# Análisis preliminar de datos genómicos de cáncer de páncreas

9/5/2020

# Tabla de contenido

## 1.- Fuente de los datos y método de análisis

Se sigue el guión de la primera práctica de la asignatura para descargar toda la información de cáncer de páncreas de GDC Portal. En resumen:

- Se selecciona páncreas en la página principal.
- Se selecciona Program = "TCGA".
- Se selecciona Sample Type = "Primary tumor" y "Solid Tissue normal".
- Save/Edit Case Set => Save as new case set => Save => Manage Sets => View Files in Repository.
- Se selecciona Experimental Strategy = "RNA-Seq".
- Se selecciona Workflow Type = "HTSeq – Counts".
- Add All Files to Cart => Cart.
- Se descargan:
    - Sample Sheet.
    - Manifest.
    - Cart.

Se descargan 182 registros de GDC Portal, con la siguiente distribución:

**Tabla 1. Distribución del número de casos de los datos descargados de GDC Portal.**

| | Primary Tumor | Solid Tissue Normal | Metastatic |
|---|---|---|---|
| Número de casos | 177 | 4 | 1 |

Aquí tengo una duda, y es cómo trabajar con los metastásicos. Se me ocurren dos opciones:

- Recodificar como "Primary Tumor".
- Eliminar su fichero .htseq.counts correspondiente, y eliminar el caso de SamplesDataFrame.

En este análisis he optado por la primera opción porque era más rápido en un análisis preliminar, aunque quizá no sea lo ideal.

El análisis se hace en R 4.0.0 y KnowSeq 1.2.0 (¡gracias Daniel por la ayuda!), en 3 scripts:

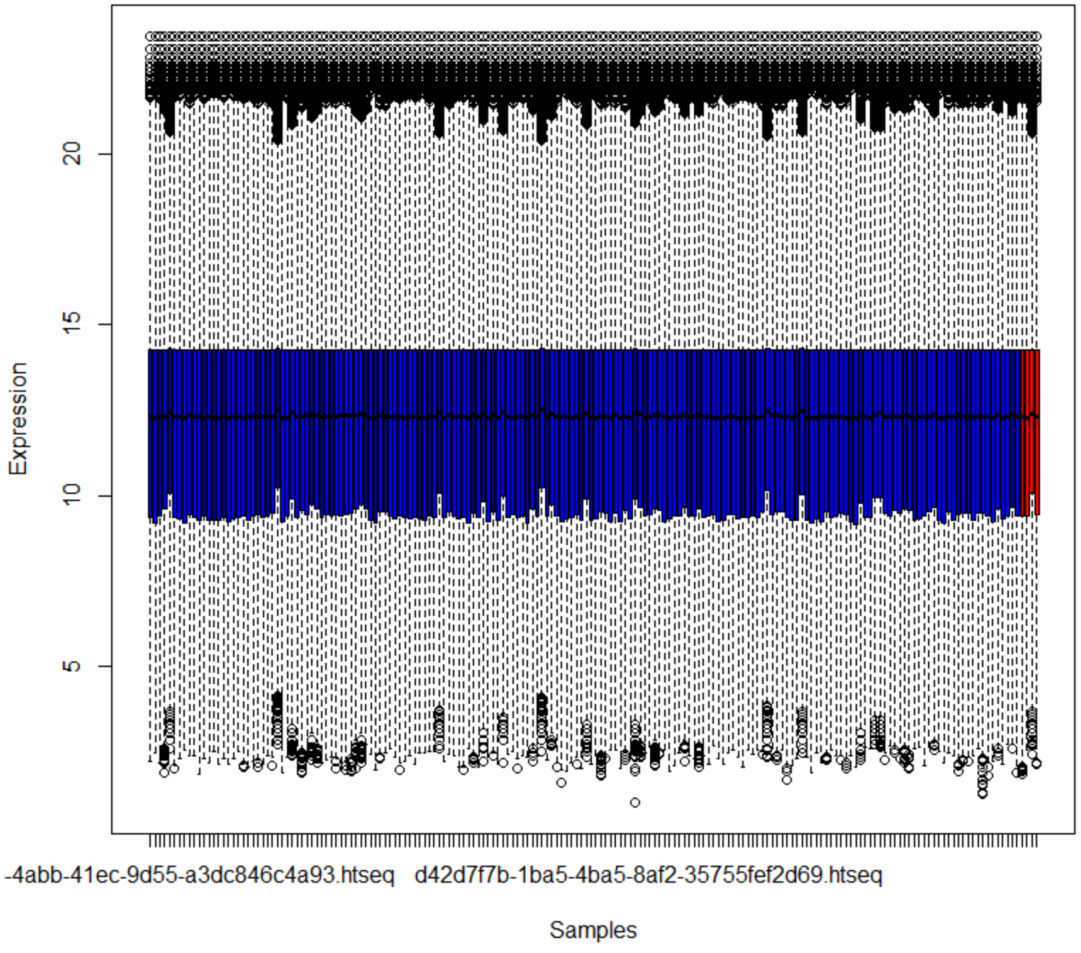
- 00_descompresion.R: Descomprime los ficheros descargados de GDC Portal para tener los ficheros .htseq.counts. Sólo hay que ejecutarlo una vez.
- 01_preprocesamiento.R: Preprocesa los datos creando SamplesDataFrame.csv.
- 02_analisis.R: Realiza el análisis de los datos.

