Informe N°3 - Modelos de Clasificación de Tumores

Aprendizaje Estadístico - FIUBA

José F. González - 100063 - jfgonzalez@fi.uba.ar

Atento a: Inga. Jemina García Revisión 8.0

1. Introducción

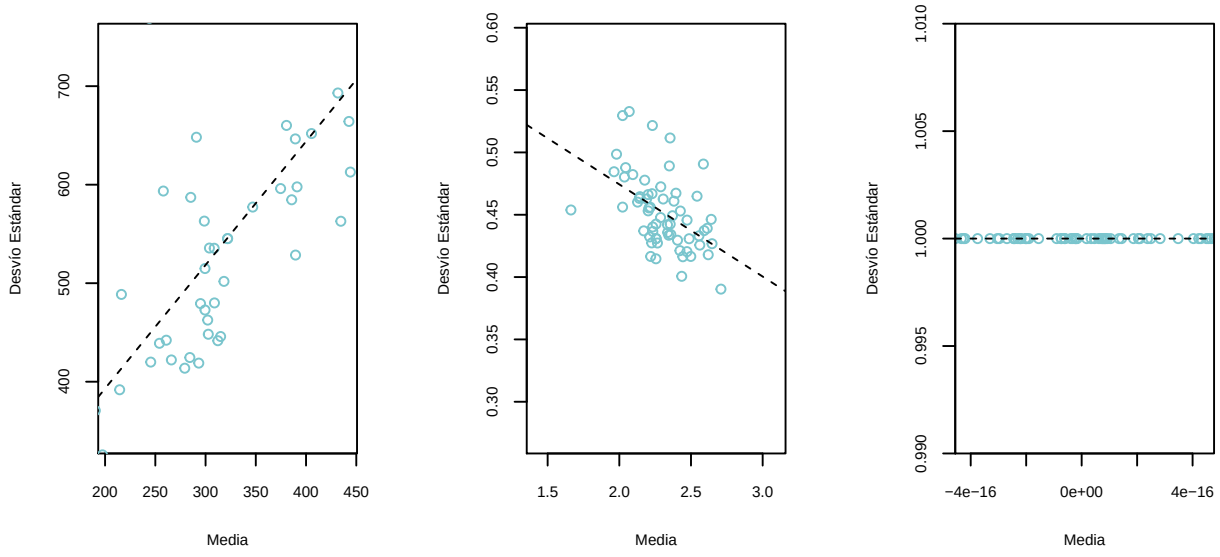
Nos interesa identificar los principales genes asociados presencia de tumores en muestras de tejido celular. Para ello disponemos el conjunto de datos `levels` con mediciones de los niveles de actividad de 2000 genes presentes en células de 62 tejidos distintos, donde 40 tejidos están etiquetados como tumores y 22 como normales. **Como criterio general buscamos las variables explicativas que maximizen la realación entre el *accuracy* de algún modelo de clasificación y la cantidad de genes predictores en una validación cruzada.**

2. Estandarización

Los niveles de actividad tienen un rango muy amplio lo que se traduce en una alta variabilidad, esto es evidente el panel izquierdo de la Figura 1 donde se muestran la media y el desvío estándar para los niveles de cada una de las muestras de tejido. Aplicando un logaritmo base diez a todos los datos se comprime el rango de datos y se obtiene un desvío estándar más representativo del conjunto como se ve en el panel central de la Figura 1, definimos así un nuevo conjunto `logLevels`.

Los métodos de reducción de variables que utilizaremos, para enfrentar la desproporción entre cantidad de predictoras y muestras, funcionan mejor con predictores estandarizados en media cero y desvío uno, por ejemplo para no afectar las penalizaciones de los métodos de reducción. Luego resultará conveniente definir un nuevo conjunto `stdLogLevels` en esta escala a cambio de perder un poco de interpretabilidad. En el panel derecho de la Figura 1 se ve el efecto de la nueva escala sobre la media y desvío de los genes de cada uno de los tejidos.

