

TRABAJO FIN DE MÁSTER

MÁSTER UNIVERSITARIO OFICIAL EN CIENCIA DE DATOS E
INGENIERÍA DE COMPUTADORES

Epidemiología y detección de biomarcadores en cáncer

Autor:

Daniel Redondo Sánchez

Tutores:

Ignacio Rojas

Luis Javier Herrera

Daniel Castillo

Granada, septiembre de 2020



**UNIVERSIDAD
DE GRANADA**

0. Índice general

Resumen / Abstract	6
1. Introducción	9
1.1. Objetivos del trabajo	9
1.2. ¿Qué es el cáncer?	9
1.2.1. Cáncer de hígado	10
1.2.2. Cáncer de colon-recto	12
1.3. Ómicas	13
1.3.1. Introducción a las ómicas	13
1.3.2. Genómica	14
1.3.3. Transcriptómica	14
2. Epidemiología del cáncer	15
2.1. Indicadores epidemiológicos	15
2.2. Fuentes de información	15
2.3. Incidencia de cáncer	16
2.3.1. Metodología	16
2.3.2. Incidencia del total del cáncer excepto piel no melanoma .	18
2.3.3. Incidencia de cáncer de hígado	20
2.3.4. Incidencia de cáncer de colon-recto	21
2.4. Mortalidad por cáncer	22
2.4.1. Metodología	22
2.4.2. Mortalidad del total del cáncer excepto piel no melanoma .	22
2.4.3. Mortalidad de cáncer de hígado	24
2.4.4. Mortalidad de cáncer de colon-recto	25
2.5. Supervivencia de cáncer	26
2.6. Prevalencia de cáncer	26
3. Machine learning aplicado a transcriptómica	29
3.1. Algoritmos de selección de características	29

3.1.1.	Mínima redundancia máxima relevancia (mRMR)	30
3.1.2.	<i>Random forest</i> (RF)	31
3.1.3.	Asociación de enfermedades (DA)	31
3.2.	Algoritmos de clasificación	31
3.2.1.	Máquinas de soporte vectorial (SVM)	32
3.2.2.	<i>Random Forest</i> (RF)	32
3.2.3.	k-vecinos más cercanos (kNN)	33
3.2.4.	Medidas de evaluación	33
4.	Detección de biomarcadores en cáncer de hígado y colon-recto	35
4.1.	Objetivos	35
4.2.	Fuente de datos	35
4.3.	Características clínicas de los tumores	37
4.3.1.	Características clínicas para cáncer de hígado	37
4.3.2.	Características clínicas para cáncer de colon-recto	40
4.4.	Metodología	43
4.4.1.	Herramientas para el análisis	43
4.4.2.	Detección de biomarcadores	44
4.4.3.	Validación cruzada en entrenamiento	45
4.4.4.	Validación en test	46
4.5.	Resultados de clasificación biclase para cáncer de hígado	46
4.5.1.	Detección de biomarcadores	46
4.5.2.	Validación cruzada en entrenamiento - SVM	49
4.5.3.	Validación cruzada en entrenamiento - RF	53
4.5.4.	Validación cruzada en entrenamiento - kNN	55
4.5.5.	Validación en test	59
4.6.	Resultados de clasificación multiclase para cáncer de hígado	60
4.6.1.	Detección de biomarcadores	60
4.6.2.	Validación cruzada en entrenamiento - SVM	62
4.6.3.	Validación cruzada en entrenamiento - RF	66
4.6.4.	Validación cruzada en entrenamiento - kNN	68
4.6.5.	Validación en test	72
4.7.	Resultados de clasificación biclase para cáncer de colon-recto	73
4.7.1.	Detección de biomarcadores	74
4.7.2.	Validación cruzada en entrenamiento	76
4.7.3.	Validación en test	79

4.8.	Resultados de clasificación multiclase para cáncer de colon-recto	80
4.8.1.	Por diagnóstico primario: sanos vs adenocarcinoma vs adenocarcinoma mucinoso	80
4.8.2.	Por código: sanos vs colon vs recto + unión rectosigmoidea	81
4.9.	Conclusiones	82
5.	biomarkeRs: una aplicación web interactiva para detección de biomarcadores	85
5.1.	Desarrollo de la aplicación	85
5.2.	Utilidades de la aplicación	85
6.	Líneas abiertas de trabajo	87
6.1.	El problema de clasificación	87
6.2.	{KnowSeq}	87
6.3.	BiomarkeRs	88
6.4.	Artículo	88
	Bibliografía	89
	Anexo I: Artículo	99

Resumen

Introducción: El cáncer es uno de los mayores problemas de salud pública del mundo con más de 17 millones de casos nuevos y 9 millones de defunciones al año.

Métodos: Este trabajo se centra en el cáncer de hígado y el cáncer de colon-recto, describiendo sus principales indicadores epidemiológicos y usando *machine learning* para analizar más de 1.100 muestras de RNA-Seq procedentes de pacientes de cáncer. Para clasificación biclase (tumor vs. tejido normal) y multiclase (varios tipos de tumor vs. tejido normal) se identifican los 10 genes más relevantes y se construyen modelos predictivos con SVM, *random forest* y kNN con validación cruzada 5-fold.

Resultados: Los mejores clasificadores biclase para cada algoritmo son validados con excelentes resultados para cáncer de hígado (F1-Score en test: 99,5 %) y cáncer de colon-recto (F1-Score: 100 %). En los mejores modelos para clasificación multiclase se obtienen medidas de evaluación ligeramente inferiores tanto en hígado (F1-Score: 91,8 %) como en colon-recto (F1-Score: **XX %**).

Se ha desarrollado una aplicación web, **BiomarkeRs**, que implementa análisis de transcriptómica y puede resultar de utilidad para personas sin conocimientos previos de programación.

Conclusiones: SVM, random forest y kNN obtienen resultados muy similares, y consiguen distinguir correctamente entre tejidos tumorales y sanos, con algunos problemas para distinguir entre diferentes tipos de cáncer. Es necesaria una validación externa e interpretaciones clínicas para establecer de forma clara una asociación gen-enfermedad.

Palabras clave: epidemiología, transcriptómica, cáncer de hígado, cáncer de colon-recto, RNA-Seq, machine learning, SVM, random forest, kNN.

Abstract

Introduction: Cancer is one of the world's largest public health problems with more than 17 million new cases and 9 million deaths every year.

Methods: This work focuses on liver cancer and colon-rectum cancer, describing their main epidemiological indicators and using machine learning to analyze more than 1,100 RNA-Seq samples from cancer patients. For binary (tumor vs. normal tissue) and multiclass (various tumor types vs. normal tissue) classification, the 10 most relevant genes are identified and predictive models are constructed with SVM, random forest and kNN with 5-fold cross-validation.

Results: The best binary classifiers are validated with excellent results for liver cancer (F1-Score in test: 99.5 %) and colon-rectum cancer (F1-Score: 100 %). Slightly lower evaluation measures are obtained in the best models for multiclass classification, both in liver (F1-Score: 91.8 %) and in colon-rectum (F1-Score: XX %).

A web application has been developed, **BiomarkeRs**, that implements transcriptomic analysis and can be useful for people without previous knowledge of programming.

Conclusions: SVM, random forest and kNN obtained very similar results, and managed to correctly distinguish between tumoral and normal tissues with some troubles distinguishing between different types of cancer. External validation and clinical interpretations are necessary to clearly establish a gene-disease association.

Keywords: epidemiology, transcriptomics, liver cancer, colorectal cancer, RNA-Seq, machine learning, SVM, random forest, kNN.

1. Introducción

1.1. Objetivos del trabajo

En el presente Trabajo Fin de Máster se analiza la epidemiología de los cánceres de hígado y colon-recto y se detectan genes que permiten identificar tumores.

- En el capítulo 1 se presenta la introducción al cáncer y a las ciencias ómicas.
- En el capítulo 2 se describe la epidemiología del cáncer de hígado y colon-recto para comprender el impacto de estos tumores.
- En el capítulo 3 se analizan técnicas de *machine learning* con aplicación en transcriptómica.
- En el capítulo 4 se detectan genes que permiten diferenciar entre personas sanas y personas con cáncer.
- En el capítulo 5 se presenta una aplicación web que permite realizar la detección de biomarcadores y crear modelos de clasificación.
- En el capítulo 6 se muestran las principales conclusiones del trabajo y se muestran líneas abiertas de trabajo.

1.2. ¿Qué es el cáncer?

El cáncer es una enfermedad en la que se produce una división incontrolada de las células [1]. Aunque generalmente se habla del cáncer como una única enfermedad, se trata en realidad de un conjunto de enfermedades multifactoriales, existiendo más de 100 tipos distintos de cáncer [2].

En general, el proceso de creación del cáncer es complejo y multifactorial: a menudo el causante no es un solo elemento, sino la combinación e interacción de distintos factores ambientales y genéticos [3]. En la mayoría de los casos el cáncer no se hereda [4].

Los factores causantes del cáncer se pueden clasificar principalmente en tres categorías:

1. Factores no modificables. Son elementos que no se pueden cambiar, como la edad o la herencia genética [5, 6].
2. Factores modificables o prevenibles. Algunos ejemplos son el tabaco, el alcohol, la dieta o la exposición a distintos carcinógenos [7].
3. Otros factores. Algunas circunstancias no se corresponden a ninguna de las categorías anteriores ya que algunos de sus aspectos no se pueden cambiar. Es el caso de factores socioeconómicos (como cobertura sanitaria en el lugar de residencia o privación económica) y factores reproductivos u hormonales (como toma de anticonceptivos, lactancia materna o terapia hormonal sustitutiva en mujeres menopáusicas) [6].

A continuación se introducen dos tipos de cáncer con los que se trabajará más adelante: el cáncer de hígado y el cáncer de colon-recto.

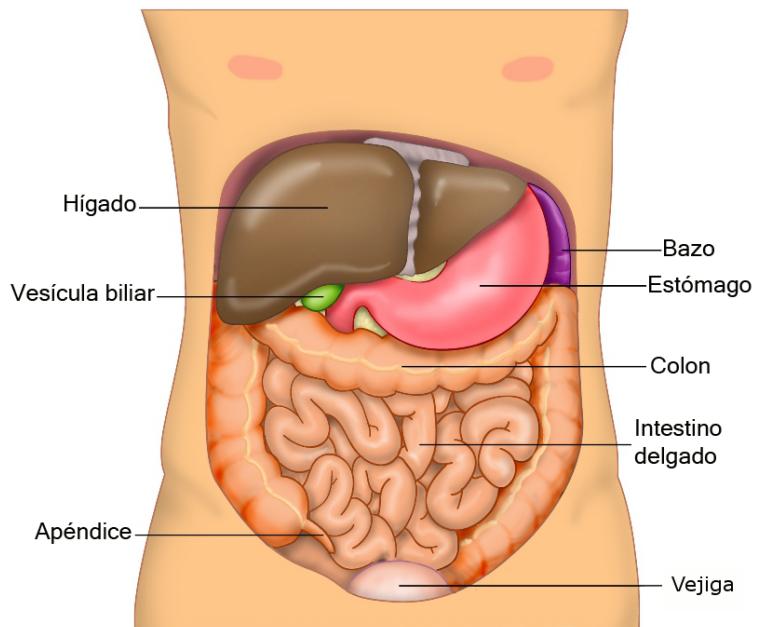
1.2.1. Cáncer de hígado

El cáncer de hígado se corresponde con el código C22 de la Clasificación Internacional de Enfermedades, Décima Revisión, integrando las neoplasias malignas de hígado y vías biliares intrahepáticas [8, 9].

Anatomía y funciones del hígado

El hígado es el órgano interno más grande y pesado del cuerpo humano, está situado en el cuadrante superior derecho del abdomen, debajo de las costillas, y está compuesto principalmente por dos lóbulos [10].

Figura 1. Anatomía del abdomen humano. Ilustración de Ties van Brussel.



Las funciones del hígado son múltiples y diversas. Las principales son procesar, particionar y metabolizar macronutrientes, regular el volumen de sangre, apoyar al sistema inmune, eliminar sustancias químicas como el alcohol y otras drogas y producir bilis para absorber grasas [11]. Es un órgano imprescindible para la vida.

Factores de riesgo

Uno de los factores de riesgo más comunes del cáncer de hígado es la presencia de cirrosis, o sustitución de células sanas de hígado por tejido cicatrizado. La cirrosis puede producirse por varias causas, siendo las más habituales el consumo excesivo de alcohol y la infección con el virus de la hepatitis B o C [12]. Otros factores de riesgo son el tabaco, la obesidad, padecer diabetes tipo II y consumir esteroides anabólicos [12, 13].

La prevención del cáncer de hígado se basa en reducir la exposición a factores de riesgo como el tabaco y el alcohol, y en vacunarse contra la hepatitis B [12].

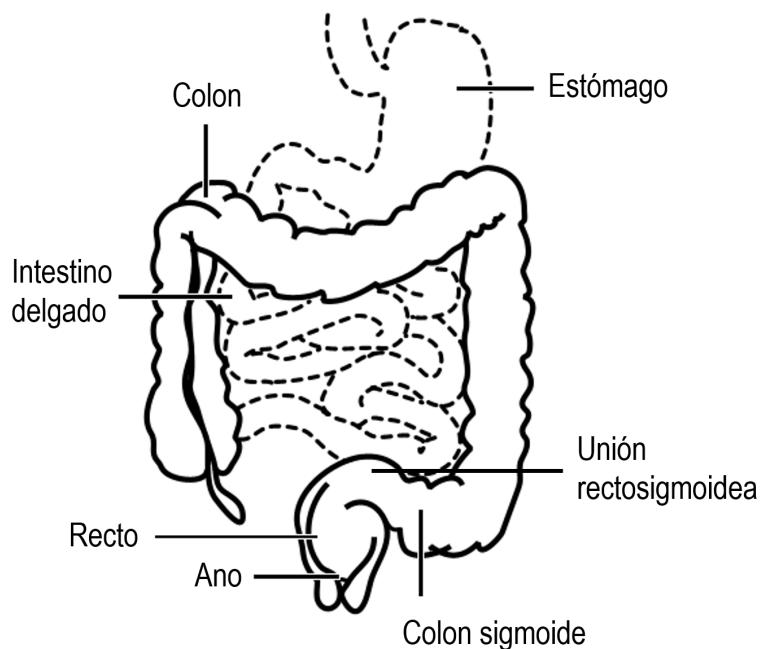
1.2.2. Cáncer de colon-recto

Las neoplasias malignas de colon, recto, unión rectosigmoidea, ano y canal anal (códigos C18-C21 según la Clasificación Internacional de Enfermedades, Décima Revisión [8, 9]) a menudo se estudian agrupadas por tener características muy similares.

Anatomía y funciones del colon-recto

El colon tiene 3 funciones principales: absorción de agua y electrolitos, producción y absorción de vitaminas y movimiento de heces hacia el recto para su eliminación por el ano [14].

Figura 2. Anatomía del intestino humano. Ilustración de Ties van Brussel.



Factores de riesgo

Los factores de riesgo del cáncer de colon-recto se pueden distinguir en factores modificables y no modificables.

Entre los factores de riesgo que son modificables destacan el sobrepeso, la inactividad física, las dietas con alto consumo de carnes rojas o procesadas y el consumo de tabaco y alcohol [15].

Algunos factores de riesgo no modificables son una edad superior a 50 años, padecer diabetes tipo 2 y tener antecedentes personales o familiares de cáncer de colon-recto, pólipos o enfermedad intestinal inflamatoria, como colitis ulcerosa y enfermedad de Crohn [15]. También existen algunos síndromes hereditarios como el síndrome de Lynch que aumentan las posibilidades de padecer cáncer de colon-recto [16].

Para intentar prevenir el cáncer de colon-recto se deben cambiar aquellos factores que son modificables: realizar ejercicio, mantener una dieta saludable y evitar el consumo de tabaco y alcohol. Además, en los últimos años se están implementando programas de cribado de cáncer de colon-recto para detectar pólipos o diagnosticar el cáncer en etapas iniciales mediante análisis como pruebas de sangre oculta en heces o colonoscopias en personas en grupos de riesgo (por ejemplo, mayores de 50 años) [17].

1.3. Ómicas

Se presenta a continuación una corta introducción a las ómicas con el objetivo de comprender los conceptos que se utilizarán más adelante.

1.3.1. Introducción a las ómicas

Los seres vivos están hechos de células. En el núcleo de cada célula se encuentran los cromosomas, estructuras que almacenan los genes del individuo. Las células humanas tienen 46 cromosomas: 23 heredados de la madre y 23 heredados del padre.

La información genética se transporta mediante los ácidos nucleicos: ácido desoxirribonucleico (DNA, por sus siglas en inglés) y ácido ribonucleico (RNA, por sus siglas en inglés). Los ácidos nucleicos son polímeros formados por nucleótidos, que están formados a su vez por un azúcar, un fosfato y una base nitrogenada. En los ácidos nucleicos hay 4 bases nitrogenadas: A, C, G y T para DNA y A, C,

G y U para RNA. La información genética se transmite del DNA al RNA, donde se traduce en la secuencia de aminoácidos de una proteína [4].

Existen muchas ómicas pero las más relevantes para este trabajo son la genómica y la transcriptómica.

1.3.2. Genómica

La genómica es la ciencia que estudia la composición, estructura y función de los genomas. Se dedica por tanto a estudiar cromosomas, mutaciones y variaciones tanto de nucleótidos concretos como de regiones del genoma. No debe confundirse con la genética, ciencia que estudia los genes de manera individual [4].

El análisis GWAS (Genome-wide association study) es un ejemplo de análisis genómico.

1.3.3. Transcriptómica

La transcriptómica estudia el transcriptoma, esto es, el conjunto de RNA presente en una célula [18]. El transcriptoma indica el nivel de expresión de genes en un determinado momento.

Los análisis de RNA-Seq y microRNA son ejemplos de análisis de transcriptómica.

Este trabajo se enmarca dentro de la transcriptómica y está basado en datos obtenidos mediante RNA-Seq, técnica en la que se miden distintas lecturas de cada gen para conocer su nivel de expresión. Por ejemplo, si una muestra tumoral tiene más expresión de un gen que la de una muestra sana se dirá que ese gen está sobreexpresado. En caso contrario se habla de un gen inhibido. La sobreexpresión o inhibición de genes puede causar desórdenes biológicos o codificaciones erróneas de proteínas que podrían conllevar el desarrollo de un cáncer.

La detección de genes como biomarcadores se ha utilizado ampliamente para intentar predecir el diagnóstico del cáncer, basándose en microarrays [19, 20], aunque el uso de microarrays está siendo reemplazado por el uso de RNA-Seq [21, 22].

2. Epidemiología del cáncer

La epidemiología se ha definido tradicionalmente como la ciencia que estudia la distribución y los determinantes de la enfermedad en los seres humanos [23]. En una definición más moderna, no limitada exclusivamente a la enfermedad, la epidemiología se define como el estudio de la aparición y distribución de los estados o acontecimientos relacionados con la salud en poblaciones específicas, incluyendo el estudio de los determinantes de estos estados y la aplicación del conocimiento al control de los problemas de la salud [24].

2.1. Indicadores epidemiológicos

Para medir en la población el impacto del cáncer se utilizan principalmente cuatro indicadores:

- **Incidencia** (casos nuevos). Mide el riesgo de presentar cáncer.
- **Mortalidad** (defunciones). Mide el riesgo de morir por cáncer.
- **Supervivencia** (porcentaje de casos vivos). Mide la historia natural del cáncer y efectividad del tratamiento.
- **Prevalencia** (casos nuevos y antiguos, vivos). Mide la carga asistencial de la enfermedad.

Además, se puede examinar la evolución de cada indicador a lo largo del tiempo, hablando así de tendencias de la incidencia, de la mortalidad, de la supervivencia o de la prevalencia.

2.2. Fuentes de información

A nivel mundial, las estadísticas de incidencia, mortalidad y prevalencia de cáncer las proporciona el *Global Cancer Observatory* (GCO), una plataforma web de la *International Agency for Research on Cancer*, de la Organización Mundial de la

Salud [25, 26]. El organismo equivalente al GCO a nivel europeo es el *European Cancer Information System* (ECIS), de reciente creación y apoyado por la Comisión Europea [27, 28]. Para conocer la supervivencia, el programa CONCORD [29] publica datos a nivel mundial y EUROCARE [30] a nivel europeo.

Aunque estos organismos proporcionan estadísticas sobre cáncer en España, también existen fuentes a nivel nacional que cuentan con datos más actualizados y con distinta metodología. La Red Española de Registros de Cáncer (REDECAN) publica periódicamente datos sobre incidencia y supervivencia de cáncer en España [31, 32], mientras que las estadísticas de mortalidad por cáncer se pueden calcular a partir de las defunciones que publica el Ministerio de Sanidad, Consumo y Bienestar Social (MSCBS) del Gobierno de España [33] y la población que proporciona el Instituto Nacional de Estadística [34].

2.3. Incidencia de cáncer

2.3.1. Metodología

Para medir de manera precisa la incidencia de cáncer en una población es necesaria la existencia de un Registro de Cáncer Poblacional. Estas entidades se dedican a registrar exhaustivamente todos los casos de cáncer diagnosticados en un área geográfica y sus datos son muy útiles para todo tipo de estudios epidemiológicos. Algunos de estos Registros cubren la población de todo un país (por ejemplo, Canadá) mientras que otros cubren regiones concretas (por ejemplo, la provincia de Granada). Desgraciadamente, muchas áreas geográficas no están cubiertas por un Registro de Cáncer Poblacional. Es el caso de España, en el que sólo el 27% de la población está cubierta por un Registro de Cáncer Poblacional [35]. Para conocer de manera estimada la incidencia de cáncer en territorios sin Registro de Cáncer Poblacional o proyectar la incidencia a años posteriores se utilizan diversos métodos matemáticos y estadísticos [25–28, 31, 35].

Con respecto a las medidas usadas para reportar la incidencia, la más sencilla y fácil de interpretar es el número nuevo de casos de cáncer, enmarcado siempre en un periodo concreto de tiempo y un área geográfica. A partir del número de casos se puede calcular la tasa bruta (TB), un indicador que tiene en cuenta el tamaño de la población y que se suele calcular por 100.000 habitantes [36].

$$TB = 100.000 \cdot \frac{\text{Número de casos nuevos}}{\text{Personas en riesgo}}$$

Para permitir comparaciones entre distintas poblaciones, o la misma población en momentos distintos, es necesario tener en cuenta la estructura de edad de la población. Para responder a esta motivación se define la tasa estandarizada por edad (TE) como aquella tasa que habría en la población de estudio si tuviese exactamente la misma estructura de edad que una población estándar predefinida [36]. La definición de la tasa estandarizada por edad para 18 grupos de edad quinquenales (0-4 años, 5-9 años, ..., 80-84 años, 85 años y más) es la siguiente:

$$TE = \sum_{i=1}^{18} \omega_i \frac{N_i}{P_i}$$

donde N_i y P_i son respectivamente el número de casos incidentes y la población en el i -ésimo grupo de edad, y ω_i es el peso que toma la población de referencia en el grupo i -ésimo, con $\sum_{i=1}^{18} \omega_i = 100.000$. Los valores de ω_i están predefinidos en base a poblaciones estándar, siendo las más utilizadas en nuestro contexto las siguientes:

- Población mundial. Propuesta por primera vez en 1960 [37] y modificada más tarde en 1966 [38], permite realizar comparaciones a nivel mundial.
- Antigua población estándar europea. Propuesta en 1976 [39] basándose en la estructura de edad de varias poblaciones escandinavas, permite comparaciones entre zonas europeas.
- Nueva población estándar europea. En el año 2013, la Oficina Europea de Estadística (EUROSTAT) realiza una revisión de la población estándar europea con el objetivo de que la población refleje fielmente el envejecimiento existente en la población europea [40]. Debido a su novedad, el uso de esta población aún no está ampliamente extendido en los organismos internacionales [28] y en ocasiones se reportan las dos tasas estandarizadas por las poblaciones estándar antigua y nueva [27].

En la Tabla 1 se muestran los pesos para cada una de las poblaciones de referencia mencionadas anteriormente.

Tabla 1. Pesos de las poblaciones estándar para el cálculo de tasas estandarizadas por edad.

Grupo de edad	Población estándar mundial	Población estándar europea 1976	Población estándar europea 2013
0-4 años	12.000	8.000	5.000
5-9 años	10.000	7.000	5.500
10-14 años	9.000	7.000	5.500
15-19 años	9.000	7.000	5.500
20-24 años	8.000	7.000	6.000
25-29 años	8.000	7.000	6.000
30-34 años	6.000	7.000	6.500
35-39 años	6.000	7.000	7.000
40-44 años	6.000	7.000	7.000
45-49 años	6.000	7.000	7.000
50-54 años	5.000	7.000	7.000
55-59 años	4.000	6.000	6.500
60-64 años	4.000	5.000	6.000
65-69 años	3.000	4.000	5.500
70-74 años	2.000	3.000	5.000
75-79 años	1.000	2.000	4.000
80-84 años	500	1.000	2.500
≥85 años	500	1.000	2.500

Para utilizar notación internacional, la tasa estandarizada por la población mundial se notará TE-M (población mundial), la tasa estandarizada por la población europea de 1976 se notará TE-aE (antigua población europea) y la de 2013 se notará TE-nE (nueva población europea).

2.3.2. Incidencia del total del cáncer excepto piel no melanoma

El cáncer de piel no melanoma se suele excluir al reportar datos de incidencia del total del cáncer debido a que es muy frecuente y cuenta con buen pronóstico, por

lo que no se suele registrar en los Registros de Cáncer Poblacionales [41, 42].