Para reproducir el análisis se puede ejecutar directamente 02_analisis.R, ya que tiene una llamada a 01_preprocesamiento.R.

## 2.- Boxplot de los datos

Se une la información de los ficheros count a una matriz, se descargan los nombres de los genes y se calcula la matriz de expresión de genes.
Se realiza a continuación un gráfico de cajas y bigotes con `dataPlot(mode = "orderedBoxplot")`. En azul se ven los casos que constituyen un tumor primario, y en rojo los casos con tejido sólido normal.

**Figura 1. Boxplot ordenado del resumen de la expresión de genes de cada muestra (azul = Primary Tumor, rojo = Solid Tissue Normal).**



## 3.- Extracción y visualización de expresión de genes

Se controla por el efecto batch un método que utiliza *surrogate variable analysis*.

Tras fijar un p-valor de 0.1 (he tenido que subirlo porque con 0.05 encontraba sólo 3 genes), se encuentran 11 genes que se consideran relevantes con respecto a los datos que tenemos de cáncer de tiroides.

Se visualizan boxplots y mapas de calor de los 11 genes detectados como relevantes tras la extracción de DEGs.

**Figura 2. Boxplot de la expresión de genes de los 11 genes (rojo = Primary Tumor, verde = Solid Tissue Normal).**

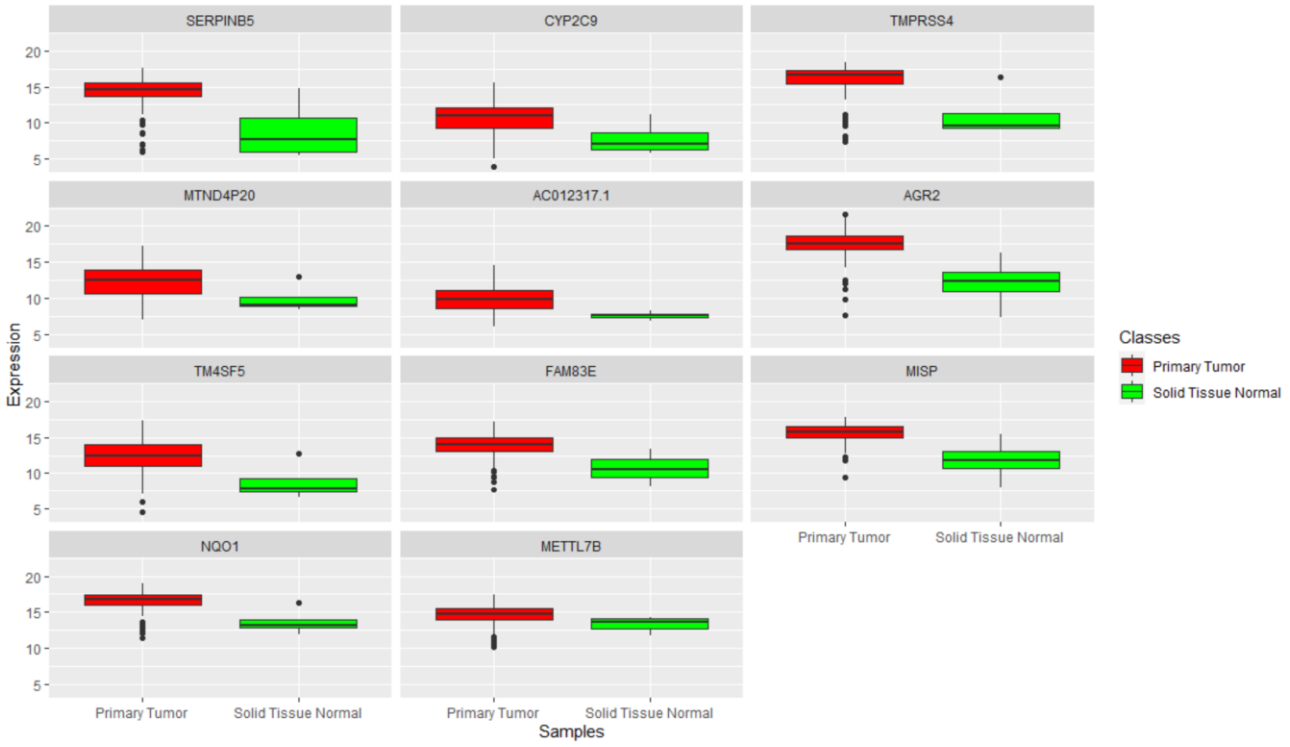
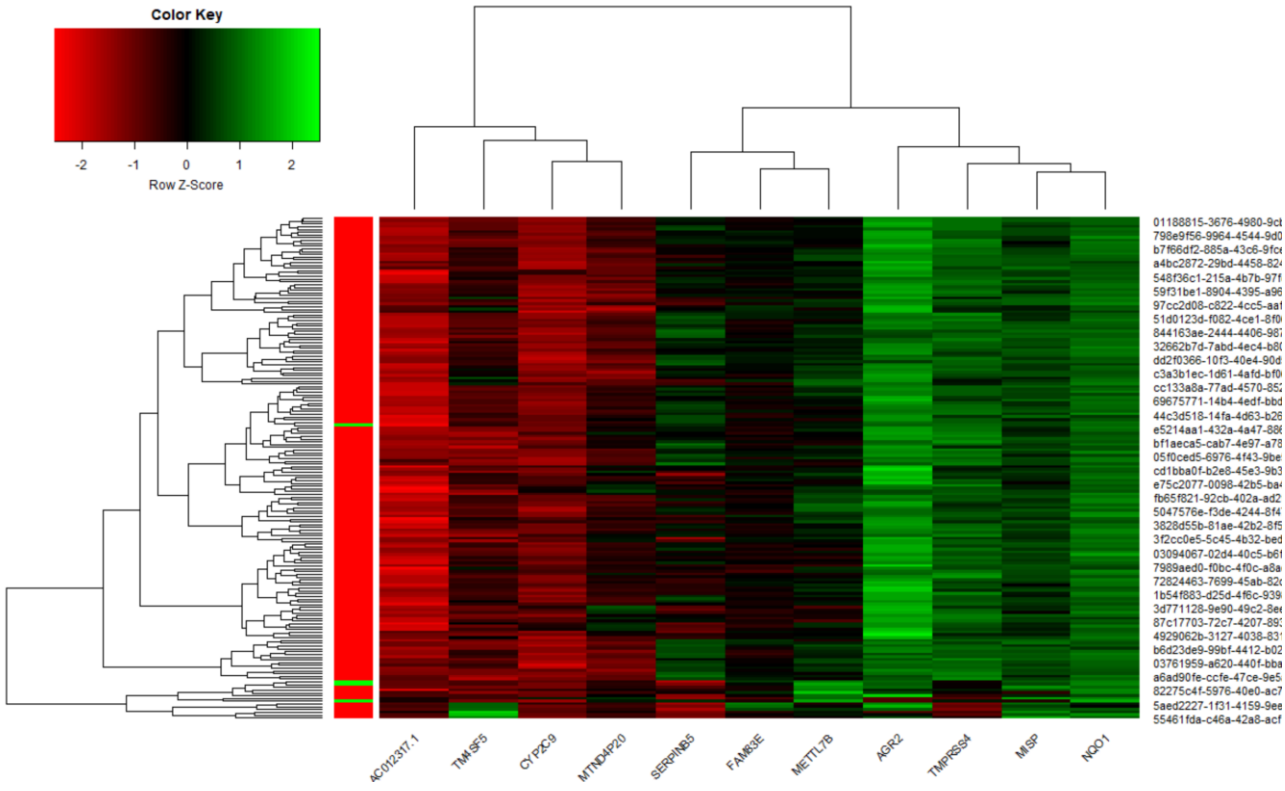