Figura 1: Variabilidad de la media y el desvío estandar en las 62 muestras de tejido. **Izquierda:** Desvío estandar y media para las 62 muestras de tejido del conjunto `levels`. **Centro:** Desvío estándar y media para las 62 muestras del conjunto `logLevels` luego de tomar logaritmo en base diez a los niveles de expresión de los genes. **Derecha:** Desvío estándar y media para las muestras de `stdLogLevels` luego de estandarizar los datos a media cero y desvío uno.



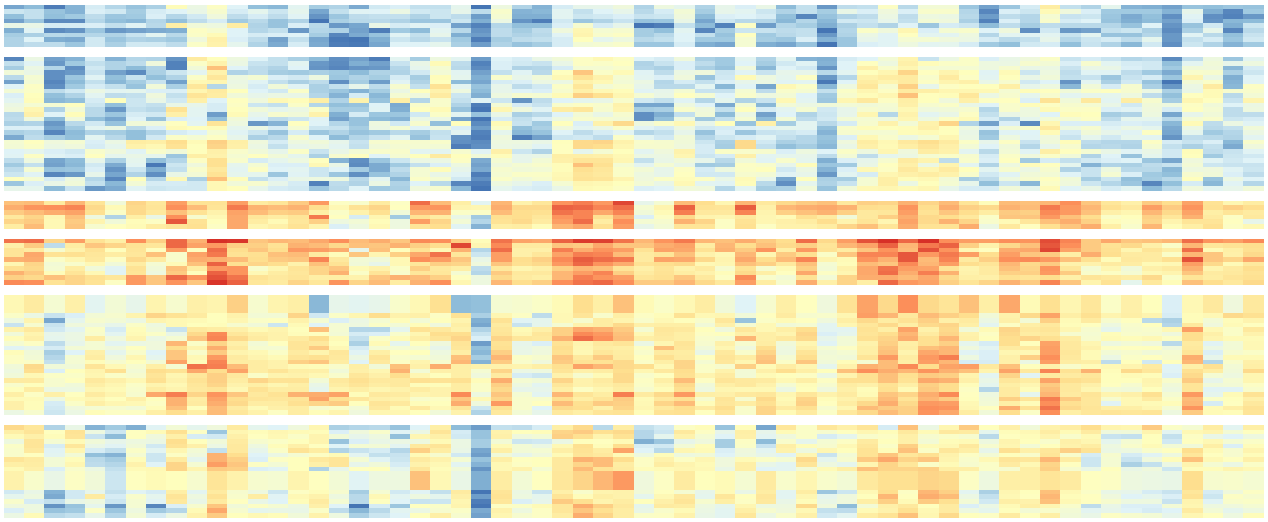


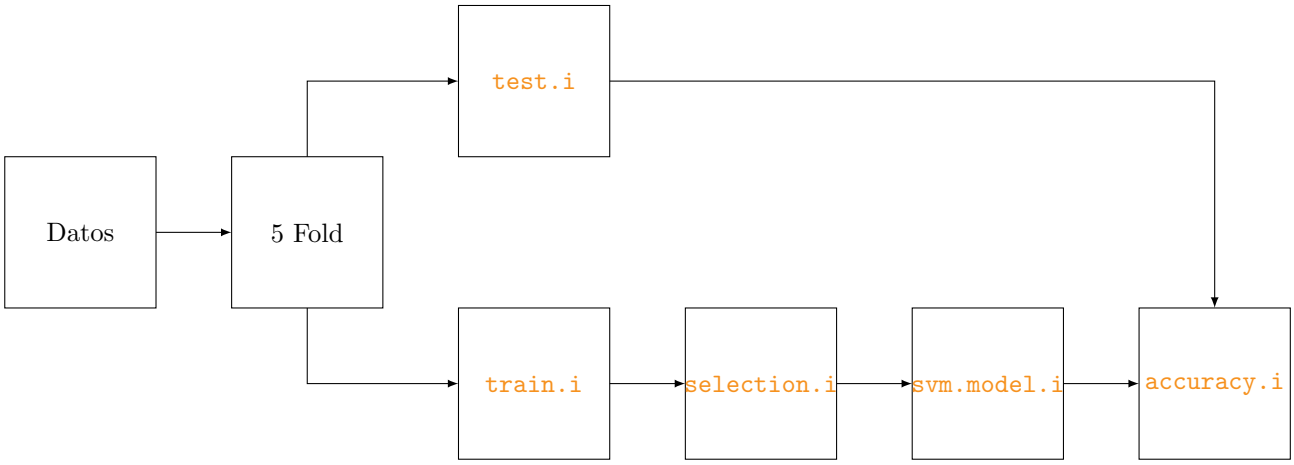
Figura 2: Mapa de niveles de actividad de los primeros 100 genes. El eje vertical corresponde a genes y el horizontal a los tejidos. Genes con mayor nivel de actividad están codificados con tonos rojos y los de menor actividad con azules. **Los genes se encuentran ordenados por un *clustering jerarquíco* según similitudes en la cadena de intesidades para facilitar la distinción de grupos correlacionados.**