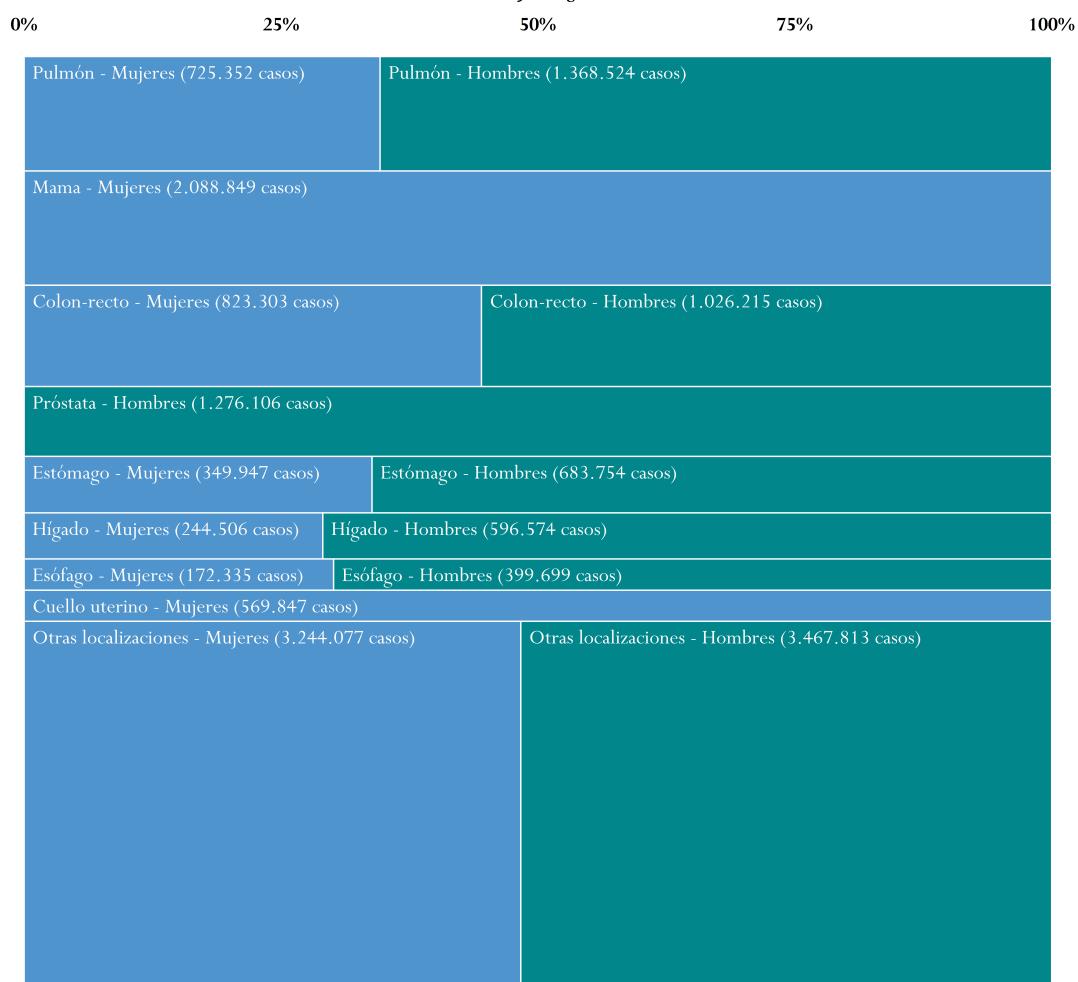
Para dar una perspectiva global del cáncer y sus diferentes tipos, en la Figura 3 se muestran las localizaciones anatómicas más frecuentes de cáncer en el mundo, así como su distribución por sexos.

Figura 3. Gráfico de mosaico con la incidencia estimada de cáncer excepto piel no melanoma en el mundo para el año 2018. Ocho localizaciones anatómicas más frecuentes en ambos sexos. Fuente: *Global Cancer Observatory*, Organización Mundial de la Salud [26].

Incidencia de cáncer en el mundo, 2018

Distribución de casos por sexo y localización anatómica.

Fuente: Global Cancer Observatory, Organización Mundial de la Salud.



El cáncer de pulmón es el más frecuente en todo el mundo, seguido por los cánceres de mama, colon-recto, próstata, estómago, hígado, esófago y cuello uterino. En la mayoría de las localizaciones anatómicas, el cáncer es más frecuente en hombres que en mujeres (Figura 3).

Tabla 2. Incidencia del total del cáncer excepto piel no melanoma en 2018, por sexo y población. Número de casos nuevos (N), tasa bruta (TB), tasa estandarizada por la población mundial (TE-M), tasa estandarizada por la antigua población europea (TE-aE) y tasa estandarizada por la nueva población europea (TE-nE).

Sexo	Población	Fuente	N	TB	TE-M	TE-aE	TE-nE
Hombres	Mundo	GCO [26]	8.818.685	229,0	204,7		
	Europa	ECIS [27]	2.059.673	572,9	302,7	436,0	651,7
	España	ECIS [27]	142.353	625,6	309,7	444,7	658,6
Mujeres	Mundo	GCO [26]	8.218.216	217,3	175,6		
	Europa	ECIS [27]	1.851.644	481,8	242,7	332,6	451,2
	España	ECIS [27]	106.647	451,1	218,4	298,5	401,7
Ambos sexos	Mundo	GCO [26]	17.036.901	223,2	187,8		
	Europa	ECIS [27]	3.911.317	525,8	266,7	374,3	531,9
	España	ECIS [27]	249.000	536,7	259,4	363,8	515,3

A nivel nacional, REDECAN ha publicado estimaciones de la incidencia en España más recientes a las mostradas en la Tabla 2, correspondientes al año 2020 [31]. En este análisis se estima el número de casos de cáncer excepto piel no melanoma en 277.394 casos (57,8% en hombres), con una tasa bruta de 588,0 por 100.000 habitantes y tasas estandarizadas de 280,3 (TE-M), 399,4 (TE-aE) y 579,8 (TE-nE).

2.3.3. Incidencia de cáncer de hígado

El cáncer de hígado es el sexto cáncer más frecuente del mundo (Figura 3), con más de 840.000 casos nuevos anuales en todo el mundo, 82.000 de ellos en Europa y 6.600 en España (Tabla 3). Es un cáncer más frecuente en hombres que en mujeres: por cada caso en mujeres hay 2,4 casos en hombres. En ambos sexos, la TE-M de España (6,5) es mayor que la de Europa (5,1) aunque mucho menor que la del mundo (9,3).

Tabla 3. Incidencia de cáncer de hígado en 2018, por sexo y población. Número de casos nuevos (N), tasa bruta (TB), tasa estandarizada por la población mundial (TE-M), tasa estandarizada por la antigua población europea (TE-aE) y tasa estandarizada por la nueva población europea (TE-nE).

Sexo	Población	Fuente	N	TB	TE-M	TE-aE	TE-nE
Hombres	Mundo	GCO [26]	596.574	15,5	13,9		
	Europa	ECIS [27]	55.825	15,5	8,0	11,7	17,7
	España	ECIS [27]	4.976	21,9	10,9	15,7	22,5
Mujeres	Mundo	GCO [26]	244.506	6,5	4,9		
	Europa	ECIS [27]	26.641	6,9	2,7	4,0	6,3
	España	ECIS [27]	1.654	7,0	2,4	3,6	6,0
Ambos sexos	Mundo	GCO [26]	841.080	11,0	9,3		
	Europa	ECIS [27]	82.466	11,1	5,1	7,4	11,3
	España	ECIS [27]	6.630	14,3	6,5	9,3	13,6

En las estimaciones publicadas por REDECAN para 2020 se estiman en España 6.595 casos de cáncer de hígado (75,4 % en hombres), con una tasa bruta de 14,0 y tasas estandarizadas de 6,5 (TE-M), 9,4 (TE-aE) y 13,9 (TE-nE) [31].

2.3.4. Incidencia de cáncer de colon-recto

El cáncer de colon-recto es el tercer cáncer más frecuente del mundo (Figura 3), con más de 1.800.000 casos nuevos anuales en todo el mundo. En Europa y España es el cáncer más frecuente en ambos性es con 510.000 casos anuales en Europa y 37.000 en España (Tabla 4). Es un cáncer ligeramente más frecuente en hombres que en mujeres: por cada caso en mujeres hay 1,2 casos en hombres. En ambos性es, la TE-M de España (81,1) es mayor que la de Europa (68,8) y la del mundo (24,2), lo que puede deberse a una mayor exposición a factores de riesgo como el tipo de dieta o la ausencia de un programa de cribado a nivel nacional [43].

Tabla 4. Incidencia de cáncer de colon-recto en 2018, por sexo y población. Número de casos nuevos (N), tasa bruta (TB), tasa estandarizada por la población mundial (TE-M), tasa estandarizada por la antigua población europea (TE-aE) y tasa estandarizada por la nueva población europea (TE-nE).

Sexo	Población	Fuente	N	TB	TE-M	TE-aE	TE-nE
Hombres	Mundo	GCO [26]	1.026.215	26,6	23,6		
	Europa	ECIS [27]	275.519	76,6	38,1	56,8	88,9
	España	ECIS [27]	23.013	101,1	45,8	68,5	107,2
Mujeres	Mundo	GCO [26]	823.303	21,8	16,3		
	Europa	ECIS [27]	236.101	61,4	25,2	37,0	56,3
	España	ECIS [27]	14.642	61,9	23,6	34,9	53,5
Ambos sexos	Mundo	GCO [26]	1.849.518	24,2	19,7		
	Europa	ECIS [27]	511.620	68,8	30,8	45,6	70,0
	España	ECIS [27]	37.655	81,1	33,9	50,4	77,5

En las estimaciones publicadas por REDECAN para 2020 se estiman en España 44.231 casos de cáncer de colon-recto (58,9 % en hombres), con una tasa bruta de 93,8 y tasas estandarizadas de 40,0 (TE-M), 59,5 (TE-aE) y 91,9 (TE-nE) [31].

2.4. Mortalidad por cáncer

2.4.1. Metodología

Los indicadores para medir la mortalidad por cáncer son los mismos que para la incidencia, cambiando número de casos por defunciones por cáncer. Es importante destacar que la mortalidad por cáncer es por definición aquella mortalidad que es causada directamente por el cáncer. En este sentido, una persona diagnosticada de cáncer que falleciese por otras causas no puede ser considerada como fallecida por cáncer, sino fallecida con cáncer.

Aunque ECIS [27] reporta mortalidad por cáncer en España para el año 2018, la mortalidad que se presenta a continuación a nivel nacional es la que proporciona el MSCBS del Gobierno de España [33], ya que se consideran datos más fiables al tratarse de datos observados basados en certificados médicos de defunción y no ser estimaciones.

2.4.2. Mortalidad del total del cáncer excepto piel no melanoma

En el mundo se producen más de 9 millones de defunciones por cáncer anualmente, siendo esta enfermedad la segunda causa de muerte a nivel global, por detrás

de las enfermedades cardiovasculares. Una de cada 6 muertes es causa directa del cáncer [44].

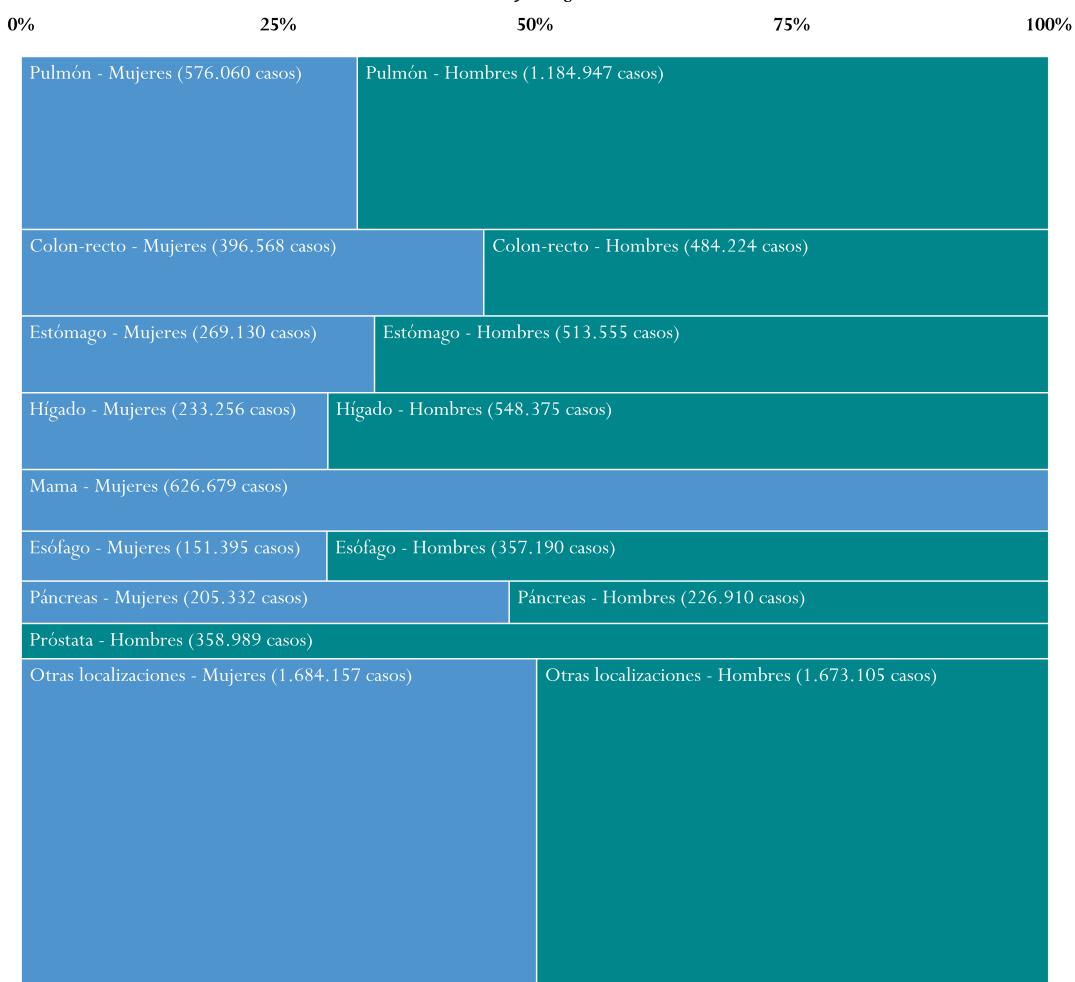
En ambos sexos, los cánceres que provocan más defunciones son los de pulmón, colon-recto y estómago, mientras que en mujeres el cáncer con más defunciones es el cáncer de mama y en hombres el de pulmón (Figura 4).

Figura 4. Gráfico de mosaico con la mortalidad estimada por cáncer excepto piel no melanoma en el mundo para el año 2018. Ocho localizaciones anatómicas con más defunciones en ambos sexos. Fuente: *Global Cancer Observatory*, Organización Mundial de la Salud [26].

Mortalidad por cáncer en el mundo, 2018

Distribución de defunciones por sexo y localización anatómica.

Fuente: *Global Cancer Observatory*, Organización Mundial de la Salud.



En España se producen más de 100.000 muertes anuales por cáncer, siendo la tasa de mortalidad inferior a la europea y a la mundial tanto en ambos sexos como en hombres y mujeres (Tabla 5).

Tabla 5. Mortalidad por total del cáncer excepto piel no melanoma en 2018, por sexo y población. Número de defunciones (N), tasa bruta (TB), tasa estandarizada por la población mundial (TE-M), tasa estandarizada por la antigua población europea (TE-aE) y tasa estandarizada por la nueva población europea (TE-nE).

Sexo	Población	Fuente	N	TB	TE-M	TE-aE	TE-nE
Hombres	Mundo	GCO [26]	5.347.295	138,9	121,9		
	Europa	ECIS [27]	1.077.986	299,8	143,2	217,4	355,4
	España	MSCBS [33]	65.610	286,4	121,3	186,8	314,1
Mujeres	Mundo	GCO [26]	4.142.577	109,5	82,7		
	Europa	ECIS [27]	851.723	221,6	86,4	128,1	201,0
	España	MSCBS [33]	42.248	177,4	63,2	94,6	151,1
Ambos sexos	Mundo	GCO [26]	9.489.872	124,3	100,5		
	Europa	ECIS [27]	1.929.709	259,4	110,8	165,8	263,9
	España	MSCBS [33]	107.858	230,8	89,4	135,4	221,2

2.4.3. Mortalidad de cáncer de hígado

Si bien el cáncer de hígado es el sexto cáncer más incidente del mundo, en mortalidad ocupa la cuarta posición, siendo la causa de 782.000 defunciones anuales (Figura 4). La muerte por cáncer de hígado es mucho más frecuente en hombres que en mujeres. En Europa el cáncer de hígado causa cerca de 80.000 muertes y en España unas 5.100, con unas tasas de mortalidad muy similares y por debajo de la tasa mundial (Tabla 6).

Tabla 6. Mortalidad por cáncer de hígado en 2018, por sexo y población. Número de defunciones (N), tasa bruta (TB), tasa estandarizada por la población mundial (TE-M), tasa estandarizada por la antigua población europea (TE-aE) y tasa estandarizada por la nueva población europea (TE-nE).

Sexo	Población	Fuente	N	TB	TE-M	TE-aE	TE-nE
Hombres	Mundo	GCO [26]	548.375	14,2	12,7		
	Europa	ECIS [27]	50.365	14,0	6,8	10,3	16,4
	España	MSCBS [33]	3.577	15,6	7,0	10,7	17,0
Mujeres	Mundo	GCO [26]	233.256	6,2	4,6		
	Europa	ECIS [27]	27.010	7,0	2,4	3,8	6,3
	España	MSCBS [33]	1.564	6,6	2,0	3,2	5,6
Ambos sexos	Mundo	GCO [26]	781.631	10,2	8,5		
	Europa	ECIS [27]	77.375	10,4	4,4	6,6	10,6
	España	MSCBS [33]	5.141	11,0	4,4	6,7	10,7

2.4.4. Mortalidad de cáncer de colon-recto

El cáncer de colon-recto es el segundo cáncer que provoca más defunciones, con más de 880.000 defunciones anuales en todo el mundo (Figura 4). La defunción por cáncer de colon-recto es ligeramente más frecuente en hombres que en mujeres. Las tasas de mortalidad en España y Europa son muy similares, por encima en ambos casos de la tasa mundial (Tabla 7).

Tabla 7. Mortalidad por cáncer de colon-recto en 2018, por sexo y población. Número de defunciones (N), tasa bruta (TB), tasa estandarizada por la población mundial (TE-M), tasa estandarizada por la antigua población europea (TE-aE) y tasa estandarizada por la nueva población europea (TE-nE).

Sexo	Población	Fuente	N	TB	TE-M	TE-aE	TE-nE
Hombres	Mundo	GCO [26]	484.224	12,6	10,8		
	Europa	ECIS [27]	131.155	36,5	16,4	25,7	44,3
	España	MSCBS [33]	9.222	40,3	15,8	25,1	44,4
Mujeres	Mundo	GCO [26]	396.568	10,5	7,2		
	Europa	ECIS [27]	115.059	29,9	10,0	15,6	26,6
	España	MSCBS [33]	6.066	25,5	7,5	11,9	20,7
Ambos sexos	Mundo	GCO [26]	880.792	11,5	8,9		
	Europa	ECIS [27]	246.214	33,1	12,8	19,9	33,8
	España	MSCBS [33]	15.288	32,7	11,2	17,7	30,9

2.5. Supervivencia de cáncer

El proyecto CONCORD (*Global surveillance of cancer survival*), coordinado por la *London School of Hygiene & Tropical Medicine*, es el mayor proyecto de investigación sobre supervivencia a nivel mundial. En su tercera y última publicación, se analizaron datos de más de 37 millones de pacientes procedentes de 322 Registros de Cáncer Poblacionales de 71 países [29]. Las tendencias de la supervivencia están aumentando con el tiempo, aunque se aprecian importantes diferencias entre territorios. Para casos de España diagnosticados durante el periodo 2010-2014, la supervivencia relativa estandarizada por edad a 5 años del diagnóstico es del 63,2% para cáncer de colon, 59,5% para cáncer de recto y 17,3% para cáncer de hígado [29].

En el contexto europeo, EUROCARE (*European Cancer Registry based study on survival and care of cancer patients*) proporciona datos para el periodo 2000-2007, estimando la supervivencia relativa estandarizada por edad a 5 años del total del cáncer en un 50,3% para hombres y 58,0% para mujeres [27, 30]. Para cáncer de colon-recto la supervivencia es similar a la media global (54,7% en hombres, 56,7% en mujeres), y para cáncer de hígado la supervivencia es muy baja (11,5% tanto para hombres como para mujeres) [27, 30].

Los datos más recientes de supervivencia para España corresponden al periodo 2008-2013 y están publicados por REDECAN [32]. Para el total del cáncer excepto piel no melanoma, la supervivencia relativa estandarizada por edad a 5 años es de 55,3% en hombres y 61,7% en mujeres. El cáncer de hígado tiene una de las supervivencias más bajas, tanto en hombres (17,9%) como en mujeres (16,2%). La supervivencia del cáncer de colon es de 63,1% en hombres y 63,9% en mujeres, ligeramente superior que la supervivencia del cáncer de recto (60,4% y 62,7% para hombres y mujeres respectivamente).

2.6. Prevalencia de cáncer

En el mundo hay más de 11 millones de casos prevalentes de cáncer a 1 año, esto es, casos de cáncer en pacientes vivos que fueron diagnosticados en el último año [26]. A 5 años el número se eleva hasta casi 39 millones de casos [26]. En la Tabla 8 se muestra la prevalencia a 1 y 5 años por sexos y población para los cánceres de hígado y colon-recto, así como para el total del cáncer excepto piel

no melanoma.

Tabla 8. Prevalencia del total del cáncer excepto piel no melanoma, cáncer de hígado y cáncer de colon-recto en 2018, por sexo y población. Número de casos prevalentes y tasas por 100.000 habitantes a 1 y 5 años.

			1 año		5 años	
			N	Tasa	N	Tasa
Total del cáncer excepto piel no melanoma	Hombres	Mundo	5.607.801	145,6	17.895.356	464,7
		Europa	1.504.232	418,4	5.086.515	1414,7
		España	105.599	464,0	356.427	1566,3
	Mujeres	Mundo	5.688.175	150,4	20.738.064	548,3
		Europa	1.434.849	373,4	5.417.680	1409,8
		España	84.409	357,0	322.341	1363,5
	Ambos sexos	Mundo	11.295.976	148,0	38.633.420	506,1
		Europa	2.939.081	395,1	10.504.195	1412,2
		España	190.008	409,5	678.768	1462,9
Hígado	Hombres	Mundo	236.669	6,1	471.525	12,2
		Europa	21.240	5,9	39.867	11,1
		España	1.924	8,5	3.618	15,9
	Mujeres	Mundo	97.621	2,6	203.685	5,4
		Europa	9.719	2,5	18.610	4,8
		España	580	2,5	1.102	4,7
	Ambos sexos	Mundo	334.290	4,4	675.210	8,8
		Europa	30.959	4,2	58.477	7,9
		España	2.504	5,4	4.720	10,2
Colon-recto	Hombres	Mundo	749.774	19,5	2.595.326	67,4
		Europa	213.233	59,3	748.455	208,2
		España	18.059	79,4	63.593	279,5
	Mujeres	Mundo	606.377	16,0	2.194.309	58,0
		Europa	178.969	46,6	655.422	170,6
		España	11.463	48,5	42.121	178,2
	Ambos sexos	Mundo	1.356.151	17,8	4.789.635	62,8
		Europa	392.202	52,7	1.403.877	188,7
		España	29.522	63,6	105.714	227,8

3. *Machine learning* aplicado a transcriptómica

3.1. Algoritmos de selección de características

Los algoritmos de selección de características o variables consisten en la elección de un subconjunto de variables relevantes que permitan:

- Obtener predicciones óptimas para un bajo número de variables.
- Proporcionar predictores menos costosos computacionalmente.
- Mejorar la interpretabilidad de los modelos resultantes y facilitar la visualización de datos.

La selección de características cuenta con interesantes aplicaciones en la genómica ya que permite encontrar conjuntos pequeños de biomarcadores que permiten diferenciar con gran precisión entre distintos estados o patologías [45, 46]. Si se considera cada gen como una variable, la selección de características consigue reducir los problemas asociados a la maldición de la dimensionalidad [47, 48].

Se distinguen principalmente 3 tipos de algoritmos de selección de características, descritos en varias referencias [46, 49]:

- Algoritmos de selección por filtrado. Utilizan técnicas estadísticas para identificar las variables más relevantes antes de diseñar el modelo predictivo. Suelen estar basados en medidas de correlación entre variables, como la información mutua, y pueden devolver un ranking de relevancia de las variables o un subconjunto óptimo de variables. Entre sus ventajas destacan un bajo costo computacional en el entrenamiento del modelo, gran interpretabilidad y facilidad de implementación. Un ejemplo de algoritmo de selección por filtrado es mRMR (mínima redundancia, máxima relevancia).
- Algoritmos de selección embebidos. Utilizan el método de entrenamiento del modelo para seleccionar simultáneamente las características más relevantes. Ejemplos de algoritmos de selección embebido son el uso de *random forest* o máquinas de soporte vectorial.

- Algoritmos de selección por envoltura. En estos métodos, el algoritmo de selección de variables está incluido en el propio modelo predictivo y es retroalimentado por él, seleccionando aquel modelo que proporciona mejor efectividad. El principal inconveniente de estos métodos es el elevado coste computacional, aunque como ventaja asegura el mejor rendimiento de entre todas las opciones que se han evaluado. Un ejemplo es MINT, una mejora de mRMR [50].

Se presentan a continuación los algoritmos de selección de variables que se utilizarán en el capítulo 4.

3.1.1. Mínima redundancia máxima relevancia (mRMR)

El método de mínima redundancia máxima relevancia (mRMR) está basado en el concepto de “información mutua” [51]. La información mutua de dos variables se cuantifica como la reducción de incertidumbre sobre una de las variables conocida la otra.

El algoritmo mRMR funciona hacia delante: partiendo del conjunto vacío de características, selecciona aquella variable que tenga alta relevancia (alta información mutua con la variable resultado) pero a su vez tenga baja redundancia (información mutua) con el resto de variables ya seleccionadas [52]. Matemáticamente, en cada paso se selecciona la variable X que maximiza la siguiente función:

$$I(X, Y) - \frac{1}{|S|} \sum_{W \in S} I(X, W)$$

siendo I la función que mide la información mutua entre dos variables, Y la variable resultado y S el conjunto de variables ya seleccionadas. El proceso se repite hasta alcanzar un cierto número prefijado de variables seleccionadas. El resultado final es un ranking de variables ordenadas en base a su importancia respecto al criterio mRMR.