**Figura 3. Mapa de calor de la expresión de genes de los 11 genes.**




## 4.- Partición entrenamiento-test y selección de características

Se crea un conjunto de entrenamiento y otro de test, con una proporción de 75%-25% y manteniendo el balanceo entre clases.

**Tabla 2. Número de casos y frecuencia relativa del reparto de casos en entrenamiento y test.**

| | **Entrenamiento** | **Test** | **Total** |
|---|---|---|---|
| | N (%) | N (%) | N (%) |
| Primary Tumor | 134 (97,8%) | 44 (97,8%) | 178 (97,8%) |
| Solid Tissue Normal | 3 (2,2%) | 1 (2,2%) | 4 (2,2%) |
| **Total** | **137 (100%)** | **45 (100%)** | **182 (100%)** |

**Selección de características**

Se emplean varios métodos de selección de características en el conjunto de entrenamiento: mRMR (mínima redundancia, máxima relevancia), RF (basado en random forest) y DA (disease association, basado en scores obtenidos en la literatura).

**Tabla 3. Orden de genes más relevantes según los distintos métodos de selección de características.**

| Ranking de importancia | mRMR | RF | DA |
|---|---|---|---|
| **1** | SERPINB5 | SERPINB5 | CYP2C9 |
| **2** | CYP2C9 | NQO1 | MTND4P20 |
| **3** | TMPRSS4 | TMPRSS4 | AC012317.1 |
| **4** | MTND4P20 | METTL7B | TM4SF5 |
| **5** | AC012317.1 | AGR2 | FAM83E |
| **6** | AGR2 | AC012317.1 | MISP |
| **7** | TM4SF5 | MISP | METTL7B |
| **8** | FAM83E | FAM83E | TMPRSS4 |
| **9** | MISP | CYP2C9 | NQO1 |
| **10** | NQO1 | MTND4P20 | AGR2 |
| **11** | METTL7B | TM4SF5 | SERPINB5 |

## 5.- Clasificador SVM

### Validación cruzada

Se realiza validación cruzada con 3-fold. Los parámetros óptimos de coste y gamma del algoritmo SVM para cada método de selección de características se presentan en la Tabla 4.

**Tabla 4. Parámetros óptimos encontrados para cada método de selección de características usando el algoritmo SVM en el conjunto de entrenamiento con validación cruzada 3-fold.**

| | Coste (c) | Gamma (g) |
|---|---|---|
| mRMR | 0.01 | 0.9 |
| RF | 0.01 | 0.9 |
| DA | 0.01 | 0.9 |

Se representa a continuación la precisión obtenida en cada fold en función del método de selección de características y el número de genes seleccionado. La función *dataPlot* se actualizó para mostrar mejor la precisión de cada fold cuando hay superposición de líneas.

