Análisis de métodos de selección de atributos con aplicación en problemas de bioinformática

Máster Universitario en Inteligencia Artificial

Curso académico
2018-2019

Alumno
Ignacio del Valle Torres

DNI
50860414P

Director de TFM
Gherardo Varando

12 créditos

Segunda Convocatoria

Diciembre 2019

Índice de Contenidos

1 RESUMEN

La selección de atributos o variables es una herramienta muy útil para el preprocesamiento de datos, especialmente en aquellos campos de estudio que contienen conjuntos de datos con decenas o cientos de miles de variables.

La utilización de las diferentes técnicas de selección de variables, las cuales han sido tradicionalmente agrupadas en tres grandes grupos: filtraje, envoltorio y embebidas, ha resultado ser muy eficiente para la resolución de distintos problemas presentes en minería de datos y en aprendizaje automático.

Al eliminar variables redundantes o no informativas, la selección de variables permite la construcción de modelos más simples y sencillos de interpretar, mejorando el rendimiento de los modelos de aprendizaje y disminuyendo además los tiempos de ejecución.

Los avances en los últimos años en biología molecular y en el campo de la genómica han provocado un crecimiento masivo de la cantidad de información a analizar por los investigadores, contribuyendo además dicho aumento al desarrollo del campo de la bioinformática. La bioinformática combina los campos de la biología, ciencias de la computación y tecnología de la información en una única disciplina. En un elevado número de problemas a resolver por esta disciplina, especialmente en aplicaciones bioinformáticas utilizadas en genética o en medicina, el número de variables es significativamente más elevado que el número de muestras, siendo por tanto muy conveniente el uso de distintas técnicas de selección de variables.

La selección de las variables más importantes (genes) es esencial para la identificación de nuevos biomarcadores o para por ejemplo diferenciar aquellos genes expresados en determinado tipo de enfermedades en contraposición a pacientes sanos. Los perfiles de expresión génica, los cuales pueden identificados mediante *microarrays* (chips) de ADN, representan el estado celular a nivel molecular, y pueden ser empleados para dicha distinción entre genes expresados en distintas patologías, incluyendo el cáncer.

Ya que los experimentos de *microarrays* generan normalmente un elevado número de variables (genes) para un reducido número de pacientes (observaciones) la clasificación de pacientes enfermos frente a sanos es normalmente muy compleja, requiriendo normalmente de un proceso de selección de variables.

En el presente trabajo de fin de máster se realiza una revisión de los principales métodos de selección de variables, haciendo hincapié en aquellos que se están aplicando en distintas aplicaciones bioinformáticas, especialmente en el análisis de *microarrays*.

Dos conjuntos de datos de *microarrays* han sido utilizados para implementar y evaluar distintas técnicas de selección de variables (pertenecientes a técnicas de filtraje, envoltorio o embebidas), con el objetivo de seleccionar aquellos métodos más eficaces para el análisis de conjuntos de datos de elevada dimensionalidad.

Los resultados obtenidos con cada uno de los dos conjuntos de datos muestran que el uso de un subgrupo de genes permite mejorar las métricas de evaluación utilizadas para evaluar los modelos de aprendizaje automático usados en la clasificación de muestras tumorales y sanas.

2 OBJETIVOS

1. Realizar una revisión bibliográfica de los principales métodos de selección de variables: filtraje, envoltorio y embebidos.

2. Investigar los principales métodos de selección de variables empleados en bioinformática, especialmente aquellos utilizados en la técnica de *microarrays* de ADN

3. Implementar y evaluar diferentes métodos de selección de variables en dos conjuntos de datos representativos de *microarrays* de ADN

3 ESTADO DEL ARTE

3.1 CONCEPTOS GENERALES DE SELECCIÓN DE VARIABLES

En aprendizaje automático, al aumentar la dimensionalidad de los datos, la cantidad de datos necesarios para la realización de un análisis fiable aumenta exponencialmente. Este fenómeno ha sido definido como la “maldición de la dimensionalidad” en problemas de optimización dinámica (Bellman, 1957).

Cuando las dimensiones del conjunto de datos son muy elevadas aumenta la dificultad en obtener resultados significativos debido a la dispersión de datos relevantes en el conjunto de datos analizado.

Una variable se puede definir como una medida individual del proceso que se está observando. Un algoritmo de aprendizaje es capaz de realizar una clasificación a partir de un conjunto de variables, y en los últimos anos el tamaño de los conjuntos de datos ha aumentado muchísimo, llegando a tener conjuntos de datos con cientos o incluso miles de variables.

La presencia de variables “ruidosas” (irrelevantes) o redundantes en nuestros datos puede jugar un papel muy importante en diferentes aspectos del modelo que vayamos a utilizar, pudiendo tener como consecuencia un elevado tiempo de entrenamiento, falta de interpretabilidad e incluso sobreentrenamiento.

El sobreentrenamiento se basa en la relación existente entre el número de observaciones y las variables presentes en el conjunto de datos. Si nos encontramos en una situación en que en que el número de variables es muy elevado comparado con el número de observaciones, el algoritmo de aprendizaje que estemos utilizando va a tener mayores probabilidades de terminar atrapado en un óptimo local.

Conjuntos de datos con muchas variables y pocas muestras tienden al sobreajuste, y un modelo sobreajustado puede identificar de forma errónea pequeñas fluctuaciones (como varianzas significativas), generando errores de clasificación.

Los algoritmos de aprendizaje pueden encontrarse afectados por ruido en los datos, el cual ha de reducirse al máximo para evitar complejidad innecesaria en los modelos inferidos y mejorar la eficacia del algoritmo.

El ruido en un conjunto de datos puede ser debido a:

- Ruido en los atributos: producido por errores en los valores de los atributos (valores ausentes o variables cuya medición ha sido errónea).

- Ruido en la clase: Muestras asignadas a más de una clase o que han sido erróneamente asignadas.

Aunque podríamos ejecutar nuestro algoritmo de aprendizaje con todas las variables de las que disponemos y dejar que dicho algoritmo decida aquellas que son importantes, estas son las principales razones por las que esto no es una buena idea:

1. La “maldición de la dimensionalidad” y el sobreajuste.
   Al aumentar la dimensionalidad del espacio de variables, el número de configuraciones puede aumentar exponencialmente, pudiendo disminuir de esta manera el número de configuraciones abarcadas por cada observación. Si tenemos más columnas que filas en nuestro conjunto de datos seremos capaces de ajustar nuestro conjunto de entrenamiento perfectamente, pero no seremos capaces de generalizar con nuevas muestras.

2. La navaja de Occam.
   Queremos construir modelos que sean lo más simples e interpretables posible, perdiendo por tanto capacidad interpretativa si utilizamos un número muy elevado número de variables.

3. Utilización de datos “basura”.
   En la mayoría de los casos vamos a contar con variables no informativas, y el uso de datos de entrada de baja calidad va a generar datos de salida de baja calidad.

Además, el uso de un elevado número de variables va a generar un modelo de elevada complejidad, que tarde mucho tiempo en ejecutarse y que sea difícil de implementar en producción.

El principal objetivo de la selección de variables consiste en ayudar a resolver los problemas anteriores mediante la selección de un subgrupo de variables que pueden ser capaces de describir nuestro conjunto de datos eliminando a la vez ruido o variables irrelevantes, siendo capaces simultáneamente de obtener buenos valores predictivos.

Las diferentes técnicas de selección de variables van a seleccionar las variables que son capaces de funcionar bien al ser usadas en combinación, incluso en el caso que alguna de ellas no posea un valor predictivo muy bueno a nivel individual.
Además, a la hora de utilizar métodos de selección de variables es muy importante realizar la selección de variables únicamente en el test de entrenamiento, con el objetivo de evitar sobreajuste.
Dicha reducción de variables se basa principalmente en la relevancia y redundancia de una variable con respecto a un objetivo o clase.

De manera más específica, podemos clasificar las distintas variables de un conjunto de datos como:

- Muy relevantes.
- No muy relevantes, pero no redundantes.
- Irrelevantes.
- Redundantes.

Una variable muy relevante va a ser siempre necesaria para una selección de variables óptima, ya que no puede ser eliminada del conjunto de datos sin afectar a su distribución original.
Una variable no muy relevante puede no ser siempre importante para un conjunto de datos, pero esto va a depender de determinadas condiciones.
Variables irrelevantes son aquellas cuya inclusión en los datos no es importante. Finalmente, las variables redundantes son poco relevantes y pueden ser sustituidas completamente por otro conjunto de variables sin afectar a la distribución del conjunto de datos original. La redundancia es analizada siempre en el caso multivariante (al examinar subgrupos de variables), mientras que la relevancia se establece para variables individuales.

El principal objetivo de la selección de variables va a ser por tanto maximizar la relevancia y minimizar la redundancia.

Para asegurarnos que se ha seleccionado el subgrupo de variables óptimas con respecto al problema que queremos resolver, los métodos de selección de variables han de evaluar un total de 2^N-1 subgrupos, donde N es el número total de variables en el conjunto de datos (se excluye del análisis el subgrupo de variables vacío).

A diferencia de los distintos métodos de reducción de dimensionalidad, la selección de variables no crea nuevas variables, ya que utiliza las ya existentes con el objetivo de reducir su número.
La evaluación directa de todos los subgrupos de variables (2^N) para un conjunto de datos determinado es un problema NP-complejo al aumentar el número de variables; por tanto, ha de utilizarse un método de selección de variables que sea computacionalmente asequible.

El proceso completo para encontrar un subgrupo óptimo de variables en un conjunto de datos consiste principalmente de los siguientes pasos:

1. Generación de un subgrupo de variables.
2. Evaluación del subgrupo de variables seleccionadas.
3. Parada en base a un criterio definido.
4. Validación de los resultados.

Los diferentes métodos de selección de variables se pueden clasificar principalmente en tres grandes grupos: métodos de filtraje, envoltorio y embebidos, los cuales son investigados en detalle en secciones posteriores.
Algunos de éstos métodos asumen una independencia entre variables completa o casi completa, aunque también se han desarrollado aproximaciones que tienen en cuenta dicha dependencia entre variables (Tabla 1).

Método	Ventajas	Inconvenientes	Ejemplos
Filtraje	Independientes del clasificador Bajo coste computacional (rápidos) Buena generalización	No interaccionan con el clasificador	Puntuación F CFS Ganancia de Información
Envoltorio	Interaccionan con el clasificador Detectan dependencia entre variables	Alto coste computacional Riesgo de sobreajuste Selección dependiente del clasificador	RFE Boruta
Embebidos	Interaccionan con el clasificador Detectan dependencia entre variables	Selección dependiente del clasificador	Árboles de decisión

Tabla 1. Breve descripción de los principales métodos de selección de atributos Adaptado de (Saeys, Inza, & Larranaga, 2007).

3.2 SELECCIÓN DE VARIABLES EN BIOINFORMÁTICA

La bioinformática se puede definir como la disciplina que combina biología, ciencias de la computación y tecnologías de la información en una única disciplina.

Una de las áreas de ciencias de la computación con una gran aplicación en bioinformática es la minería de datos, principalmente el descubrimiento de patrones y relaciones a partir de un conjunto de datos.
La minería de datos consiste en una serie de pasos principales: la integración de los datos, el preprocesamiento de estos, la construcción de un modelo inductivo y finalmente la toma de decisiones a partir de los resultados de dicho modelo.
El preprocesamiento de datos tiene como objetivo la obtención de datos más precisos, y es en esta sección donde se emplean las técnicas de selección de atributos. El conjunto de datos seleccionado es finalmente usado para la construcción y entrenamiento de modelos con el objetivo de encontrar nueva información.

Uno de los objetivos de la minería de datos consiste en el uso de técnicas de clasificación a partir de información específica. En bioinformática, las técnicas de clasificación pueden ser utilizadas para por ejemplo la predicción de enfermedades o la clasificación de ciertas secuencias de ácido desoxirribonucleico (ADN) o de ácido ribonucleico (ARN).
Sin embargo, uno de los principales problemas en el empleo de técnicas de clasificación en bioinformática consiste la elevada dimensionalidad de los datos (el elevado número de variables o genes en cada observación).
Por ejemplo, si estamos comparando el genoma de pacientes afectados por una enfermedad con el genoma de pacientes no afectados, cada paciente (u observación o instancia) va a tener alrededor de miles de genes analizados (variables).
Y tal y como se ha discutido previamente, con un elevado número de variables la presencia de variables redundantes o irrelevantes aumenta exponencialmente.

Por tanto, la selección de variables es una herramienta muy útil que puede ser aplicada en diferentes campos de la bioinformática, tal y como se detalla las secciones posteriores (Hosseini, Nematbaksh, & Nadimi, 2017).

3.2.1 SELECCIÓN DE VARIABLES EN ANÁLISIS DE SECUENCIAS

El análisis de secuencias consta de distintas etapas, incluyendo el aislamiento de una secuencia determinada (bien de ADN, ARN o péptidos que componen una proteína) y la interpretación de la secuencia una vez aislada. La obtención de este tipo de información puede ser utilizada por ejemplo con el objetivo de ampliar el conocimiento de las secuencias de genes relevantes para una determinada enfermedad.

La selección de atributos ha sido empleada en el análisis de secuencias en base a dos tipos de análisis distintos:

1. Análisis de contenido: referido a la función de la secuencia que estamos analizando, como por ejemplo su función biológica. Se están utilizando de esta manera técnicas de selección de atributos para la identificación de regiones promotoras génicas (las cuales van a controlar si un gen se expresa en un momento determinado) o en la predicción de dianas de microARNs (Kim, Nam, Rhee, Lee, & Zhang, 2006).

2. Análisis de señal: enfocado en la identificación de regiones determinadas de la secuencia, como elementos estructurales o de regulación. Se utilizan técnicas de regresión para identificar regiones reguladores de determinados genes, utilizándose técnicas de selección de atributos para extraer las regiones que maximizan el ajuste al modelo de regresión (Keles, van der Laan, & Eisen, 2002).

3.2.2 SELECCIÓN DE VARIABLES EN CLASIFICACIÓN DE pre-microARNs

MicroARNs (miARNs) son secuencias de ARN compuestas por entre 21-23 nucleótidos que no codifican para proteínas, las cuales juegan un importante papel en la regulación de sus diferentes genes diana.

Participan en un elevado número de procesos biológicos, incluyendo proliferación celular, formación de órganos y en la progresión de diferentes enfermedades como el cáncer.
Debido a la dificultad en su detección mediante técnicas experimentales, se están utilizando distintas técnicas computacionales para su aislamiento (Xuan et al., 2011).
Los precursores de miARNs (pre-miARNs) presentan una secuencia en forma de horquilla de entre 60-70 nucleótidos, siendo dicha secuencia un atributo importante para identificación computacional de miARNs. El equipo de (Xuan et al., 2011) ha utilizado un algoritmo genético con el objetivo de seleccionar los atributos que mejor identifican pre-miARNs) construyendo un modelo de clasificación capaz de distinguir "auténticos" de pseudo pre-miARN.

3.2.3 SELECCIÓN DE VARIABLES EN ESPECTROMETRÍA DE MASAS

La técnica de espectrometría de masas se emplea para el análisis de proteínas en distintas muestras, permitiendo generar un perfil proteómico del cual podemos extraer información de diversos procesos biológicos o enfermedades.
Un análisis proteómico puede llegar a contener alrededor de 15000 variables, siendo utilizadas actualmente diferentes técnicas de selección de atributos para seleccionar las más informativas.
Se están utilizando diferentes métodos de filtraje (como el test-t) para realizar selección de atributos en este tipo de aproximaciones (Liu, Li, & Wong, 2002).

3.2.4 SELECCIÓN DE VARIABLES EN ANÁLISIS DE MICROARRAYS

Con la publicación de la primera secuencia del genoma humano en 2001, el campo de la bioinformática ha experimentado un rápido y exponencial crecimiento, atrayendo un considerable interés de la comunidad científica tanto en el área de las ciencias de la computación como en los campos de la biología o la medicina.