Es un método muy conocido que ha sido utilizado ampliamente en ciencias -ómicas [53–57]. En R está implementado mediante la función `KnowSeq::featureSelection(mode = 'mrmr')` [58], que a su vez utiliza `praznik::MRMR` [59].

3.1.2. *Random forest* (RF)

Uno de los resultados del modelo de clasificación *random forest*, detallado en la sección 3.2.2. es un ranking de variables según su importancia. Este método de selección de variables se trata por tanto de un método embebido.

La importancia de una variable en el modelo se puede cuantificar usando la reducción media en precisión del modelo al aleatorizar los valores de la variable manteniendo su distribución [60, 61]. También se puede usar la reducción media de otras medidas de entropía como el índice de Gini [62].

Este algoritmo en R está implementado en `KnowSeq::featureSelection(mode = 'rf')` [58] que a su vez utiliza `randomForest::randomForest` [63].

3.1.3. Asociación de enfermedades (DA)

El método de selección de características mediante asociación de enfermedades (DA, por sus siglas en inglés: *Disease Association*) permite encontrar aquellos genes que están asociados en la literatura científica con una determinada enfermedad, en base a evidencias tales como rutas metabólicas afectadas, expresión de RNA o fármacos. Utiliza para ello la plataforma *targetValidation* de *Open Targets* [64], que contiene para cada gen una puntuación midiendo la asociación gen-enfermedad en el rango de 0 (no hay asociación) a 1 (la asociación es total). El método DA obtiene esas puntuaciones y las ordena, obteniendo un ranking de genes en base a su asociación con la enfermedad.

El método DA está implementado en R en `KnowSeq::featureSelection(mode = 'da')` [58], que utiliza a su vez la REST API de *targetValidation* [64].

3.2. Algoritmos de clasificación

Dado un conjunto de datos con varias variables, siendo una de ellas la clase, el problema de clasificación consiste en encontrar un modelo que prediga la clase basándose en el resto de variables. Se presentan a continuación los algoritmos de clasificación que se utilizarán en el capítulo 4.

3.2.1. Máquinas de soporte vectorial (SVM)

Las máquinas de soporte vectorial (SVM) son uno de los algoritmos más populares de la última década debido a su alta eficacia en problemas con grandes cantidades de datos. Consiste en un caso particular de *kernel machines*, una familia de métodos en el que el espacio de características se transforma para ser representados en un espacio más complejo, de alta dimensión o incluso infinita. Para ello, utilizan *kernels*: funciones que devuelven el producto escalar entre las transformaciones de dos argumentos, sin tener que especificar de forma explícita la transformación realizada. Resumidamente, en los algoritmos SVM la transformación de los datos en el nuevo espacio permite que las clases sean linealmente separables mediante un hiperplano que separa las clases dejando el máximo margen posible entre ellas [65]. Aquellas instancias que limitan el margen de ese hiperplano son llamadas vectores soporte. Aunque en un principio el algoritmo SVM fue ideado para problemas binarios [65], se ha generalizado para problemas multi-clase [66].

Uno de los *kernel* más comunes para SVM es el de base radial, que viene determinado por dos parámetros: coste (c) y gamma (γ). El coste c mide la permisividad que se permite a la existencia de clases mal clasificadas por el modelo, que podrían ser ejemplos extraños (*outliers*). Un coste alto puede sobreajustar el modelo (*overfitting*), mientras que un coste bajo puede reducir la precisión del modelo en el conjunto de entrenamiento. El parámetro γ mide el nivel de influencia de cada vector soporte para la construcción del hiperplano separador. Para cada problema se pueden encontrar valores de coste y gamma que optimizan los resultados del algoritmo SVM con base radial, aunque debido a que la búsqueda exhaustiva suele ser costosa computacionalmente se suele realizar una búsqueda de los mejores parámetros dentro de unos posibles valores de c y γ , técnica que se denomina tuning en rejilla.

Los algoritmos SVM están implementados en varios paquetes de R, siendo el más común `{e1071}` [67].

3.2.2. Random Forest (RF)

Random forest (RF) es uno de los algoritmos de machine learning más usados en la actualidad y se puede aplicar a tareas de clasificación y regresión. Para clasificación es un método en el que se crean varios árboles de decisión sin correlación

entre sí para elegir la clase más votada por los árboles como la clase predicha. Para conseguir esta ausencia de correlación, cada vez que se considera una división en cada árbol se obliga a que la variable que dividirá las instancias tenga que pertenecer a un subconjunto de las variables seleccionado aleatoriamente [60, 61]. Debido a este método de construcción de árboles, es un algoritmo cuya principal desventaja es la ausencia de interpretabilidad.

El algoritmo de RF es una mejora del método de *bagging* en árboles de decisión, que consiste en crear árboles basados en una selección aleatoria sin reemplazamiento de las instancias del conjunto de entrenamiento, reduciendo así la varianza de las predicciones [68].

El paquete de R `{randomForest}` implementa el algoritmo RF [63].

3.2.3. k-vecinos más cercanos (kNN)

El algoritmo de los k-vecinos consiste en asignar a cada dato sin clasificar la clase más común entre sus k-vecinos más cercanos [69]. Aunque la distancia euclídea es la más usual, se pueden utilizar otras distancias para el cálculo de los k-vecinos más cercanos [70] y es recomendable normalizar los atributos para que todos tengan el mismo peso. Además, contando con un conjunto de entrenamiento se puede hallar el parámetro k óptimo para el problema (*tuning*).

En este trabajo se utilizará la implementación de kNN realizada en `{caret}` [71].

3.2.4. Medidas de evaluación

Tras realizar las predicciones con el algoritmo, sus resultados se suelen presentar en una matriz de confusión del siguiente modo:

		Predicción	
		Positivo	Negativo
Clase real	Positivo	Verdaderos positivos (TP)	Falsos negativos (FN)
	Negativo	Falsos positivos (FP)	Verdaderos negativos (TN)

Partiendo de la matriz de confusión, para evaluar la efectividad de un algoritmo de clasificación se pueden utilizar varios indicadores que miden diferentes aspectos.

CAPÍTULO 3. MACHINE LEARNING APLICADO A TRANSCRIPTÓMICA

34

- La precisión (*accuracy* en inglés, no confundir con el término *precision* que indica el valor predictivo positivo: $TP/(TP + FP)$) mide la proporción de predicciones correctas entre el número total de predicciones:

$$\text{Precisión} = \frac{TP + TN}{TP + FP + TN + FN}$$

- La sensibilidad mide la proporción de positivos reales que han sido correctamente identificados como positivos.

$$\text{Sensibilidad} = \frac{TP}{TP + FN}$$

- La especificidad mide la proporción de negativos reales que han sido correctamente identificados como negativos.

$$\text{Especificidad} = \frac{TN}{TN + FP}$$

- El F1-Score es la medida de evaluación más adecuada para problemas no equilibrados. Ante este tipo de problemas, muchos clasificadores suelen estar sesgados hacia la clase mayoritaria. F1-Score busca un equilibrio entre valor predictivo positivo ($TP/(TP + FP)$) y sensibilidad ($TP/(TP + FN)$).

$$\text{F1-score} = \frac{2 * TP}{2 * TP + FP + FN}$$

Estas medidas inicialmente definidas para problemas biclase se pueden extender con facilidad a problemas multiclase [72].

4. Detección de biomarcadores en cáncer de hígado y colon-recto

4.1. Objetivos

Hay dos objetivos principales:

1. Describir las características clínicas de los pacientes de cáncer de hígado y colon-recto.
2. Predecir en base a unos pocos genes si una persona padece o no cáncer de hígado o cáncer de colon-recto. En clasificación multiclasé se predice además el diagnóstico primario de cáncer.

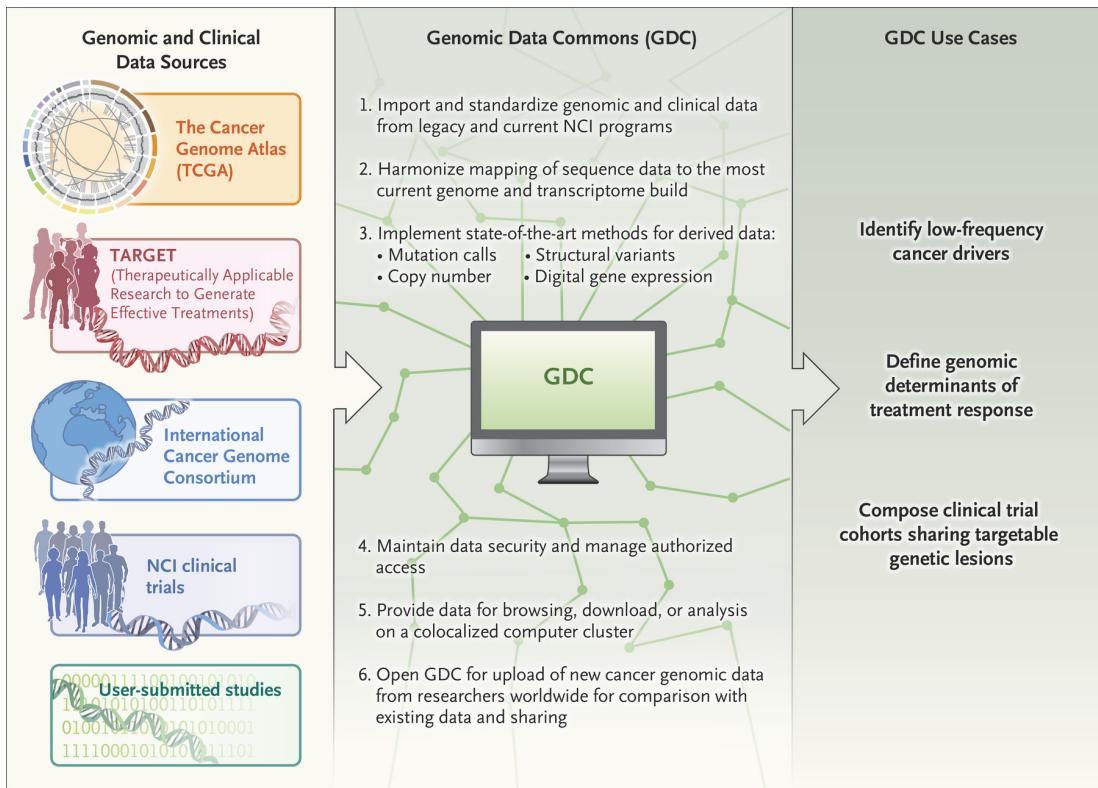
4.2. Fuente de datos

La fuente de los datos es GDC (*Genomic Data Commons*) Portal, una plataforma web sobre cáncer del Instituto Nacional del Cáncer de Estados Unidos (*National Cancer Institute*) [73, 74]. GDC Portal fue desarrollado por el Instituto Nacional del Cáncer de Estados Unidos, la Universidad de Chicago, el Instituto de Ontario para la Investigación del Cáncer y la empresa *Leidos Biomedical Research*. Su principal fortaleza reside en la integración y armonización de diversas fuentes heterogéneas, creando así un sistema de información amplio y robusto [75].

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

36

Figura 5. Diagrama de funcionalidad y utilidad de GDC. Extraído de Grossman et al. [75].



A día 22 de Junio, GDC Portal contenía información sobre unos 84.000 casos, 23.000 genes y más de 3 millones de mutaciones de genes [73]. Algunos de estos datos son abiertos mientras que para otros es necesario solicitar acceso. La información de la que dispone es muy variada y se puede distinguir en tres grandes categorías:

- Información clínica como la edad del sujeto, su sexo o el estadio del cáncer del que ha sido diagnosticado.
- Información genética y transcriptómica proveniente de diversos proyectos de investigación.
- Imágenes de tejidos tumorales y sanos.

Para el presente trabajo se han descargado de GDC Portal todos los datos que cumplen las siguientes condiciones:

- Son datos transcriptómicos del programa Cancer Genoma Atlas (TCGA), dirigido por dos organismos estadounidenses: el Instituto Nacional del Cáncer (NCI) y el Instituto Nacional para la Investigación del Genoma Humano (NHGRI) [76].
- Contienen información sobre tumores o tejidos sanos de cáncer de hígado o colon-recto. Se han excluido metástasis y tumores recurrentes.
- El tipo de estrategia experimental es RNA-Seq y el tipo de flujo de trabajo es HTSeq - Counts. Esta información es de acceso abierto ya que no permite la identificación de un individuo.

Para cáncer de hígado se han descargado datos sobre 462 tejidos de los cuales 404 son cancerosos (87,4 %) y 58 son sanos (12,6 %). Para cáncer de colon-recto se han descargado datos sobre 695 tejidos: 644 con cáncer (92,7 %) y 51 sanos (7,3 %). Es relevante destacar que las muestras de tejido sano se corresponden a tejido adyacente al tumor no afectado por el cáncer que es extraído del mismo paciente. Como no a todos los pacientes con cáncer se les ha secuenciado también tejido sano, existe un gran desequilibrio entre el número de muestras de las dos principales clases del problema.

4.3. Características clínicas de los tumores

A continuación se describe la información clínica de aquellas personas diagnosticadas con cáncer, descargada de la plataforma GDC Portal [73].

4.3.1. Características clínicas para cáncer de hígado

En la Tabla 9 se muestra la distribución de casos de cáncer de hígado según algunas variables clínicas de interés. Los casos se recogieron entre los años 1995 y 2013, con el 67,6 % de los casos recogidos entre los años 2010 y 2013. La mayoría de los casos son hombres (65,3 %), están diagnosticados en estadios iniciales (70,3 % en estadios I y II) y son carcinomas hepatocelulares (89,6 %). La edad media de diagnóstico es de 60,1 años (mediana: 61,7 años) con un rango de edad que comprende de los 16 a los 87 años. Más de la mitad de los casos son caucásicos (52,5 %), que es la raza más común seguida por asiáticos (39,9 %) y afroamericanos (4,7 %). Aproximadamente dos de cada tres personas estaban vivas en el momento del último contacto realizado (63,6 %).

**CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE
38 HÍGADO Y COLON-RECTO**

Tabla 9. Características clínicas de los casos de cáncer de hígado. Distribución de casos y porcentaje según sexo, grupo de edad, raza, diagnóstico primario, estadio y estado vital.

	Número de casos	Porcentaje
Total	404	100 %
Sexo		
Hombre	264	65,3 %
Mujer	140	34,7 %
Grupo de edad		
≤ 40 años	34	8,4 %
41-50 años	40	9,9 %
51-60 años	106	26,2 %
61-70 años	127	31,4 %
71-80 años	77	19,1 %
> 80 años	16	4,0 %
Desconocido	4	1,0 %
Raza		
Caucásico	212	52,5 %
Asiático	161	39,9 %
Afroamericano	19	4,7 %
Desconocido	10	2,5 %
Indio americano o nativo de Alaska	2	0,5 %
Diagnóstico primario		
Carcinoma hepatocelular	362	89,6 %
Colangiocarcinoma	33	8,2 %
Otros	8	2,0 %
Estadio		
Estadio I	189	46,8 %
Estadio II	95	23,5 %
Estadio III	82	20,3 %
Estadio IV	7	1,7 %
Desconocido	31	7,7 %
Estado vital		
Vivo	257	63,6 %
Fallecido	146	36,1 %
Desconocido	1	0,2 %

En la Tabla 10 se muestran tablas de contingencia del estado vital según sexo, grupo de edad y estadio. Se han realizado pruebas de chi cuadrado (χ^2) [77] para evaluar la independencia o no del estado vital con respecto a las distintas variables, aplicando la corrección de Yates [78] cuando fue necesario.

El código completo del análisis se muestra en el fichero *analisis_higado/01_analisis_datos_clinicos.R* del repositorio de GitHub asociado al trabajo [79].

Tabla 10. Características clínicas de los casos de cáncer de hígado. Distribución de estado vital según sexo, grupo de edad, estadio y diagnóstico primario.

	Vivos	Fallecidos	p-valor
Número de tumores	257	146	
Sexo			0,033
Hombre	178 (67,7 %)	85 (32,3 %)	
Mujer	79 (83,2 %)	16 (16,8 %)	
Grupo de edad			0,018
\leq 40 años	24 (70,6 %)	10 (29,4 %)	
41-50 años	27 (67,5 %)	13 (32,5 %)	
51-60 años	68 (64,2 %)	38 (35,8 %)	
61-70 años	89 (70,6 %)	37 (29,4 %)	
71-80 años	42 (54,5 %)	35 (45,5 %)	
> 80 años	5 (31,3 %)	11 (68,8 %)	
Estadio			< 0,001
Estadio I	138 (73,4 %)	50 (26,6 %)	
Estadio II	64 (67,4 %)	31 (32,6 %)	
Estadio III	39 (47,6 %)	43 (52,4 %)	
Estadio IV	3 (42,9 %)	4 (57,1 %)	
Diagnóstico primario			0,126
Carcinoma hepatocelular	233 (64,4 %)	129 (35,6 %)	
Colangiocarcinoma	17 (51,5 %)	16 (48,5 %)	
Otros	7 (87,5 %)	1 (12,5 %)	

Para las variables con datos faltantes se ha realizado un análisis de casos completos. La mortalidad entre los casos es el doble en hombres (32,3 %) que en mujeres

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

(16,8 %), y aumenta conforme aumenta la edad, desde 29,4 % en menores de 41 años hasta 68,8 % en mayores de 80 años. El estadio es uno de los principales factores pronósticos del cáncer, algo que se refleja en la gran diferencia existente en la mortalidad entre estadios. En los estadios iniciales (I-II) la mortalidad está cerca al 30 % y en los más avanzados (III-IV) más cerca del 55 %. Se detecta una dependencia entre el estado vital y todas las variables consideradas excepto el diagnóstico primario, con p-valores < 0,05.

4.3.2. Características clínicas para cáncer de colon-recto

620 tumores de colon-recto de los 644 considerados tienen disponible información clínica. En la Tabla 11 se muestra la distribución de casos de cáncer de colon-recto según algunas variables de interés. Los casos se recogieron entre los años 1998 y 2013, con la mayoría de los casos recogidos entre los años 2007 y 2011. La proporción entre hombres y mujeres es similar (53,1 % y 46,9 % respectivamente) y los estadios más comunes son los intermedios (estadio II: 36,3 % y estadio III: 28,7 %). La edad media de diagnóstico es de 66,7 años (mediana: 68,1 años), con un rango de edad de entre 31 y 90 años. Aproximadamente la mitad de los pacientes son caucásicos (47,3 %), que es la raza más común seguida por afroamericanos (10,5 %) y asiáticos (2,1 %). Se desconoce la raza del 40,0 % de las personas. El principal diagnóstico primario es el adenocarcinoma (83,4 %) seguido por el adenocarcinoma mucinoso (12,7 %). Aproximadamente cuatro de cada cinco personas estaban vivas en el momento del último contacto realizado (79,2 %).

Tabla 11. Características clínicas de los casos de cáncer de colon-recto. Distribución de casos y porcentaje según sexo, grupo de edad, estadio, raza, diagnóstico primario y estado vital.

	Número de casos (Porcentaje)
Total	620 (100 %)
Sexo	
Hombre	329 (53,1 %)
Mujer	291 (46,9 %)
Grupo de edad	
≤ 40 años	16 (2,6 %)
41-50 años	59 (9,5 %)
51-60 años	101 (16,3 %)
61-70 años	173 (27,9 %)
71-80 años	171 (27,6 %)
> 80 años	98 (15,8 %)
Desconocido	2 (0,3 %)
Raza	
Caucásico	293 (47,3 %)
Asiático	13 (2,1 %)
Afroamericano	65 (10,5 %)
Desconocido	248 (40,0 %)
Indio americano o nativo de Alaska	1 (0,2 %)
Diagnóstico primario	
Adenocarcinoma	517 (83,4 %)
Adenocarcinoma mucinoso	79 (12,7 %)
Otros	24 (3,9 %)
Estadio	
Estadio I	105 (16,9 %)
Estadio II	225 (36,3 %)
Estadio III	178 (28,7 %)
Estadio IV	89 (14,4 %)
Desconocido	23 (3,7 %)
Estado vital	
Vivo	491 (79,2 %)
Muerto	129 (20,8 %)

En la Tabla 12 se muestran tablas de contingencia del estado vital según sexo, gru-

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

po de edad y estadio. Se han realizado pruebas de chi cuadrado (χ^2) [77] para evaluar la independencia o no del estado vital con respecto a las distintas variables, aplicando la corrección de Yates [78] cuando fue necesario. El código completo del análisis se muestra en el fichero *análisis_cr/01_análisis_datos_clínicos.R* del repositorio de GitHub asociado al trabajo [79].

Tabla 12. Características clínicas de los casos de cáncer de colon-recto. Distribución de estado vital según sexo, grupo de edad, estadio y diagnóstico primario.

	Vivos	Fallecidos	p-valor
Número de tumores	491	129	
Sexo			0,993
Hombre	260 (79,0 %)	69 (21,0 %)	
Mujer	231 (79,4 %)	60 (20,6 %)	
Grupo de edad			< 0,001
≤ 40 años	14 (87,5 %)	2 (12,5 %)	
41-50 años	51 (86,4 %)	8 (13,6 %)	
51-60 años	85 (84,2 %)	16 (15,8 %)	
61-70 años	150 (86,7 %)	23 (13,3 %)	
71-80 años	120 (70,2 %)	51 (29,8 %)	
> 80 años	70 (71,4 %)	28 (28,6 %)	
Estadio			< 0,001
Estadio I	98 (93,3 %)	7 (6,7 %)	
Estadio II	192 (85,3 %)	33 (14,7 %)	
Estadio III	139 (78,1 %)	39 (21,9 %)	
Estadio IV	48 (53,9 %)	41 (46,1 %)	
Diagnóstico primario			0,25
Adenocarcinoma	409 (79,1 %)	108 (20,9 %)	
Adenocarcinoma mucinoso	60 (75,9 %)	19 (24,1 %)	
Otros	22 (91,7 %)	2 (8,3 %)	

Para las variables con datos faltantes se ha realizado un análisis de casos completos (exclusión del análisis de los casos con tenían datos faltantes). La mortalidad es muy similar en hombres y mujeres, sin existir diferencias significativas (p-valor: 0,993). Se detecta una dependencia entre el estado vital con las variables de grupo de edad y estadio con p-valores < 0,001. En mayores de 70 años la mortalidad

está cerca del 30 %, el doble de la mortalidad existente en otros grupos de edad. Hay grandes diferencias de mortalidad en función del estadio diagnosticado, que pasa del 6,7 % en el estadio I al 46,1 % en el estadio IV. La mortalidad es baja (8,3 %) cuando el diagnóstico primario no es adenocarcinoma ni adenocarcinoma mucinoso.

4.4. Metodología

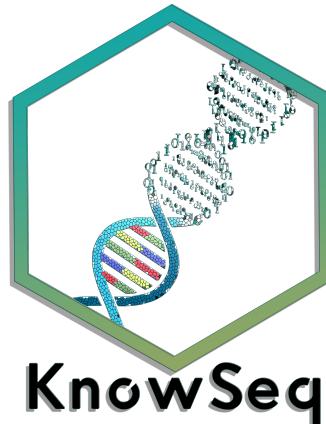
4.4.1. Herramientas para el análisis

Para el análisis se ha utilizado el software estadístico R (v.4.0.1) [80] y *{KnowSeq}* (v.1.2.0) [58], paquete de R que ha sido desarrollado por los tutores del presente trabajo y en el que el autor ha contribuido con algunas mejoras y actualizaciones (<https://github.com/CasedUgr/KnowSeq/commits?author=danielredondo>):

- Corregir un fallo que hacía que el método de selección de características RF no funcionase correctamente ante problemas de clasificación multiclas.
- Mantener los parámetros óptimos de SVM (coste y gamma) y kNN (k) encontrados tras optimización de los parámetros (*tuning*) en conjunto de entrenamiento para aplicarlos directamente en el modelo para el conjunto de test.
- Pequeñas modificaciones en la documentación de funciones.

{KnowSeq} está además disponible en Docker con el comando *Docker run -it casedugr/knowseq* y en *Bioconductor*, la plataforma de código abierto en R más relevante para el análisis de datos de genómica y transcriptómica [81].

Figura 6. Logo de *{KnowSeq}*.



Todo el código de los análisis está disponible en las carpetas `analysis_higado` y `analysis_cr` del repositorio de GitHub asociado al trabajo [79]. Además, en el fichero `session_info.txt` se muestran todos los paquetes de R utilizados y sus versiones como resultado de ejecutar `devtools::session_info()`.

Además, para la visualización de datos y resultados del presente trabajo se utilizan técnicas de visualización de datos para una mejor comprensión de los resultados empleando diagramas de Sankey, diagramas de cajas, y mapas de calor.

4.4.2. Detección de biomarcadores

Preprocesamiento

Tras descargar los ficheros de GDC Portal y descomprimirlos con R (ficheros `analysis_higado/02_descompresion` y `analysis_cr/02_descompresion` en GitHub [79]), se preprocesa la información para trabajar con matrices y data.frames.

Extracción de genes diferencialmente expresados (DEG)

Se extraen los genes diferencialmente expresados (DEG, por sus siglas en inglés: *Differentially Expressed Genes*) utilizando un p-valor de $p = 0,001$ y controlando por el efecto batch usando el método SVA (*Surrogate Variable Analysis*) [82]. El efecto batch es un sesgo que se puede producir al procesar las muestras biológicas por tandas, y puede ocurrir por varias razones como las condiciones del laboratorio, diferencias entre personal, o incluso la hora del día a la que se procesan las muestras [83, 84]. Para la extracción de DEG se fija además un LFC (*log fold change*) de 1, que se puede entender como que un gen sólo se considera un DEG

si una de las dos expresiones de genes (ya sea de tumores o tejidos sanos) es al menos el doble que la otra. Finalmente, los resultados se extraen a una matriz de DEGs.

DUDA: ¿La extracción de genes relevantes (DEG) no debería hacerse sólo con datos de entrenamiento? He visto en la literatura que también se hace como en este trabajo (extracción de DEGs antes de partición train-test) <https://www.nature.com/articles/s41598-020-60595-1> pero no entiendo bien el por qué. ¿Cómo se puede justificar?

Partición entrenamiento-test

Se realiza una partición entrenamiento-test de manera aleatoria con equilibrio de clases: el 75 % de las muestras de cada clase pasan a formar parte del conjunto de entrenamiento y el 25 % restante del conjunto de test.