**Figura 4. Precisión de cada fold en el conjunto de entrenamiento usando el algoritmo SVM con los mejores genes seleccionados por la técnica de mRMR.**

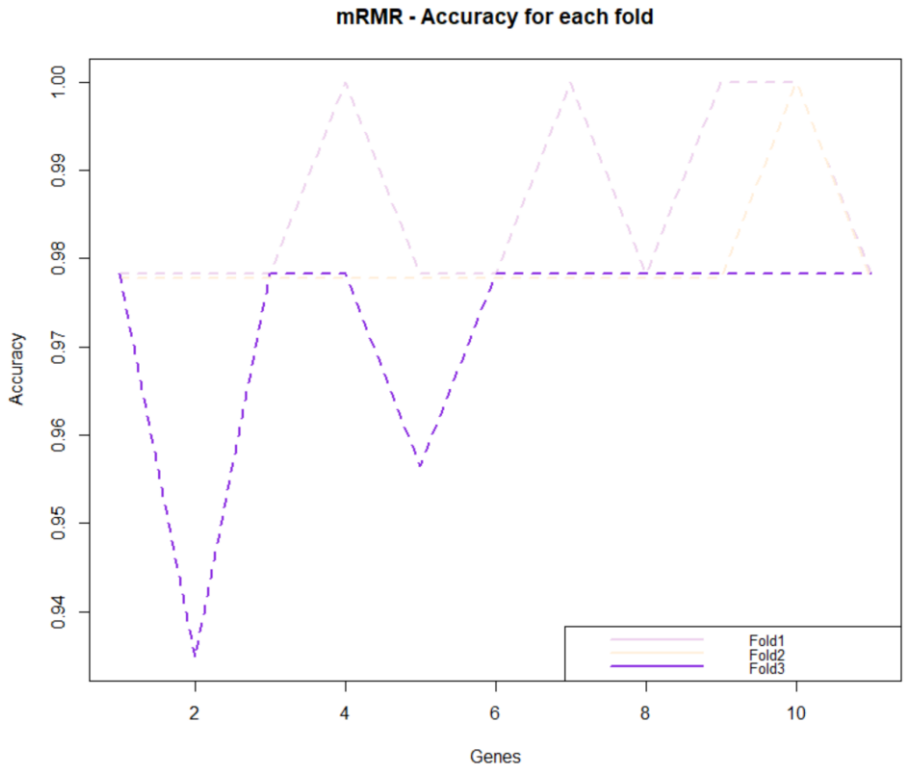



**Figura 5. Precisión de cada fold en el conjunto de entrenamiento usando el algoritmo SVM con los mejores genes seleccionados por la técnica de random forest.**

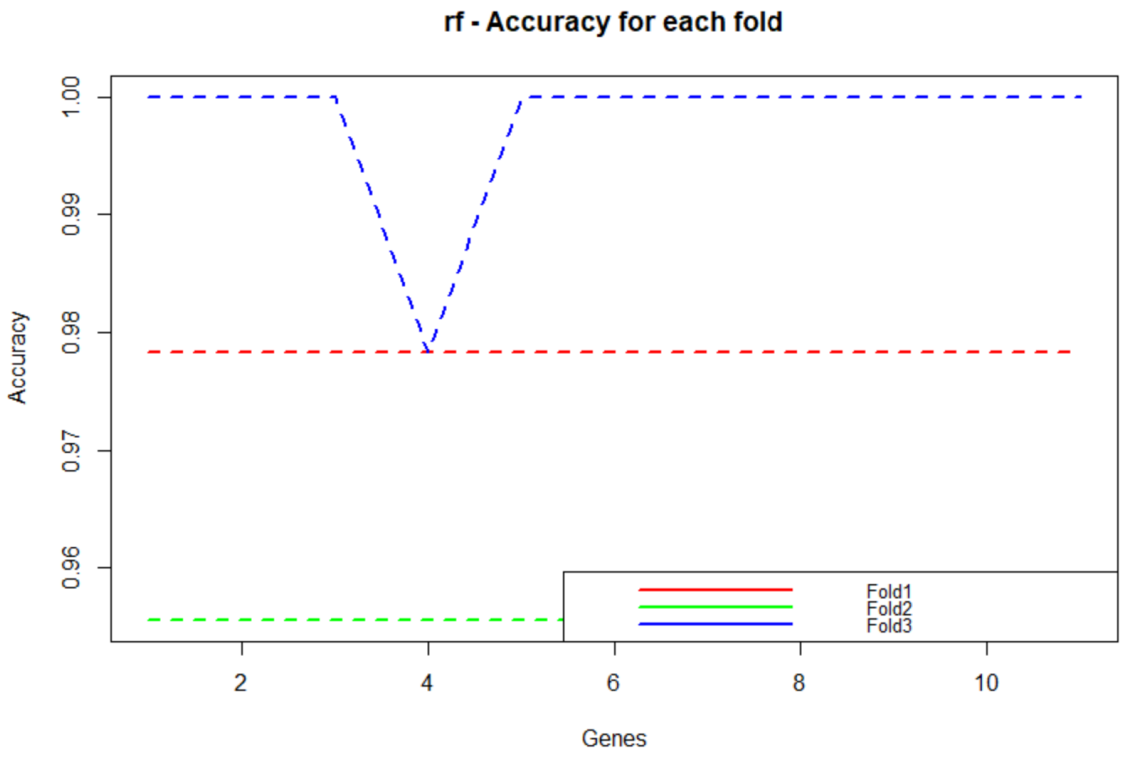


**Figura 6. Precisión de cada fold en el conjunto de entrenamiento usando el algoritmo SVM con los mejores genes seleccionados por la técnica de DA.**