La tecnología de *microarrays* se emplea para recoger información de tejidos y muestras celulares con el objetivo de analizar diferencias de expresión en genes los cuales pueden ser útiles para el diagnóstico de enfermedades.

El análisis de este tipo de datos ha supuesto un reto para los investigadores que trabajan en algoritmos de aprendizaje durante los últimos 20 años, debido principalmente a su elevada dimensionalidad (entre 2000 y 25000 variables o genes) en oposición a un reducido número de muestras (a menudo menos de 100 pacientes).

Un problema clásico de clasificación de datos de *microarrays* consiste en el diagnóstico de pacientes sanos y enfermos de cáncer a partir de su perfil de expresión génica (problema de clasificación binaria). Sin embargo, existen también conjuntos de datos de *microarrays* en los que el principal objetivo consiste en el diagnóstico diferencial de diferentes tipos de tumores (problema multiclase), complicando de esta manera la tarea a ejecutar.

Al aparecer en este tipo de datos la ya definida maldición de la dimensionalidad, no sólo aumenta el tiempo de entrenamiento, sino que también se favorece la aparición de falsos positivos, siendo dichos falsos positivos un claro problema a evitar en el diagnóstico médico (Verónica Bolón-Canedo & Alonso-Betanzos, 2019).

Se ha demostrado además que la mayoría de los genes medidos en un experimento de *microarrays* no son fundamentales para la precisión en la clasificación, ya que un subgrupo de genes con elevada capacidad discriminatoria es capaz de asegurar el diagnóstico de la enfermedad.

La técnica basada en los *microarrays* de ADN es por tanto ampliamente utilizada para estudiar la expresión de un elevado número de genes simultáneamente.

Con la activación de un gen la maquinaria celular empieza a copiar determinadas secuencias de dicho gen, produciendo el ARN mensajero (ARNm), secuencia que se utiliza como molde para la generación de proteínas.

En la técnica de *microarrays* se extraen las moléculas de ARNm presentes en cada célula; las cuales son marcadas y usadas como molde para la generación del ADN copia (ADNc) complementario al ARNm (Figura 1).

El ADNc de células normales y controles es marcado con distintos fluoróforos, los cuales son añadidos al chip del *microarray*. Los ADNc que representan los ARNm celulares se hibridan (unen) a los ADNcs sintéticos presentes en el chip.

La hibridación de bases complementarias entre la muestra y las secuencias de genes en el chip genera una señal lumínica que es medida por un escáner, identificando de esta manera los genes expresados en la muestra.

Si un gen es muy activo va a producir una gran cantidad de moléculas de ARNm, y por tanto más ADNc marcados, los cuales van a hibridarse al ADN en el chip de microarray generando una señal fluorescente de elevada intensidad.
Genes menos activos producen una menor cantidad de ARNm, y por tanto menos ADNc, resultando en una intensidad de fluorescencia muy débil.
Si no hay señal de fluorescencia quiere decir que no ha habido unión o hibridación, y por tanto el gen no se encuentra activado.

Si co-hibridamos muestras tumorales (marcadas en rojo) con muestras normales (marcadas en verde), van a competir por los ADNc sintéticos presentes en el chip: si la señal es roja quiere decir que ese gen específico se encuentra más expresado en las células tumorales y viceversa si la señal es verde. Si la señal es amarilla dicho gen presenta niveles de expresión similares tanto en controles como en muestras tumorales. De esta manera, al determinar las sondas con mayor actividad se determina el ARNm más activo y por tanto los genes más relevantes en la muestra.

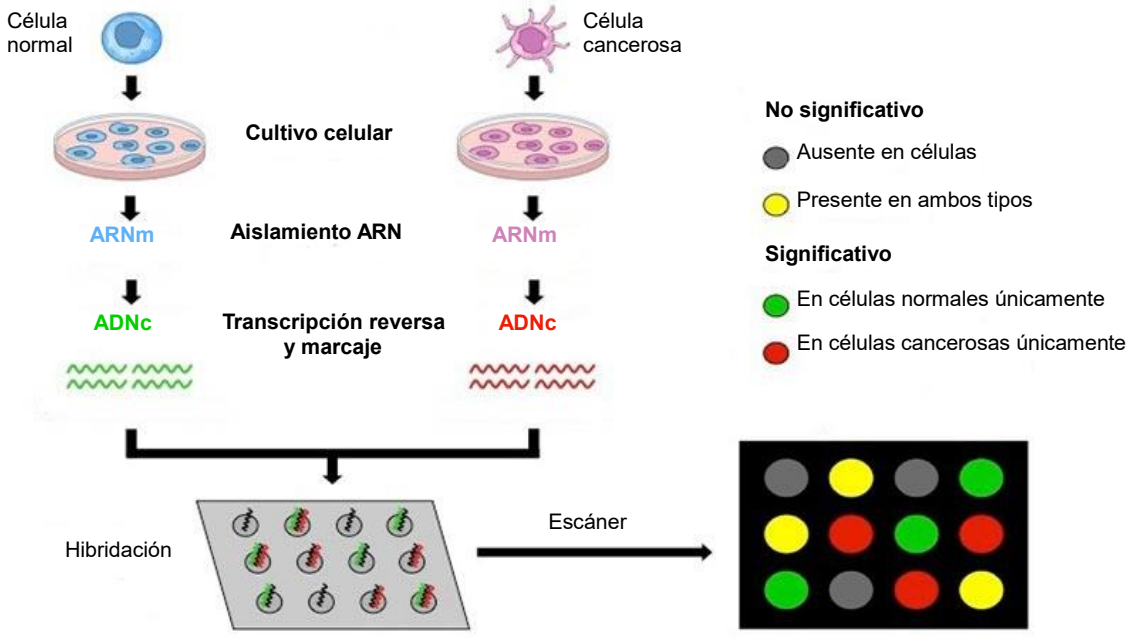

Figura 1. Diagrama representativo de un experimento de *microarrays* basado en dos colores. Adaptado de (Aryal, 2018).

La generación del perfil de expresión génica de pacientes es en la actualidad una técnica de rutina en la investigación biomédica, pero al testarse miles de sondas para cada muestra, la dimensionalidad del conjunto de datos es muy elevada, siendo este el principal problema del análisis de datos de *microarrays*.
Ya que la dimensión del conjunto de datos es muy elevada, las técnicas de clasificación son computacionalmente muy intensivas y complejas, viéndose además afectada la precisión del algoritmo de clasificación.

El cáncer es una de principales causas de muerte a nivel global, siendo responsable de más de 8 millones de muertes anuales de acuerdo con los datos obtenidos por la organización mundial de la salud en 2014.
No consiste además en una única enfermedad, existiendo más de 100 tipos de cánceres identificados en la actualidad; esperándose además que el número de nuevos casos de cáncer aumente hasta un 70% en las siguientes dos décadas (R. K. Singh & Sivabalakrishnan, 2015).
Las técnicas de selección de variables pueden aportar información relevante acerca del conjunto de datos de pacientes que estamos analizando. Si aplicamos la selección de variables a distintos tipos de cáncer y entre las variables más relevantes para el conjunto de datos aparecen por ejemplo varios genes comunes a distintos tipos de cáncer podríamos llegar a inferir la relevancia de dichos genes en la progresión de ciertos tipos de dicha enfermedad.

Se han diseñado un elevado número de experimentos de *microarrays* con el objetivo de investigar los diversos mecanismos genéticos del cáncer, habiéndose empleado en los últimos 10 años diferentes técnicas de aprendizaje automático con los siguientes objetivos:

1. Clasificación de diversos tipos de cánceres.
2. Distinción entre tejido canceroso y no canceroso.
3. Identificación de subtipos de cáncer más agresivos.

Si todos los genes son aplicados a la clasificación de tumores la eficiencia va a encontrarse muy perjudicada por la presencia de genes redundantes. Por tanto, eliminar redundancia y seleccionar genes (variables) en estos conjuntos de datos de datos puede

mejorar la eficacia de aprendizaje y la precisión en el modelo de clasificación, ayudando a mejorar el diagnóstico, tratamiento y predicción de casos de cáncer.

Por todos estos motivos el análisis de *microarrays* requiere normalmente de la utilización de técnicas de selección de atributos para reducir el conjunto de variables y mejorar la precisión del algoritmo de clasificación empleado.

Con respecto a su estructura, el conjunto de datos de un microarray es representado normalmente en forma tabular, donde cada fila representa un gen, cada columna una muestra o medida en el tiempo, y en cada entrada de la tabla se almacenan los valores de expresión medidos para cada gen.

La selección de atributos es una técnica ampliamente utilizada en el análisis de datos de *microarrays* debido principalmente a su simplicidad, pero presenta como dificultad que las interacciones y correlaciones entre variables son casi siempre ignoradas.
Para subsanar este problema se están proponiendo continuamente procedimientos más sofisticados, los cuales seleccionan los subgrupos óptimos de variables con respecto a un criterio determinado en lugar de filtrar directamente las variables a priori menos interesantes (Tabla 2).

Sin embargo, estos métodos tienden generalmente al sobreajuste; los subgrupos de variables obtenidos pueden ser óptimos para el conjunto de datos de aprendizaje, pero puede que no funcionen tan bien en un conjunto de datos independiente.
Además, suelen estar basados en algoritmos computacionalmente costosos, siendo de esta manera no muy sencillos de implementar.

Método	Tipo	Descripción
Test-t	Filtraje	Encuentra variables con la mayor diferencia del valor medio entre grupos y la mínima variabilidad dentro de cada grupo
CFS	Filtraje	Encuentra variables con elevada correlación con la clase pero no correlacionadas entre ellas
Ganancia de información	Filtraje	Mide cómo de común es una variable en una clase comparada con el resto de las clases
Algoritmos genéticos	Envoltorio	Encuentran el conjunto de variables más pequeño para el que el criterio de optimización no se ve disminuido
Búsqueda secuencial	Envoltorio	Algoritmo basado en búsqueda heurística que encuentra las variables con el criterio de optimización más elevado mediante la incorporación cada vez de una variable al conjunto de datos
SVM-RFE	Embebido	Construye un clasificador SVM eliminando variables de manera simultánea a su creación de acuerdo a su importancia
Bosques aleatorios	Embebido	Crea un número determinado de árboles de decisión utilizando distintos algoritmos para mejorar su precisión
LASSO	Embebido	Construye un modelo lineal que asigna el valor cero a un elevado número de coeficientes de variables, seleccionando aquellas con un valor distinto de cero

Tabla 2. Algunos de los métodos de selección de variables utilizados en el análisis de datos de *microarrays* y sus principales características. Adaptado de (Hira & Gillies, 2015).

En el caso de los experimentos de *microarrays*, y debido al elevado número de variables con respecto al número de muestras, los métodos de filtraje han sido muy utilizados debido a su bajo coste computacional y su baja tendencia al sobreajuste.
La selección de variables en estos conjuntos de datos se ha realizado principalmente en base a dos objetivos distintos: para le predicción de clases o para la identificación de biomarcadores.
En el primer caso se utilizan normalmente técnicas de clasificación supervisada, mientras que si lo que queremos es encontrar genes informativos, ignoramos el rendimiento, teniendo que evaluar los genes de manera individual.

3.3 CARACTERÍSTICAS PRINCIPALES DE LOS DATOS DE MICROARRAYS

3.3.1 TAMAÑO REDUCIDO DE LOS CONJUNTOS DE DATOS

Se trata del primer problema que encontramos al analizar conjuntos de datos de *microarrays*, ya que la mayoría de ellos presentan menos de 100 muestras. Dicho reducido tamaño es muy importante si tenemos en cuenta que la estimación del error se encuentra muy afectada al contar con tamaños de muestras reducidas.

Sin una estimación apropiada del error no vamos a conseguir métodos robustos de clasificación, provocando que se estén publicando un relativamente elevado número de trabajos con resultados erróneos. De esta manera, al analizar nuevamente datos de diferentes publicaciones de estudios realizados con el objetivo de predecir la prognosis de pacientes de cáncer se ha encontrado que la mayoría de ellos no eran capaces de clasificar pacientes mejor que si se hiciera de manera aleatoria (Michiels, Koscielny, & Hill, 2005).

Es por ello muy importante la utilización del método más apropiado de validación para estimar el error de clasificación.

3.3.2 CLASES NO BALANCEADAS (ASIMÉTRICAS)

Otro problema muy común en experimentos de *microarrays*; se da cuando una o varias clases presentan un número de instancias significativamente superior al resto de clases presentes en el conjunto de datos.

Normalmente, y para complicar el análisis, la clase que nos suele interesar es aquella con menos instancias. Por ejemplo, si estamos trabajando en un experimento de diagnóstico de cáncer utilizando *microarrays*, la clase cáncer suele ser normalmente aquella con menos muestras, ya que es más sencillo contar con pacientes sanos.

En estos casos, los algoritmos de aprendizaje que utilizan clasificadores tienen una predisposición hacia las clases que presentan un número superior de instancias, ya que las reglas que predicen correctamente instancias pertenecientes a la clase minoritaria se suelen ignorar (se tratan como ruido), favoreciéndose la predicción de las clases mayoritarias.

Por ello, instancias pertenecientes a clases minoritarias son a menudo clasificadas de manera incorrecta. Este problema es de especial importancia cuando el desequilibrio de clases es más acusado en el conjunto de datos de validación que en el conjunto de datos de entrenamiento. Este problema es también común en conjuntos de datos con múltiples clases presentes.

Para solventar este problema se suelen utilizar diferentes técnicas de preprocesamiento, como técnicas de muestreo que generan un subgrupo del conjunto de datos original mediante la eliminación de instancias o bien técnicas que amplían el conjunto de datos replicando instancias o creando otras nuevas a partir de las preexistentes. Existen también métodos híbridos que utilizan una combinación de ambas técnicas de muestreo.

Una de las técnicas desarrolladas para resolver el problema de desequilibrio de clases es la denominada SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), en la que la clase minoritaria es aumentada mediante la introducción de ejemplos sintéticos (artificiales). Aunque esta técnica ha sido aplicada a conjuntos de datos de *microarrays*, los autores del trabajo concluyeron que no es capaz de atenuar la predisposición en la clasificación de la clase mayoritaria de la mayoría de los clasificadores.

En los últimos años, los clasificadores de ensamblaje han sido desarrollados para dar una posible solución al desequilibrio de clases en los problemas de clasificación, siendo en muchas ocasiones combinados con técnicas de preprocesamiento como la ya citada SMOTE.

Estos algoritmos basados en métodos de ensamblaje han sido capaces de mejorar los resultados obtenidos mediante la utilización de técnicas de preprocesamiento, siendo en la actualidad ampliamente utilizados para resolver el problema de desbalanceo de clases en conjuntos de datos de *microarrays*.

3.3.3 COMPLEJIDAD DE LOS DATOS

La complejidad de los datos es una medida propuesta recientemente con el objetivo de representar características de datos considerados complejos en tareas de clasificación, como el solapamiento de clases, su divisibilidad o la linealidad de los límites de decisión. Estas medidas han sido aplicadas en análisis de expresión génica, demostrando que una baja complejidad se corresponde con un menor error de clasificación.

En particular, medidas de solapamiento de clases como F1 (porcentaje máximo de discriminación de Fisher) se enfocan en la efectividad de una única dimensión en separar las clases. Examinan el rango y dispersión de los valores del conjunto de datos en cada una de las clases y comprueban el solapamiento entre clases distintas.

3.3.4 VARIACIÓN EN LOS DATOS

Este es otro problema que se suele dar cuando el conjunto de datos es dividido en grupo de entrenamiento y de validación.
Se da cuando en el conjunto de datos de validación se produce un cambio en la distribución de un único atributo, en una combinación de ellos o en los límites de la clase. Como resultado, la presunción que tanto el conjunto de datos de entrenamiento como el de validación siguen la misma distribución, no se cumple para determinadas aplicaciones o escenarios, obstaculizando de esta manera el proceso de selección de variables y la posterior clasificación. Si la distribución del conjunto de validación difiere significativamente de la del entrenamiento y no se realiza una selección de variables correcta, ciertos clasificadores pueden llegar a asignar de manera incorrecta las instancias a cada una de las clases.
Esta variación en los datos también puede aparecer si realizamos al utilizar validación cruzada, la cual divide el conjunto de datos en diferentes subgrupos, aunque en este caso existen métodos de partición que se pueden utilizar para resolver este problema.