Genes más relevantes

Se utiliza el conjunto de entrenamiento para seleccionar los diez genes más relevantes mediante los tres métodos de selección de características detallados en el capítulo anterior: mRMR, RF y DA.

4.4.3. Validación cruzada en entrenamiento

Se realiza validación cruzada 5-fold para tres algoritmos de clasificación: SVM, RF y kNN. Los parámetros de los clasificadores se optimizan (*tuning*) usando los 10 genes más relevantes de cada método de selección de características (mRMR, RF y DA). Para SVM con kernel radial se realiza optimización de los dos parámetros utilizando una rejilla con los valores especificados en la Tabla 13.

Tabla 13. Valores de los parámetros que se utilizan para el tuning en rejilla de los parámetros óptimos del algoritmo SVM.

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

46

Coste	0	0,01	0,02	0,025	0,03	0,04	0,05	0,06
	0,07	0,08	0,09	0,1	0,25	0,5	0,75	0,9
Gamma	0,01	0,05	0,1	0,25	0,5			
	0,75	1	1,5	2	5			

Posteriormente, para cada algoritmo se usan como variables los k genes más relevantes encontrados cada uno de los tres métodos de selección de características con $k = 1, 2, \dots, 10$. Por tanto, para cada algoritmo de clasificación (SVM, RF, kNN) se crean 30 modelos distintos. Como el problema a resolver tiene un gran desequilibrio de clases, la medida de evaluación de los modelos escogida es el F1-Score medio de los 5 folds. El mejor modelo (aquel con mayor F1-Score) es posteriormente validado en el conjunto de test.

4.4.4. Validación en test

El mejor modelo encontrado en la fase previa para cada clasificador (SVM, RF, kNN) se utiliza para predecir el tipo de muestra del conjunto de test. Se comparan los resultados predichos con los reales, mostrando las matrices de confusión y los indicadores de evaluación F1-Score y precisión.

4.5. Resultados de clasificación biclase para cáncer de hígado

El código completo del análisis está disponible en el fichero *análisis_higado/03_análisis_biclase.R* del repositorio de GitHub asociado al trabajo [79].

4.5.1. Detección de biomarcadores

Para la clasificación biclase en cáncer de hígado se cuenta con 404 tumores y 58 muestras de tejido sano (Tabla 14).

Tabla 14. Distribución de tipos de muestra para el análisis de cáncer de hígado biclase.

4.5. RESULTADOS DE CLASIFICACIÓN BICLASE PARA CÁNCER DE HÍGADO

47

	Número de casos	Porcentaje
Tumores	404	87,4 %
Tejido sano	58	12,6 %
Total	462	100 %

De los 24.645 genes presentes en los datos se extraen 2.274 genes que presentan en su expresión diferencias significativas entre las muestras de tumores y las de tejido sano. En la Tabla 15 se muestra la partición entrenamiento - test realizada y en la Figura 7 se representa en un diagrama de Sankey.

Tabla 15. Distribución entrenamiento-test según tipo de muestra y proporción entre clases (tumores/tejido sano) para el análisis de hígado biclase.

	Total	Entrenamiento	Test
Tumores	404 (100 %)	303 (75,0 %)	101 (25,0 %)
Tejido sano	58 (100 %)	44 (75,9 %)	14 (24,1 %)
Proporción tumores/sanos	6,97	6,89	7,21

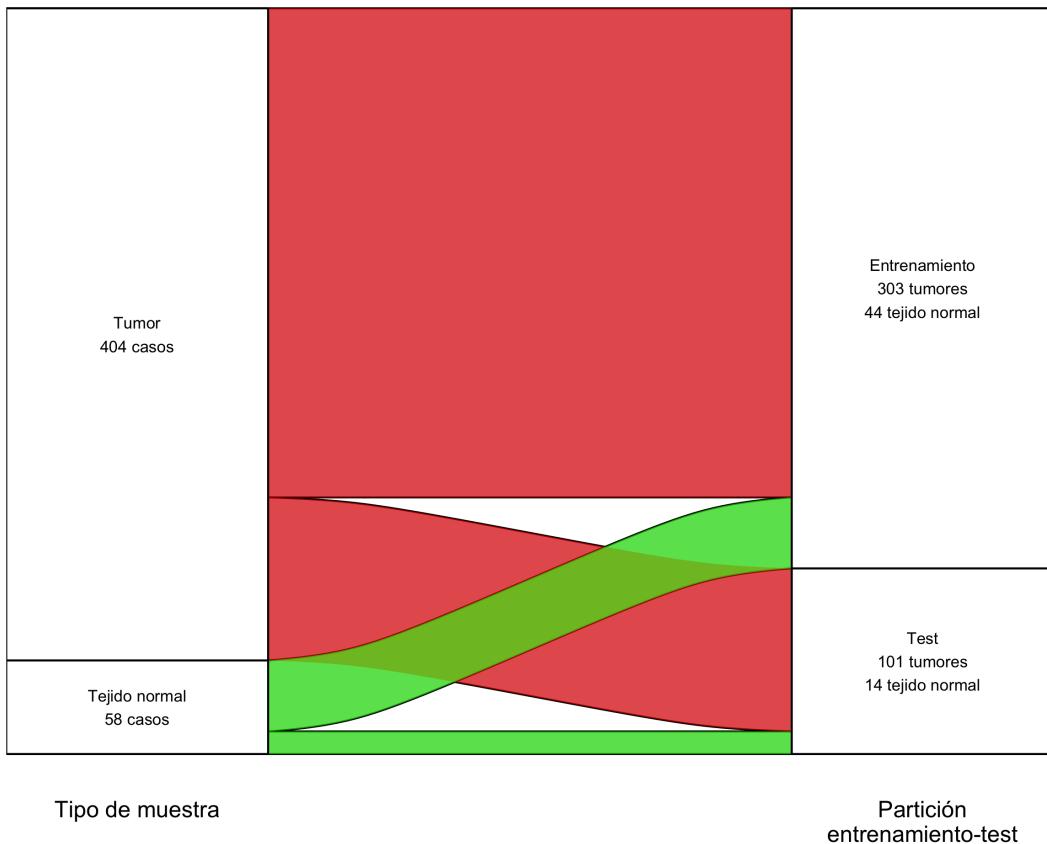
Figura 7. Diagrama de Sankey mostrando la partición entrenamiento-test realizada según tipo de muestra para el análisis de hígado biclase.

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

48

Partición en conjuntos de entrenamiento y test

Reparto 75% - 25% con balanceo de clases



A continuación se muestran los diez genes más relevantes encontrados por cada método de selección de características:

Tabla 16. Diez genes más relevantes según los distintos métodos de selección de características para el análisis de hígado biclase.

Ranking	mRMR	RF	DA
1	ANGPTL6	ANGPTL6	TERT
2	THY1	PTH1R	RSPO3
3	ADAMTS13	ADAMTS13	HOXA13
4	CELSR3	BMPER	SIX1
5	CCNE1	PRC1	TOP2A
6	CDH13	CLEC4G	GPC3
7	C14orf180	VIPR1	SSX1
8	GABRD	CLEC4M	BUB1B
9	AP000439.2	OIT3	RET
10	CEP152	GABRD	ESR1

Se observa que tres genes han sido valorados como relevantes para los algoritmos de mRMR y RF: ANGPTL6, ADAMTS13 y GABRD.

4.5.2. Validación cruzada en entrenamiento - SVM

En la Tabla 17 se muestran los parámetros óptimos de SVM obtenidos tras el tuning.

Tabla 17. Parámetros óptimos de SVM encontrados para los 10 genes más relevantes de cada método de selección de características.

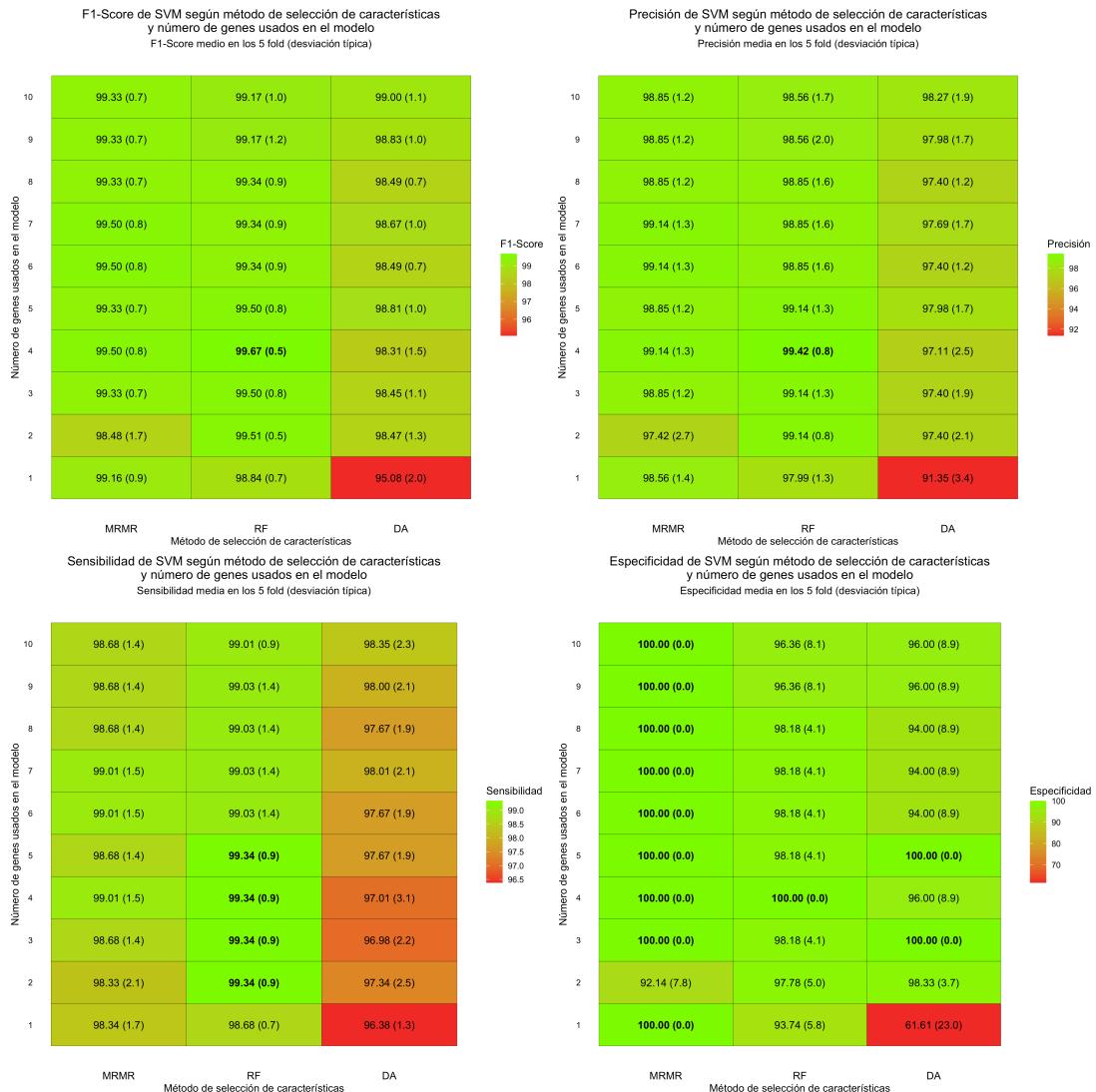
Parámetro	mRMR	RF	DA
Coste	0,05	0,75	0,1
Gamma	0,05	0,1	0,06

Para estos parámetros óptimos, en la Figura 8 se muestran las medidas de evaluación medias obtenidas en las 5 fold.

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

50

Figura 8. Mapa de calor con valores medios y desviación típica de los 5-fold de F1-Score, precisión, sensibilidad y especificidad de SVM según método de selección de características y número de genes usados.

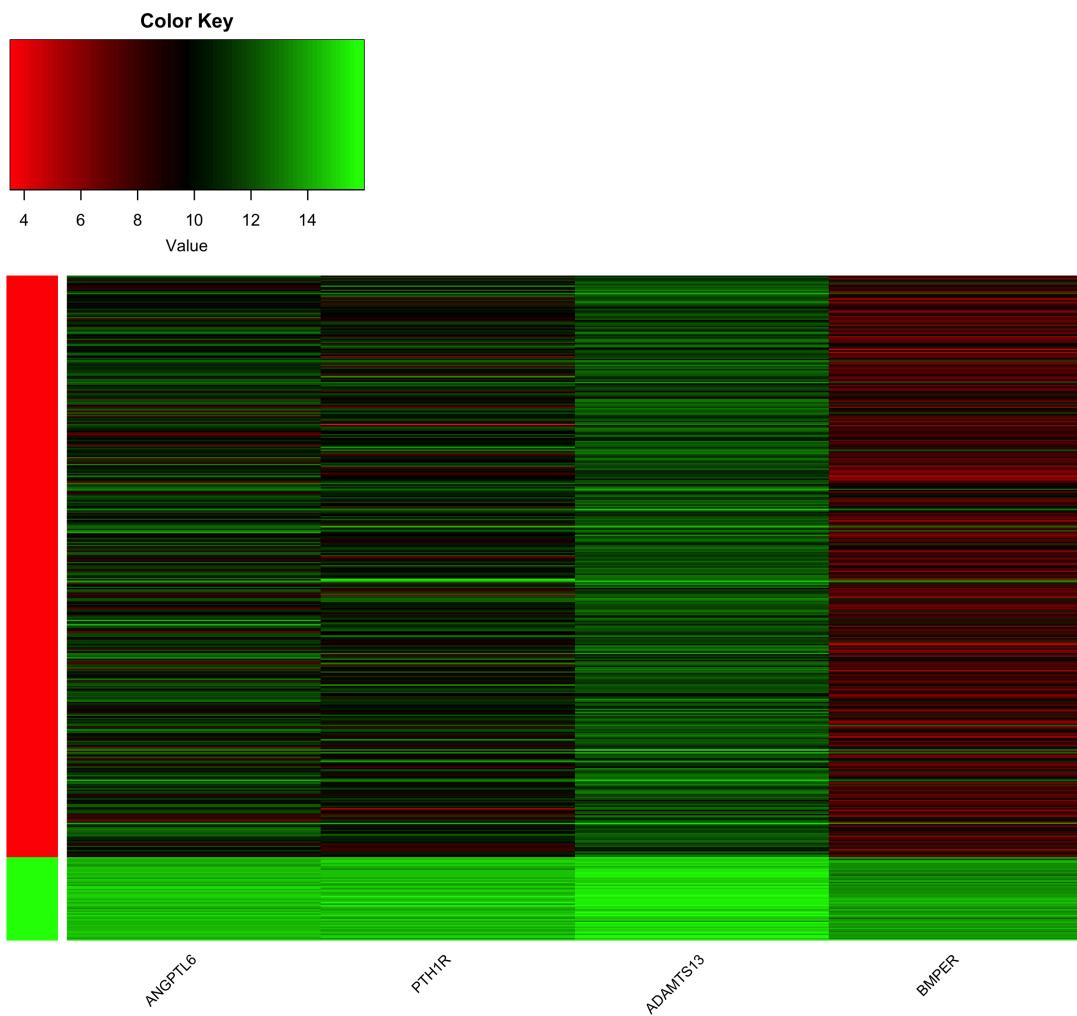


El mejor F1-Score para SVM se obtiene con 4 genes seleccionados con RF (99,67 %), configuración que obtiene también los mejores valores de precisión (99,42 %), sensibilidad (99,34 %) y especificidad (100 %). Los 4 genes seleccionados son: ANGPTL6, PTH1R, ADAMTS13 y BMPER. Se analiza la expresión de genes de estos 4 genes, utilizando para ello un mapa de calor (Figura 9) y diagramas de caja (Figura 10).

4.5. RESULTADOS DE CLASIFICACIÓN BICLASE PARA CÁNCER DE HÍGADO

51

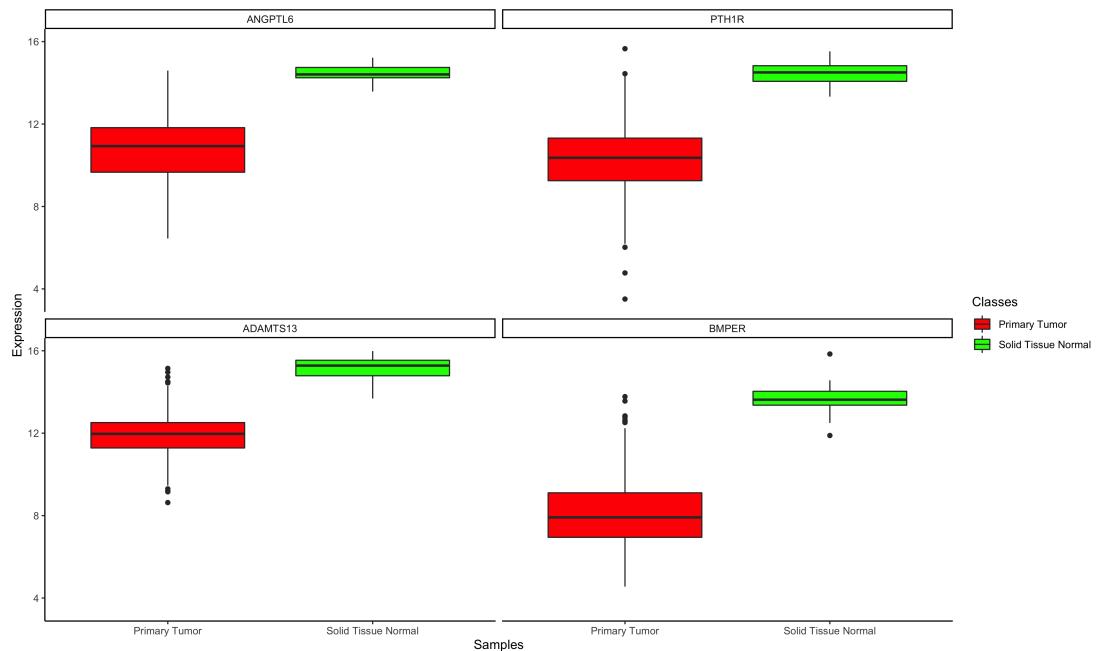
Figura 9. Mapa de calor de expresión de genes por tipo de muestra (columna izquierda: rojo = tumor, verde = tejido sano) en los 4 genes más relevantes encontrados en el mejor modelo de SVM con RF como método de selección de características.



En el mapa de calor se observa que en general los genes seleccionados están sobreexpresados en tejidos sanos, e inhibidos en tejidos tumorales.

**CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE
52 HÍGADO Y COLON-RECTO**

Figura 10. Diagrama de caja de expresión de genes por tipo de muestra en los 4 genes más relevantes encontrados en el mejor modelo de SVM con RF como método de selección de características.

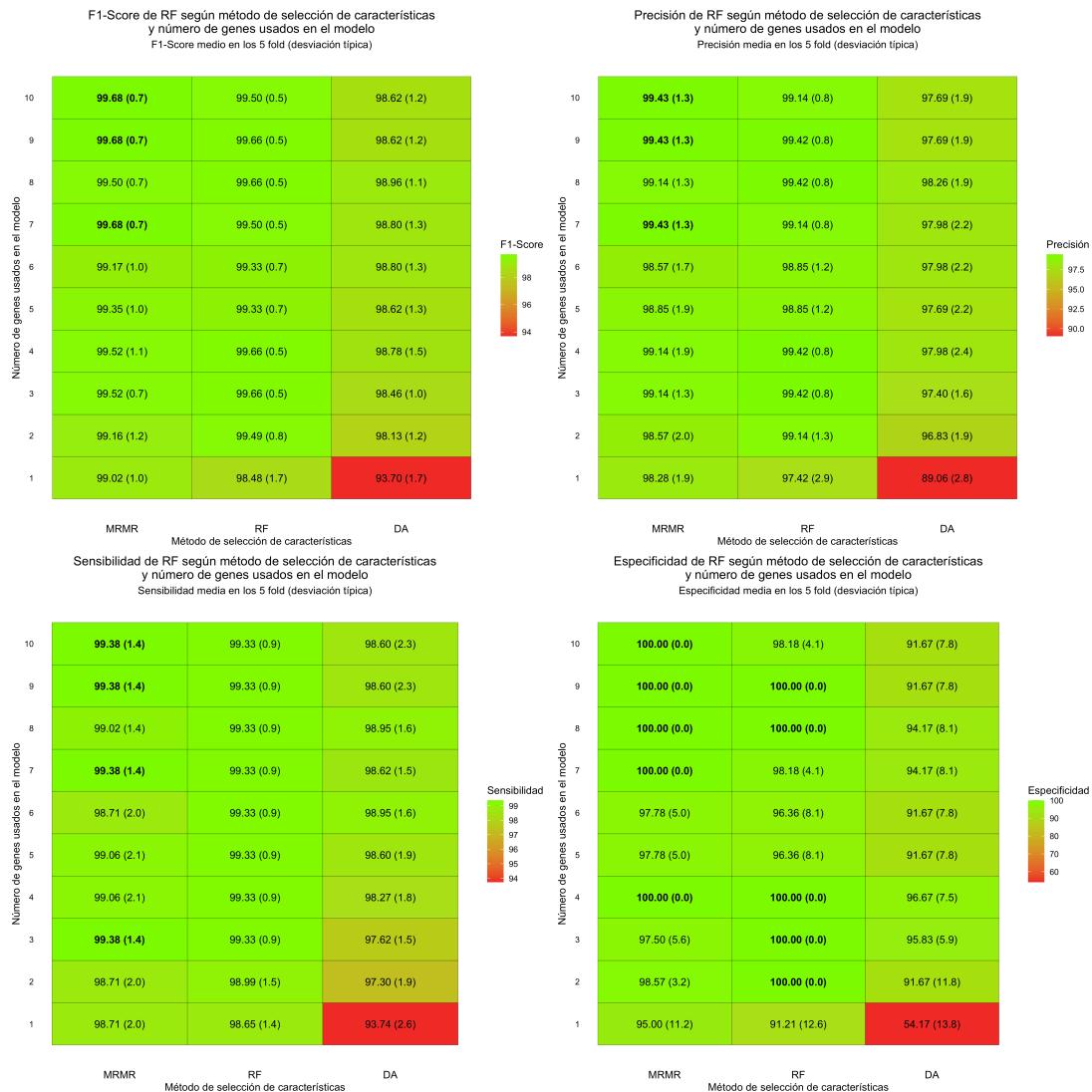


El diagrama de caja aporta cierta información adicional al mapa de calor, ya que en algunos genes se detectan outliers y se observa mejor el posible solapamiento entre la expresión de genes para tumores y tejidos sanos.

4.5.3. Validación cruzada en entrenamiento - RF

En la Figura 11 se muestran las medidas de evaluación medias obtenidas en las 5 fold.

Figura 11. Mapa de calor con valores medios y desviación típica de los 5-fold de F1-Score, precisión, sensibilidad y especificidad de RF según método de selección de características y número de genes usados.



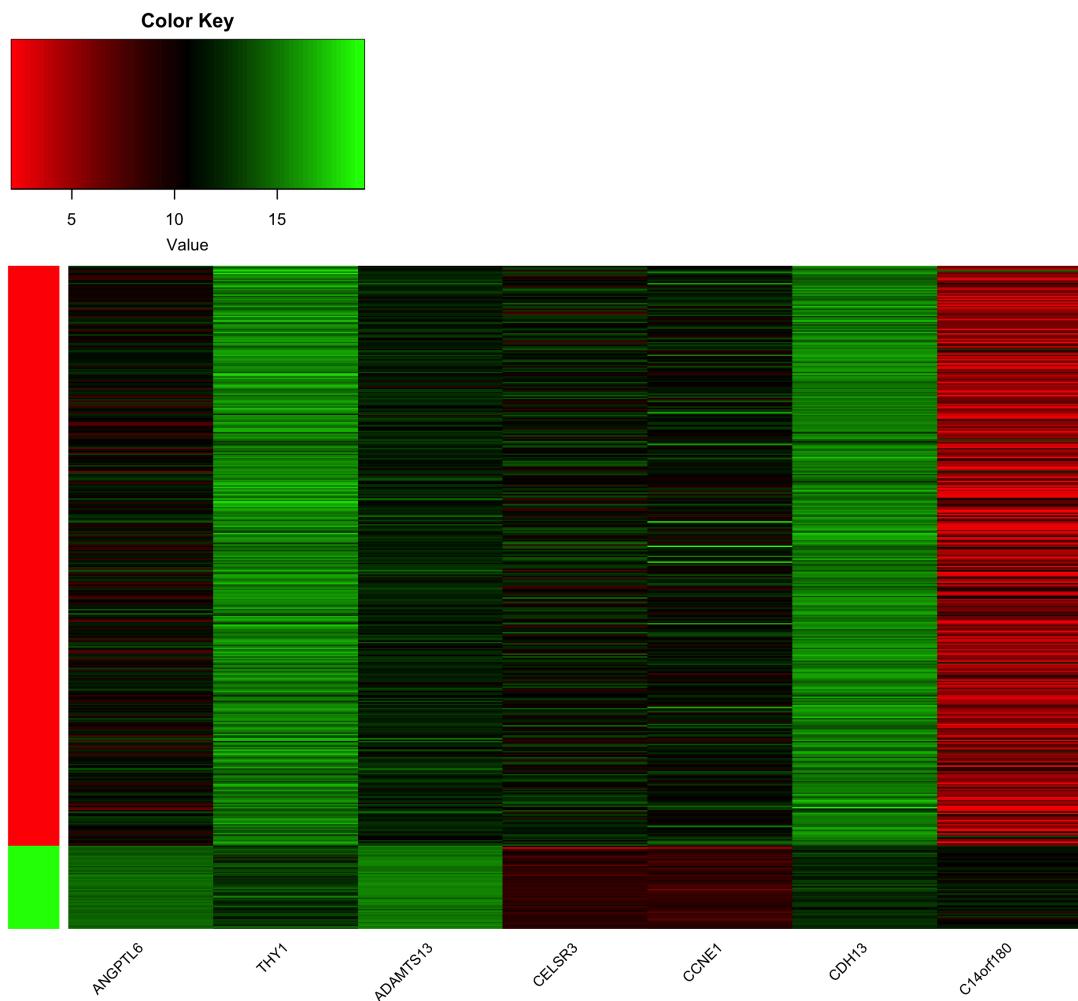
Para random forest el mejor F1-Score (99,68 %) se obtiene con el método mRMR considerando 7, 9 o 10 genes. Se selecciona aquel método que tiene menor número

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

54

de genes: mRMR con 7 genes. Los 7 genes seleccionados son: ANGPTL6, THY1, ADAMTS13, CELSR3, CCNE1, CDH13 y C14orf180. Se analiza su expresión por tipo de muestra, utilizando para ello un mapa de calor (Figura 12) y diagramas de caja (Figura 13).