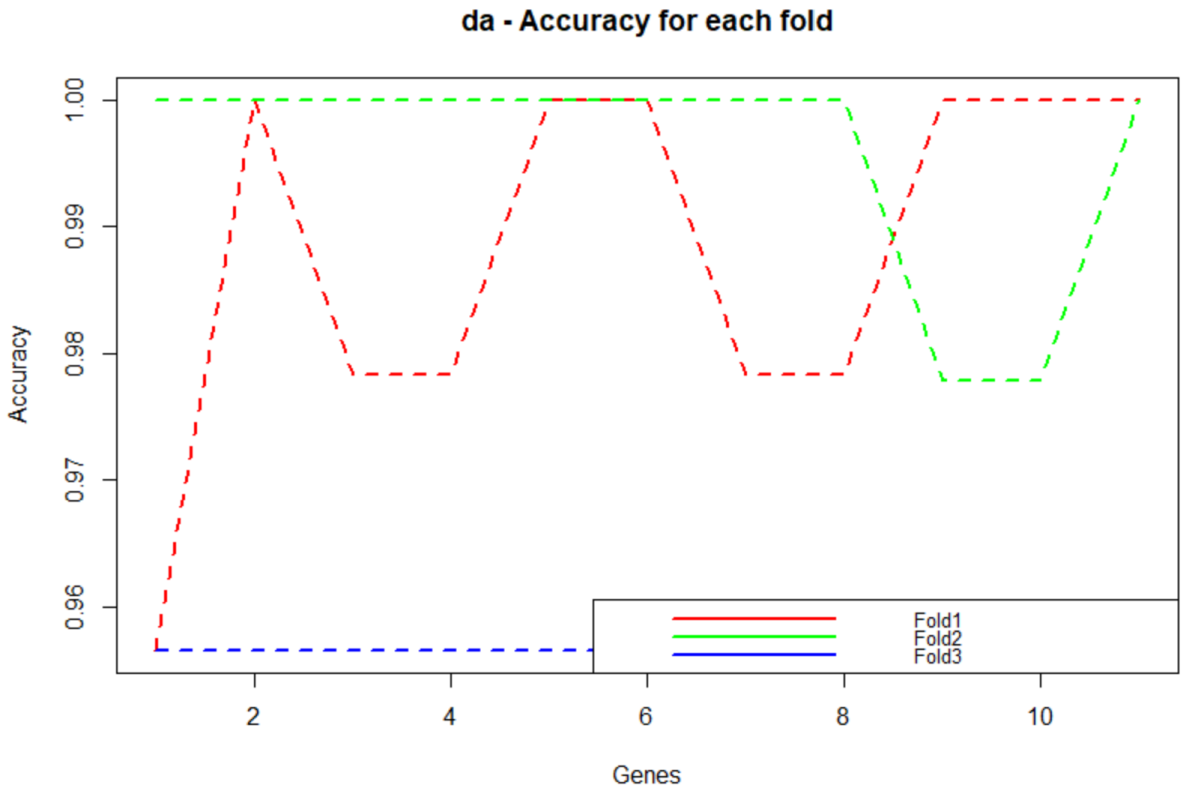



**Tabla 5. Precisión media obtenida en el conjunto de entrenamiento usando el algoritmo SVM con validación cruzada 3-fold, según número de genes seleccionado y técnica de selección de características.**

| Número de genes | mRMR | RF | DA |
|---|---|---|---|
| 1 | 97.8 | 97.8 | 97.1 |
| 2 | 96.4 | 97.8 | 98.6 |
| 3 | 97.8 | 97.8 | 97.8 |
| 4 | 98.5 | 97.1 | 97.8 |
| 5 | 97.1 | 97.8 | 98.6 |
| 10 | 99.3 | 97.8 | 97.8 |

Pero el mejor modelo no debería elegirse en base a la precisión exclusivamente, sobre todo en este caso en el que hay desbalanceo de clases. Analizamos la especificidad de los modelos:

**Tabla 6. Especificidad media obtenida en el conjunto de entrenamiento usando el algoritmo SVM con validación cruzada 3-fold, según número de genes seleccionado y técnica de selección de características.**