3.3.5 VALORES ANÓMALOS

En determinados conjuntos de datos de *microarrays* existen muestras que son asignadas (etiquetadas) de manera incorrecta o identificadas como probablemente contaminadas, las cuales han de ser designadas como valores anómalos, ya que pueden ejercer un efecto negativo en la selección de genes informativos para la clasificación de muestras.
Existen métodos especialmente desarrollados para la detección de valores anómalos, siendo de esta manera la identificación de valores anómalos un paso de preprocesamiento necesario en el análisis de microarrays si queremos realizar predicciones lo más exactas posible.

3.4 MÉTODOS DE FILTRAJE

Los métodos de filtraje se emplean generalmente para un preprocesamiento de los datos, ya que son independientes de cualquier algoritmo de aprendizaje. En este tipo de aproximación las variables son seleccionadas en función de las características del conjunto de datos sin la intervención de ningún algoritmo de aprendizaje.

Los métodos de filtraje pueden emplearse con el objetivo de evaluar variables individuales o subgrupos completos de variables. Como se discutirá a continuación, pueden ser del tipo univariante o multivariante, siendo los distintos criterios empleados clasificados como medidas de información, distancia, consistencia, similitud o estadísticas. Cada una de las medidas va a ser además más apropiada para una tarea determinada: bien para regresión, clasificación o agrupamiento (Tabla 3).

Los métodos de filtraje se basan en la estimación de propiedades intrínsecas de los datos, generando soluciones más generales y funcionando por tanto muy bien en una amplia gama de clasificadores.

Una categoría especifica de los métodos de filtraje son las aproximaciones basadas en clasificación, las cuales evalúan cada atributo individualmente. Una vez todos los atributos han sido evaluados, son ordenados en base a su puntuación. A continuación, los mejores *k* atributos son seleccionados, siendo *k* especificado por el usuario.

Aunque este tipo de aproximación es muy rápida y permite trabajar con miles de atributos, no existe un consenso para elegir la dimensión del espacio de atributos, siendo muy complicado seleccionar el número *k* de variables que van a ser seleccionadas. Y lo que es más importante, variables muy informativas al ser combinadas con otras pueden llegar a ser descartadas si presentan una correlación débil con la clase objetivo.

Los diferentes métodos de filtraje son sencillos, rápidos de implementar y no muy costosos computacionalmente, permitiendo la eliminación de variables no informativas. Es importante señalar además que presentan una capacidad de generalización bastante elevada.

A la hora de seleccionar variables hay que tener en cuenta finalmente el criterio utilizado para definir la importancia de una variable; las variables irrelevantes van a ser de esta manera aquellas que no tengan influencia en la variable objetivo o en las etiquetas de la clase en un problema de clasificación.

La metodología utilizada en las técnicas de filtraje para la selección de variables se realiza por lo general en dos pasos: en el primer paso, el método estadístico utilizado ordena las variables basándose en un criterio determinado (cada variable se ordena independientemente de las otras variables en base únicamente a su relación con la variable objetivo).

En el segundo paso las variables con el ranking más alto son utilizadas en el modelo de regresión o clasificación. La determinación del número de variables a utilizar es arbitraria, siendo normalmente elegida por el usuario (Figura 2).

Figura 2. Diagrama representativo del funcionamiento de los métodos de filtraje. Adaptado de (Vaushik, 2016).

Los métodos de filtraje pueden ser agrupados en dos clases principales: los univariante evalúan (y generan normalmente una clasificación) de una variable individual, mientras que los métodos de filtraje multivariante evalúan un subgrupo de variables completo.
Las principales diferencias entre los métodos univariante y multivariante se pueden encontrar en la siguiente tabla (Tabla 3).

Modelo de búsqueda	Ventajas	Inconvenientes	Ejemplos
Univariante	Rápidos y escalables Independientes del clasificador	Ignoran dependencia entre variables Eliminan variables con poder discriminatorio dentro de un grupo	Chi-cuadrado t-test Ganancia de información
Multivariante	Independientes del clasificador Coste computacional reducido	Más lentos y menos escalables que las univariante Pueden incluir variables redundantes	CFS FCBF

Tabla 3. Ventajas y desventajas de los métodos de filtraje. Adaptado de (Aziz, Verma, & Srivastava, 2017)

La generación de los subgrupos de variables para para filtraje multivariante va a depender de la estrategia de búsqueda empleada, las cuales suelen clasificarse en selección hacia delante, eliminación hacia atrás, selección bidireccional y selección de heurística de subgrupos de variables.

Como principal desventaja de este tipo de métodos podemos señalar que no tienen en cuenta la redundancia entre variables o la interacción entre variables (una variable puede no ser muy buena para predecir una variable, y por tanto eliminado al usar alguno de estos métodos, pero si puede serlo en combinación con otra variable).

Con respecto a su utilización en análisis bioinformáticos, y más específicamente en técnicas de *microarrays*, los métodos de filtraje son muy utilizados este tipo de análisis debido a su ya descrita sencillez y rapidez en la implementación.

Entre las distintas pruebas estadísticas utilizadas en esta categoría de selección de variables se pueden destacar las descritas en la Tabla 4.

Nombre	Tipo	Criterio	Aplicación
Ganancia de información	Univariante	Información	Clasificación
Correlación	Univariante	Estadístico	Regresión
Chi-cuadrado	Univariante	Estadístico	Clasificación
mRmR	Multivariante	Información	Clasificación, regresión
CFS	Multivariante	Estadístico	Clasificación, regresión
FCBF	Multivariante	Información	Clasificación
Relief y ReliefF	Univariante	Distancia	Clasificación, regresión
Puntuación de Fisher	Univariante	Estadístico	Clasificación
MCFS	Multivariante	Similitud	Agrupación
ReliefC	Univariante	Distancia	Agrupación

Tabla 4. Características de los principales métodos de filtraje utilizados en selección de variables. Adaptado de (Jović, Brkić, & Bogunović, n.d.).

En los siguientes apartados se describen en más detalle las principales características de los principales métodos de filtraje empleados en análisis bioinformáticos.

3.4.1 FILTRAJE BASADO EN CORRELACIÓN

La correlación es una técnica muy sencilla usada para analizar la relación entre una variable y la variable a predecir u objetivo. Se trata de una medida estadística que cuantifica la relación lineal existente entre dos variables; cuanto mayor sea la correlación, más linealmente relacionadas se encontrarán ambas variables.
De esta manera, si dos variables están correlacionadas podremos predecir una a partir de la otra.

Si dos variables predictoras se encuentran muy correlacionadas entre ellas, van a proporcionar esencialmente información redundante acerca de la variable objetivo. Para la generación de buenos modelos de aprendizaje automático vamos a buscar en general variables con una correlación muy alta con la variable objetivo pero que además no se encuentren correlacionadas entre ellas.

El coeficiente de correlación de Pearson entre dos variables se define por la siguiente fórmula:

$$r_{xy} = \frac{\sum x_i y_i - n\acute{x}\acute{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum yi}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

Los valores de dicho coeficiente de correlación de Pearson oscilan entre [-1;1], donde -1 implica una correlación negativa perfecta (cuando una variable aumenta, la otra disminuye), +1 implica correlación positiva perfecta, y 0 ausencia de correlación lineal entre las dos variables (Hall, 1999).
Basándonos en correlación podemos utilizar dos procedimientos diferentes para realizar selección de variables, el primero basado en fuerza bruta, encontrando únicamente variables correlacionadas sin tener en cuenta ninguna otra característica del conjunto de datos y el segundo basado en la identificación de 2 o más variables correlacionadas entre sí.

Se consideran por regla general variables con una correlación elevada aquellas con un coeficiente superior (en valor absoluto) a 0.8. Sin embargo, hay que tener en cuenta que este método sirve para relaciones lineales, ya que, si la relación entre las variables es

no lineal, el coeficiente de correlación de Pearson puede ser muy cercano a cero, aunque exista una correspondencia entre ambas variables.
Es por tanto preferible realizar inicialmente una gráfica (si las dimensiones del conjunto de datos lo permiten) con el objetivo de explorar la distribución de los datos. Basarnos por tanto únicamente en el valor del coeficiente de correlación para analizar la relación entre dos variables puede llevar a resultados equivocados.

3.4.2 FILTRAJE BASADO EN INFORMACIÓN MÚTUA O GANANCIA DE INFORMACIÓN

Es uno de los métodos de selección de variables más utilizados, siendo un método de filtraje univariante que genera una clasificación ordenada de todas las variables, sobre las que se aplica un umbral determinado (Battiti, 1994). Se seleccionan normalmente aquellas variables que presentan un valor más positivo.

La información mutua mide la dependencia mutua entre dos variables (investiga como de bien podemos conocer una variable a partir de lo que sabemos de otra).
Determina la similitud entre la distribución conjunta *p(X, Y)* y los productos de las distribuciones individuales *p(X)p(Y)*. Es capaz de medir la cantidad de información que la presencia o ausencia de una variable contribuye a la correcta predicción de la variable objetivo. Se define a partir de la siguiente fórmula

$$I(X;Y) = \int_X \int_Y p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy$$

El resultado es un valor positivo, igual a cero si las dos variables son independientes. Cuanto mayor sea su valor, mayor será la dependencia entre ambas variables.
Al ser una medida arbitraria y no encontrarse normalizada (su valor no se encuentra en un rango determinado), a la hora de seleccionar el umbral para la selección de variables se suelen utilizar las diez o veinte primeras en la clasificación proporcionada por la información mutua.

3.4.3 FILTRAJE BASADO EN LA PUNTUACIÓN DE FISHER

La puntuación de Fisher mide la dependencia entre dos variables, y es utilizado en conjuntos de datos con variables categóricas y problemas de clasificación binaria (han de ser además variables finitas y no continuas) (Gu, Li, & Han, 2011).

Basándose en la distribución chi-cuadrada, nos permite comparar la distribución de las clases de la variable objetivo con las categorías de la variable, devolviendo un p-valor que indica la diferencia entre ambas distribuciones.
Cuanto más reducido sea el *p*-valor, más significativa será la variable a la hora de predecir la variable objetivo.
Sin embargo, en conjunto de datos muy grandes, la gran mayoría de las variables van a tener unos *p*-valores muy pequeños, aparentando tener un gran poder predictivo, siendo dicho resultado debido al efecto del tamaño de las muestras.

3.4.4 FILTRAJE BASADO EN ANÁLISIS DE VARIANZA (ANOVA)

El análisis de varianza (ANOVA) es un método apropiado para variables continuas donde la variable objetivo es binaria; el ANOVA compara la distribución de la variable cuando la variable objetivo es 1, frente a la distribución de la variable cuando la variable objetivo es 0 (Albon, 2018).
Si las variables son categóricas se calcula una estadística chi-cuadrado entre cada variable y la variable objetivo. Si las variables son cuantitativas, se calcula el valor F entre cada variable y la variable a predecir.
El cálculo del valor F (puntuación F) mediante el ANOVA prueba la significación entre la variable objetivo y cada una de las variables (calcula si las medias para cada grupo son significativas)
Asume una relación lineal entre las variables y la variable objetivo y que los predictores se encuentran normalmente distribuidos.

3.4.5 FILTRAJE BASADO EN ReliefF

El algoritmo Relief, desarrollado por Kira y Rendell en 1992 (Kira & Rendell, 1992), utiliza una aproximación basada en métodos de filtraje para selección de atributos; siendo muy adecuado si lo que queremos es detectar interacciones entre variables.

Fue inicialmente diseñado para ser aplicado en problemas de clasificación binaria con variables discretas o numéricas. Funciona asignando una puntuación a cada variable, las cuales son a continuación ordenadas y seleccionadas en base a dicha puntuación para la selección de atributos. Dichas puntuaciones pueden ser también utilizadas como pesos de variables en procesos de modelado posteriores.

Es sistema de puntuación utilizado por Relief está basado en la identificación de diferencias entre valores de variables entre pares de vecinos cercanos de una instancia. Si el valor de diferencia entre valores de variables pertenece a un par de instancias de la misma clase (denominado "*hit*"), la puntuación de la variable disminuye, aumentando en caso contrario.

El algoritmo ReliefF (Kononenko, 1994) es una extensión introducida a partir del algoritmo Relief, el cual como ya se ha descrito funciona extrayendo una muestra al azar del conjunto de datos para localizar a continuación su vecino más cercano de la misma y opuesta clase.
Los valores de los atributos de los vecinos más cercanos son comparados con la muestra extraída y utilizados para actualizar los valores de relevancia para cada atributo.

La selección de variables basada en ReliefF funciona asignando diferentes pesos a cada una de las variables en base a su relevancia a la categoría.
La intuición detrás del funcionamiento del algoritmo es que un atributo que sea importante en el conjunto de datos debería diferenciarse entre muestras de diferentes clases y tener el mismo valor para muestras dentro de la misma clase.

El algoritmo ReliefF añade a estas características la capacidad de tratar con problemas multiclase, siendo además más robusto. Puede ser implementado además en conjuntos de datos "ruidosos" o incompletos.

Este método puede ser aplicado prácticamente en cualquier tipo de problema, presenta una variación muy baja y es capaz de detectar dependencias entre variables que otros métodos no son capaces de identificar.

3.4.6 FILTRAJE BASADO EN TIPO DE VARIABLES

También podemos implementar otros métodos de selección de variables basados en el tipo de variables que nos podamos encontrar en el conjunto de datos: eliminación de variables constantes, cuasi-constantes o duplicadas.

- Variables constantes: son aquellas que presentan un único valor (el mismo) para todas las observaciones del conjunto de datos. Estas variables no proporcionan información que permita al modelo de aprendizaje automático la predicción o selección de una variable objetivo.
- Variables cuasi-constantes: un único valor es compartido por la gran mayoría de observaciones en el conjunto de datos. Se suelen eliminar variables que presentan el mismo valor para alrededor del 99% de las observaciones
- Variables duplicadas: las podemos definir como aquellas que presentan el mismo valor para todas las observaciones en el conjunto de datos.

3.5 MÉTODOS ENVOLTORIO

Los métodos envoltorio incorporan un algoritmo de aprendizaje de manera similar a una caja negra, utilizando la capacidad predictiva del algoritmo para evaluar la utilidad relativa de un subgrupo de variables.

El algoritmo de selección de variables utiliza el método de aprendizaje como una subrutina, con el coste computacional que conlleva la utilización del algoritmo de aprendizaje para evaluar cada subgrupo de variables. A pesar de ello, suelen presentar un mejor rendimiento que los métodos de filtraje.

Los métodos envoltorio son también conocidos como algoritmos avariciosos de búsqueda, considerando de esta manera la selección de un conjunto de variables como un problema de búsqueda. Son avariciosos porque realizan una búsqueda entre todas las combinaciones posibles de variables para determinar aquella que produce el algoritmo de aprendizaje con el mejor rendimiento (Figura 3).

La ventaja de este tipo de métodos consiste en que garantizan la selección del mejor conjunto de variables, y su utilización junto a validación cruzada garantiza en general un modelo muy robusto al sobreajuste (Tabla 5).

Su principal desventaja radica en que encontrar testar todas las posibles combinaciones de variables es computacionalmente muy costoso.

Utilizan un clasificador específico para seleccionar el conjunto de variables, con la ventaja que este método va a seleccionar el mejor conjunto de variables para maximizar el rendimiento de ese algoritmo determinado de aprendizaje. Sin embargo, dicho conjunto de variables puede no ser el óptimo si decidimos utilizar un algoritmo de aprendizaje diferente.