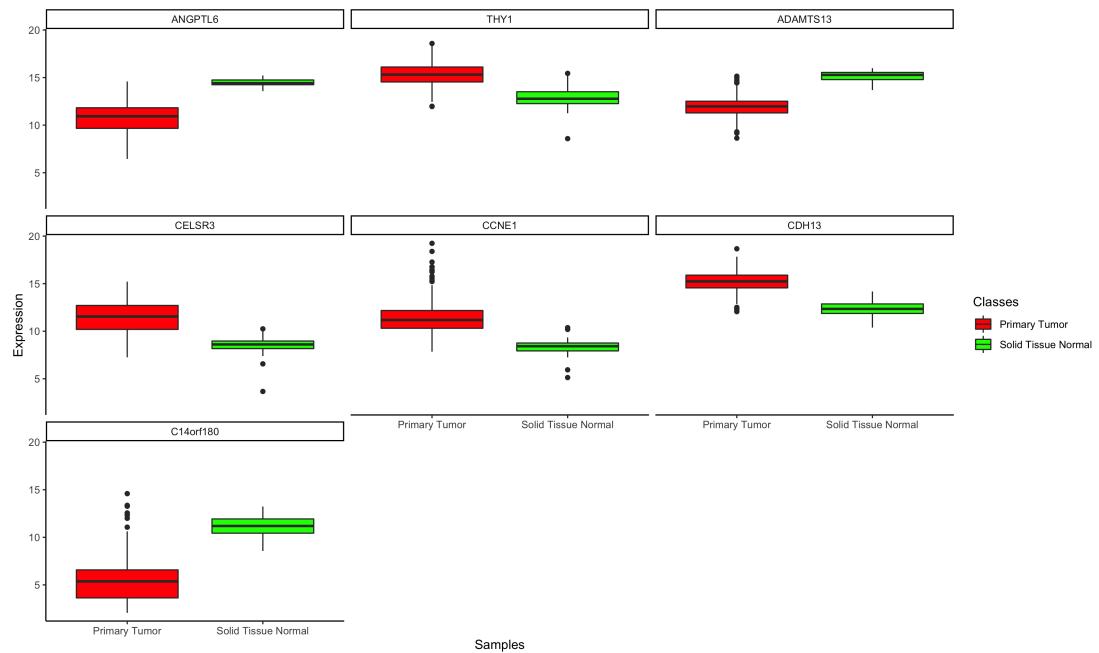
Figura 12. Mapa de calor de expresión de genes por tipo de muestra (columna izquierda: rojo = tumor, verde = tejido sano) en los 4 genes más relevantes encontrados en el mejor modelo de RF con RF como método de selección de características.



4.5. RESULTADOS DE CLASIFICACIÓN BICLASE PARA CÁNCER DE HÍGADO

55

Figura 13. Diagrama de caja de expresión de genes por tipo de muestra en los 4 genes más relevantes encontrados en el mejor modelo de RF con RF como método de selección de características.



4.5.4. Validación cruzada en entrenamiento - kNN

El número óptimo de vecinos para los 10 genes más relevantes es 7 para mRMR y DA y 5 para RF (Tabla 18).

Tabla 18. Número óptimo de vecinos encontrado para los 10 genes más relevantes de cada método de selección de características.

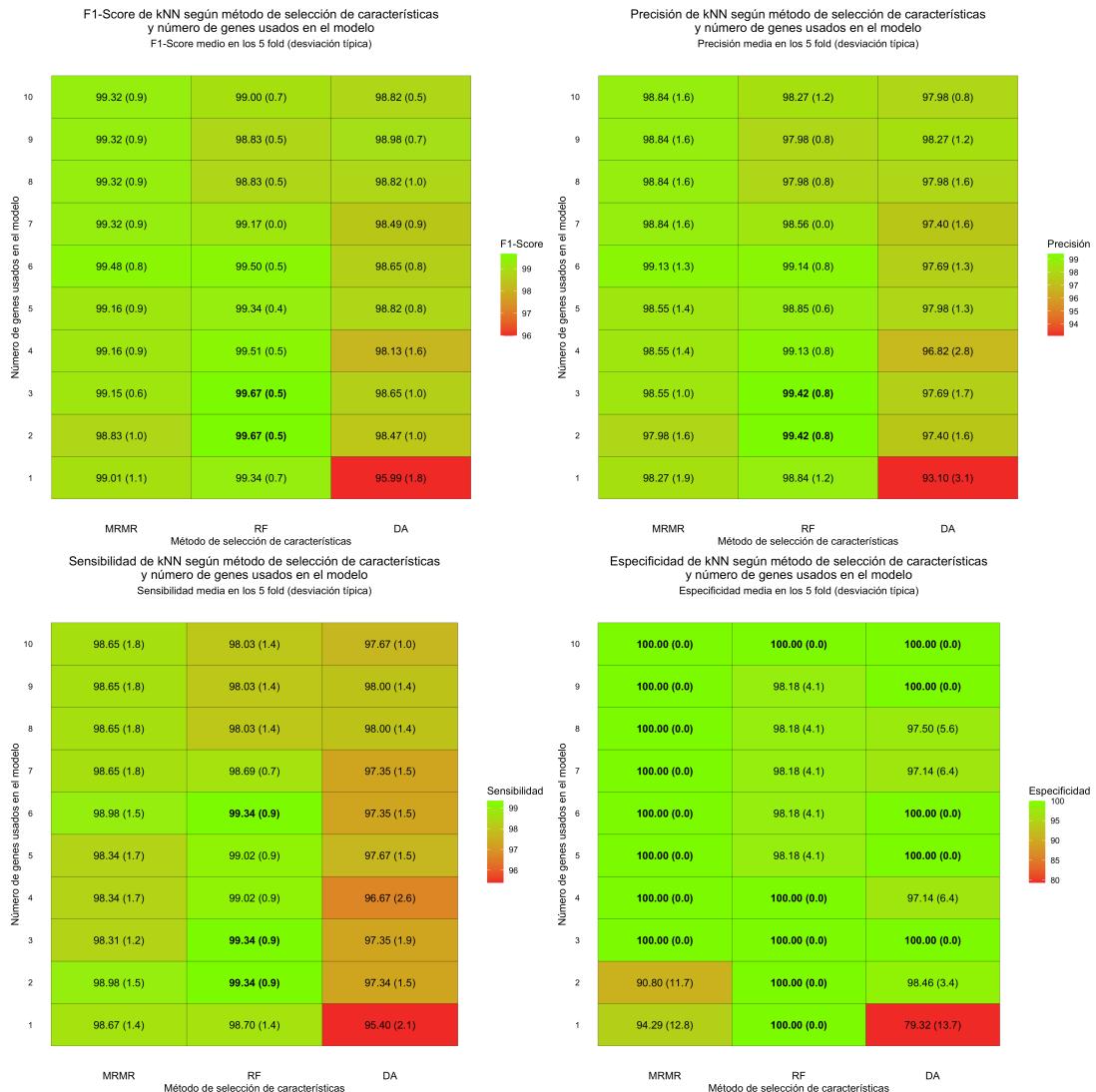
	mRMR	RF	DA
k	7	5	7

En la Figura 14 se muestran las medidas de evaluación medias obtenidas en las 5 fold.

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

56

Figura 14. Mapa de calor con valores medios y desviación típica de los 5-fold de F1-Score, precisión, sensibilidad y especificidad de kNN según método de selección de características y número de genes usados.

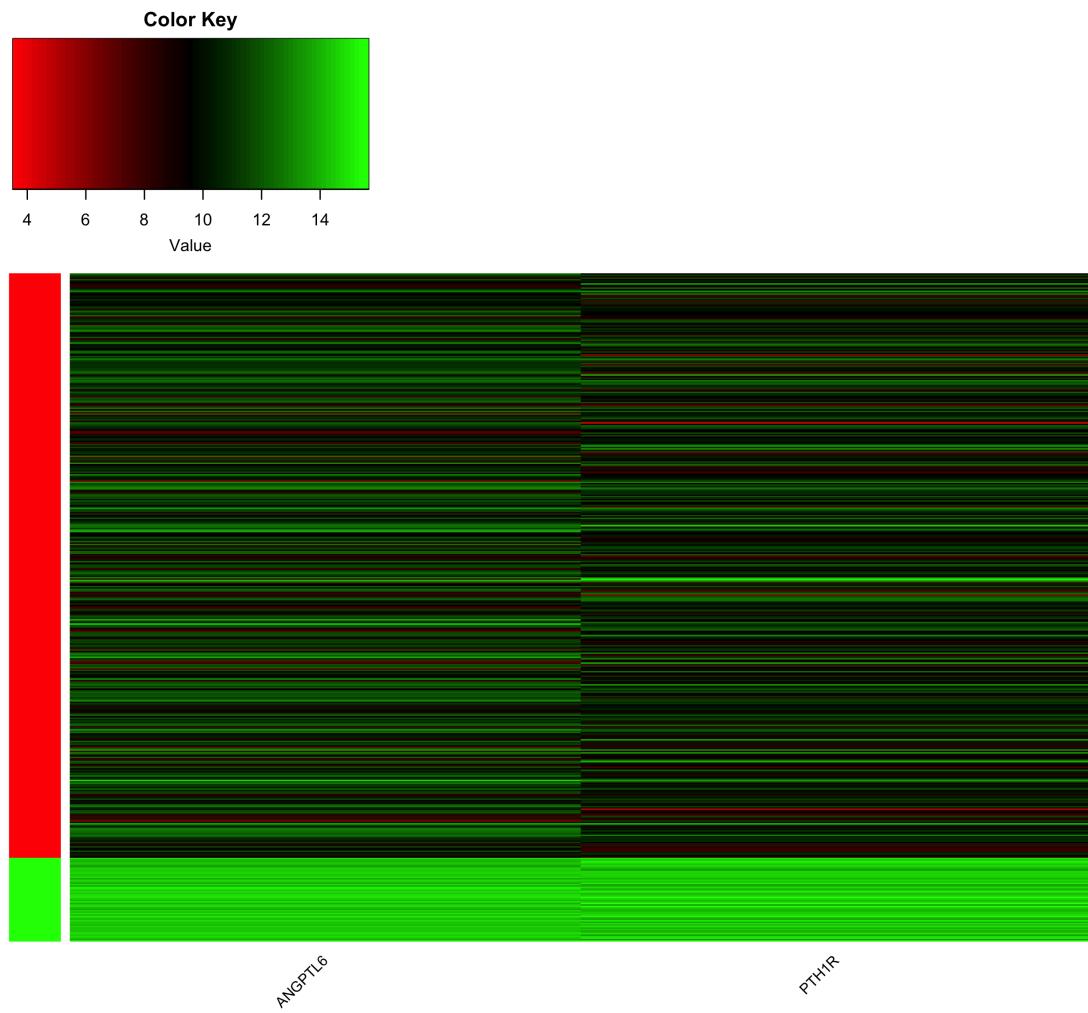


Para kNN, el mejor F1-Score (99,68 %) se obtiene con el método RF para 2 y 3 genes. Se selecciona aquel método que tiene menor número de genes: RF con 2 genes. Los 2 genes seleccionados son: ANGPTL6 y PTH1R. Se analiza su expresión de genes, utilizando para ello un mapa de calor (Figura 15) y diagramas de caja (Figura 16).

4.5. RESULTADOS DE CLASIFICACIÓN BICLASE PARA CÁNCER DE HÍGADO

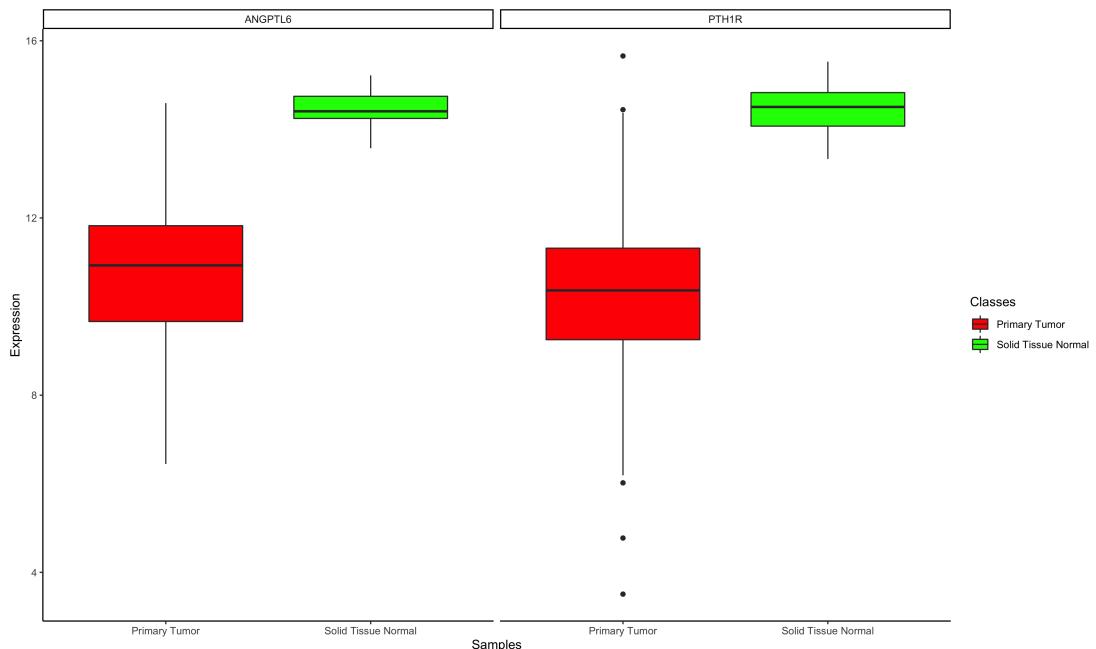
57

Figura 15. Mapa de calor de expresión de genes por tipo de muestra (columna izquierda: rojo = tumor, verde = tejido sano) en los 2 genes más relevantes encontrados en el mejor modelo de kNN con RF como método de selección de características.



**CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE
58 HÍGADO Y COLON-RECTO**

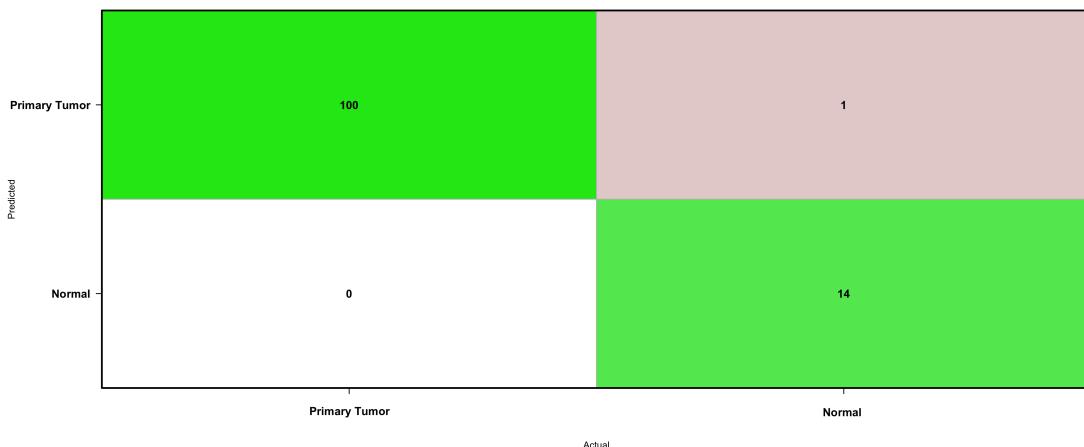
Figura 16. Diagrama de caja de expresión de genes por tipo de muestra en los 2 genes más relevantes encontrados en el mejor modelo de kNN con RF como método de selección de características.



4.5.5. Validación en test

Al validar en el conjunto de test el mejor modelo encontrado para SVM (4 genes con RF), RF (7 genes con mRMR) y kNN (2 genes con RF) se obtiene la misma clasificación: una predicción perfecta a falta de un falso positivo, esto es, una muestra de tejido normal que ha sido clasificada como tumor. El F1-Score en test es de 99,5 % con una precisión de 99,1 %.

Figura 17. Matriz de confusión de los mejores modelos encontrados de SVM, RF y kNN en el conjunto de test.



En la Tabla 19 se muestra un resumen de los mejores modelos obtenidos y su F1-Score y precisión en conjunto de entrenamiento y conjunto de test.

Tabla 19. Resumen de clasificación biclase para cáncer de hígado. Mejor modelo encontrado para SVM, RF y kNN con biomarcadores seleccionados, parámetros optimizados para cada algoritmo, F1-Score y precisión (Acc) en conjunto de entrenamiento y conjunto de test.

		Biomarcadores	Parámetros	F1 train	Acc train	F1 test	Acc test
SVM	4 genes RF		$c = 0.75$ $\gamma = 0.1$	99,67	99,42	99,5	99,13
RF	7 genes mRMR		—	99,68	99,43	99,5	99,13
kNN	2 genes RF		$k = 5$	99,67	99,42	99,5	99,13

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

4.6. Resultados de clasificación multiclase para cáncer de hígado

El código completo del análisis se muestra en el fichero *analisis_higado/04_analisis_multiclas.R* del repositorio de GitHub asociado al trabajo [79].

4.6.1. Detección de biomarcadores

Para la clasificación multiclase en cáncer de hígado se van a considerar los dos principales tipos de diagnósticos primarios presentes en los datos clínicos de cáncer de hígado: carcinoma hepatocelular y colangiocarcinoma. Se excluyen los tumores con otros diagnósticos primarios (8 casos). En la Tabla 20 se muestra la distribución de muestras.

Tabla 20. Distribución de tipos de muestra para el análisis de cáncer de hígado multiclase.

	Número de casos	Porcentaje
Carcinoma hepatocelular	363	80,0 %
Colangiocarcinoma	33	7,3 %
Tejido sano	58	12,8 %
Total	454	100 %

Se extraen 8.533 genes que presentan en su expresión diferencias significativas entre las muestras de tumores y las de tejido sano. En la Tabla 21 se muestra la partición entrenamiento - test realizada y en la Figura 18 se representa en un diagrama de Sankey.

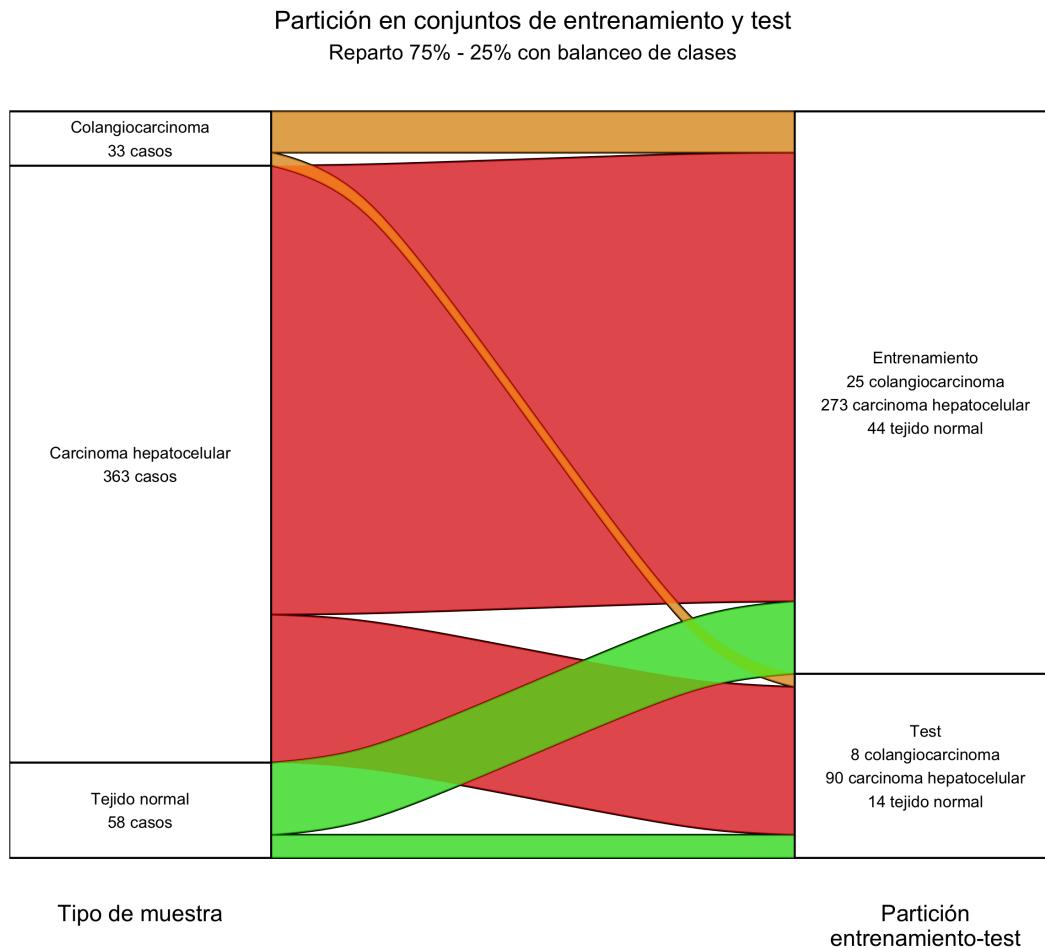
Tabla 21. Distribución entrenamiento-test según tipo de muestra y proporción entre clases para el análisis de hígado multiclase.

4.6. RESULTADOS DE CLASIFICACIÓN MULTICLASE PARA CÁNCER DE HÍGADO

61

	Total	Entrenamiento	Test
Carcinoma hepatocelular	363 (100 %)	273 (75,2 %)	90 (24,8 %)
Colangiocarcinoma	33 (100 %)	25 (75,8 %)	8 (24,2 %)
Tejido sano	58 (100 %)	44 (75,9 %)	14 (24,1 %)
Proporción carc. hepat./sanos	6,3	6,2	6,4
Proporción colang./sanos	0,6	0,6	0,6

Figura 18. Diagrama de Sankey mostrando la partición entrenamiento-test realizada según tipo de muestra para el análisis de hígado multiclase.



A continuación se muestran los diez genes más relevantes encontrados por cada método de selección de características:

**CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE
62 HÍGADO Y COLON-RECTO**

Tabla 22. Diez genes más relevantes según los distintos métodos de selección de características para el análisis de hígado multiclase.

Ranking	mRMR	RF	DA
1	ANGPTL6	ANGPTL6	WWTR1
2	FTLP3	GABRD	BIRC3
3	PLXDC1	CDH13	CDH1
4	RAB25	STAB2	ROS1
5	WDR66	BMPER	POLQ
6	AP2B1	ECM1	FGFR2
7	CDH13	ADAMTS13	KLF6
8	PTPN13	GDF2	CBFB
9	SLC31A1	CLEC4G	FGFR3
10	ADAMTS13	SPDL1	CLTCL1

Se observa que tres genes han sido valorados como relevantes para los algoritmos de mRMR y RF: ANGPTL6, CDH13 y ADAMTS13.

4.6.2. Validación cruzada en entrenamiento - SVM

En la Tabla 23 se muestran los parámetros óptimos de SVM obtenidos tras el tuning.

Tabla 23. Parámetros óptimos de SVM encontrados para los 10 genes más relevantes de cada método de selección de características.

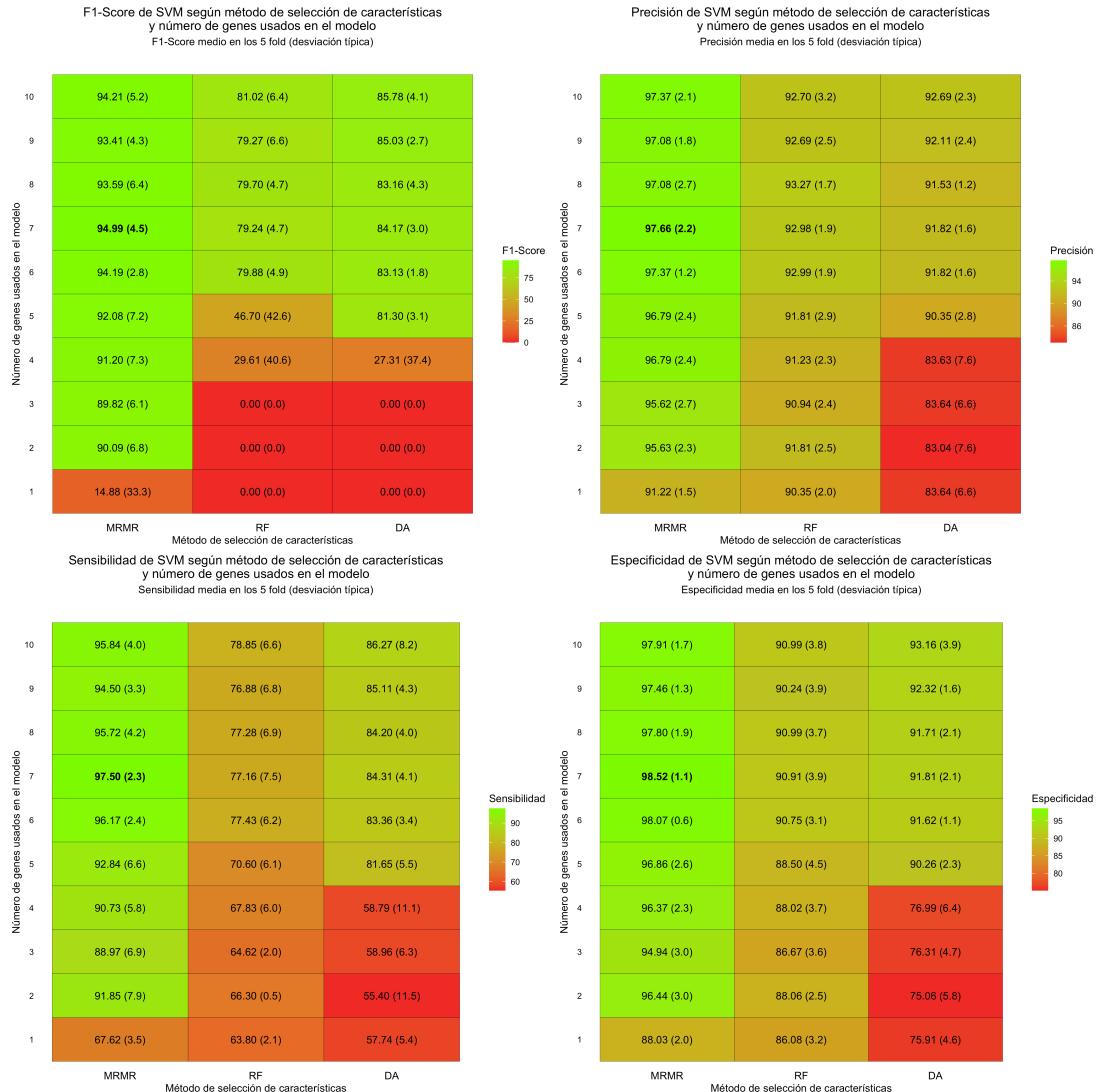
Parámetro	mRMR	RF	DA
Coste	1	5	2
Gamma	0,025	0,08	0,02

Para estos parámetros óptimos, en la Figura 19 se muestran las medidas de evaluación medias obtenidas en las 5 fold.