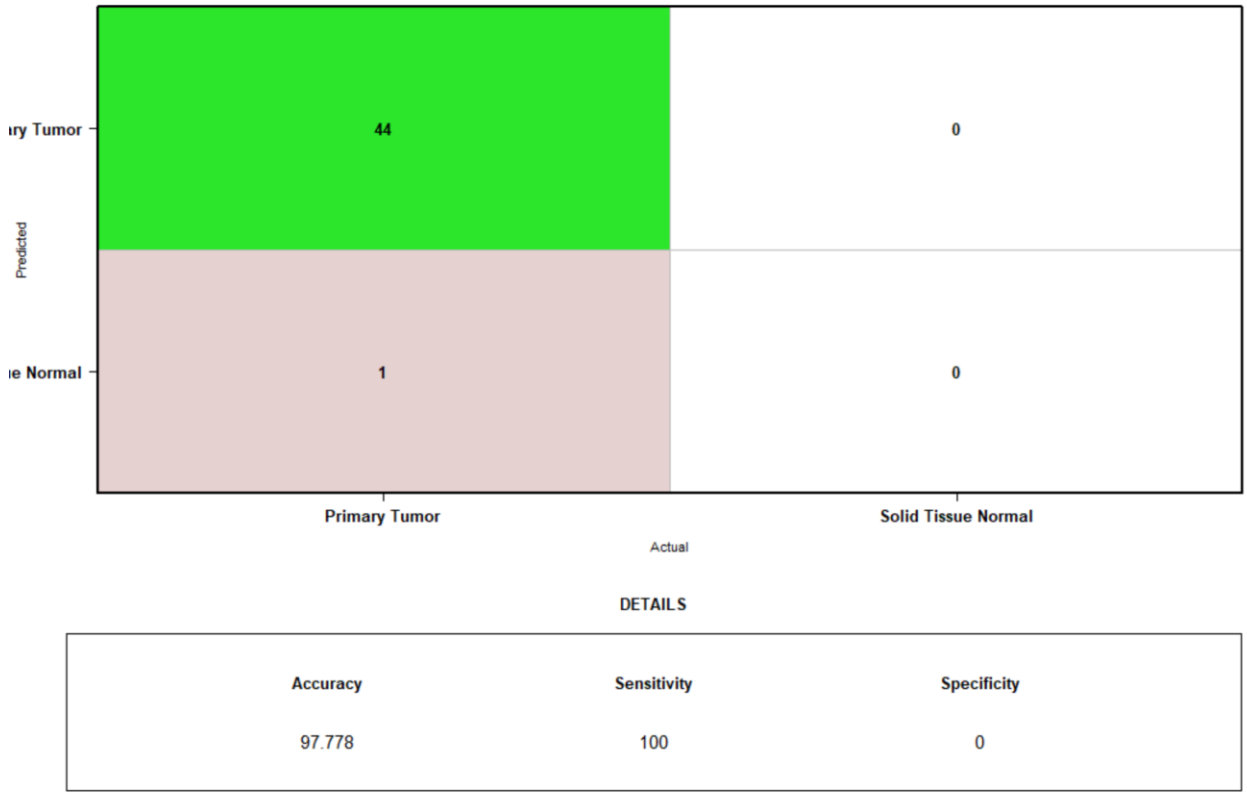
| Número de genes | mRMR | RF | DA |
|---|---|---|---|
| 1 | 0 | 0 | 0.33 |
| 2 | 0.33 | 0 | 0.33 |
| 3 | 0 | 0 | 0 |
| 4 | 0.33 | 0 | 0 |
| 5 | 0 | 0 | 0.33 |
| 10 | **0.67** | 0 | 0.33 |

El clasificador definitivo es entonces aquel que utiliza los 10 genes seleccionados mediante mRMR (sólo no se usa un gen, el menos relevante según mRMR).

**Resultado de clasificador definitivo en conjunto de test**

El mejor modelo encontrado en validación cruzada (10 genes seleccionados mediante mRMR), se evalúa ahora en el conjunto de test, obteniendo los siguientes resultados.

**Figura 7. Matriz de confusión, precisión, sensibilidad y especificidad del conjunto de test al usar el clasificador SVM con el mejor clasificador encontrado tras validación cruzada (mRMR, 10 genes).**




Este algoritmo no consigue clasificar perfectamente el conjunto de test, equivocándose en el caso de tejido normal, clasificándolo como tumor (falso positivo).

## 6.- Clasificador kNN

**Validación cruzada**

Se procede análogamente para el clasfiicador kNN:

**Tabla 7. Parámetro k óptimo encontrado para cada método de selección de características usando el algoritmo SVM en el conjunto de entrenamiento con validación cruzada 3-fold.**

| | k |
|---|---|
| mRMR | 23 |
| rf | 23 |
| da | 23 |

**Figura 8. Precisión de cada fold en el conjunto de entrenamiento usando el algoritmo SVM con los mejores genes seleccionados por la técnica de mRMR.**