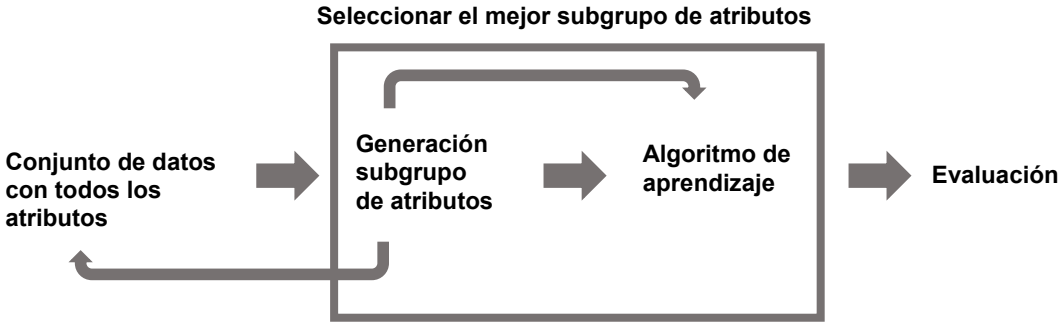

Figura 3. Diagrama representativo del funcionamiento de los métodos envoltorio. Adaptado de (Vaushik, 2016).

Estos métodos añaden o eliminan una variable cada vez, y dicha variable es seleccionada en base únicamente al rendimiento del clasificador. El algoritmo continúa añadiendo o eliminando variables hasta que se alcanza un número de variables determinado u otro criterio predefinido previamente.

Modelo de búsqueda	**Ventajas**	**Inconvenientes**	**Ejemplos**
Algoritmos de selección secuencial	Interacción sencilla con el clasificador Consideran dependencia entre variables	Riesgo de sobreajuste Pueden quedar atrapados en un óptimo local Dependen del clasificador	SFS SBE
Algoritmos de selección evolutiva	Interaccionan con el clasificador Modelan dependencia entre variables Mayor precisión que los métodos de filtraje	Coste computacional elevado Dependen del clasificador Riesgo elevado de sobre ajuste	Algoritmos genéticos Optimización de colonia de hormigas Colonia de abejas artificial

Tabla 5. Ventajas y desventajas de los métodos envoltorio. Adaptado de (Aziz et al., 2017).

Existen los siguientes tipos principales de métodos envoltorio:

3.5.1 SELECCIÓN HACIA DELANTE (SFS)

En este tipo de modelos de búsqueda comenzamos sin tener ninguna variable en el modelo y en cada iteración seguimos agregando variables que mejoran nuestro modelo hasta que la adición de una nueva variable no mejora el rendimiento de este (Pudil, Novovičová, & Kittler, 1994).
Se utilizan por tanto con el objetivo de reducir un espacio con *d* variables hasta un espacio con *k* variables, donde $k < d$.
Como criterio establecido para evaluar el rendimiento se pueden utilizar por ejemplo el área bajo la curva para clasificación o regresión r cuadrada para regresión.

En este tipo de implementaciones el criterio de parada es un número de variables arbitrario, terminándose la búsqueda al llegar a dicho número de variables.

3.5.2 ELIMINACIÓN HACIA ATRÁS (SBE)

Para su implementación partimos del conjunto de datos con todas las variables y vamos eliminando aquella variable que menos contribuye al rendimiento del clasificador en cada iteración hasta que no observamos ninguna mejora al eliminar otra variable.
Se utilizan con el objetivo de reducir un espacio con *d* variables hasta un espacio con *k* variables, donde $k < d$.
Se van a evaluar primero todas las *n* variables, a continuación, las *n-1*, *n-2* y así sucesivamente todas las combinaciones de variables.
Como el método anterior es computacionalmente muy costoso, y si el número de variables es muy grande su uso puede no ser posible.
Al igual que en los métodos descritos anteriormente, como criterio establecido para evaluar el rendimiento se pueden usar por ejemplo la métrica área bajo la curva para clasificación o la regresión r cuadrada para regresión.

3.5.3 ALGORITMOS GÉNETICOS

Los algoritmos genéticos (Holland, 1975) representan un grupo de metaheurísticas de optimización que han conseguido resultados muy satisfactorios en multitud de aplicaciones, incluyendo la selección de atributos.

El siguiente diagrama muestra su utilización para la selección de atributos (Figura 4

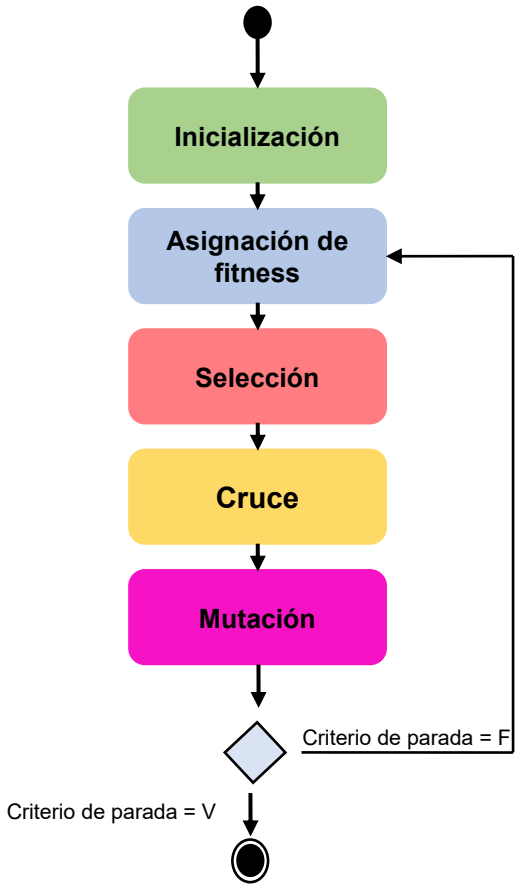

Figura 4. Algoritmos genéticos en selección de variables. Adaptado de (Gandhi, 2018)

En un algoritmo genético la población se compone de cromosomas, los cuales se encuentran generalmente codificados en forma de una secuencia binaria donde cada dígito representa un gen.

En cada iteración se asignan una serie de operaciones a cada cromosoma, con el objetivo de mejorar las cualidades de los individuos que van a formar la siguiente población (generación).
Se utilizan los siguientes operadores consecutivamente en cada iteración hasta llegar al criterio de parada:

- Selección
- Cruce
- Mutación

En el caso de los experimentos de *microarrays*, los métodos envoltorio no han sido muy utilizados, debido principalmente su elevado coste computacional.
Al aumentar el número de variables va a aumentar exponencialmente el espacio de los subgrupos de variables, resultando un problema muy complejo al estar trabajando con miles de variables, como es el caso de los *microarrays*.
Además, dichos métodos presentan un elevado riesgo de sobreajuste debido al reducido tamaño de las muestras en dichos experimentos.

3.5.4 BORUTA

El algoritmo Boruta (Kursa, Jankowski, & Rudnicki, 2010) se encuentra construido alrededor del algoritmo de clasificación de bosques aleatorios, teniendo como principal objetivo la captura de variables importantes con respecto a una variable objetivo.
Aunque podríamos utilizar test univariante para la selección de variables, como se ha señalado dichos test asumen la independencia entre las variables, y en general los conjuntos de datos (incluyendo aquellos utilizados en biomedicina) presentan relaciones, bien conocidas u ocultas entre las variables del conjunto de datos.
Podríamos entonces utilizar métodos multivariante, ajustando de esta forma nuestro modelo obteniendo los pesos asignados a cada una de nuestras variables. De esta manera podemos eliminar recursivamente las variables que no mejoran nuestro modelo en cada iteración hasta alcanzar un criterio de parada que hemos predefinido.
Este tipo de aproximación es sin duda más lento y computacionalmente más costoso, pero presenta la ventaja de ser multivariante, ponderando las relaciones entre las variables, seleccionando finalmente aquel conjunto de variables que son capaces de explicar en su conjunto la mayoría de la varianza en el conjunto de datos. Presenta sin

embargo el problema de la arbitrariedad en el criterio de parada, o la decisión de cuántas variables eliminar en cada iteración.
Podríamos incluir validación cruzada para parar de eliminar variables al llegar a la situación óptima, pero este método no deja de consistir en la maximización del rendimiento de un regresor o un clasificador mediante el uso de un conjunto de datos al que hemos eliminado un elevado número de variables. Podemos, por tanto, haber eliminado un número excesivo de variables importantes en nuestros datos.

La principal aportación de Boruta en este campo consiste en que trata de capturar todas aquellas variables importantes o interesantes que tengamos en nuestros datos con respeto a una variable objetivo.
Su utilización resulta especialmente apropiada para el análisis de datos biomédicos, donde se recogen regularmente medidas de miles de variables (como genes, proteínas o metabolitos) y no tenemos además ninguno tipo de información acerca de su importancia con respecto a la variable objetivo, o donde debemos establecer el límite de importancia de dichas variables.

La forma de funcionamiento del algoritmo Boruta es la siguiente:

1. Primero duplicamos el conjunto de datos, introduciendo aleatoriedad en las variables con la generación de variables duplicadas, repartiendo dichos valores en cada columna. Estas variables se denominan variables sombra (*shadow*).
2. Entrenamos un clasificador en el conjunto de datos y calculamos la importancia de cada variable. Cuanto mayor sea su valor de relativa importancia, mejor o más importante será la variable. Los métodos de ensamblaje basados en árboles, como los clasificadores de bosques aleatorios o *Gradient Boosted Trees* funcionan muy bien, no solo porque son capaces de capturar complejas relaciones no lineales entre las variables, sino que además son muy apropiados en casos donde el número de variables excede el de muestras. Además, está generalmente aceptado que tienden menos al sobreajuste que otro tipo de algoritmos de entrenamiento.
3. Una vez entrenado nuestro conjunto de datos con las variables duplicadas, el algoritmo comprueba, para cada una de las variables

reales si presentan una importancia superior que la mejor de las variables sombra. Lo que estamos comparando es si la variable presenta una puntuación-Z (número de desviaciones estándar de la media) superior al de la puntuación-Z máxima de su variable sombra. Si la importancia es superior, la variable es almacenada en un vector ("*hit*") y continuamos con la siguiente iteración hasta llegamos a un conjunto predeterminado de iteraciones.

En cada iteración se comprueba por tanto si una variable se comporta mejor de lo esperado al azar. Utilizamos para ello la distribución binomial para comparar el número de veces que una variable se comportó mejor que las variables sombra.

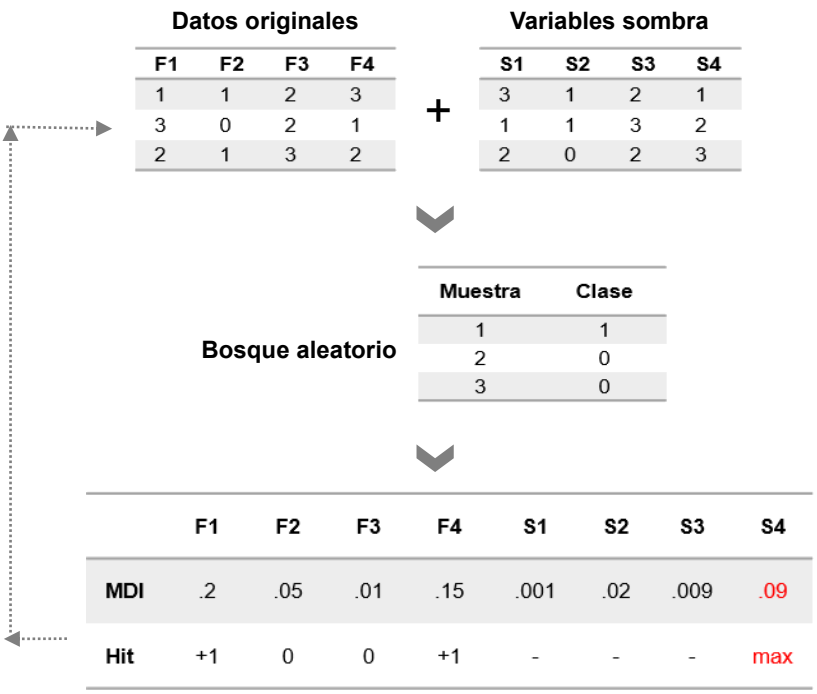

Figura 4. Forma de funcionamiento del algoritmo Boruta. Adaptado de (Homola, 2015)

3.5.5 ELIMINACIÓN RECURSIVA DE ATRIBUTOS (RFE)

La eliminación recursiva de atributos se basa en construir repetidamente un modelo y elegir la variable más eficiente o la menos eficiente, eliminar dicha variable y repetir el proceso con el resto de las variables. El proceso se repite hasta que se han apartado todas las variables del conjunto de datos.

Las variables son entonces ordenadas de acuerdo con la posición en la que han sido eliminadas. Es por tanto un proceso de optimización avaricioso que pretende encontrar el subgrupo de variables con la mayor eficiencia.

La estabilidad de este tipo de métodos depende del tipo de modelo que se utilice para la ordenación de variables en cada iteración.

3.6 MÉTODOS EMBEBIDOS

Estos métodos ejecutan la selección de variables en el proceso de entrenamiento, siendo normalmente específicos de determinados procesos de aprendizaje.
La búsqueda del subgrupo óptimo de variables se encuentra integrada en la construcción del clasificador, pudiendo ser definida como una búsqueda el espacio compuesto por los subgrupos de variables y las hipótesis.

Los métodos embebidos combinan características de los métodos de filtraje y envoltorio, aprendiendo por tanto las variables que mejor contribuyen a la precisión del modelo de manera simultánea a su creación (Figura 5).

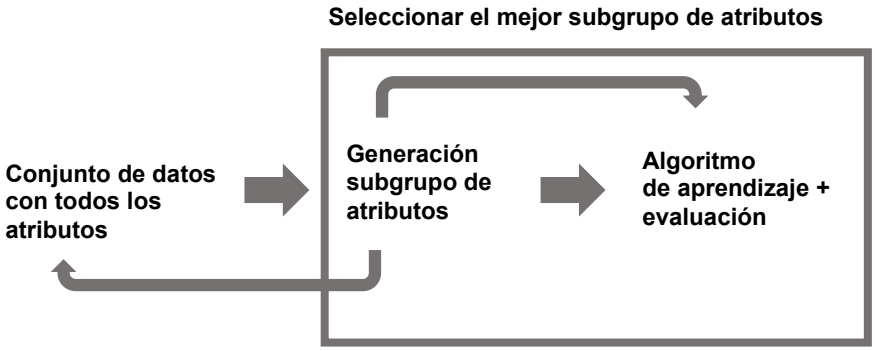

Figura 5. Diagrama representativo del funcionamiento de los métodos embebidos. Adaptado de (Vaushik, 2016).

Una de las principales ventajas de este tipo de aproximaciones consiste en que estos métodos son capaces de capturar dependencias con un coste computacional menor que los métodos envoltorio (Tabla 6).

Modelo de búsqueda	Ventajas	Inconvenientes	Ejemplos
Embebidos	Interaccionan con el clasificador Mayor rendimiento y precisión que los métodos de filtraje Modelan la relación entre variables con menor coste computacional que los envoltorio Menor tendencia al sobreajuste que los métodos envoltorio	Selección de variables dependiente del clasificador	Árboles de decisión SVM-RFE Bosques aleatorios LASSO

Tabla 6. Ventajas y desventajas de los métodos embebidos. Adaptado de (Aziz et al., 2017).

3.6.1 REGULARIZACIÓN

Son métodos basados en modelos de regularización con funciones objetivas que minimizan los errores de ajuste, obligando a los coeficientes de las variables a ser lo más cercanos posibles a cero.
La regularización consiste en añadir una penalización a diferentes parámetros del modelo para reducir la libertad del mismo. El modelo tendrá entonces menor tendencia a ajustar ruido del conjunto de entrenamiento.
Al aplicar una penalización a los coeficientes que multiplican a cada uno de los predictores en el modelo lineal conseguimos evitar el sobreajuste. De esta manera, el modelo tendrá menor probabilidad de ajustarse a el ruido del conjunto de entrenamiento y mejorará la capacidad de generalización de este.