4.6. RESULTADOS DE CLASIFICACIÓN MULTICLASE PARA CÁNCER DE HÍGADO

63

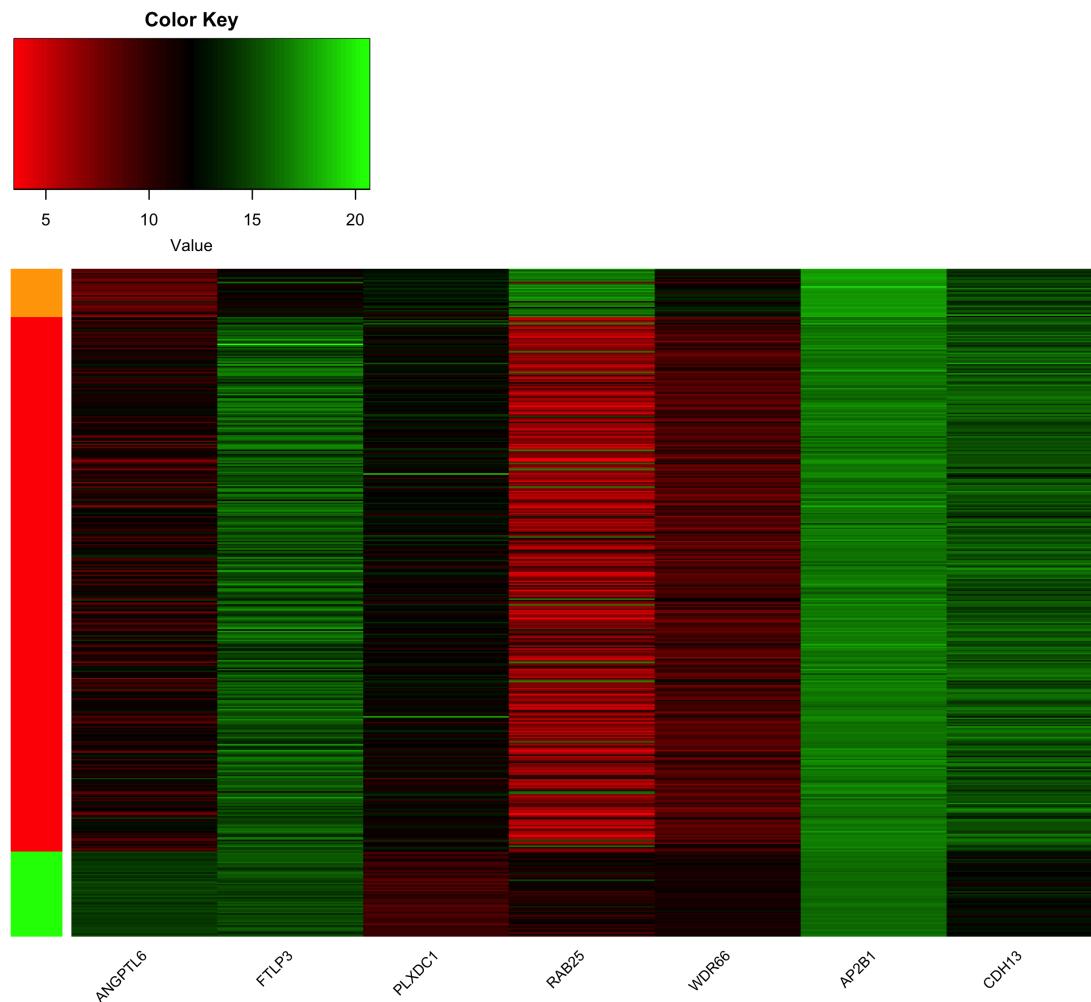
Figura 19. Mapa de calor con valores medios y desviación típica de los 5-fold de F1-Score, precisión, sensibilidad y especificidad de SVM según método de selección de características y número de genes usados.



El mejor F1-Score para SVM se obtiene con 7 genes seleccionados con mRMR (94,99 %), configuración que obtiene también los mejores valores de precisión (97,66 %), sensibilidad (97,50 %) y especificidad (98,52 %). Los 7 genes seleccionados son: ANGPTL6, FTLP3, PLXDC1, RAB25, WDR66, AP2B1 y CDH13. Se analiza la expresión de genes de estos 7 genes, utilizando para ello un mapa de calor (Figura 20) y diagramas de caja (Figura 21).

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

Figura 20. Mapa de calor de expresión de genes por tipo de muestra (columna izquierda: rojo = carcinoma hepatocelular, naranja = colangiocarcinoma, verde = tejido sano) en los 4 genes más relevantes encontrados en el mejor modelo de SVM con RF como método de selección de características.

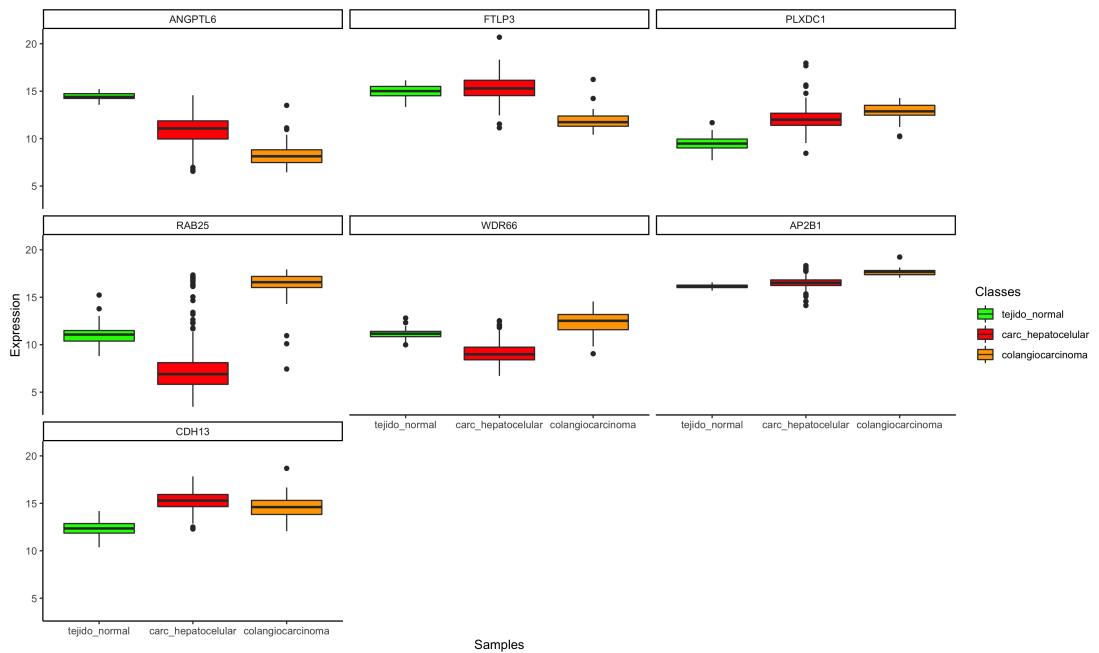


En el mapa de calor se observa que hay grandes diferencias entre la expresión de genes de los distintos tipos de tejidos.

4.6. RESULTADOS DE CLASIFICACIÓN MULTICLASE PARA CÁNCER DE HÍGADO

65

Figura 21. Diagrama de caja de expresión de genes por tipo de muestra en los 4 genes más relevantes encontrados en el mejor modelo de SVM con RF como método de selección de características.

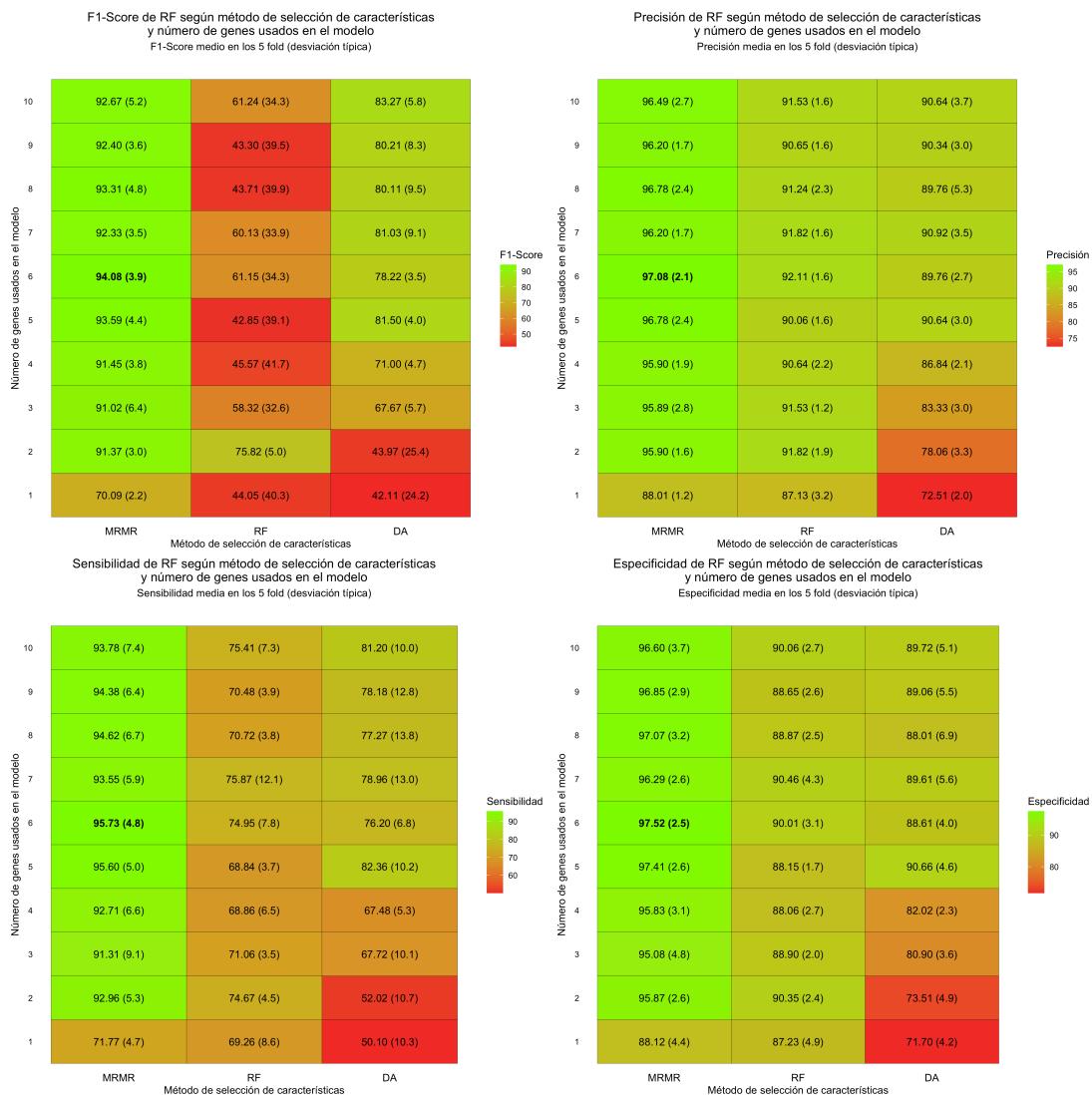


CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

4.6.3. Validación cruzada en entrenamiento - RF

En la Figura 22 se muestran las medidas de evaluación medias obtenidas en las 5 fold.

Figura 22. Mapa de calor con valores medios y desviación típica de los 5-fold de F1-Score, precisión, sensibilidad y especificidad de RF según método de selección de características y número de genes usados.



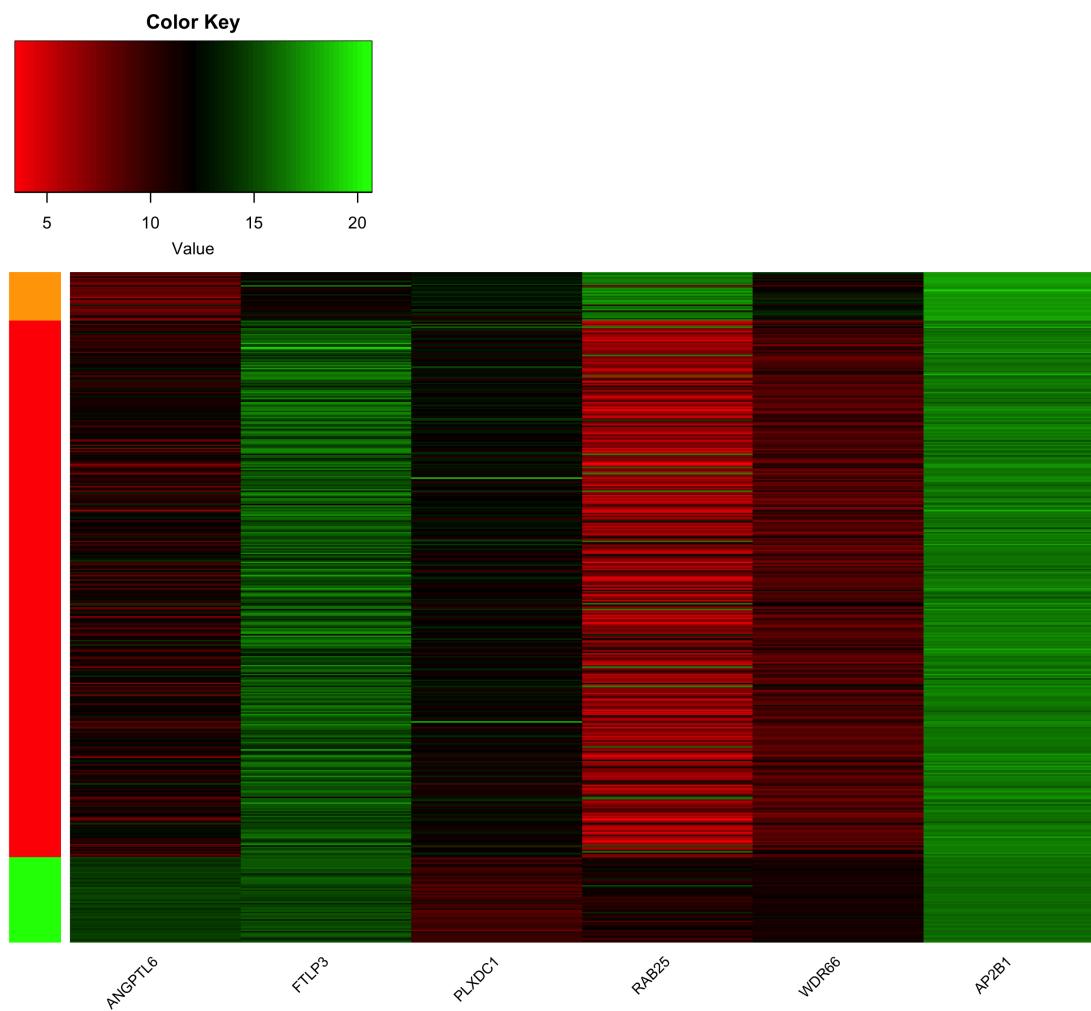
Para random forest, el mejor F1-Score (94,08 %) se obtiene con el método mRMR, considerando 6 genes. Los 6 genes seleccionados son: ANGPTL6, FTLP3, PLXDC1,

4.6. RESULTADOS DE CLASIFICACIÓN MULTICLASE PARA CÁNCER DE HÍGADO

67

RAB25, WDR66 y AP2B1. Se analiza su expresión de genes, utilizando para ello un mapa de calor (Figura 23) y diagramas de caja (Figura 24).

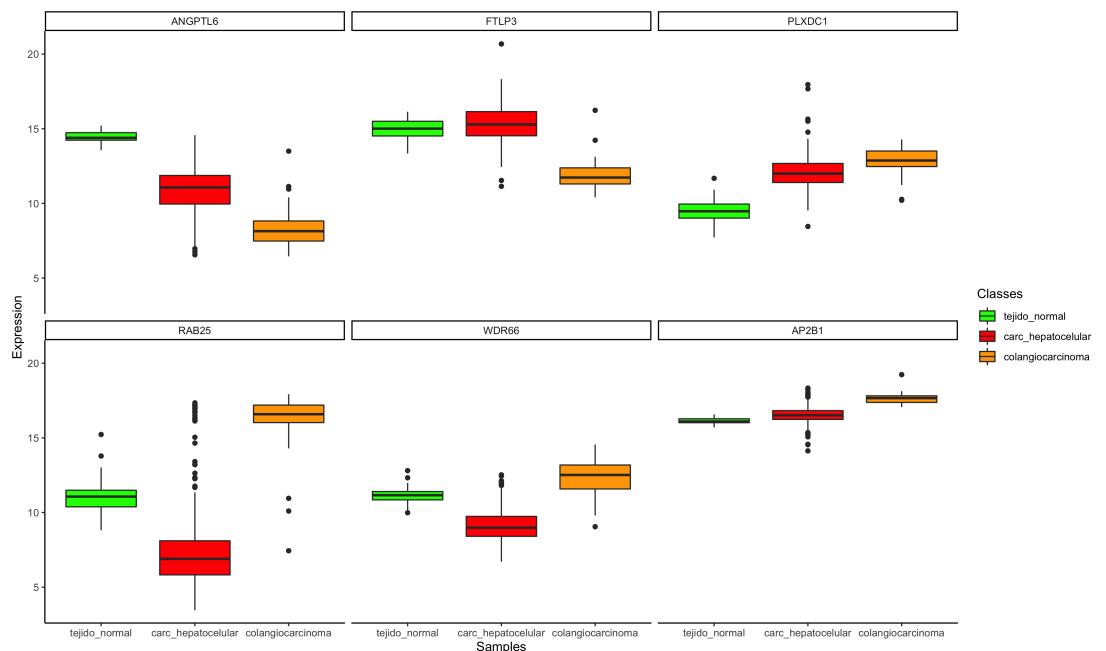
Figura 23. Mapa de calor de expresión de genes por tipo de muestra (columna izquierda: rojo = carcinoma hepatocelular, naranja = colangiocarcinoma, verde = tejido sano) en los 6 genes más relevantes encontrados en el mejor modelo de RF con mRMR como método de selección de características.



CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

68

Figura 24. Diagrama de caja de expresión de genes por tipo de muestra en los 6 genes más relevantes encontrados en el mejor modelo de RF con mRMR como método de selección de características.



4.6.4. Validación cruzada en entrenamiento - kNN

En la Tabla 24 se muestra el número óptimo de vecinos para los 10 genes más relevantes para mRMR, RF y DA.

Tabla 24. Número óptimo de vecinos encontrado para los 10 genes más relevantes de cada método de selección de características.

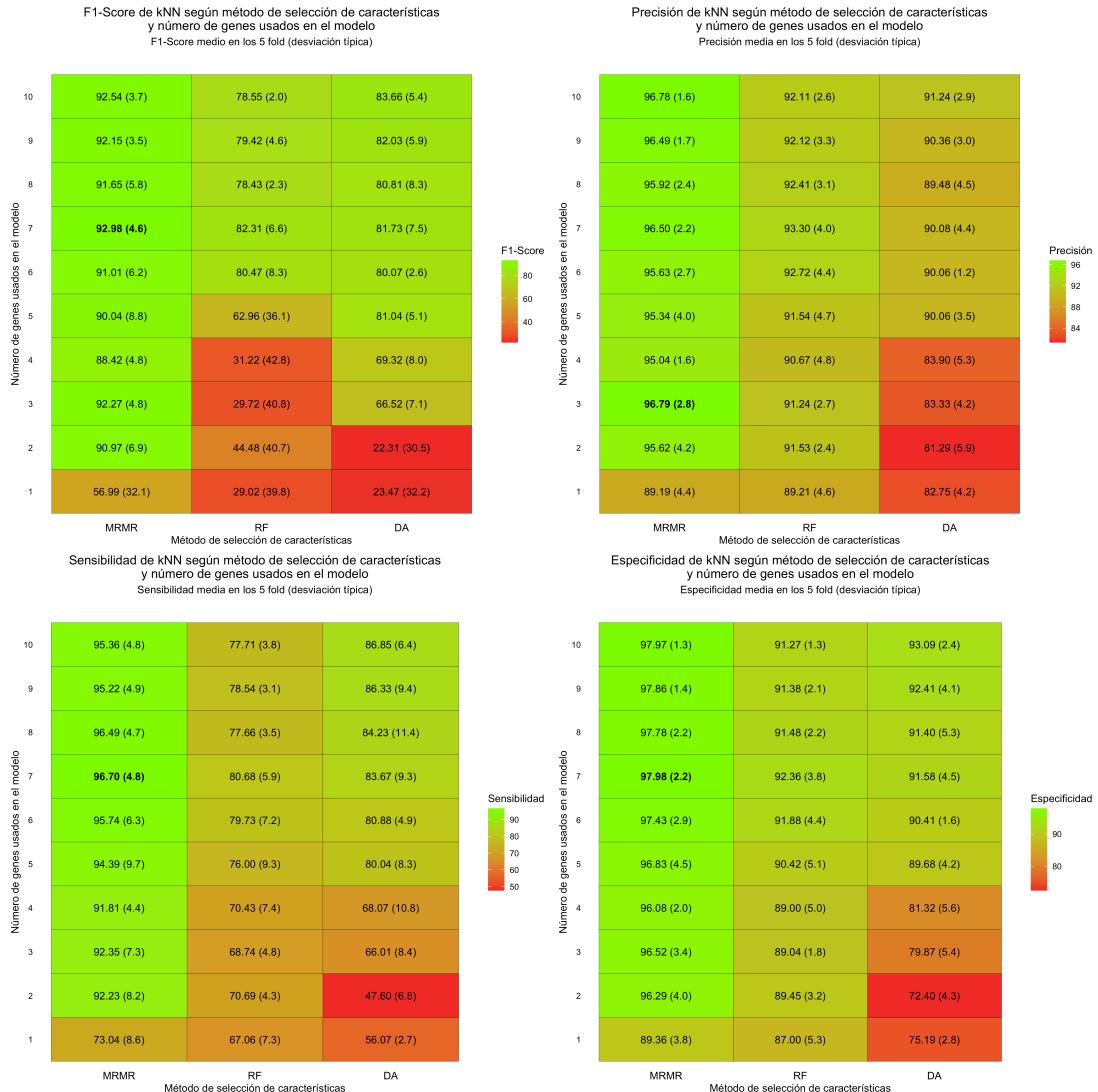
	mRMR	RF	DA
k	7	9	13

En la Figura 25 se muestran las medidas de evaluación medias obtenidas en las 5 fold.

4.6. RESULTADOS DE CLASIFICACIÓN MULTICLASE PARA CÁNCER DE HÍGADO

69

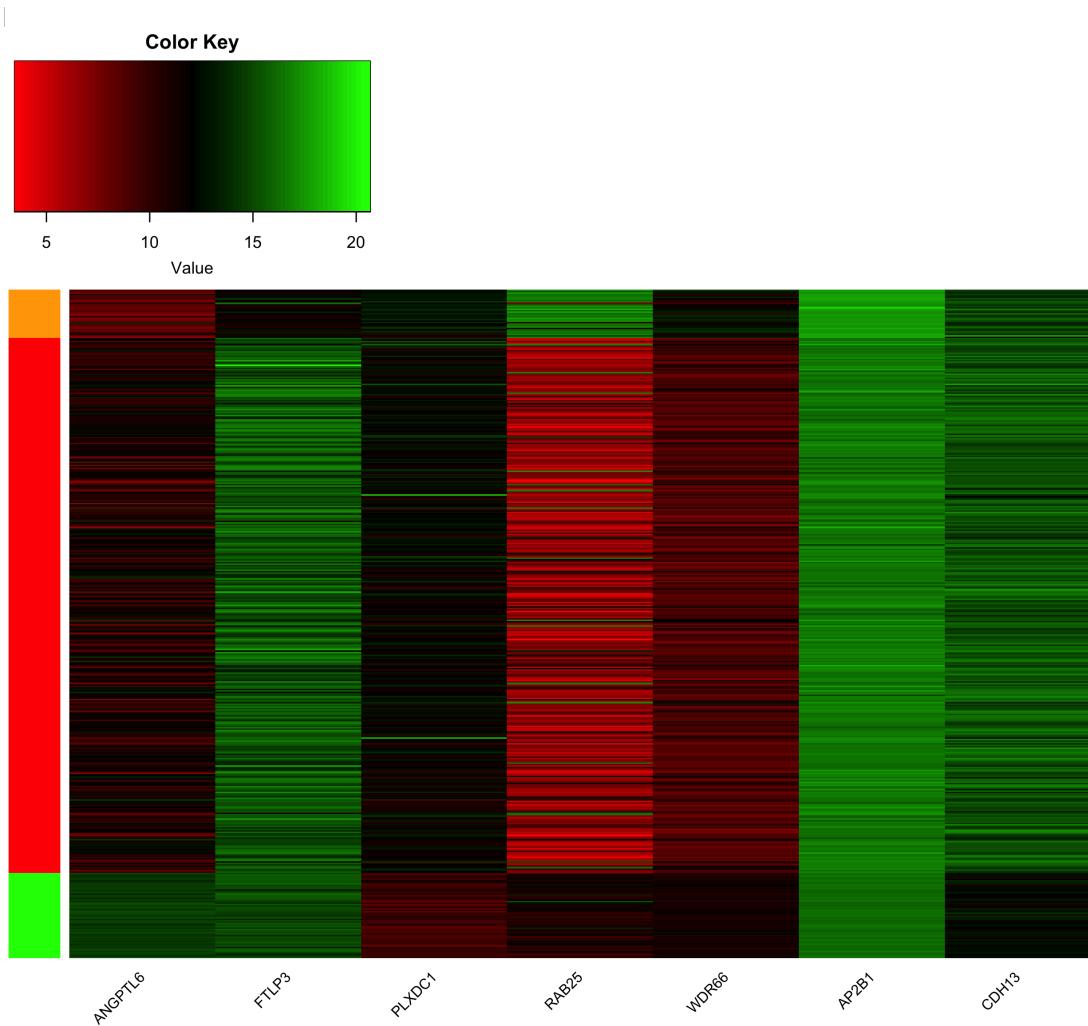
Figura 25. Mapa de calor con valores medios y desviación típica de los 5-fold de F1-Score, precisión, sensibilidad y especificidad de kNN según método de selección de características y número de genes usados.