3. Reducción de variables

El mapa de calor de de la Figura 2 muestra los logaritmos de niveles de actividad, en los distintos genes del eje vertical, como tonos rojos para genes con mayor expresión y tonos azules para genes con menor expresión. Los genes fueron agrupados en un cluster jerárquico según los patrones de su cadena de intesidad revelando estructuras similares en las expresiones de distintos genes. Esto se interpreta como procesos de regulación entre ciertos grupos de genes, lo que sugiere efectos de multicolinealidad fuertes. Analizando la matriz de correlación de `stdLogLevels` se cuenta un 9 % de las correlaciones entre genes son mayor a 0,7 y un 2 % son correlación mayor a 0,8. Los casos de alta correlación son generalmente interpretados como genes con regulación directa entre ellos.

Para evaluar los distintos métodos de reducción utilizaremos el algoritmo descripto en la Figura 3 tomando 5 subconjuntos de datos sobre los cuales se realiza una reducción de variables y luego se construye un modelo de SVM lineal que se evalua sobre el conjunto de entrenamiento correspondiente mediante la medida *accuracy*. **De esta forma logramos que los métodos de reducción se evaluen sobre datos distintos a los que fueron contruidos. El rendimiento global de cada reducción se evalua computando el *accuracy* promedio de los 5 ensayos.**

Figura 3: Algoritmo para la evaluación de los distintos métodos de reducción. Se realiza la selección de variables sobre 5 subconjuntos de entrenamiento, se construye un modelo SVM y se computa la medida *accuracy* sobre el conjunto de evaluación correspondiente.



3.1. Selección por *Lasso*

El primer enfoque para reducir dimensión y seleccionar variables simultaneamente es penalizar por *LASSO* utilizando el algoritmo propuesto en la Figura 3 y minimizando por CV el error de clasificación. **El rendimiento global es de un *accuracy* promedio 73.8 % para los distintos modelos SVM. En total se seleccionaron 45 genes, donde 15 de ellos fueron elegidos más de una vez en distintos *folds* y se los muestra en la Figura 4.**