Para modelos lineales existen en general tres tipos de regularización:

- Regularización L1 (Lasso)
- Regularización L2 (Ridge)
- Regularización L1/L2 (Red elástica)

Por lo general, cuanto mayor sea la penalización, mayor será la generalización (se incrementa penalización según va aumentando la complejidad del modelo). Sin embargo, si la penalización es muy elevada el modelo puede perder poder predictivo.

El algoritmo de aprendizaje tiene como objetivo minimizar la diferencia entre el valor predicho y el valor real de la observación durante el ajuste del modelo, a los que hay que sumar el componente de regularización.
Dicho componente de regularización es una penalización a los coeficientes que el modelo lineal añade a las variables, tal y como podemos observar en la función de coste de regularización:

$$\frac{1}{2m}\sum(y - ypred)^2 + \lambda\sum\phi$$

Si queremos minimizar la ecuación, al incrementar lambda (al aumentar la penalización) tenemos que disminuir los coeficientes.

3.6.1.1 LASSO

El método LASSO (*least absolute shrinkage and selection operator*, por sus siglas en inglés) fue desarrollado por Robert Tibshirani en 1996 (Tibshirani, 1996) para ser aplicado en modelos de regresión lineal.
Su implementación consigue una mayor exactitud e interpretabilidad del modelo reduciendo el valor de la mayoría de los coeficientes a cero.
Funciona utilizando el término de penalización para reducir los coeficientes de determinadas variables a cero con el objetivo de realizar selección de variables.
Sin embargo, al ser utilizado en conjuntos de datos multidimensionales con un número reducido de muestras, donde existe un número elevado de información ruidosa, el método realiza el mismo grado de reducción en todas las variables. De esta manera, las variables redundantes son reducidas a cero, y también lo son variables relacionadas a estas, generándose una estimación tendenciosa.
Tibshirani demostró que LASSO es más estable y exacto que métodos tradicionales en selección de variables, como el mínimo cuadrado parcial, la selección de variables de subgrupos o la regresión ridge.
LASSO funciona combinando la parsimonia y estabilidad de predicción de los dos métodos principales de penalización, la regresión ridge y la regresión de subgrupo, los cuales fueron desarrollados con el objetivo de mejorar las deficiencias en los estimadores de la regresión ordinaria de mínimos cuadrados (OLS).

La regularización LASSO permite reducir algunos de los coeficientes a cero, permitiendo que alguno de los theta en la fórmula sean cero. El regularizador LASSO fuerza los pesos de muchas variables al valor cero.
De esta manera, una determinada variable será multiplicada por cero a la hora de estimar la variable objetivo, no contribuyendo a la predicción final. Por tanto, dicha variable puede ser eliminada al no contribuir a la predicción.

3.6.1.2 RIDGE

La regresión ridge (Hoerl & Kennard, 1970) reduce los coeficientes continuamente hasta cero, pero mantiene todos los predictores en el modelo. Permite por tanto mejorar los errores de predicción al reducir en tamaño los coeficientes de regresión que sean demasiado grades para reducir el sobre ajuste, pero no realiza selección de variables y por tanto no produce un modelo más interpretable.

Añade el cuadrado de la magnitud de los coeficientes como término de penalización a la función de coste:

$$\frac{1}{2m}\sum(y - pred)^2 + \lambda\sum\phi^2$$

De esta manera, la regularización Ridge se diferencia de la Lasso en que lo que trata de minimizar es el cuadrado de theta.
Al aumentar el valor de lambda los coeficientes de regresión se aproximan a cero (pero no llegan nunca a ser cero). Por lo tanto, este tipo de regularización no es apropiada para la selección de variables, pero sí para la optimización del modelo.
Es importante indicar que, al aumentar la penalización, el número de variables eliminadas va a aumentar (no es por tanto aconsejable utilizar una penalización muy alta o baja).
De todas formas, si la penalización es muy alta y variables importantes son eliminadas, vamos a observar un claro descenso en el rendimiento del algoritmo, lo cual nos llevará inmediatamente a darnos cuenta de que debemos disminuir la regularización.

3.6.2 ARBOLES DE DECISIÓN Y BOSQUES ALEATORIOS

Los árboles de decisión son algoritmos de aprendizaje que pueden ser utilizados tanto en clasificación como en regresión, siendo ampliamente utilizados debido principalmente a los siguientes motivos: presentan una capacidad predictiva muy elevada, bajo sobreajuste y sencilla interpretabilidad.
Su interpretabilidad viene dada por la sencillez en derivar la importancia de cada variable en el árbol de decisión (es relativamente simple cuantificar la contribución de cada variable en la decisión).

Los algoritmos de bosques aleatorios (*Random forests* en inglés) suelen estar formados por cientos de árboles de decisión construidos a partir de observaciones y variables aleatorias del conjunto de datos. Al contener un subgrupo de variables y observaciones se garantiza que los árboles no se encuentran correlacionados y son por tanto menos propensos al sobreajuste.
Funcionan dividiendo el conjunto de datos de forma jerárquica en compartimentos cada vez más pequeños, hasta llegar a compartimentos donde las observaciones son todas de la misma clase.
En cada nodo del árbol, el conjunto de datos se divide en dos compartimentos, albergando cada uno de ellos observaciones que son más similares entre ellas que con respecto al otro compartimento (Figura 6).

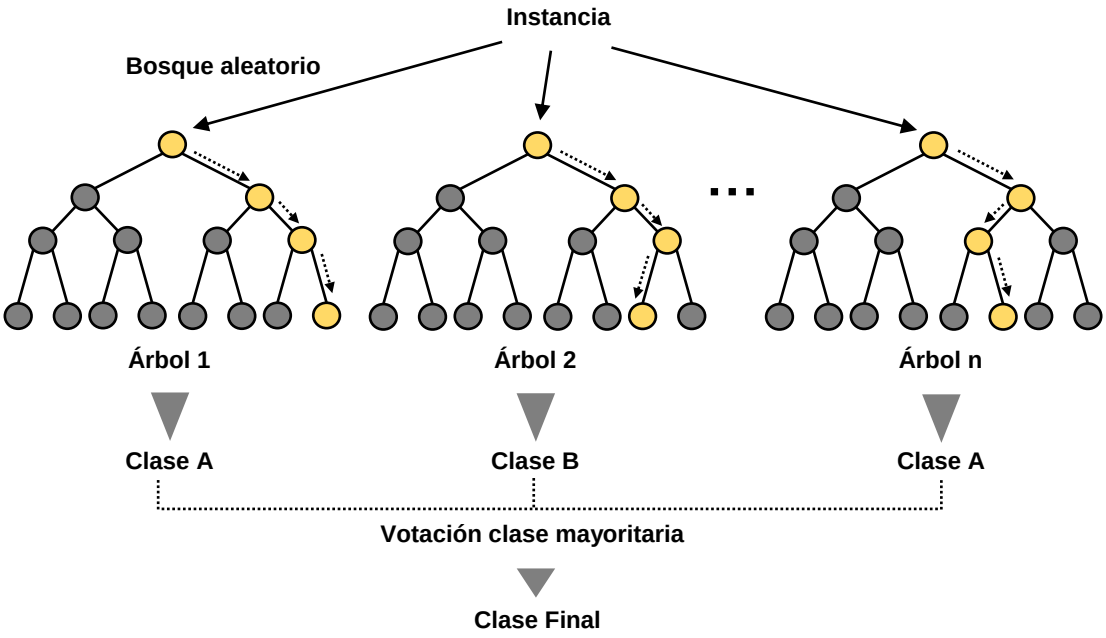

Figura 6. Diagrama del funcionamiento de los bosques aleatorios. Adaptado de (Koehrsen, 2017)

La importancia de cada variable se deriva de la pureza de cada uno de los compartimentos, siendo la disminución de la impureza para cada variable promediada a lo largo de cada uno de los árboles para determinar la importancia de la variable.
En problemas de clasificación la medida de la impureza (en cada decisión de dividirse un nodo) se determina mediante la impureza gini o la ganancia de información. Para regresión se utiliza la varianza como medida de impureza.

Por tanto, al realizar el entrenamiento de un árbol podemos cuantificar la disminución de impureza a la que contribuye cada variable. Cuanto más disminuya la impureza una variable más relevante será.
La contribución a la impureza de cada variable puede ser promediada a lo largo de todos los distintos árboles de decisión para determinar la importancia final de cada atributo.

Con respecto a las limitaciones de la utilización de árboles aleatorios se puede destacar:

- Variables correlacionadas van a tener igual o similar importancia
- La importancia de variables correlacionadas es menor que la importancia real
- Los árboles tienden al sobreajuste con la existencia de variables categóricas con muchas categorías (mostrarán estas variables como importantes cuando en realidad interfieren en la generalización del modelo).

El procedimiento a seguir a la hora de realizar selección de variables al utilizar bosques aleatorios (RF) sería el siguiente:

1) Construir un RF.
2) Determinar la importancia de cada variable de acuerdo con el RF.
3) Seleccionar las variables con mayor importancia.

Como alternativa podemos utilizar además la eliminación recursiva de variables de la siguiente manera:

1) Construir un RF.
2) Determinar la importancia de cada variable de acuerdo con el RF.
3) Eliminar la variable menos importante.

4) Repetir la construcción del RF con el resto de las variables hasta cumplir un criterio determinado

Este último método que utiliza eliminación recursiva de variables va a funcionar mejor cuando tenemos un número muy elevado de variables correlacionadas a seleccionar. Si la variable eliminada se encuentra correlacionada con otra variable en el conjunto de datos, la importancia de la variable con la que está correlacionada aumentará.

Existen además distintos algoritmos que se pueden implementar para la creación de los árboles de decisión:

- **ID3**: Únicamente puede utilizarse con atributos categóricos; selecciona el mejor atributo en cada nodo en función del valor de la métrica ganancia de información. Una vez construido el árbol los nodos que pueden generar un sobreajuste del modelo son eliminados (podados).

- **C4.5**: Admite además tributos numéricos. Para crear las aristas en un nodo que contiene un atributo numérico, C4.5 crea automáticamente una serie de puntos de corte en el rango de valores numéricos del atributo. C4.5 divide entonces el conjunto de datos en aquellos trozos cuyos valores en el atributo se encuentran en cada intervalo resultante de la división por dichos puntos de corte. Se extraen todas las reglas de decisión posibles del árbol construido y se ordenan por su eficacia, podando aquellas reglas que empeoran la eficacia global del árbol de decisión.

- **CART**: Produce árboles binarios, pudiendo ser usado tanto en regresión como en clasificación.

Resolver el problema de selección de atributos en aplicaciones de bioinformática (análisis de *microarrays*) presenta numerosas ventajas ya que determinadas variables pueden ser muy costosas (o innecesarias) de medir, procesar e incluso almacenar dependiendo del proceso biológico que estemos estudiando.
Por ejemplo, y tal y como se ha descrito, la selección de atributos puede utilizarse para disminuir el coste de aplicación de un determinado modelo de diagnóstico disminuyendo el número de genes que son medidos.

3.6.3 MÁQUINAS DE SOPORTE VECTORIAL BASADAS EN ELIMINACIÓN RECURSIVA DE VARIABLES (SVM-RFE)

Una máquina de soporte vectorial (SVM) es un algoritmo de aprendizaje supervisado muy eficiente en la construcción de un clasificador.
Su principal objetivo consiste en la creación de una frontera de decisión entre clases, permitiendo la predicción de etiquetas de uno o más vectores de atributos.

Esta frontera de decisión, conocida como hiperplano, se encuentra orientada de tal manera que se encuentra lo más alejada posible de punto más cercano de cada una de las clases.
Los puntos más cercanos se denominan vectores de soporte. Dichos vectores de soporte son los puntos que definen el hiperplano, y la regla de clasificación de la SVM depende ellos.

El modelo de SVM representa el conjunto de datos en el espacio y los divide de acuerdo con las diferentes clases mediante el uso del hiperplano más adecuado (Figura 7).
Nuevos puntos son clasificados en base a dicho hiperplano y valores predichos son comparados con los reales para evaluar la precisión del modelo.

Dado un conjunto de entrenamiento, su hiperplano óptimo viene determinado por $wx^T + b = 0$. El objetivo de entrenar una SVM consiste en encontrar los valores de w y b que hacen que el hiperplano separe los datos y maximice 1/ ||w||2.

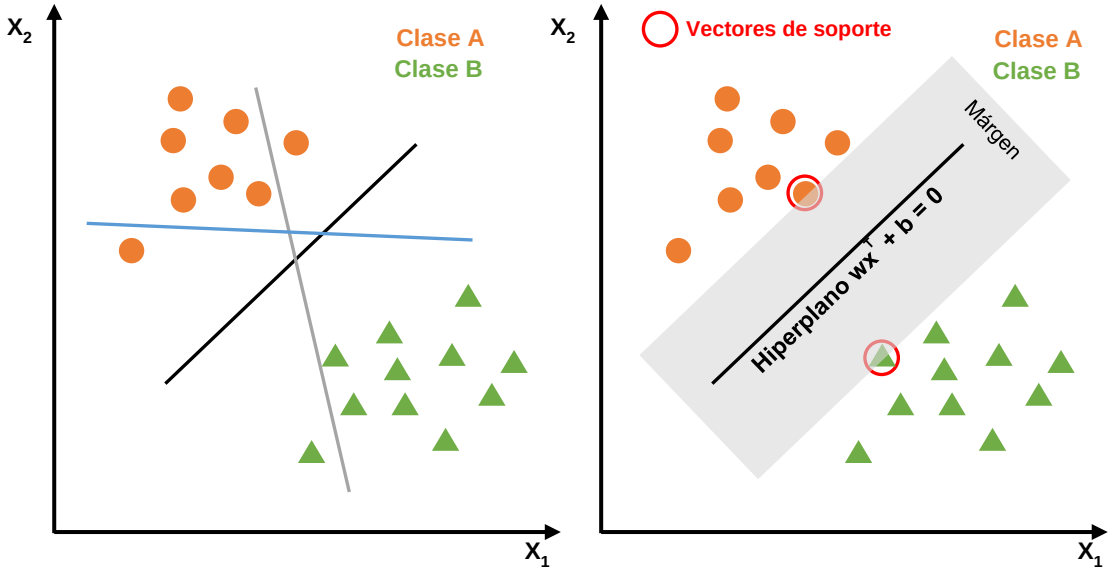

Figura 7. Algoritmos de máquinas de soporte vectorial. Adaptado de (Huang et al., 2018)

El algoritmo SVM fue propuesto inicialmente para construir un clasificador lineal por Vapnik en 1963 (Vapnik, 1963); posteriormente se propuesto alternativas utilizando el denominado método del kernel, el cual permite modelar modelos no lineales de mayores dimensiones.

En un problema no lineal, una función basada en kernel puede ser utilizada para añadir dimensiones adicionales a los datos, creando un problema lineal en un espacio multidimensional.

El objetivo de la eliminación recursiva de variables (RFE) consiste en la selección de variables considerando recursivamente conjuntos de variables cada vez más reducidos. Esta estrategia asume que las variables se encuentran normalizadas o estandarizadas previamente.

El concepto de RFE está basado en entrenar un modelo que contenga determinados parámetros (también llamados pesos o coeficientes), como regresión lineal o máquinas de soporte vectorial de manera repetitiva.

La primera vez que entrenamos el modelo incluimos todas las variables; cuando encontramos la variable con el parámetro más pequeño (indicando que es la menos importante), dicha variable es eliminada del conjunto de datos. Dicho proceso es repetido de manera recursiva en el conjunto de datos que está siendo reducido hasta que el número de variables deseadas es alcanzado.

Funcionan por lo tanto testando todas las posibles combinaciones de variables, desde una variable hasta *N*, siendo *N* el número de variables. Testan todas las variables individualmente para a continuación testar todas las posibles combinaciones de dos variables, tres variables, y así sucesivamente hasta que se alcanzan N variables.