Para kNN, el mejor F1-Score (92,98 %) se obtiene con el método mRMR para 7 genes. Los 7 genes seleccionados son: ANGPTL6, FTLP3, PLXDC1, RAB25, WDR66, AP2B1 y CDH13. Se analiza su expresión de genes, utilizando para ello un mapa de calor (Figura 26) y diagramas de caja (Figura 27).

**CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE
70 HÍGADO Y COLON-RECTO**

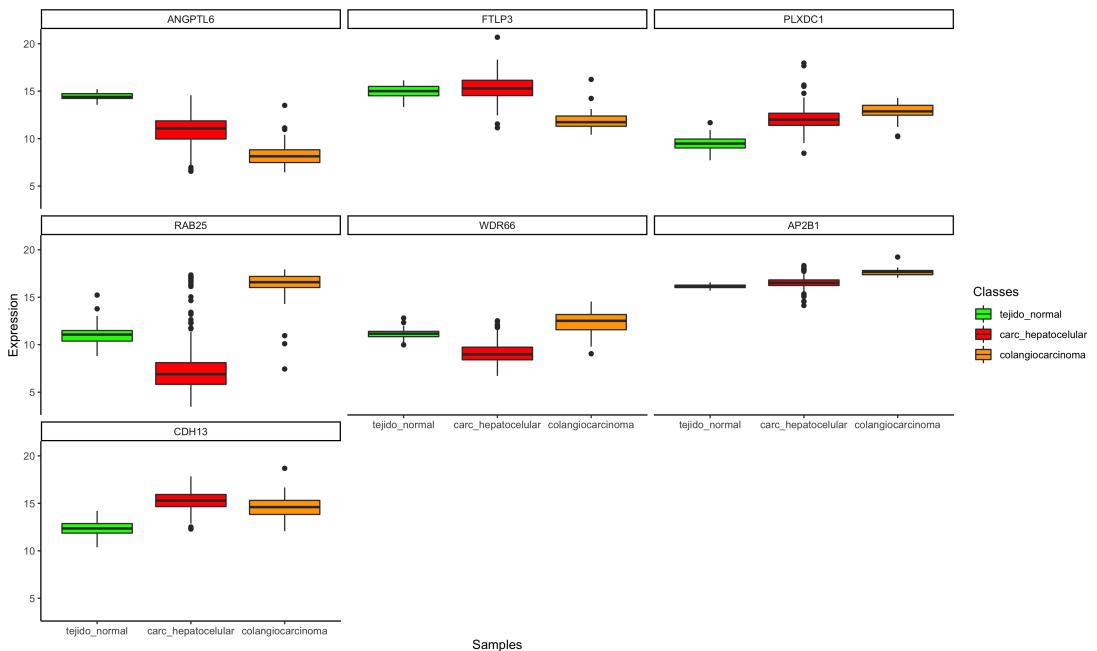
Figura 26. Mapa de calor de expresión de genes por tipo de muestra (columna izquierda: rojo = carcinoma hepatocelular, naranja = colangiocarcinoma, verde = tejido sano) en los 7 genes más relevantes encontrados en el mejor modelo de kNN con mRMR como método de selección de características.



4.6. RESULTADOS DE CLASIFICACIÓN MULTICLASE PARA CÁNCER DE HÍGADO

71

Figura 27. Diagrama de caja de expresión de genes por tipo de muestra en los 7 genes más relevantes encontrados en el mejor modelo de kNN con mRMR como método de selección de características.



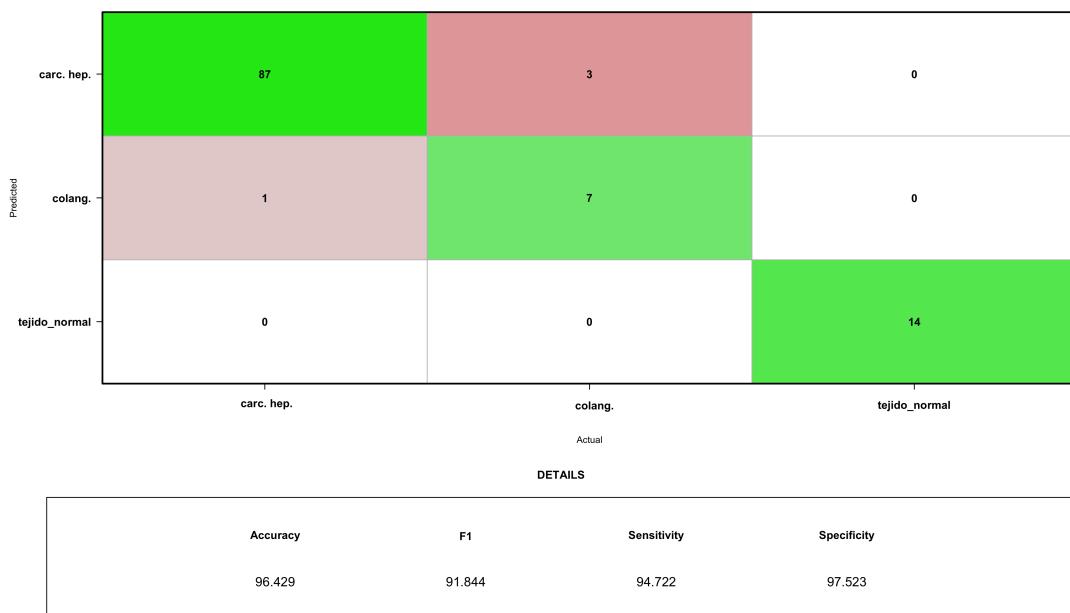
CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

72

4.6.5. Validación en test

Al validar en el conjunto de test el mejor modelo encontrado para SVM y kNN (ambas con 7 genes con mRMR), se obtiene la misma clasificación, descrita en la Figura 28, con F1-Score en test es de 91,8 % y una precisión de 96,4 %.

Figura 28. Matriz de confusión de los mejores modelos encontrados de SVM y kNN en el conjunto de test.



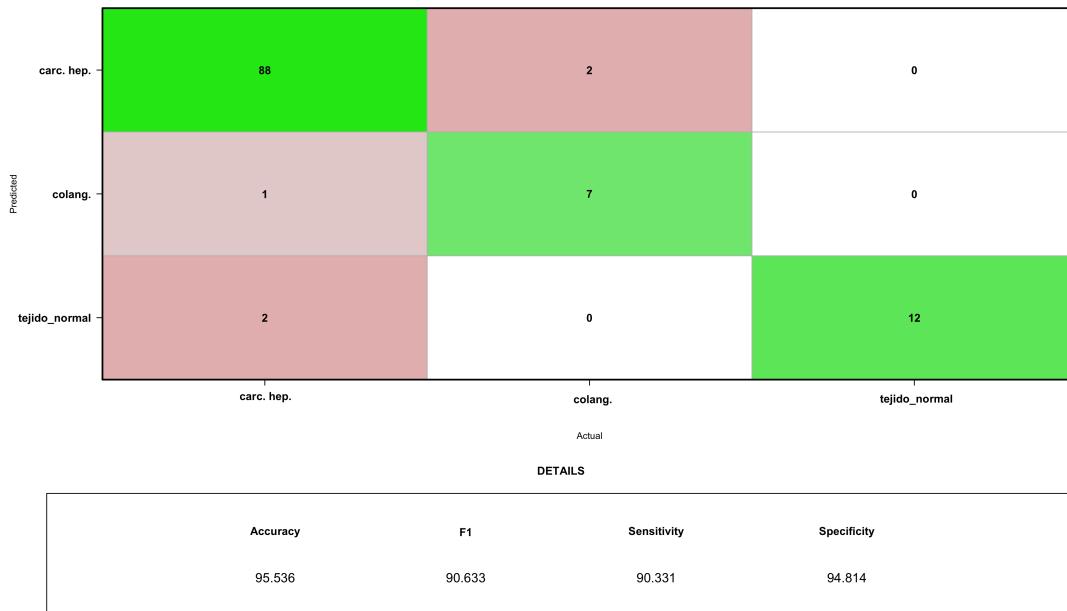
El modelo distingue bien entre tejido tumoral y tejido sano, pero comete 4 errores distinguiendo carcinomas hepatocelulares de colangiocarcinomas. Estos errores de clasificación pueden ser debidos a posibles similitudes biológicas entre ambos tipos de cáncer.

La validación en el conjunto de test del mejor modelo encontrado para RF (6 genes con mRMR) se muestra en la Figura 29, siendo los resultados ligeramente inferiores a los obtenidos con SVM y kNN.

Figura 29. Matriz de confusión del mejor modelo encontrado de RF en el conjunto de test.

4.7. RESULTADOS DE CLASIFICACIÓN BICLASE PARA CÁNCER DE COLON-RECTO

73



En la Tabla 25 se muestra un resumen de los mejores modelos obtenidos y su F1-Score y precisión en conjunto de entrenamiento y conjunto de test.

Tabla 25. Resumen de clasificación multiclase para cáncer de hígado. Mejor modelo encontrado para SVM, RF y kNN con biomarcadores seleccionados, parámetros optimizados para cada algoritmo, F1-Score y precisión (Acc) en conjunto de entrenamiento y conjunto de test.

		Biomarcadores	Parámetros	F1 train	Acc train	F1 test	Acc test
SVM	7 genes mRMR		$c = 1$ $\gamma = 0.025$	94,99	97,66	91,84	96,43
RF	6 genes mRMR		—	94,08	97,08	90,63	95,54
kNN	7 genes mRMR		$k = 7$	92,98	96,79	91,84	96,43

4.7. Resultados de clasificación biclase para cáncer de colon-recto

El código completo del análisis se muestra en el fichero *analisis_cr/03_analisis_biclase.R* del repositorio de GitHub asociado al trabajo [79].

**CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE
74 HÍGADO Y COLON-RECTO**

4.7.1. Detección de biomarcadores

Para la clasificación biclase en cáncer de colon-recto se cuenta con 644 tumores y 51 muestras de tejido sano.

Tabla 26. Distribución de tipos de muestra para el análisis de cáncer de colon-recto biclase.

	Número de casos	Porcentaje
Tumor	644	92,7 %
Tejido sano	51	7,3 %
Total	695	100 %

Se extraen 6.836 genes que presentan en su expresión diferencias significativas entre las muestras de tumor y las de tejido sano. En la Tabla 27 se muestra la partición entrenamiento - test realizada y en la figura 30 se representa en un diagrama de Sankey.

Tabla 27. Distribución entrenamiento-test según tipo de muestra y proporción entre clases para el análisis de colon-recto biclase.

	Total	Entrenamiento	Test
Tumores	644 (100 %)	483 (75,0 %)	161 (25,0 %)
Tejido sano	51 (100 %)	39 (76,5 %)	12 (23,5 %)
Proporción tumores/sanos	12,6	12,4	13,4

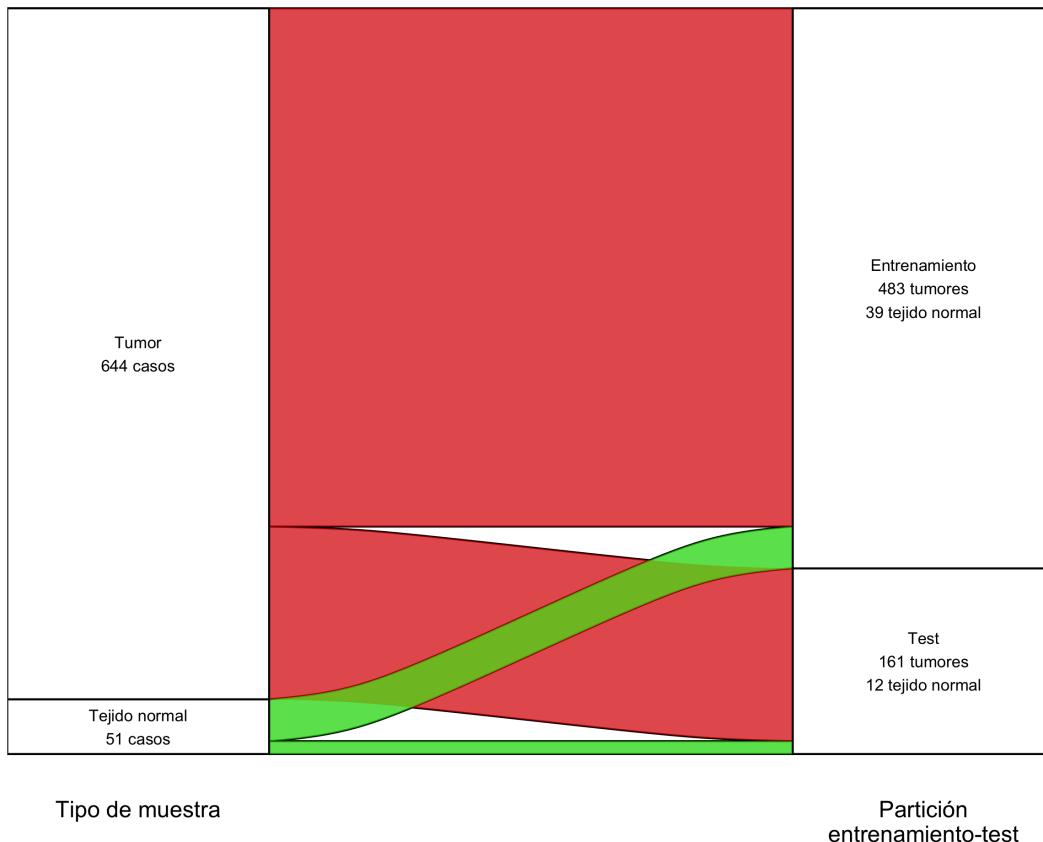
Figura 30. Diagrama de Sankey mostrando la partición entrenamiento-test realizada según tipo de muestra para el análisis de colon-recto biclase.

4.7. RESULTADOS DE CLASIFICACIÓN BICLASE PARA CÁNCER DE COLON-RECTO

75

Partición en conjuntos de entrenamiento y test

Reparto 75% - 25% con balanceo de clases



CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

A continuación se muestran los diez genes más relevantes encontrados por cada método de selección de características:

Tabla 28. Diez genes más relevantes según los distintos métodos de selección de características para el análisis de colon-recto biclase.

Ranking	mRMR	RF	DA
1	BEST4	VSTM2A	ETV4
2	MET	CA7	SCN7A
3	EPOP	COL11A1	RSPO2
4	RXRG	GLP2R	PHOX2B
5	C5orf34	SLC39A10	SALL4
6	DHRS7C	ENC1	POU5F1B
7	NKX2-3	ESM1	TNFRSF17
8	ESM1	CEMIP	SCN9A
9	SGCG	CA2	TLX1
10	MDFI	KRT80	WT1

Se observa que el gen ESM1 es el único gen común a dos algoritmos de selección de características: mRMR y RF.

4.7.2. Validación cruzada en entrenamiento

En las Tablas 29 y 30 se muestran los parámetros óptimos de SVM y kNN obtenidos tras el tuning.

Tabla 29. Parámetros óptimos de SVM encontrados para los 10 genes más relevantes de cada método de selección de características.

Parámetro	mRMR	RF	DA
Coste	0,05	0,05	0,05
Gamma	0,06	0,07	0,06

Tabla 30. Número óptimo de vecinos encontrado para los 10 genes más relevantes de cada método de selección de características.

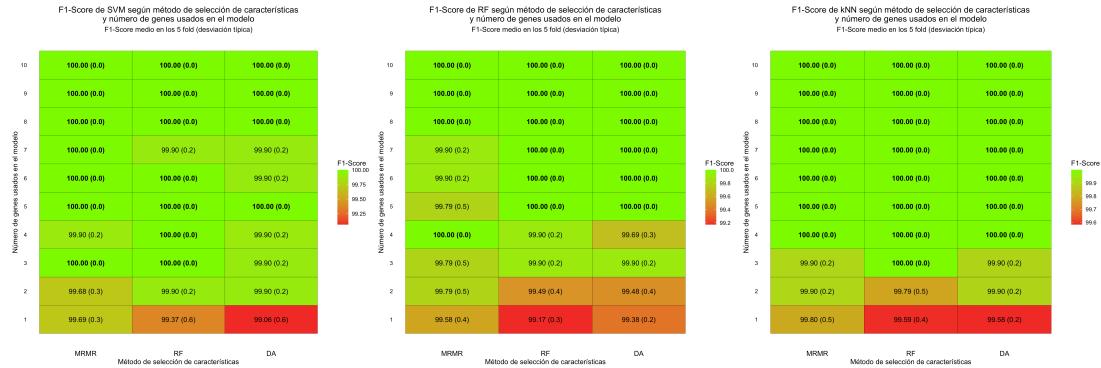
	mRMR	RF	DA
k	23	23	23

4.7. RESULTADOS DE CLASIFICACIÓN BICLASE PARA CÁNCER DE COLON-RECTO

77

En la Figura 31 se muestran para SVM, RF y kNN el F1-Score medio obtenido en las 5 fold.

Figura 31. Mapa de calor con valores medios y desviación típica de los 5-fold de F1-Score de SVM, RF y kNN según método de selección de características y número de genes usados.



Se alcanzan F1-Score de 100 % en muchos casos. En caso de igualdad de F1-Score, se seleccionará como mejor modelo aquel que emplee el menor número de genes. Por tanto, los mejores modelos son:

- Para SVM: mRMR ó RF con 3 genes.
- Para RF: mRMR con 4 genes.
- Para kNN: RF con 3 genes.

Se analiza la expresión de 3 genes más relevantes de RF y los 4 más relevantes de mRMR con diagramas de caja (Figuras 32 y 33).

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

Figura 32. Diagrama de caja de expresión de genes por tipo de muestra en los 3 genes más relevantes con RF como método de selección de características.

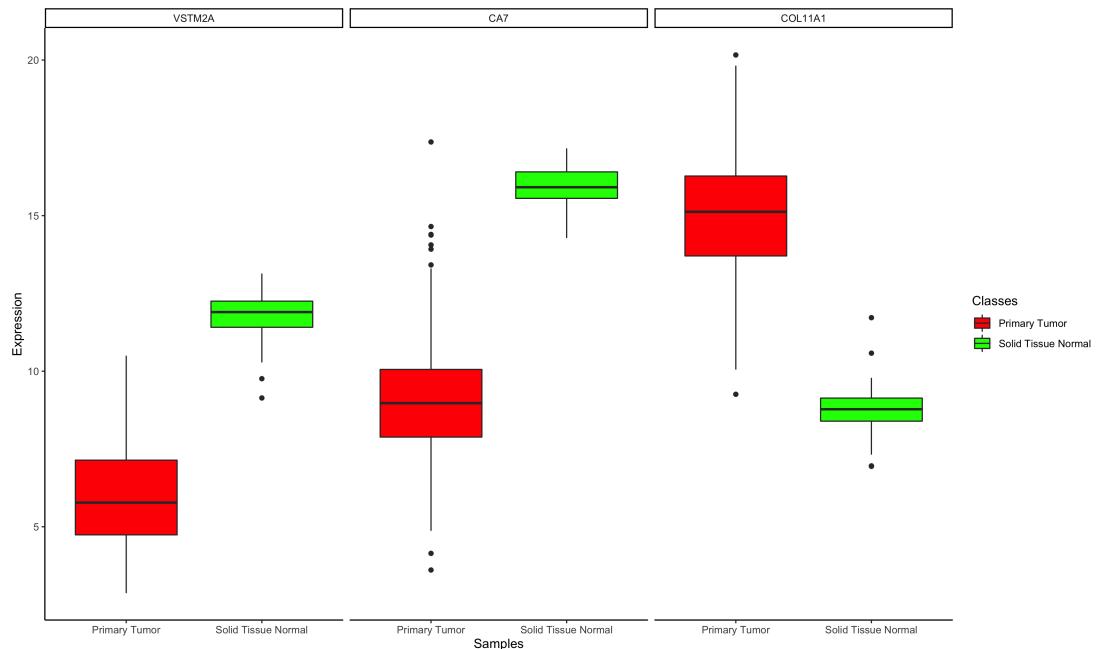
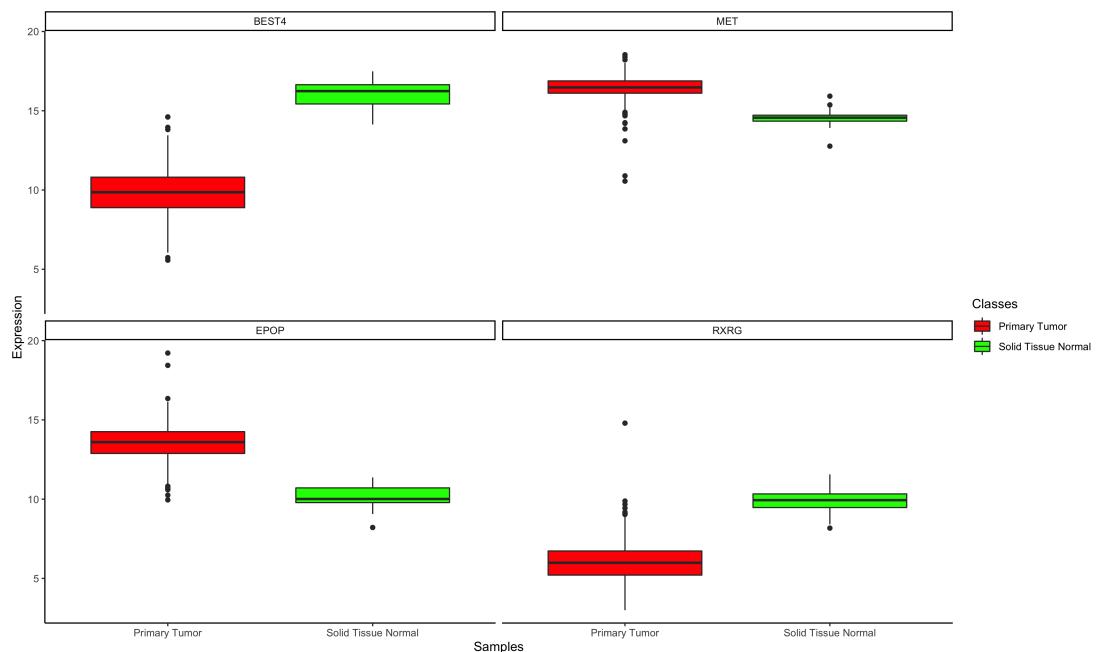


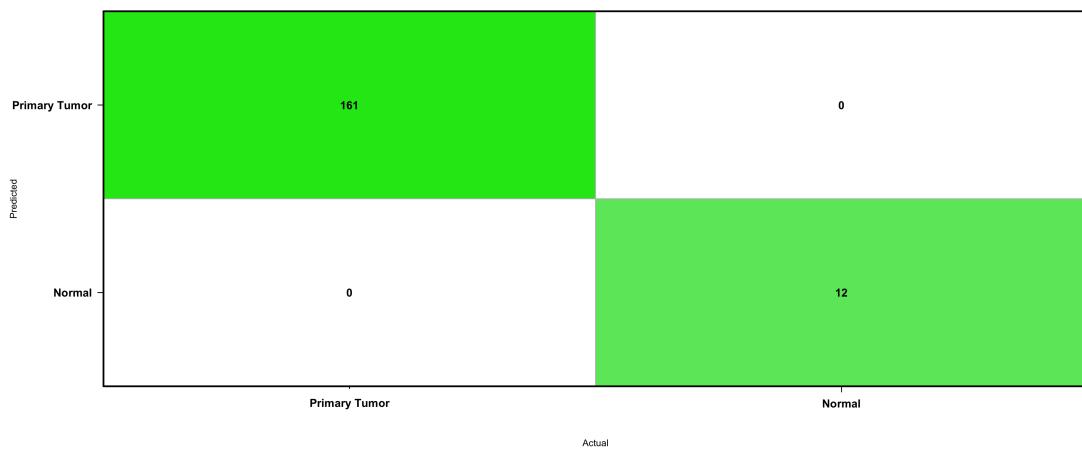
Figura 33. Diagrama de caja de expresión de genes por tipo de muestra en los 4 genes más relevantes con mRMR como método de selección de características.



4.7.3. Validación en test

Al validar en el conjunto de test los mejor modelo encontrado para SVM (dos modelos: 3 genes con mRMR y 3 genes con RF), RF (4 genes con mRMR) y kNN (3 genes con RF) se obtiene la misma clasificación: una predicción perfecta en test (F1-Score 100 %, precisión 100 %) como se muestra en la Figura 34.

Figura 34. Matriz de confusión de los mejores modelos encontrados de SVM, RF y kNN en el conjunto de test.



En la Tabla 31 se muestra un resumen de los mejores modelos obtenidos y su F1-Score y precisión en conjunto de entrenamiento y conjunto de test.

Tabla 31. Resumen de clasificación biclase para cáncer de colon-recto. Mejor modelo encontrado para SVM, RF y kNN con biomarcadores seleccionados, parámetros optimizados para cada algoritmo, F1-Score y precisión (Acc) en conjunto de entrenamiento y conjunto de test.