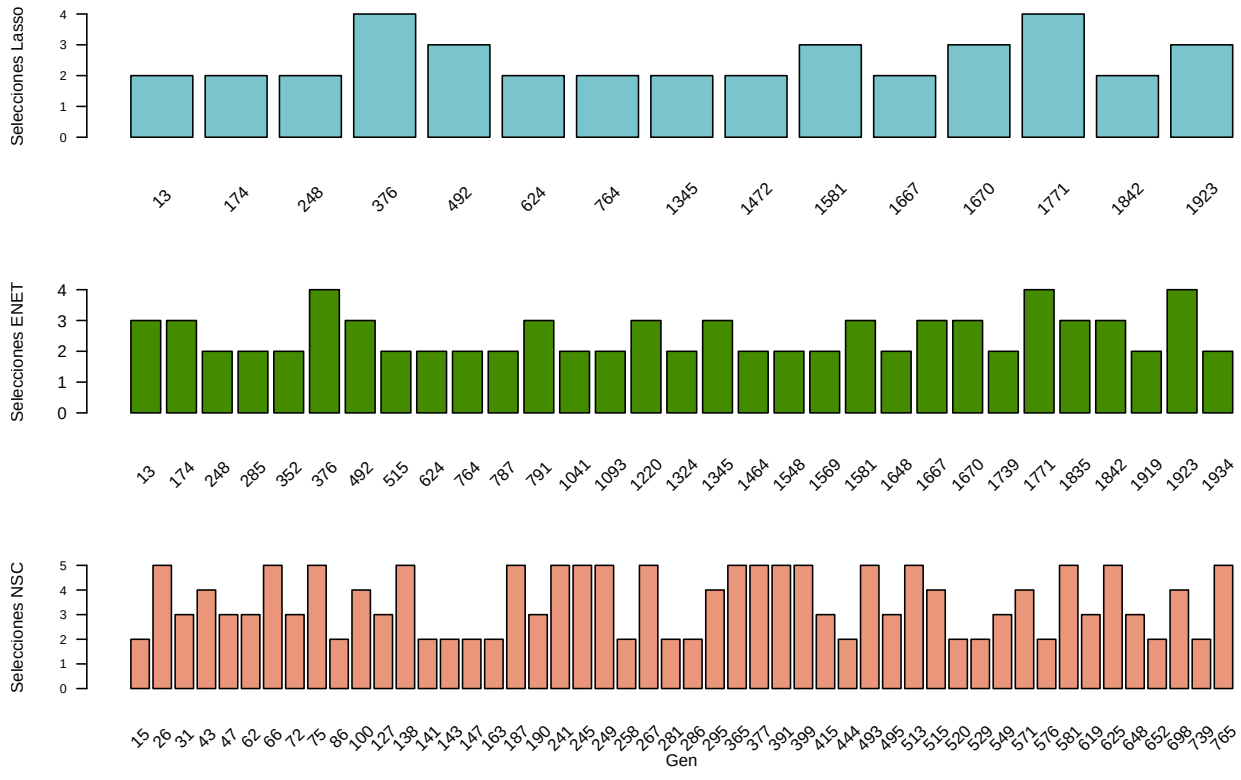
3.2. Selección por Elastic Net 0.75

Bajo efectos de colinealidad *Lasso* elige arbitrariamente entre muchos posibles conjuntos colineales de predictoras omitiendo algunos genes que puedan estar relacionados con la salida. *Elastic Net* combina las penalizaciones de *Ridge* y *Lasso* obteniendo un método que es más inclusivo con las variables correlacionadas y da mejores resultados a costa de aumentar la cantidad de genes a 59 y tener que ajustar dos parametros α y λ por CV. **En la Figura 4 se muestran 31 genes que fueron seleccionados más de una vez en los distintos *folds*, el *accuracy* promedio obtenido por el método es de 75.4 %.**

3.3. Selección por *Nearest Shrunken Centroids*

NSC es un método desarrollado en el contexto de identificar clases de genes y selección automática de genes que caracterizan la clase. Es una modificación del nearest centroids aplicando una contracción de los centroides de cada clase hacía cero, ajustado por CV reduciendo error de clasificación, eliminando los que alcanzan el cero y clasificando sobre la nueva disposición por proximidad a los centroides que sobreviven. En total selecciona 157 genes, de los cuales 50 se seleccionan más de una vez y se muestran en el panel inferior de la Figura 4. **El modelo SVM con esta selección alcanza un 85.1 % de *accuracy* promedio.**

Figura 4: Resultados del algoritmo de reducción. En el el eje vertical la cantidad de veces que un gen del eje horizontal es seleccionado en los distintos *folds*. **Superior:** Genes que son seleccionados más de una vez por *Lasso*. El modelo SVM con esta selección alcanza el 83.8 % de *accuracy* promedio. **Centro:** Genes seleccionados más de una vez por *Elastic Net 0.75*. La clasifición de SVM alcanza el 78.8 % de *accuracy* promedio. **Inferior:** Los primeros 50 genes seleccionados más de una vez por *NSC*. El modelo SVM logra un 84 % de *accuracy* promedio.



