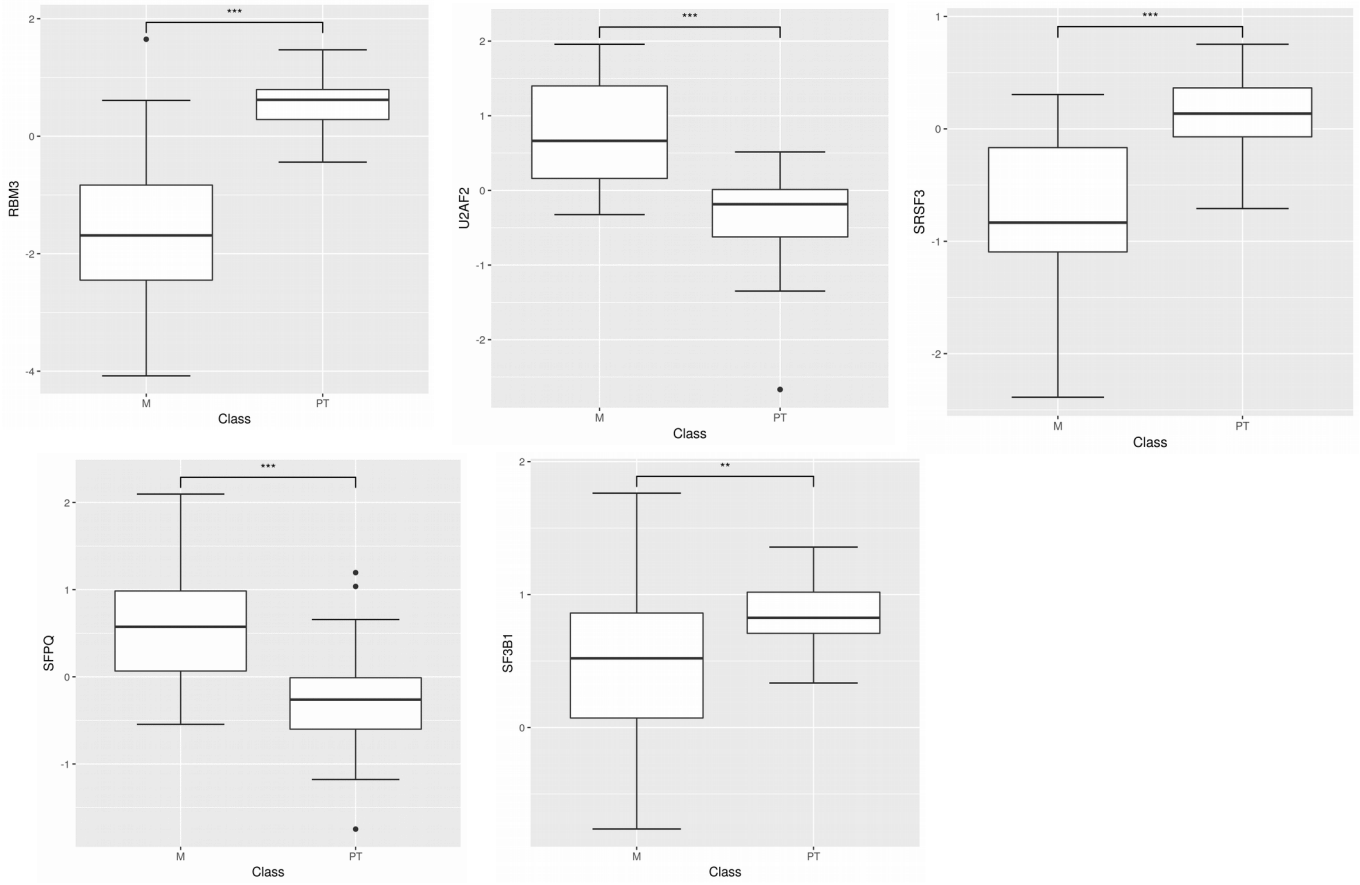
Esta es la matriz de confusión obtenida por RandomForest. Para este caso, el desbalance entre clases es pequeño, y el error obtenido en las dos clases no es elevado.

	M	PT	class.error
M	22	5	0.19
PT	1	48	0.02

RandomForest le asignó las siguientes importancias individuales a los factores considerados.

RBM3, 100
U2AF2, 92
SRSF3, 90.666
SFPQ, 78
SF3B1, 40.222
SRRM1, 20.444
SRRM4, 0

Según este ranking, los cuatro factores más importantes para diferenciar a los tumores primarios de los metastásicos fueron RBM3, U2AF2, SRSF3 y SFPQ. Al factor SF3B1 se le dio importancia, pero mucho menor que los otros cuatro factores.



Conclusiones

- El conjunto de factores utilizado fue efectivo para discriminar las clases de muestras en el dataset internacional "*Grasso*".

- Los factores SF3B1 y RBM3 están más elevados en los tumores primarios.

- Las muestras de cáncer de próstata metastásico se caracterizaron por tener valores elevados de U2AF2 y SFPQ (este menos que U2AF2), y valores pequeños de RBM3 y SRSF3. Siendo U2AF2 el factor más importante para diferenciar esta clase.