	Biomarcadores	Parámetros	F1 train	Acc train	F1 test	Acc test
SVM	3 genes mRMR	$c = 0,05$ $\gamma = 0,06$	100	100	100	100
	3 genes RF	$c = 0,05$ $\gamma = 0,07$	100	100	100	100
RF	4 genes mRMR	—	100	100	100	100
kNN	3 genes RF	$k = 23$	100	100	100	100

**CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE
80 HÍGADO Y COLON-RECTO**

4.8. Resultados de clasificación multiclase para cáncer de colon-recto

Nota: Para esta parte he hecho 2 análisis distintos y debería elegir uno de ellos. Los resumo mucho aquí abajo. Por coherencia con hígado multiclase, quizá debería usar el primero que os muestro (por diagnóstico primario), que además tiene mejores resultados, pero quería comentaros brevemente los resultados de los dos análisis. Y otra cosa... en este análisis he aumentado el número de genes a 20, porque los máximos F1 se suelen encontrar entre 13 y 18 genes. Si limito sólo a 10 genes, el mejor F1 se obtiene siempre para 10 genes.

Mostrar mejor el primero. Importante discutir las similitudes biológicas entre los adenocarcinomas.

Para las conclusiones, comentario de Dani relevante: si tras hacer un análisis de este tipo, la clasificación no es buena, se puede proponer como trabajo futuro combinar expresión con microRNA o alteraciones somáticas para mejorarla.

4.8.1. Por diagnóstico primario: sanos vs adenocarcinoma vs adenocarcinoma mucinoso

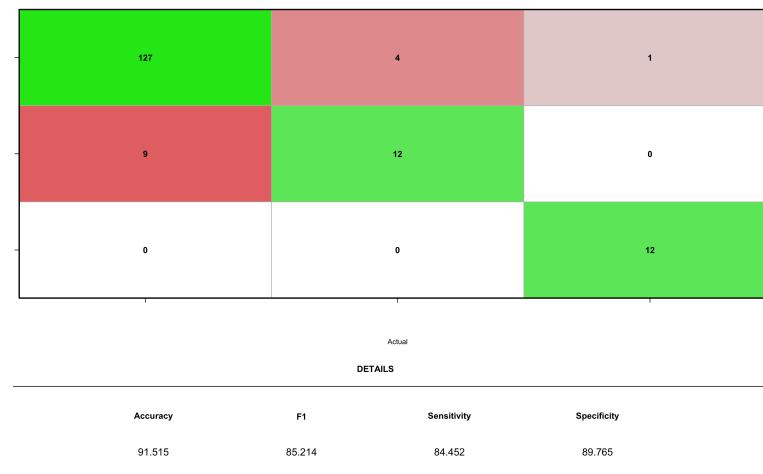
Los dos principales diagnósticos primarios son: adenocarcinoma y adenocarcinoma mucinoso. El resto tienen menos de 10 casos.

	N
Adenocarcinoma, NOS	530
Adenocarcinoma mucinoso	87
Tejido normal	51

La mejor matriz de confusión (RF con 17 genes elegidos con mRMR) es la siguiente:

4.8. RESULTADOS DE CLASIFICACIÓN MULTICLASE PARA CÁNCER DE COLON-RECTO

81



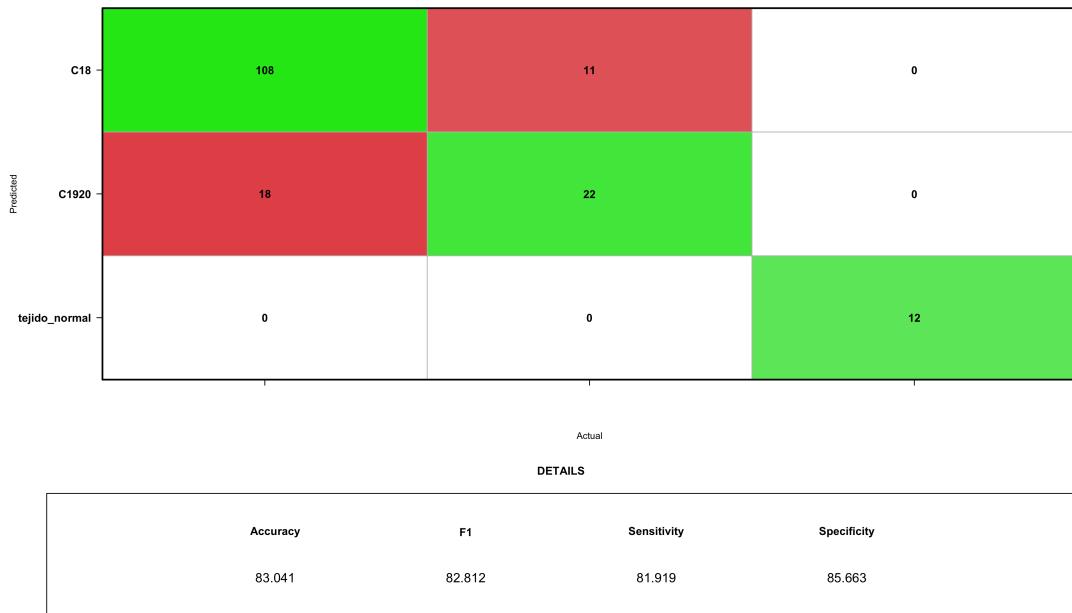
4.8.2. Por código: sanos vs colon vs recto + unión recto-sigmoidea

	N
Colon (C18)	476
Recto + Unión rectosigmoidea (C19-C20)	163
Tejido normal	51

La mejor matriz de confusión (kNN con 14 genes elegidos con RF) es la siguiente:

CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE HÍGADO Y COLON-RECTO

82



(Probé también sanos vs C18 vs C19 vs C20, pero los resultados empeoraban mucho).

4.9. Conclusiones

mRMR y RF han demostrado ser buenos métodos de selección de características, siempre mejores que DA, aunque carecen del enfoque médico en el que se basa este último método.

Los 3 algoritmos de clasificación propuestos (SVM, RF y kNN) han obtenido resultados muy similares, tanto para problemas biclase como multiclas. Una de las ventajas de trabajar con varios algoritmos de clasificación es que se puede escoger aquel que utiliza el menor número de genes, por las ventajas que ello conlleva (menor coste computacional, mayor interpretabilidad).

Pese a haber encontrado genes que separan correctamente las distintas clases, para considerar que existen evidencias claras gen-enfermedad es necesario contar con una interpretación clínica [85]. Este razonamiento está en concordancia con el debate existente sobre el uso de p-valores en la ciencia [86–88]. De igual forma, pese a los excelentes resultados de los modelos encontrados en el conjunto de

test, la validez externa de los métodos está por probar [89] y sería recomendable comprobar su efectividad en distintos conjuntos de datos.

**CAPÍTULO 4. DETECCIÓN DE BIOMARCADORES EN CÁNCER DE
HÍGADO Y COLON-RECTO**

84

5. biomarkeRs: una aplicación web interactiva para detección de biomarcadores

La aplicación ya está traducida al inglés, y he implementado transiciones de carga como propuso Daniel. La desarrollaré completamente en Agosto.

Capítulo por escribir, podéis pasar al siguiente porque sólo hay notas sobre lo que puedo escribir.

5.1. Desarrollo de la aplicación

Shiny, versión, documentación breve sobre Shiny, reactividad, incluye CSS, comentar que es accesible a todo el mundo (Shiny Server, o shinyapps.io), subir a una URL y compartir....

El fichero `shiny\app.R` del repositorio de GitHub asociado al trabajo [79] contiene el código de R desarrollado para crear la aplicación web.

5.2. Utilidades de la aplicación

Útil para realizar análisis genéticos para personas sin apenas conocimientos previos de programación. Capturas de pantalla con ejemplos. Escribir pequeño manual de uso. Quizá grabar vídeo mostrando la aplicación (y subir GIF a README del repositorio de GitHub).

**CAPÍTULO 5. BIOMARKERS: UNA APLICACIÓN WEB INTERACTIVA
86 PARA DETECCIÓN DE BIOMARCADORES**

6. Líneas abiertas de trabajo

6.1. El problema de clasificación

Para el problema de clasificación de muestras de tejidos existen varias líneas abiertas de trabajo:

- Otros métodos de clasificación. Además de los algoritmos de clasificación SVM, RF y kNN, se pueden emplear otros métodos. Por ejemplo, se puede clasificar usando SVM con núcleos de base no radial o crear un *ensemble* de algoritmos para permitir una estructura más flexible del modelo que se adapte mejor a los datos.
- Otros métodos de selección de características. Para lograr un equilibrio entre resultados numéricos y clínicos, se pueden utilizar algoritmos de selección de características que usan evidencias existentes de relaciones gen-enfermedad y añaden reducción de la redundancia. Es el caso de DARED, un algoritmo en el que están trabajando los tutores de este trabajo y que supone una mejora con respecto al método DA empleado en el capítulo 4.

6.2. {KnowSeq}

{KnowSeq} es un paquete de R en continua evolución y abierto a la colaboración mediante GitHub [58]. Algunas mejoras propuestas por el autor que se pueden implementar a corto plazo son las siguientes:

- Un tuning preciso de los parámetros de los algoritmos de clasificación. Para realizar el tuning de los parámetros óptimos de SVM y kNN se han utilizado los 10 genes más relevantes. Posteriormente, se seleccionaba como mejor modelo aquel que con menor número de genes obtenía el mejor F1-Score. Quizá para este número de genes, menor de 10, existan otros parámetros que puedan optimizar los resultados del algoritmo. Tras implementar este cambio, se podrían obtener parámetros óptimos para cada método de selección de características y número de genes.

- Nuevos gráficos. Dentro de la función `KnowSeq::dataPlot` se pueden implementar los mapas de calor planteados en el presente trabajo para medir F1-Score, precisión, sensibilidad y especificidad (por ejemplo, Figura 8). En el caso de estas figuras, permiten detectar fácilmente los indicadores más altos debido a los colores empleados y al resaltado en negrita del valor más alto (realizado automáticamente con el código).

Conforme vaya haciendo estos cambios en `KnowSeq`, puedo ir pasándolos a la parte de análisis donde nombro los cambios que he realizado en el paquete.

6.3. BiomarkeRs

Siguiendo la iniciativa de la aplicación web `BiomarkeRs` se debe continuar haciendo esfuerzos para acercar el análisis de datos transcriptómicos a personas sin conocimientos previos de programación. En este sentido, el uso de tutoriales o cursos cortos para enseñar a usar la plataforma pueden ser de utilidad. Además, la aplicación web debería aprovechar todo el potencial presente en `{KnowSeq}` y actualizarse con las nuevas funcionalidades que se implementen en el paquete. El hecho de que sea una aplicación de código abierto puede ayudar a esta actualización.

6.4. Artículo

Finalmente, como consecuencia de este trabajo se ha desarrollado un artículo presente en el [Anexo I: Artículo científico](#) que será enviado a una revista del campo de la bioinformática. *¿Debería nombrar la revista a la que se puede enviar? ¿De qué creéis que debería escribir el artículo? ¿Cáncer de hígado o colon-recto? ¿Sólo biclase o también multiclas?*

Bibliografía

- [1] American Cancer Society. What is Cancer? Disponible en: <https://www.cancer.org/cancer/cancer-basics/what-is-cancer.html> [Consultado 18/06/2020].
- [2] National Cancer Institute. What is Cancer? Disponible en: <https://www.cancer.gov/about-cancer/understanding/what-is-cancer> [Consultado 18/06/2020].
- [3] Lucia Migliore and Fabio Coppedè. Genetic and environmental factors in cancer pathogenesis. *Mutation Research*, 512:135–153, 2012.
- [4] Benjamin A Pierce. *Genética: Un enfoque conceptual*. Editorial Médica Panamericana, 3^a edition, 2010.
- [5] World Health Organization. *World Cancer Report 2014*. 2014.
- [6] World Health Organization. *World Cancer Report. Cancer research for cancer prevention*. 2020.
- [7] V. J. Cogliano, R. Baan, K. Straif, Y. Grosse, B. Lauby-Secretan, F. El Ghissassi, V. Bouvard, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet, and C. P. Wild. Preventable Exposures Associated With Human Cancers. *JNCI Journal of the National Cancer Institute*, 103(24):1827–1839, 2011.
- [8] World Health Organization (WHO). *ICD-10: International Statistical Classification of diseases and related health problems: 10th revision*. 1990.
- [9] Ministerio de Sanidad Consumo y Bienestar Social. Edición electrónica de la CIE-10-ES Diagnósticos. Disponible en: https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_10_mc.html [Consultado 21/06/2020].
- [10] Sherif R. Z. Abdel-Misih and Mark Bloomston. Liver Anatomy. *Surgical Clinics of North America*, 90(4):643–653, 2010.
- [11] Elijah Trefts, Maureen Gannon, and David H. Wasserman. The liver. *Current Biology*, 27(21):R1147–R1151, 2017.
- [12] American Cancer Society. Liver Cancer Risk Factors. Disponible en: <https://www.cancer.org/cancer/liver-cancer/causes-risks-prevention/risk-factors.html> [Consultado 18/06/2020], 2019.

- [13] Jorge A. Marrero, Robert J. Fontana, Sherry Fu, Hari S. Conjeevaram, Grace L. Su, and Anna S. Lok. Alcohol, tobacco and obesity are synergistic risk factors for hepatocellular carcinoma. *Journal of Hepatology*, 42(2):218–224, 2005.
- [14] Laura L. Azzouz and Sandeep Sharma. *Physiology, Large Intestine*. 2020.
- [15] American Cancer Society. Colorectal Cancer Risk Factors. Disponible en: <https://www.cancer.org/cancer/colon-rectal-cancer/causes-risks-prevention/risk-factors.html> [Consultado 20/06/2020].
- [16] Henry T. Lynch and Albert de la Chapelle. Hereditary Colorectal Cancer. *New England Journal of Medicine*, 348(10):919–932, 2003.
- [17] B. Levin, D. A. Lieberman, B. McFarland, R. A. Smith, D. Brooks, K. S. Andrews, C. Dash, F. M. Giardiello, S. Glick, T. R. Levin, P. Pickhardt, D. K. Rex, A. Thorson, and S. J. Winawer. Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *CA: A Cancer Journal for Clinicians*, 58(3):130–160, 2008.
- [18] Thomas M. Schmidt. *Encyclopedia of Microbiology*. Academic Press, 4th editio edition, 2019.
- [19] Zne-Jung Lee. An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer. *Artificial Intelligence in Medicine*, 42(1):81–93, jan 2008.
- [20] Rosalia Maglietta, Annarita D'Addabbo, Ada Piepoli, Francesco Perri, Sabino Liuni, Graziano Pesole, and Nicola Ancona. Selection of relevant genes in cancer diagnosis based on their prediction accuracy. *Artificial Intelligence in Medicine*, 40(1):29–44, may 2007.
- [21] Rory Stark, Marta Grzelak, and James Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.
- [22] Marcel C. Van Verk, Richard Hickman, Corné M.J. Pieterse, and Saskia C.M. Van Wees. RNA-Seq: revelation of the messengers. *Trends in Plant Science*, 18(4):175–179, apr 2013.

- [23] B MacMahon and TF Pugh. *Epidemiology: Principles and Methods*. 1970.
- [24] Miquel Porta, editor. *A Dictionary of Epidemiology*. Fifth edit edition, 2008.
- [25] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018.
- [26] International Agency for Research on Cancer and World Health Organization. Global Cancer Observatory, Cancer Today. Disponible en: <https://gco.iarc.fr/today/home> [Consultado 21/06/2020].
- [27] European Comission. ECIS - European Cancer Information System. Disponible en: <https://ecis.jrc.ec.europa.eu> [Consultado 21/06/2020].
- [28] J. Ferlay, M. Colombet, I. Soerjomataram, T. Dyba, G. Randi, M. Bettio, A. Gavin, O. Visser, and F. Bray. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *European Journal of Cancer*, 103:356–387, 2018.
- [29] Claudia Allemani, Tomohiro Matsuda, Veronica Di Carlo, Rhea Harewood, Melissa Matz, Maja Nikšić, Audrey Bonaventure, Mikhail Valkov, Christopher J Johnson, Jacques Estève, Olufemi J Ogunbiyi, Gulnar Azevedo e Silva, Wan-Qing Chen, Sultan Eser, Gerda Engholm, Charles A Stiller, Alain Monnereau, Ryan R Woods, Otto Visser, Gek Hsiang Lim, Joanne Aitken, Hannah K Weir, and Michel P Coleman. Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet*, 391(10125):1023–1075, mar 2018.
- [30] Roberta De Angelis, Milena Sant, Michel P Coleman, Silvia Francisci, Paolo Baili, Daniela Pierannunzio, Annalisa Trama, Otto Visser, Hermann Brenner, Eva Ardanaz, Magdalena Bielska-Lasota, Gerda Engholm, Alice Nenneke, Sabine Siesling, Franco Berrino, and Riccardo Capocaccia. Cancer survival in Europe 1999–2007 by country and age: results of EUROCARE-5—a population-based study. *The Lancet Oncology*, 15(1):23–34, jan 2014.
- [31] Red Española de Registros de Cáncer (REDECAN). Estimaciones de la incidencia del cáncer en España, 2020. Disponible

- en: https://funca.cat/redecan/redecan.org/es/Informe_incidencia_REDECAN_2020.pdf [Consultado 18/06/2020].
- [32] Marcela Guevara, Amaia Molinuevo, Diego Salmerón, Rafael Marcos-Gragera, María Dolores Chirlaque, José Ramón Quirós, Araceli Alemán, Dolores Rojas, Consol Sabater, Matilde Chico, Rosario Jiménez, Arantza López de Munain, Visitación de Castro, María José Sánchez, Josefina Perucha, Carmen Sánchez-Contador, Jaume Galceran, Eva Ardanaz, and Nerea Larrañaga. Supervivencia de Cáncer en España, 2002-2013. Disponible en: https://funca.cat/redecan/redecan.org/es/Informe_Supervivencia_REDECAN_2020.pdf [Consultado 26/06/2020], 2019.
- [33] Ministerio de Sanidad Consumo y Bienestar Social. Estadísticas de defunciones según la causa de muerte. Disponible en: <https://pestadistico.inteligenciadegestion.mscbs.es/> [Consultado 21/06/2020].
- [34] Instituto Nacional de Estadística (INE). Estadísticas de cifras de población. Disponible en: <http://ine.es/> [Consultado 21/06/2020].
- [35] Daniel Redondo-Sánchez. Modelización Matemática de la Estimación de Incidencia de Cáncer. 2019.
- [36] IARC. *Registros de Cáncer: Principios y Métodos*. 1995.
- [37] Segi M. Cancer mortality for selected sites in 24 countries (1950–57). *Sendai, Japan: Department of Public Health, Tohoku University of Medicine.*, 1960.
- [38] Waterhouse PAH Doll R, Payne P. Cancer incidence in five continents, Volume I. *Geneva: Union Internationale Contre le Cancer.*, 1966.
- [39] JAH Waterhouse, CS Muir, P Correa, and J Powell. Cancer incidence in five continents, Volume III. *Lyon: IARC*, page 3:456, 1976.
- [40] EUROSTAT. Revision of the European standard population: Report of the Eurostat's task force. Technical report, Luxembourg: European Union., 2013.
- [41] Randy Gordon. Skin Cancer: An Overview of Epidemiology and Risk Factors. *Seminars in Oncology Nursing*, 29(3):160–169, aug 2013.
- [42] Vishal Madan, John T Lear, and Rolf-Markus Szeimies. Non-melanoma skin cancer. *The Lancet*, 375(9715):673–685, feb 2010.

- [43] Miroslav Zavoral, Stepan Suchanek, Filip Zavada, Ladislav Dusek, Jan Mužik, Bohumil Seifert, and Premysl Fric. Colorectal cancer screening in Europe. *World Journal of Gastroenterology*, 15(47):5907, 2009.
- [44] World Health Organization. Cancer factsheet. Disponible en: <https://www.who.int/news-room/fact-sheets/detail/cancer> [Consultado 29/06/2020], 2018.
- [45] Eric P Xing, Michael I Jordan, and Richard M Karp. Feature Selection for High-Dimensional Genomic Microarray Data.
- [46] Khawla Tadist, Said Najah, Nikola S. Nikolov, Fatiha Mrabti, and Azeddine Zahi. Feature selection methods and genomic big data: a systematic review. *Journal of Big Data*, 6(1):79, dec 2019.
- [47] Richard Bellman. *Dynamic Programming*. 1957.
- [48] Richard Bellman. *Adaptive Control Processes: A Guided Tour*. 1961.
- [49] Ignacio Rojas, Luis Javier Herrera Maldonado, and Daniel Castillo Secilla. Apuntes de la asignatura Biología Computacional con Big Data omics e Ingeniería Biomédica. Máster de Ciencia de Datos e Ingeniería de Computadores. Universidad de Granada. 2020.
- [50] Dan He, Irina Rish, David Haws, and Laxmi Parida. MINT: Mutual Information Based Transductive Feature Selection for Genetic Trait Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(3):578–583, may 2016.
- [51] Daphne Koller and Mehran Sahami. Toward Optimal Feature Selection. In *ICML*, 1996.
- [52] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, aug 2005.
- [53] Chris Ding and Hanchuan Peng. Minimum Redundancy Feature Selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205, 2005.

- [54] J Yang, Z Zhu, S He, and Z Ji. Minimal-redundancy-maximal-relevance feature selection using different relevance measures for omics data classification. *IEEE symposium on computational intelligence in bioinformatics and computational biology (CIBCB)*, pages 246–51, 2013.
- [55] Juan Manuel Gálvez, Daniel Castillo, Luis Javier Herrera, Belén San Román, Olga Valenzuela, Francisco Manuel Ortúño, and Ignacio Rojas. Multiclass classification for skin cancer profiling based on the integration of heterogeneous gene expression series. *PLOS ONE*, 13(5):e0196836, may 2018.
- [56] Daniel Castillo, Juan Manuel Galvez, Luis J. Herrera, Fernando Rojas, Olga Valenzuela, Octavio Caba, Jose Prados, and Ignacio Rojas. Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level. *PLOS ONE*, 14(2):e0212127, feb 2019.
- [57] Juan Manuel Galvez, Daniel Castillo, Luis J. Herrera, Olga Valenzuela, Octavio Caba, Jose Carlos Prados, Francisco Manuel Ortuno, and Ignacio Rojas. Towards Improving Skin Cancer Diagnosis by Integrating Microarray and RNA-seq Datasets. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2020.
- [58] Daniel Castillo-Secilla, Juan Manuel Galvez, Francisco Carrillo-Perez, Marta Verona-Almeida, Francisco Manuel Ortuno, Luis Javier Herrera, and Ignacio Rojas. KnowSeq: KnowSeq R/Bioc package: Beyond the traditional Transcriptomic pipeline. Disponible en: <https://github.com/CasedUgr/KnowSeq>, 2020.
- [59] Miron B Kursa. praznik: Tools for Information-Based Feature Selection. Disponible en: <https://CRAN.R-project.org/package=praznik>, 2020.
- [60] Leo Breiman. Random Forest. *Machine Learning*, (45):5–32, 2001.
- [61] Leo Breiman. Manual on setting up, using, and understanding random forest. Technical report, 2002.
- [62] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forest of randomized trees. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 431–439, 2013.

- [63] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. Disponible en: <https://CRAN.R-project.org/doc/Rnews/>, 2002.
- [64] Open Targets. Target Validation: Open Targets Platform. Disponible en: <https://www.targetvalidation.org> [Consultado 07/07/2020].
- [65] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*, pages 144–152, New York, New York, USA, 1992. ACM Press.
- [66] Kai-Bo Duan and S. Sathiya Keerthi. Which Is the Best Multiclass SVM Method? An Empirical Study. pages 278–285. 2005.
- [67] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. Disponible en: <https://CRAN.R-project.org/package=e1071>, 2019.
- [68] Leo Breiman. Bagging Predictors. *Machine Learning*, 24:123–140, 1996.
- [69] N. S. Altman. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, 46(3):175–185, aug 1992.
- [70] Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke, and Chih-Fong Tsai. The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, 5(1):1304, dec 2016.
- [71] Max Kuhn. caret: Classification and Regression Training. Disponible en <https://CRAN.R-project.org/package=caret>, 2020.
- [72] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, aug 2018.
- [73] National Cancer Institute and National Institutes of Health. GDC Portal. Disponible en: <https://portal.gdc.cancer.gov/> [Consultado 22/06/2020].
- [74] National Cancer Institute. National Cancer Institute. Disponible en: <https://www.cancer.gov> [Consultado 22/06/2020].

- [75] Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.
- [76] National Cancer Institute. The Cancer Genome Atlas Program. Disponible en: <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga> [Consultado 05/07/2020].
- [77] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, jul 1900.
- [78] F. Yates. Contingency Tables Involving Small Numbers and the χ^2 Test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217, 1934.
- [79] Daniel Redondo-Sánchez. Repositorio GitHub: Epidemiología y detección de biomarcadores en cáncer. Disponible en: https://github.com/danielredondo/TFM_ciencia_de_datos.
- [80] R Core Team. R: A Language and Environment for Statistical Computing, 2020.
- [81] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.
- [82] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, mar 2012.

- [83] Wilson Wen Bin Goh, Wei Wang, and Limsoon Wong. Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology*, 35(6):498–507, jun 2017.
- [84] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, oct 2010.
- [85] Yotam Drier and Eytan Domany. Do Two Machine-Learning Based Prognostic Signatures for Breast Cancer Capture the Same Biological Processes? *PLoS ONE*, 6(3):e17795, mar 2011.
- [86] S J Evans, P Mills, and J Dawson. The end of the p value? *British heart journal*, 60(3):177–80, sep 1988.
- [87] S N Goodman. Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of internal medicine*, 130(12):995–1004, jun 1999.
- [88] Robert Matthews. Storks Deliver Babies ($p = 0.008$). *Teaching Statistics*, 22(2):36–38, jun 2000.
- [89] Allan Steckler and Kenneth R McLeroy. The importance of external validity. *American journal of public health*, 98(1):9–10, jan 2008.

Anexo I: Artículo

Anexo I: Artículo científico

Escribir en Agosto