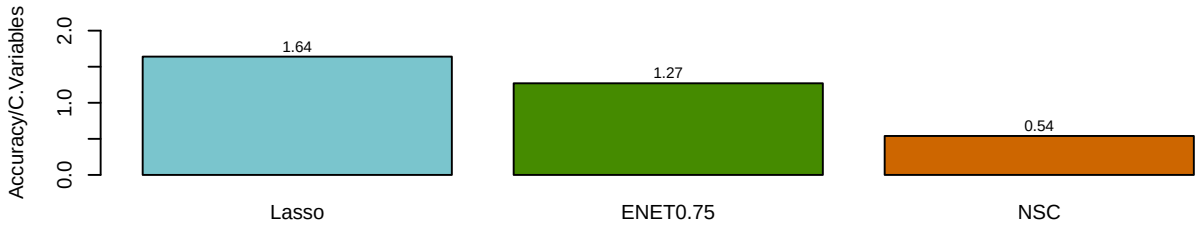


Figura 5: Comparación de la proporción entre el *accuracy* promedio del modelo SVM y la cantidad de variables explicativas de las selecciones de variables de *Lasso*, *Elastic Net* y *NSC*

Evaluamos la relación entre cantidad de variables y desempeño del clasificador como el cociente entre el *accuracy* promedio del modelo SVM y la cantidad total de genes utilizados en los distintos folds. La mejor relación se obtiene con los genes seleccionados por *Lasso* como se muestra en la Figura 5.

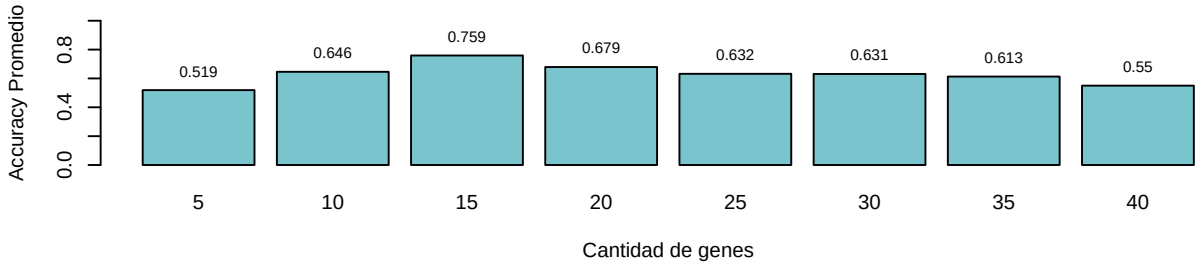
4. Clasificación

Para evaluar los distintos modelos de clasificación proponemos el algoritmo de la Figura 7. Utilizaremos el conjunto de genes seleccionado por *Lasso* que nos da la mejor relación entre *accuracy* y cantidad de variables para construir un método de calsificación. **Pero entrenaremos tan solo con los genes que fueron seleccionados más de una vez en `selection.i`, del algoritmo de la Figura 3, resultando en un conjunto final de tan solo 15 genes que se muestran en el panel superior de la Figura 4** Esto se justifica por la Figura 6 donde se realizaron ajustes de un modelo logístico con distintas cantidades de predictoras, al superar las 15 se empiezan a agregar los genes que tienen una sola aparición en el algoritmo de la Figura 3 y el rendimiento decrece.

4.1. Clasificación por SVM

Volvemos a clasificar por SVM lineal pero ahora en lugar de realizar la selección de *Lasso* para cada fold usamos siempre la misma selección de 15 genes que ya fueron identificados. En promedio el método alcanza un *accuracy* promedio de 85.4 % para los 5 folds del algoritmo de la Figura 7.