Se selecciona entonces el mejor subconjunto de variables a partir de todos los posibles subconjuntos de variables mediante la optimización de una métrica para un algoritmo de aprendizaje determinado. Por ejemplo, si el clasificador es regresión logística y el conjunto de datos contiene 3 variables, el algoritmo evaluará las siguientes combinaciones:

- Todas las posibles combinaciones con una variable
- Todas las posibles combinaciones con dos variables
- Las tres variables juntas

Finalmente, seleccionará aquella que genera el mejor rendimiento (por ejemplo, la precisión del clasificador) del clasificador de regresión logística.
Como los métodos anteriores, es una técnica computacionalmente muy costosa, y si el número de variables es muy grande su uso puede no ser posible

Basándose en las SVM se ha propuesto (Guyon, Weston, Barnhill, & Vapnik, 2002) un método de una máquina de soporte vectorial basado en la eliminación recursiva de atributos (SVM-RFE) para su utilización en selección de variables.
Realiza selección de variables entrenando iterativamente un clasificador SVM con un grupo determinado de variables y eliminando la variable menos importante determinada por el SVM.

El algoritmo SVM-RFE entrena una SVM con un kernel lineal y elimina el atributo con el menor valor *w d*el hiperplano de decisión dado por la SVM entrenada.

La idea es que la orientación del hiperplano que separa las clases modelado por la SVM puede ser utilizada para la selección de variables informativas. De esta manera, si el plano es ortogonal a la dimensión de una variable determinada, entonces dicha variable va a ser relevante (y viceversa).

El método SVM-RFE puede entonces eliminar las variables menos informativas y seleccionar las más relevantes basándose en los pesos de los clasificadores.

Por poner un ejemplo relacionado con la selección de atributos en conjuntos de datos bioinformáticos, se han utilizado SVM-RFE para seleccionar atributos con el objetivo de predecir proteínas esenciales (aquellas vitales para la supervivencia celular). El método ha consistido en la eliminación de atributos que compartían propiedades biológicas con otros atributos de acuerdo con el coeficiente de correlación de Pearson (Zhong, Wang, Peng, Zhang, & Li, 2015).

Su implementación en conjuntos de datos de *microarrays* ha tenido como resultado menores tiempos de ejecución y mejor clasificación que otros; habiendo sido utilizado también en problemas de clasificación multiclase.

3.7 OTROS ALGORITMOS USADOS PARA SELECCIÓN DE VARIABLES EN MICROARRAYS-TENDENCIAS

La tendencia actual consiste en la utilización no solo de los métodos clásicos ya mencionados para la selección de variables, sino también en el empleo de nuevas aproximaciones, como los métodos híbridos, de ensamblaje o de agrupación.

Los métodos híbridos combinan normalmente dos o más algoritmos de selección de atributos de manera secuencial. De esta manera, se ha propuesto una aproximación que incorpora un filtro mRMR (mínima redundancia-máxima relevancia) basado en información mutua en SVM-RFE para minimizar la redundancia entre los genes seleccionados. Dicha aproximación ha sido capaz de mejorar la exactitud del clasificador utilizando un menor número de genes comparado con ambos métodos en solitario.

Leung y Hung (Yukyee Leung & Yeungsam Hung, 2010) propusieron un método de filtros y envoltorios múltiples (MFMW). La idea consiste en que los métodos de filtraje son muy rápidos, pero sus predicciones son inexactas, mientras que los envoltorio maximizan la exactitud del clasificador a expensas de un elevado coste computacional. El algoritmo MFMW está basado en aproximaciones híbridas ya descritas que maximizan la exactitud de un clasificador determinado con respecto a un conjunto de genes filtrado.

La desventaja principal de los métodos híbridos que combinan filtraje y envoltorio consiste en que la exactitud del clasificador depende de la elección del método de filtraje y envoltorio seleccionados. MFMW resuelve dicho problema haciendo uso de múltiples filtros y envoltorios para mejorar la exactitud y solidez en la clasificación.

Los métodos de selección de variables basados en ensamblaje se basan en la suposición que combinar el resultado de múltiples test va a ser mejor que la utilización de un único test.
Aunque los métodos de aprendizaje por ensamblaje han sido aplicados principalmente a clasificación, ha sido recientemente cuando se han empezado a aplicar a la selección de genes en conjuntos de datos de *microarrays*.
De esta manera, ha sido por ejemplo propuesto un ensamblaje de filtros, a partir de la observación de la variabilidad de resultados proporcionados por diferentes filtros al ser aplicados a distintos conjuntos de datos de *microarrays*.
De esta manera, un determinado filtro podía arrojar resultados muy buenos en la clasificación de un determinado conjunto de datos, y a la vez proporcionar muy malos resultados en otro conjunto de datos, aun perteneciendo ambos al mismo dominio.
El ensamblaje propuesto obtiene una predicción de clasificación para cada uno de los diferentes filtros que componen el ensamblaje, y combina finalmente todas las predicciones mediante votación. Experimentos realizados en 10 conjuntos de *microarrays* distintos mostraron que el método de ensamblaje obtenía el error de clasificación más bajo para cada uno de los cuatro clasificadores testados.

Métodos de agrupamiento (*clustering*) para *microarrays* han sido también propuestos recientemente. La mayoría de las técnicas de selección de genes se basan en la presunción de la independencia entre genes (como se ha detallado, una aproximación clásica se basa en la ordenación de los genes individualmente).
Sin embargo, es conocido que los genes interactúan unos con otros mediante redes de regulación génica. Para abordar esta situación, Lovato et al han presentado un esquema de selección de variables basado en el modelo denominado *Counting Grid* (Lovato, Bicego, Cristani, Jojic, & Perina, 2012), el cual puede medir y tener en cuenta la relación e influencia de los genes entre ellos.

3.7.1 ESTABILIDAD DE LOS MÉTODOS DE SELECCIÓN DE ATRIBUTOS

Dependiendo de cada aplicación particular, distintos algoritmos de selección de atributos pueden ser aplicados, siendo seleccionado aquel que mejor cumpla determinados criterios; aunque bien es cierto que un problema que se suele pasar por alto es la estabilidad de los algoritmos de selección de atributos que utilicemos (Chandrashekar & Sahin, 2014).

La estabilidad de un algoritmo de selección de variables puede definirse como la consistencia del algoritmo en la generación de un robusto subgrupo de variables cuando nuevas muestras de entrenamiento son incorporadas, o cuando algunas de ellas son eliminadas.

Si el algoritmo selecciona un subgrupo de variables diferentes en cada perturbación creada en el conjunto de datos original, dicho algoritmo no resulta consistente para realizar la selección de variables.

Se han realizado estudios donde se introduce inestabilidad en el conjunto de datos, generada tras aplicar un algoritmo de entrenamiento determinado. Al emplearse técnicas de envoltorio con el objetivo de analizar dicha inestabilidad, se han introducido además medidas de estabilidad junto con posibles soluciones para aliviar dicho problema.

Se ha desarrollado también un algoritmo de fusión multicriterio, el cual utiliza múltiples algoritmos de selección de variables para puntuar las variables, las cuales son combinadas para obtener un subgrupo robusto basado en la combinación de múltiples clasificadores para mejorar la precisión.

3.7.2 SELECCIÓN POR ESTABILIDAD

La selección por estabilidad es un método relativamente nuevo para selección de atributos, basado en la selección de subgrupos en combinación con un determinado algoritmo de selección (el cual puede ser regresión, SVMs o cualquier otro método similar).

La idea principal consiste en aplicar un algoritmo de selección de atributos en diferentes subgrupos del conjunto de datos y con distintos subgrupos de variables.

Tras repetir el proceso un determinado número de veces, los resultados de la selección pueden ser agregados, seleccionando por ejemplo aquella variable que haya sido

seleccionada como importante más veces cuando ha sido estudiada en cada subgrupo del conjunto de datos.

Esperaríamos entonces que aquellas variables más importantes presenten puntuaciones cercanas al 100%, ya que han sido siempre seleccionadas cuando ha sido posible.

Variables menos importantes, pero de cualquier manera relevantes presentarán puntuaciones superiores a cero, ya que habrán sido seleccionadas cuando variables más importantes no se encontrarán presentes en el subgrupo a analizar.

Finalmente, variables irrelevantes tendrán puntuaciones iguales o lo más cercanas a cero posible, ya que nunca habrán formado parte de las variables seleccionadas.

4 MATERIALES Y MÉTODOS

Con el objetivo de evaluar diferentes métodos de selección de atributos se han utilizado dos diferentes conjuntos de datos de *microarrays* disponibles en el paquete *datamicroarray*, el cual puede ser instalado en lenguaje R a partir del siguiente enlace: https://github.com/ramhiser/datamicroarray

Como conjuntos de datos a evaluar se han considerado dos conjuntos de datos de *microarrays* de carácter binario, ya que son mucho más comunes en la literatura que los problemas de clasificación multiclase. Un experimento representativo de *microarrays* consiste en diferenciar si un determinado paciente presenta cáncer o no, siendo por tanto la gran mayoría de los conjuntos de datos presentes en la literatura de carácter binario.

Los dos conjuntos de datos seleccionados han sido los siguientes:

Nombre	Muestras	Variables	Clases
snc	60	7128	2
próstata	102	12600	2

El conjunto de datos **snc** (acrónimo para sistema nervioso central) se ha utilizado con el objetivo de predecir la respuesta a terapia tumores infantiles embrionarios del sistema nervioso central (Pomeroy et al., 2002). Se compone de 60 biopsias de pacientes obtenidas antes de recibir cualquier tipo de tratamiento sobre las que se ha analizado la expresión de 7128 genes. De los 60 pacientes, 21 de ellos fallecieron (clase 0) y 39 sobrevivieron (clase 1) transcurridos 24 meses.

El conjunto de datos **próstata** se compone de 102 biopsias de tejido de próstata de este mismo número de pacientes obtenidas tras cirugía (D. Singh et al., 2002). Dichas biopsias se dividen en 52 muestras de cáncer de próstata (clase 1) y 50 muestras no cancerosas (clase 0), sobre las que se ha analizado la expresión de 12600 genes.

Ambos conjuntos de datos han sido preprocesados, comprobando la ausencia de valores nulos y sus valores escalados entre 0 y 1 mediante el uso de *MinMaxScaler*

(https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html) como paso previo a la selección de atributos y posterior evaluación.

Tras el preprocesamiento, a cada uno de los dos conjuntos de datos se les han aplicado distintos métodos de selección de atributos (filtraje, envoltorio o embebidos) implementados en Python (Tabla 7).

Método	Tipo	Implementación
Puntuación F	Filtraje	http://featureselection.asu.edu/
Puntuación Fisher	Filtraje	http://featureselection.asu.edu/
reliefF	Filtraje	http://featureselection.asu.edu/
Boruta	Envoltorio	https://github.com/scikit-learn-contrib/boruta_py
RFE	Envoltorio	https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html
SFS	Envoltorio	http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/
Arboles de decisión	Embebido	https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html
L1 mediante Regresión Logística y LASSO (L1)	Embebido	https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression

Tabla 7. Métodos de selección de variables empleados en el siguiente trabajo y dónde acceder a ellos para su implementación.

A continuación, los conjuntos de datos sobre los que se han aplicado los diferentes métodos de selección de atributos han sido evaluados con los siguientes algoritmos de aprendizaje:

- Clasificador de soporte vectorial (SVC).
- Clasificador bayesiano ingenuo (NB).
- Clasificador de bosques aleatorios (RF).

Los principales parámetros utilizados para cada uno de los clasificadores han sido los siguientes (Tabla 8).

Clasificador	Parámetros	Implementación
SVC	kernel = linear C = 1	https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
NB	Parámetros por defecto	https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
RF	n_estimators = 100 max_depth = 5	https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Tabla 8. Parámetros empleados para los distintos clasificadores y dónde acceder a ellos para su implementación.

4.1 MÉTRICAS DE EVALUACIÓN

Con el objetivo de evaluar el comportamiento de los distintos métodos de selección de atributos tras la aplicación de un determinado clasificador, se han utilizado las siguientes métricas:

- **Exactitud**: corresponde a la fracción total de predicciones que el modelo realiza correctamente.
- **Precisión**: indica la proporción correcta de predicciones positivas. Correspondería a la proporción de pacientes diagnosticados con cáncer que tienen cáncer en realidad.
- **Sensibilidad (*Recall*)**: define el número de elementos identificados correctamente como positivos con respecto al total de verdaderos positivos. Sería el porcentaje de pacientes correctamente diagnosticados con cáncer.
- **F1**: Media armónica de precisión y sensibilidad.
- **Área bajo la curva ROC**: cálculo del área bajo la curva ROC a partir de las puntuaciones de las predicciones.
- **Tiempo de ejecución:** tiempo de ejecución de un determinado proceso.

Debido principalmente a los elevados tiempos de ejecución de determinados métodos de selección de atributos aplicados a conjuntos de datos con unas dimensiones tan elevadas, cada uno de los dos conjuntos de datos ha sido analizado de dos maneras distintas:

I) VALIDACION *HOLD-OUT*

- División del conjunto de datos en entrenamiento y test, dedicando un 33% a test
- Evaluación de cada conjunto de datos con los clasificadores SVC, NB y RF (sin selección de atributos)
- Seleccionar 10 variables con cada uno de los siguientes métodos de selección de atributos
    1. Puntuación F
    2. Fisher
    3. reliefF
    4. Boruta
    5. RFE
    6. Selección de variables hacia delante (sfs)
    7. RF
    8. Regresión logística y LASSO
- Evaluación de cada conjunto de datos con las 10 variables seleccionadas con cada uno de los clasificadores (SVC, NB y RF).

II) VALIDACION CRUZADA

- División del conjunto de datos mediante validación cruzada (10 bolsas)
- Evaluación de cada conjunto de datos con los clasificadores SVC, NB y RF (sin selección de atributos)
- Seleccionar 10 variables con cada uno de los siguientes métodos de selección de atributos
    1. Puntuación F
    2. RF
    3. Regresión logística y LASSO
- Evaluación de cada conjunto de datos con las 10 variables seleccionadas con cada uno de los clasificadores (SVC, NB y RF).

El código empleado y los resultados obtenidos pueden ser accedidos libremente a partir de la siguiente dirección de github: https://github.com/idelvalle/TFM

5 RESULTADOS

5.1 ANÁLISIS DEL CONJUNTO DE DATOS SISTEMA NERVIOSO CENTRAL (SNC)

5.1.1 SNC CON VALIDACION *HOLD-OUT*

El conjunto de datos SNC ha sido inicialmente analizado utilizando validación *hold-out*, dedicando un 33% de las muestras para el conjunto de datos test.
Cada uno de los distintos métodos de selección de variables ha sido evaluado para la selección de 10 variables.
El número de variables (10) ha sido elegido de manera arbitraria, con el principal objetivo de averiguar si seleccionando un número tan reducido de variables con respecto al total un clasificador es capaz de asignar las muestras a cada una de las dos clases.
La tabla 9 muestra los tiempos de ejecución para la selección de 10 variables de cada uno de los distintos métodos de selección de variables empleados.

Método	Tiempo de ejecución
Puntuación F	7.9ms
Fisher	9.83ms
reliefF	41.6ms
Boruta	59s
RFE	**4min 39s**
SFS	**1h 32min 35s**
RF	165ms
Regresión logística y LASSO	30.3µs

Tabla 9. Tiempos de ejecución para la selección de 10 variables para cada uno de los distintos métodos de selección de variables implementados en el conjunto de datos SNC.

Como se puede observar en la tabla 9, los métodos envoltorio eliminación recursiva de atributos RFE y especialmente la búsqueda secuencial hacia delante (SFS) presentan unos tiempos de ejecución claramente superiores al resto de los métodos que se han utilizado (destacados en rojo).

Variables seleccionadas

A continuación, se compararon las variables seleccionadas por cada uno de los métodos, para analizar si existían coincidencias en la selección de variables por cada uno de los métodos utilizados.
En la siguiente tabla las variables en rojo representan aquellas presentes en al menos 7 de los 8 métodos de selección de variables empleados y aquellas en verde representan las seleccionadas por al menos 5 métodos (Tabla 10).