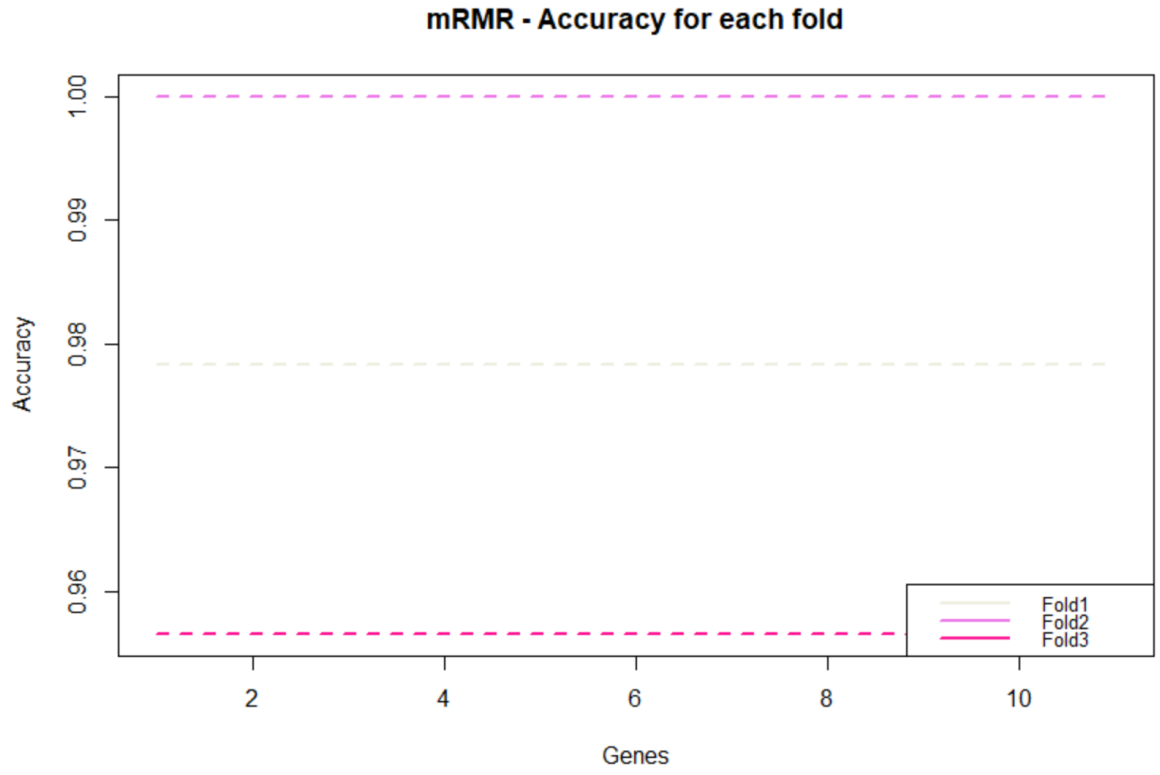



**Figura 9. Precisión de cada fold en el conjunto de entrenamiento usando el algoritmo kNN con los mejores genes seleccionados por la técnica de RF.**

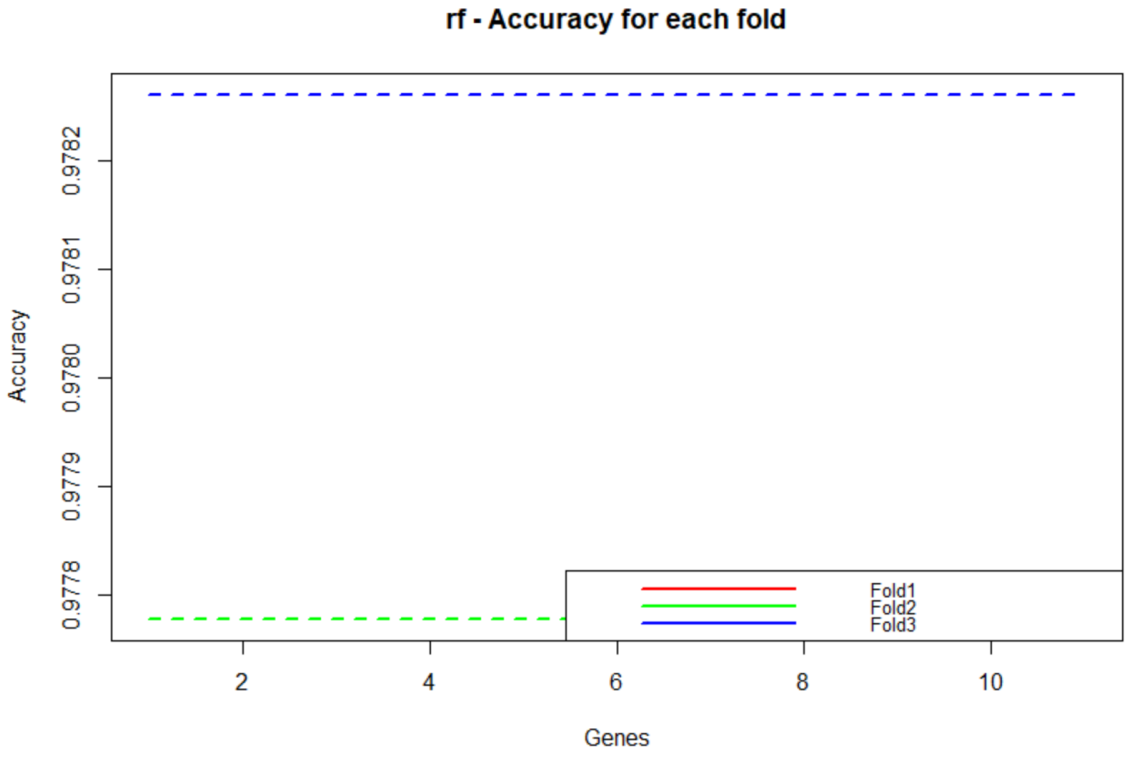


**Figura 10. Precisión de cada fold en el conjunto de entrenamiento usando el algoritmo kNN con los mejores genes seleccionados por la técnica de DA.**