Figura 6: Efecto de agregar más genes del conjunto elegido por *Lasso*. La predicción mejora hasta los 15 primeros que coinciden con los que fueron elegidos más de una vez y se muestran en la Figura 4. Los que fueron elegidos sólo una vez representan una particularidad del *fold* al que pertenecian. Los resultados son obteniddos utilizando un modelo logístico en el algoritmo de la Figura 7.



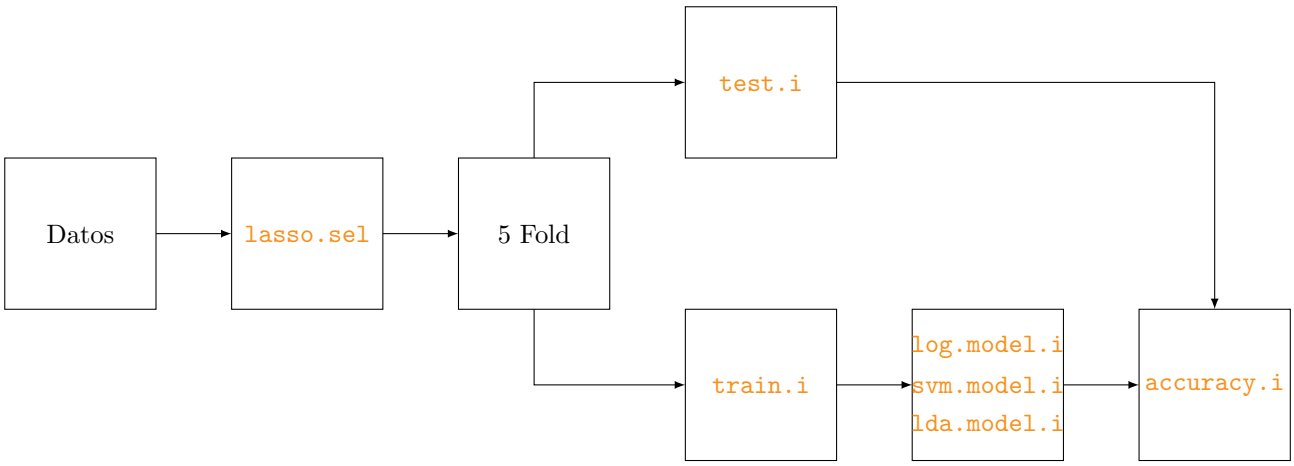


Figura 7: Algoritmo para la evaluación de los distintos métodos de clasificación. Se comienza utilizando las variables seleccionadas por *Lasso* y se realizan 5 subconjuntos de muestras. Luego se construyen distintos modelos de clasificación y se computa el *accuracy* promedio.

4.2. Clasificación por Regresión Logística

Repetimos el proceso clasificando cosntruyendo modelos logísticos para los 5 subconjuntos de entrenamiento y evaluando sobre los 5 de evaluación. En promedio este método alcanza un *accuracy* promedio de 75.8 %. En nivel de corte adoptado para clasificar como tumor es cuando la probabilidad sea mayor al 80 %, se mantiene este valor constante.

4.3. Clasificación por LDA

En promedio el método alcanza un *accuracy* promedio de 82.2 %. Nuevamente se utilizó un nivel de probabilidad 80 % para clasificar como tumor.

4.4. Clasificación por QDA

En promedio el método alcanza un *accuracy* promedio de 67.5 % con un nivel de probabilidad 80 % para clasificar como tumor. Se lo utiliza como un representante de un límite de decisión cuadrático. Su desempeño inferior al los modelos lineales puede indicar un exceso de flexibilidad o que se violen algunas suposiciones del modelo como la normalidad de las muestras. Sería interesante realizar un test de normalidad.