Puntuación F	Fisher	reliefF	Boruta	RFE	SFS	RF	LASSO
U07563	**U07563**	**U07563**	U66711	**U35139**	BioB.5.at	D83243	**J02611**
U34038	U34038	U21858	**J02611**	X97544	BioC.5.at	M24248	**U07563**
Z11899	Z11899	M14328	**U35139**	**U07563**	BioDn.3.at	M27281	**U35139**
U35048	U35048	Z11899	Z11899	U66406	BioB.5.st	**U07563**	U39318
J02611	**J02611**	M32886	K03008	M33197	BioB.M	U68031	X76302
U35139	**U35139**	U25165	U78876	**J02611**	BioC.3.at	X57129	X77197
U82306	U82306	Z50022	M26692	U66711	CreX.5	X74295	L10333
U78876	U78876	U26403	M27281	U18383	BioB.3.at	AB006781	U66711
L05628	L05628	X13293	**U07563**	U39318	BioDn.5.at	D43682	Z11899
X57129	X57129	X94703	M18728	U39226	CreX.3.at	X79200	K03189

Tabla 10. Variables seleccionadas en el conjunto de datos SNC por cada uno de los métodos de selección de variables implementados.

De todos los métodos utilizados, la variable U07563 ha sido seleccionada por 7 de los 8 métodos, mientras que las variables J02611 y U35139 lo han sido por al menos 5 de los distintos métodos de selección.
Además, la selección de variables realizada por la búsqueda hacia delante (SFS) y árboles de decisión seleccionan variables que no son compartidas por el resto de los métodos.

Resultados evaluación con distintos clasificadores

A continuación, cada uno de los tres clasificadores seleccionados (SVC, NB y RF) ha sido evaluado antes y después de la selección de las 10 variables por cada uno de los distintos métodos de selección de variables descritos previamente.
En las siguientes tablas se presentan los tiempos de ejecución y los valores de las diferentes métricas en cada uno de los casos (antes y después de la selección de variables), resaltándose en negrita aquellos valores que igualan o mejoran los resultados de cada clasificador sin selección de variables (Tablas 11-13).

Resultados Clasificador de Soporte Vectorial

En el caso de este clasificador, la utilización de regresión logística en combinación con la penalización L1 (LASSO) seleccionando 10 variables es el único algoritmo de aprendizaje que consigue mejorar las métricas obtenidas por dicho clasificador en ausencia de la selección de variables.

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
SVC	79.9ms	0.7	0.69	0.7	0.69	0.58
SVC-F	**1.19ms**	0.65	0.66	0.65	0.65	0.57
SVC-Fisher	**799us**	0.65	0.66	0.65	0.65	0.57
SVC-reliefF	**895µs**	0.65	0.62	0.65	0.62	**0.65**
SVC-Boruta	**966µs**	0.40	0.43	0.40	0.41	0.45
SVC-RFE	**1.8ms**	**0.70**	**0.69**	**0.70**	0.66	0.40
SVC-SFS	**855µs**	0.65	0.42	0.65	0.51	0.53
SVC-RF	**3ms**	0.65	0.64	0.65	0.64	**0.58**
SVC-LR	**954µs**	**0.75**	**0.75**	**0.75**	**0.73**	0.55

Tabla 11. Métricas obtenidas en el conjunto de datos SNC con el clasificador SVC con y sin selección de variables empleando validación *hold-out*.

Resultados Clasificador Bayesiano Ingenuo

Para este clasificador la selección de 10 variables no consigue mejorar las métricas obtenidas en ausencia de la selección de variables.

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
NB	4.21ms	0.75	0.82	0.75	0.70	0.62
NB-F	**1.29ms**	0.70	0.70	0.70	**0.70**	0.57
NB-Fisher	**834µs**	0.70	0.70	0.70	**0.70**	0.57
NB-reliefF	**1.07ms**	0.60	0.60	0.60	0.60	**0.63**
NB-Boruta	**832µs**	0.55	0.53	0.55	0.54	0.44
NB-RFE	**731µs**	0.60	0.54	0.60	0.55	0.42
NB-SFS	163ms	0.55	0.59	0.55	0.56	0.60
NB-RF	**1.96ms**	0.70	0.70	0.70	**0.70**	0.53
NB-LR	**1.32ms**	0.70	0.69	0.70	0.69	0.60

Tabla 12. Métricas obtenidas en el conjunto de datos SNC con el clasificador NB con y sin selección de variables empleando validación *hold-out*.

Resultados Clasificador de Bosques Aleatorios

En este caso se consiguen ligeras mejoras de las métricas F1 y ROC-AUC, principalmente con la selección de variables usando los métodos de filtraje y el método embebido utilizando la penalización LASSO.

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
RF	151ms	0.65	0.61	0.65	0.58	0.51
RF-F	**137ms**	**0.65**	**0.64**	**0.65**	**0.64**	**0.54**
RF-Fisher	**140ms**	**0.65**	**0.64**	**0.65**	**0.64**	**0.54**
RF-reliefF	**135ms**	0.60	0.57	0.60	**0.58**	**0.58**
RF-Boruta	183ms	0.50	0.46	0.50	0.48	0.24
RF-RFE	164ms	0.60	0.41	0.60	0.49	0.34
RF-SFS	**2.45ms**	0.50	0.50	0.50	0.50	0.43
RF-RF	173ms	0.55	0.49	0.55	0.51	**0.51**
RF-LR	168ms	**0.70**	**0.69**	**0.70**	**0.69**	0.47

Tabla 13. Métricas obtenidas en el conjunto de datos SNC con el clasificador RF con y sin selección de variables empleando validación *hold-out*.

Para el conjunto de datos SNC empleando validación *hold-out*, de los tres clasificadores empleados el clasificador SVC con 10 variables seleccionadas utilizando regresión logística y penalización LASSO es claramente el que aporta una mejora significativa con respecto al uso del clasificador sin selección de variables.
Es importante señalar que la selección de 10 variables empleando RFE y especialmente SFS conlleva unos tiempos de ejecución significativamente superiores comparado con el resto de los métodos, siendo este el principal motivo por el que no se ha realizado validación cruzada en este apartado.

5.1.2 SNC CON VALIDACION CRUZADA (10 bolsas)

En esta sección de los resultados se han seleccionado 3 métodos de selección de variables: puntuación F (filtraje), arboles de decisión y regresión logística (ambos embebidos) con penalización LASSO.
Se han seleccionado 10 variables con cada uno de los métodos y se han comparado los resultados de las métricas de evaluación con cada uno de los tres clasificadores empleados previamente (SVC, NB y RF) antes y después de la selección de las 10 variables empleando validación cruzada con 10 bolsas.
A continuación, se presentan los resultados de las métricas de evaluación y tiempos de ejecución para la selección de 10 variables usando validación cruzada para cada uno

de los distintos clasificadores. En negrita se representan los valores que igualan o mejoran los resultados de cada clasificador sin selección de variables (Tabla 14-16).

Resultados Clasificador de Soporte Vectorial

La utilización de regresión logística en combinación con la penalización L1 (LASSO) es el único algoritmo de aprendizaje que consigue mejorar (aunque no en gran medida) las métricas obtenidas por dicho clasificador en ausencia de la selección de variables.

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
SVC	1.84s	0.68	0.66	0.68	0.64	0.62
SVC-F	**255ms**	0.66	0.64	0.66	0.62	0.61
SVC-RF	7.43s	0.6	0.49	0.6	0.53	0.61
SVC-LR	**335ms**	**0.7**	**0.7**	**0.7**	**0.67**	**0.67**

Tabla 14. Métricas obtenidas en el conjunto de datos SNC con el clasificador SVC con y sin selección de variables empleando validación cruzada.

Resultados Clasificador Bayesiano Ingenuo

Para este clasificador la selección de 10 variables empleando puntuación F consigue mejorar levemente las métricas obtenidas en ausencia de la selección de variables.

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
NB	340ms	0.66	0.61	0.66	0.61	0.65
NB-F	**253ms**	**0.67**	**0.7**	**0.67**	**0.66**	0.58
NB-RF	7.29s	0.57	0.53	0.57	0.53	0.64
NB-LR	**340ms**	0.64	**0.66**	0.64	**0.62**	**0.71**

Tabla 15. Métricas obtenidas en el conjunto de datos SNC con el clasificador NB con y sin selección de variables empleando validación cruzada.

Resultados Clasificador de Bosques Aleatorios

De manera similar a los resultados obtenidos con el clasificador SVC, la utilización de regresión logística en combinación con la penalización L1 (LASSO) es el único algoritmo de aprendizaje que consigue mejorar (aunque no en gran medida) las métricas obtenidas por dicho clasificador en ausencia de la selección de variables.

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
RF	13.1s	0.67	0.51	0.67	0.56	0.6
RF-F	**12.31s**	0.63	**0.6**	0.63	**0.59**	**0.61**
RF-RF	21.8s	0.6	**0.55**	0.6	**0.56**	**0.6**
RF-LR	**12.3s**	**0.68**	**0.67**	**0.68**	**0.65**	**0.67**

Tabla 16. Métricas obtenidas en el conjunto de datos SNC con el clasificador RF con y sin selección de variables empleando validación cruzada.

Para concluir el análisis del conjunto de datos SNC, la utilización del clasificador SVC en combinación con la selección de 10 variables empleando regresión logística y LASSO resulta ser el mejor clasificador con respecto al uso del mismo clasificador en ausencia de selección de variables.

5.2 ANÁLISIS DEL CONJUNTO DE DATOS PRÓSTATA

5.2.1 PRÓSTATA CON VALIDACION *HOLD-OUT*

El conjunto de datos PRÓSTATA ha sido inicialmente analizado de manera similar al SNC, utilizando validación *hold-out* y dedicando un 33% de las muestras para el conjunto de datos test.
Tal y como se presentó en el apartado anterior, cada uno de los distintos métodos de selección de variables ha sido evaluado con la selección arbitraria de 10 variables.
La siguiente tabla muestra los tiempos de ejecución para la selección de 10 variables de cada uno de los distintos métodos de selección de variables empleados (Tabla 17).
Como se puede observar en la tabla 17, los métodos RFE y especialmente el SFS (con un tiempo de ejecución de más de 7 horas para la selección de variables), presentan unos tiempos de ejecución claramente superiores al resto de métodos que se han utilizado (destacados en rojo).

Método	Tiempo de ejecución
Puntuación F	46ms
Fisher	25ms
reliefF	85.9ms
Boruta	6min 46s
RFE	**37min 47s**
SFS	**7h 27min 25s**
RF	457ms
Regresión logística y L1	29.8µs

Tabla 17. Tiempos de ejecución para la selección de 10 variables para cada uno de los distintos métodos de selección de variables implementados en el conjunto de datos PRÓSTATA.

Variables seleccionadas

A continuación, y tal y como se realizó con el conjunto de datos SNC, se compararon las variables seleccionadas por cada uno de los métodos.
En la siguiente tabla (Tabla 18), las variables en rojo representan aquellas presentes en al menos 7 de los 8 métodos de selección de variables empleados y aquellas en verde representan las seleccionadas por al menos 5 métodos.

Puntuación F	Fisher	reliefF	Boruta	RFE	SFS	RF	LASSO
V6185	**V6185**	**V6185**	V4365	V10125	V0	V306	V288
V8965	V8965	**V5890**	V306	**V6185**	V1	V329	**V5890**
V4365	V4365	V8729	**V8850**	V3333	V2	**V5890**	**V6185**
V5890	**V5890**	V9937	V299	**V5890**	V3	**V6185**	V6866
V6866	V6866	V9850	V10956	**V10553**	V4	V7520	V7623
V10553	**V10553**	**V10553**	**V5890**	V10234	V5	**V8850**	V10215
V8850	**V8850**	V6866	**V6185**	V9034	V6	V9034	**V10553**
V12148	V12148	V9172	V8965	V11858	V9	V9172	V10875
V10494	V10494	**V8850**	V10138	V5661	V22	V10494	V10956
V10138	V10138	V9184	V12148	V11137	V204	V12148	V11858

Tabla 18. Variables seleccionadas en el conjunto de datos PRÓSTATA por cada uno de los métodos de selección de variables implementados.

Para este conjunto de datos existen 2 variables que son seleccionadas con la mayoría de los métodos empleados (V6185 y V5890), mientras que otras dos (V10553 y V8850) han sido seleccionadas por al menos 5 de los métodos empleados.
Además, el método SFS selecciona 10 variables que no son compartidas por el resto de los métodos de selección de variables.

Resultados evaluación con distintos clasificadores

Cada uno de los tres clasificadores empleados (SVC, NB y RF) ha sido evaluado antes y después de la selección de 10 variables por cada uno de los distintos métodos de selección de variables descritos previamente.
En las siguientes tablas se presentan los tiempos de ejecución y los valores de las diferentes métricas en cada uno de los casos. En negrita se representan los valores que igualan o mejoran los resultados de cada clasificador sin selección de variables (Tabla 19-21).

Resultados Clasificador de Soporte Vectorial

En el caso de este clasificador, la selección de variables empleando Boruta consigue mejorar levemente las métricas obtenidas en ausencia de selección de variables.
Sin embargo, hay que señalar que para la implementación de Boruta en este caso no se han conseguido ajustar los parámetros para que se seleccionen 10 variables, sino 22, no siendo por tanto comparable dicho resultado con el resto de las técnicas de selección de variables.
Por otra parte, los métodos de filtraje empleados consiguen igualar las métricas en ausencia de selección de variables a la vez que mejorar sensiblemente los tiempos de ejecución.

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
SVC	398ms	0.88	0.89	0.88	0.88	0.96
SVC-F	**1.02ms**	**0.88**	**0.89**	**0.88**	**0.88**	0.93
SVC-Fisher	**992µs**	**0.88**	**0.89**	**0.88**	**0.88**	0.93
SVC-reliefF	**1.13ms**	0.85	0.86	0.85	0.85	0.94
SVC-Boruta	**1.27ms**	**0.91**	**0.91**	**0.91**	**0.91**	**0.96**
SVC-RFE	**1.1ms**	0.85	0.85	0.85	0.85	**0.96**
SVC-SFS	**1.3ms**	0.68	0.75	0.68	0.65	0.85
SVC-RF	**1.01ms**	**0.88**	**0.89**	**0.88**	**0.88**	0.92
SVC-LR	**1.06ms**	0.85	0.85	0.85	0.85	0.94

Tabla 19. Métricas obtenidas en el conjunto de datos PRÓSTATA con el clasificador SVC con y sin selección de variables empleando validación *hold-out*.

Resultados Clasificador Bayesiano Ingenuo

En el caso del clasificador bayesiano ingenuo, todos los métodos de selección de variables empleados consiguen mejorar las métricas obtenidas en ausencia de selección de variables. La selección de 10 variables mediante RFE arroja los mejores resultados, aunque hay que tener en cuenta el elevado tiempo que tarda dicha técnica de selección de atributos en seleccionar 10 variables (Tabla 20).

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
NB	11.8ms	0.56	0.58	0.56	0.53	0.56
NB-F	**782µs**	**0.88**	**0.89**	**0.88**	**0.88**	**0.95**
NB-Fisher	**1.28ms**	**0.88**	**0.89**	**0.88**	**0.88**	**0.95**
NB-reliefF	**816µs**	**0.91**	**0.91**	**0.91**	**0.91**	**0.95**
NB-Boruta	**935µs**	**0.91**	**0.91**	**0.91**	**0.91**	**0.96**
NB-RFE	**742µs**	**0.94**	**0.95**	**0.94**	**0.94**	**0.94**
NB-SFS	463ms	**0.59**	**0.61**	**0.59**	**0.56**	**0.74**
NB-RF	**823µs**	**0.91**	**0.91**	**0.91**	**0.91**	**0.95**
NB-LR	**746µs**	**0.88**	**0.89**	**0.88**	**0.88**	**0.96**

Tabla 20. Métricas obtenidas en el conjunto de datos PRÓSTATA con el clasificador NB con y sin selección de variables empleando validación *hold-out*.