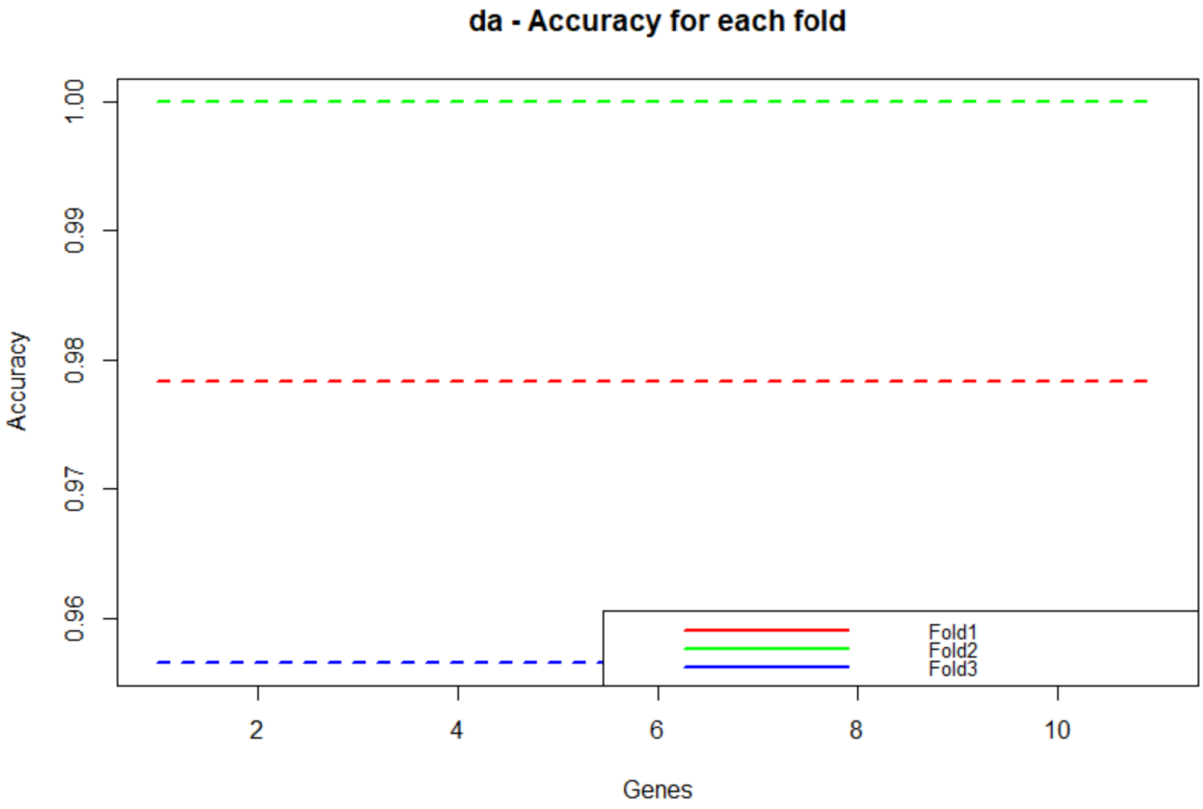



**Tabla 8. Precisión media obtenida en el conjunto de entrenamiento usando el algoritmo SVM con validación cruzada 3-fold, según número de genes seleccionado y técnica de selección de características.**

| Número de genes | mRMR | RF | DA |
|---|---|---|---|
| 1 | 97.8 | 97.8 | 97.8 |
| 2 | 97.8 | 97.8 | 97.8 |
| 3 | 97.8 | 97.8 | 97.8 |
| 4 | 97.8 | 97.8 | 97.8 |
| 5 | 97.8 | 97.8 | 97.8 |
| 10 | 97.8 | 97.8 | 97.8 |

Parece ser que clasifican todos los casos como tejido tumoral. Al ver la especificidad se confirma:

**Tabla 9. Especificidad media obtenida en el conjunto de entrenamiento usando el algoritmo SVM con validación cruzada 3-fold, según número de genes seleccionado y técnica de selección de características.**

| Número de genes | mRMR | RF | DA |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 |

kNN no es un buen clasificador para estos datos. No se elige un mejor clasificador y no se aplica en el conjunto de test (porque clasifica todos los casos como tumor).

## 7.- Enfermedades relacionadas

Se recuperan todas las enfermedades vinculadas a los 10 genes más relevantes del método mRMR, que ha sido el mejor clasificador encontrado tras la validación cruzada. Luego, se seleccionan todas aquellas enfermedades que contienen los términos “cancer” (6 enfermedades), “neopl” (10 enfermedades) y “panc” (0 enfermedades).

El listado de enfermedades es el siguiente: "cancer", "urinary bladder cancer", "skin cancer", "lung cancer", "breast cancer", "thyroid cancer", "neoplasm", "epithelial neoplasm", "stomach neoplasm", "skin neoplasm", "head and neck malignant neoplasia", "lung neoplasm", "thyroid neoplasm", "breast neoplasm", "colonic neoplasm", "liver neoplasm".

Aunque se detectan enfermedades como “cancer” y “neoplasm”, no aparece el cáncer de páncreas como tal entre las enfermedades relacionadas con los genes.