4.5. Clasificación por *Bagging* de Árboles de Decisión

Bagging es un método que aplimos a árboles de decisión para reducir su varianza construyendo B árboles de decisión y promediando sus predicciones. Aplicandolo en el algoritmo de la Figura 7 y **el método alcanza un *accuracy* promedio de 80.8 %.** Suele evitar el sobreajuste, lo que podría ser útil si trabajasemos con muchos más genes en lugar de unos pocos. *Bagging* es un caso particular de *Random Forest* que utiliza un subconjunto aleatorio de los genes en cada corte de árbol, este último es un método común en problemas de clasificación de tumores, especialmente cuando hay muchas categorías de cancer y alta dimensionalidad. Decidimos utilizar todos los genes en cada corte ya que las selección de variables redujo los efectos de collinealidad y no habría mucha ventaja en muestrear genes.

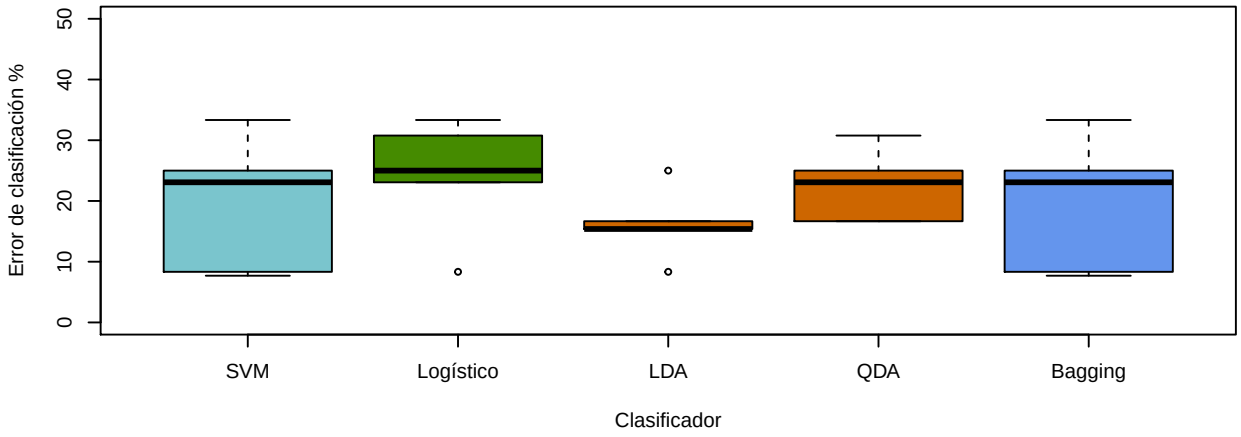


Figura 8: Comparación de los errores de clasificación de los distintos modelos.

5. Resultados

En la Figura 8 se muestran los resultados de los distintos clasificadores. Todos los métodos tienen un buen desempeño y todos validan la selección de variables. Logístico y LDA tienen la ventaja de ser muy interpretables y proporcionan una probabilidad de si un tejido presenta o no tumor. Si hubiese efectos de colinealidad remantes entre los genes seleccionados el método *Bagging* debería tener mejor rendimiento respecto de los demás, lo que no sucede.

En conclusión los genes seleccionados en el panel superior de la Figura 4 por *Lasso* es un conjunto escaso, interpretable y representativo del problema, el método apropiado de clasificación dependerá del contexto, si se busca predecir con precisión sobre nuevos datos SVM tendrá un buen desempeño, pero en una investigación médica los modelos logístico y LDA son más interpretables y dan una noción de contribución positiva o negativa al resultado.

6. Bibliografía

1. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America. 1999.
2. R. Tibshirani, T. Hastie, B. Narasimhan, R. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression. Departments of Health, Research and Policy, and Statistics, Statistics and Health, Research and Policy,and Medicine and Biochemistry, Stanford University, Stanford.
3. S. Wang, Ji Zhu, Improved centroids estimation for the nearest shrunken centroid classifier. Department of Statistics, University of Michigan.
4. K. Myungsook, K. Nyunsu, Nearest Shrunken Centroid as Feature Selection of Microarray Data. Computer Science Department, California Letheran University.