Resultados Clasificador de Bosques Aleatorios

La utilización de RF con 10 variables seleccionadas mediante árboles de decisión y regresión logística consiguen mejorar las métricas obtenidas con todas las variables del conjunto de datos**.**

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
RF	143ms	0.88	0.89	0.88	0.88	0.93
RF-F	**131ms**	**0.88**	**0.89**	**0.88**	**0.88**	**0.95**
RF-Fisher	**135ms**	**0.88**	**0.89**	**0.88**	**0.88**	**0.95**
RF-reliefF	**131ms**	**0.88**	**0.89**	**0.88**	**0.88**	**0.96**
RF-Boruta	469ms	**0.91**	**0.91**	**0.91**	**0.91**	**0.97**
RF-RFE	473ms	0.85	0.85	0.85	0.85	**0.96**
RF-SFS	**2.86ms**	0.79	0.82	0.79	0.79	0.84
RF-RF	456ms	**0.91**	**0.91**	**0.91**	**0.91**	**0.96**
RF-LR	475ms	**0.91**	**0.91**	**0.91**	**0.91**	**0.94**

Tabla 21. Métricas obtenidas en el conjunto de datos PRÓSTATA con el clasificador RF con y sin selección de variables empleando validación *hold-out*.

Al igual que sucedía con el conjunto de datos SNC, es importante señalar que la selección de 10 variables empleando RFE y especialmente SFS conlleva unos tiempos de ejecución muy elevados, siendo este el principal motivo por el que no se ha realizado validación cruzada en este apartado.

5.2.2 PRÓSTATA CON VALIDACION CRUZADA (10 bolsas)

En esta sección se ha procedido de manera similar a la empleada con el conjunto de datos SNC, presentando las siguientes tablas los tiempos de ejecución y los valores de las diferentes métricas para la selección de 10 variables empleando validación cruzada en cada uno de los casos. En negrita se representan los valores que igualan o mejoran los resultados de cada clasificador sin selección de variables (Tabla 22-24).

Resultados Clasificador de Soporte Vectorial

En este caso se obtiene una ligera mejora de las diferentes métricas empleando las variables obtenidas usando la puntuación F.

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
SVC	8.18s	0.91	0.92	0.91	0.91	0.98
SVC-F	**627ms**	**0.92**	**0.93**	**0.92**	**0.92**	0.97
SVC-RF	11.7s	0.9	**0.92**	0.9	0.9	0.97
SVC-LR	**971ms**	**0.91**	**0.92**	**0.91**	**0.91**	0.96

Tabla 22. Métricas obtenidas en el conjunto de datos PRÓSTATA con el clasificador SVC con y sin selección de variables empleando validación cruzada.

Resultados Clasificador Bayesiano Ingenuo

Tal y como sucedía al emplear validación *hold-out*, la selección de variables mejora radicalmente las métricas obtenidas empleando todas las variables.
Dicha mejora es ligeramente superior al resto empleando las variables seleccionadas por la puntuación F (presenta un tiempo de ejecución menor).

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
NB	796ms	0.62	0.62	0.62	0.57	0.63
NB-F	**586ms**	**0.93**	**0.94**	**0.93**	**0.93**	**0.96**
NB-RF	11.2s	**0.93**	**0.94**	**0.93**	**0.93**	**0.96**
NB-LR	943ms	**0.93**	**0.94**	**0.93**	**0.93**	**0.96**

Tabla 23. Métricas obtenidas en el conjunto de datos PRÓSTATA con el clasificador NB con y sin selección de variables empleando validación cruzada.

Resultados Clasificador de Bosques Aleatorios

Para este clasificador de nuevo la puntuación F consigue resultados ligeramente superiores al resto de técnicas de selección de variables.

Algoritmo	Tiempo	Exactitud	Precisión	Sensibilidad	F1	ROC-AUC
RF	14.9s	0.89	0.9	0.89	0.89	0.93
RF-F	**13.2s**	**0.93**	**0.94**	**0.93**	**0.93**	**0.97**
RF-RF	25.6s	**0.92**	**0.93**	**0.92**	**0.92**	**0.97**
RF-LR	**13.6s**	**0.92**	**0.93**	**0.92**	**0.92**	**0.97**

Tabla 24. Métricas obtenidas en el conjunto de datos PRÓSTATA con el clasificador RF con y sin selección de variables empleando validación cruzada.

En el caso del conjunto de datos PRÓSTATA, la utilización del clasificador NB en combinación con la selección de 10 variables empleando puntuación F resulta ser el mejor clasificador con respecto al uso del mismo clasificador en ausencia de selección de variables.

6 DISCUSIÓN

Las diferentes técnicas de selección de variables son herramientas que resulta ser muy útil en un número elevado de escenarios que impliquen el uso de algoritmos de aprendizaje. La clave para su utilización consiste principalmente en tener un objetivo claro y entender el método que mejor funciona para alcanzarlo.

Al seleccionar las mejores variables para mejorar el rendimiento del modelo, es relativamente sencillo verificar si un método particular funciona mejor frente a otras alternativas.

El presente trabajo fin de máster ha consistido en una revisión bibliográfica de las principales técnicas de selección de variables, las cuales se han agrupado tradicionalmente en técnicas de filtraje, envoltorio y embebidas. En la revisión se técnicas se han analizado principalmente aquellas más utilizadas en distintas aplicaciones bioinformáticas.

Uno de los campos de la bioinformática que ha experimentado un rápido crecimiento en los últimos años ha sido el estudio de enfermedades complejas como el cáncer. El desarrollo y expansión de la técnica de análisis de *microarrays* de ADN está incrementando la capacidad para la predicción y el tratamiento de diferentes tipos de cáncer.

El reducido tamaño de muestras, la alta dimensionalidad y la presencia de clases no balanceadas son los principales problemas que se presentan a la hora de analizar los conjuntos de datos de *microarrays* de ADN.

Para la sección de implementación del presente trabajo se han seleccionado dos conjuntos de datos de *microarrays* ligeramente distintos. En la literatura, la mayoría de los trabajos publicados corresponden a conjuntos de datos con dos clases o binarios (si el paciente sobrevive o no o si el paciente tiene cáncer o no). Por esta razón se han seleccionado dos conjuntos de datos binarios para evaluar diferentes técnicas de selección de variables.

El conjunto de datos SNC contiene 60 muestras (21 pertenecientes a una clase y 39 a otra clase) sobre las que se han analizado más de 7000 genes. Por otro lado, el conjunto de datos PRÓSTATA contiene 102 muestras (50 pertenecientes a una clase y 52 a la otra) sobre las que se han analizado más de 12000 genes.

Sobre ambos conjuntos de datos se ha evaluado la eficiencia de tres clasificadores ampliamente utilizados en el análisis de este tipo de datos: SVC, NB y RF y de diferentes técnicas de selección de variables pertenecientes a los métodos de filtraje, envoltorio y

embebidos. Para la selección de variables se ha decidido arbitrariamente un número de 10 variables para evaluar la eficiencia de los distintos clasificadores con dichas variables.

Ambos conjuntos de datos se han analizado inicialmente utilizando validación *hold-out* destinando un 33% al conjunto de datos test. Dicho análisis ha permitido evaluar los tiempos de ejecución de cada una de las técnicas de selección de variables seleccionadas. Para ambos conjuntos de datos los métodos de filtraje han resultado ser los más rápidos y los métodos envoltorio (especialmente RFE y SFS) han resultado ser los más lentos para la selección de diez variables.

Los tiempos de ejecución tan elevados de dichos métodos envoltorio han impedido la utilización de validación cruzada en este apartado con el *hardware* empleado para la realización de este trabajo.

Si se investigan las diez variables seleccionadas por cada uno de los métodos para ambos conjuntos de datos, las técnicas de filtraje y la regresión logística (embebida) en combinación LASSO son las técnicas que seleccionan más variables en común entre ellas. El resto de las técnicas, especialmente la RFE seleccionan diez variables que no son compartidas con el resto de las técnicas.

Con respecto al conjunto de datos SNC, y empleando validación cruzada, el clasificador SVC en combinación con penalización L1 resulta ser el algoritmo que arroja las mejores métricas con respecto a los resultados en ausencia de selección de variables.

La selección de 10 variables no consigue sin embargo mejorar mucho las métricas obtenidas (hay que señalar además que este es un conjunto de datos con clases no balanceadas y la exactitud ha resultado no ser una buena métrica para evaluar este tipo de conjunto de datos). Además, el conjunto de datos SNC ha sido analizado por otros grupos empleando otras técnicas de selección de variables y tampoco se han conseguido mejoras significativas empleando dichas técnicas (V. Bolón-Canedo, Sánchez-Maroño, Alonso-Betanzos, Benítez, & Herrera, 2014).

En el caso del conjunto de datos PRÓSTATA, y utilizando también validación cruzada, el clasificador NB utilizando validación cruzada con las diez variables seleccionadas por la puntuación F resulta ser el algoritmo que arroja las mejores métricas. En este conjunto de datos la selección de variables es capaz de al menos igualar las métricas obtenidas en ausencia de selección de variables. Dicho conjunto de datos presenta unas métricas bastante buenas en ausencia de selección de variables, especialmente empleando los clasificadores SVC y RF.

Para futuros trabajos donde se quieran aplicar las técnicas de selección de variables a conjuntos de datos con características similares a los dos conjuntos de datos estudiados en el presente trabajo, se pueden aplicar los siguientes criterios en base a los resultados obtenidos.

La elección del clasificador puede variar en gran medida los resultados; el clasificador SVC parece funcionar bastante bien con este tipo de datos, presentando además unos tiempos de ejecución muy razonables.

Los métodos de filtraje funcionan muy bien y son además muy rápidos en comparación con el resto de las técnicas analizadas, aunque la regresión logística en combinación con la penalización LASSO también arroja buenos resultados. En el caso de conjuntos de datos con clases balanceadas, merecería la pena la aplicación de técnicas apropiadas como SMOTE, descrita en el presente trabajo.

Finalmente, sería conveniente encontrar el número de variables seleccionadas óptimo para encontrar los mejores resultados, no limitándonos a seleccionar únicamente el número de variables seleccionadas en el presente trabajo.

El rapidísimo incremento en los últimos años de los conjuntos de datos con elevada dimensionalidad y reducido número de muestras está provocando que se estén desarrollando nuevas técnicas de selección de variables (como las denominadas híbridas). Sin embargo, no existe una única técnica universal, igual que no existe un clasificador que sirva para todos los conjuntos de datos. La selección de los parámetros óptimos va a depender por tanto del conjunto de datos con el que estemos trabajando, requiriendo de una evaluación minuciosa de los parámetros, técnicas y algoritmos a emplear.

7 CONCLUSIONES

1. Los conjuntos de datos empleados en bioinformática presentan características propias que se benefician del empleo de técnicas de selección de variables.

2. El número de variables seleccionadas resulta crítico para obtener los mejores resultados, siendo muy importante la asignación óptima de dicho parámetro.

3. Las técnicas de filtraje empleadas funcionan muy bien (tanto en rapidez como en métricas) en los conjuntos de datos de microarrays utilizados, aunque hay que tener en cuenta sus limitaciones.

4. Las técnicas envoltorio requieren de unos tiempos de ejecución muy elevados con respecto al resto de técnicas de selección de variables testadas.

5. Con respecto al clasificador empleado, es mejor probar al menos tres distintos para optimizar los resultados.

6. No hay dos conjuntos de datos exactamente iguales, con lo cual es mejor invertir un tiempo razonable en la optimización de los diferentes parámetros mediante preferiblemente validación cruzada.

ANEXO-BIBLIOGRAFÍA

Albon, C. (2018). *Python Machine Learning Cookbook*. O'Reilly Media.

Aryal, S. (2018). DNA Microarrays. Retrieved November 23, 2019, from Microbe Notes website: https://microbenotes.com/dna-microarray/

Aziz, R., Verma, C. K., & Srivastava, N. (2017). Dimension reduction methods for microarray data: a review. *AIMS Bioengineering*, *4*(1), 179–197.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, *5*(4), 537–550.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, *282*, 111–135.

Bolón-Canedo, Verónica, & Alonso-Betanzos, A. (2019). *Microarray Bioinformatics*. Springer.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Gandhi, R. (2018). Genetic Algorithms. Retrieved October 12, 2019, from medium.com website: https://medium.com/datadriveninvestor/genetic-algorithms-9f920939f7cc

Gu, Q., Li, Z., & Han, J. (2011). Generalized Fisher Score for Feature Selection. *UAI'11 Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 266–273. Barcelona.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*.

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*. University of Waikato.

Hira, Z. M., & Gillies, D. F. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics*, *2015*, 1–13.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67.

Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory*

analysis with applications to biology, control, and artificial intelligence. University Michigan Press.

Homola, D. (2015). BorutaPy – an all relevant feature selection method. Retrieved October 12, 2019, from danielhomola.com website: http://danielhomola.com/2015/05/08/borutapy-an-all-relevant-feature-selection-method/

Hosseini, M., Nematbaksh, N., & Nadimi, M. (2017). Feature selection techniques in bioinformatics. *National Conference on Emerging Trends in Electrical, Electronics and Computer Engineering*, (July).

Huang, S., Nianguang, C. A. I., Penzuti Pacheco, P., Narandes, S., Wang, Y., & Wayne, X. U. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics and Proteomics*, *15*(1), 41–51.

Jović, A., Brkić, K., & Bogunović, N. (n.d.). *A review of feature selection methods with applications*.

Keles, S., van der Laan, M., & Eisen, M. B. (2002). Identification of regulatory elements using a feature selection method. *Bioinformatics*, *18*(9), 1167–1175.

Kim, S. K., Nam, J. W., Rhee, J. K., Lee, W. J., & Zhang, B. T. (2006). miTarget: MicroRNA target gene prediction using a support vector machine. *BMC Bioinformatics*.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. *ML92 Proceedings of the Ninth International Workshop on Machine Learning*, 249–256.

Koehrsen, W. (2017). Random Forest Simple Explanation. Retrieved October 12, 2019, from medium.com website: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

Kononenko, I. (1994). Estimating Attributes: Analysis and Extensions of RELIEF. *Machine Learning: ECML-94*, 171–182. Springer.

Kursa, M., Jankowski, A., & Rudnicki, W. (2010). Boruta - A System for Feature Selection. *Fundam. Inform.*, *101*, 271–285.

Liu, H., Li, J., & Wong, L. (2002). A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. *Genome Informatics. International Conference on Genome Informatics*, *13*, 51–60.

Lovato, P., Bicego, M., Cristani, M., Jojic, N., & Perina, A. (2012). Feature Selection Using Counting Grids: Application to Microarray Data. In G. Gimel'farb, E. Hancock, A. Imiya, A. Kuijper, M. Kudo, S. Omachi, … K. Yamada (Eds.), *Structural, Syntactic, and Statistical Pattern Recognition* (pp. 629–637). Berlin,

Heidelberg: Springer Berlin Heidelberg.

Michiels, S., Koscielny, S., & Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, *365*(9458), 488–492.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., … Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, *415*(6870), 436–442.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*(11), 1119–1125.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., … Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, *1*(2), 203–209.

Singh, R. K., & Sivabalakrishnan, M. (2015). Feature Selection of Gene Expression Data for Cancer Classification: A Review. *Procedia Computer Science*, *50*, 52–57.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Vapnik, V. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*.

Vaushik, S. (2016). Feature Selection methods with example (Variable selection methods). Retrieved October 13, 2019, from analytycsvidhya.com website: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

Xuan, P., Guo, M. Z., Wang, J., Wang, C. Y., Liu, X. Y., & Liu, Y. (2011). Genetic algorithm-based efficient feature selection for classification of pre-miRNAs. *Genetics and Molecular Research*, *10*(2), 588–603.

Yukyee Leung, & Yeungsam Hung. (2010). A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *7*(1), 108–117.

Zhong, J., Wang, J., Peng, W., Zhang, Z., & Li, M. (2015). A feature selection method for prediction essential protein. *Tsinghua Science and Technology*, *20*(5), 491–499.