eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Máster y Doctorado en

Ingeniería **C**omputacional y **S**istemas **I**nteligentes

Konputazio Zientziak eta Adimen Artifiziala Saila –
Departamento de Ciencias de la Computación e Inteligencia Artificial

Máster I.C.S.I.

Master Thesis

Clasificadores supervisados para predecir la abstinencia a 12 meses en fumadores tratados con vareniclina

Borja Santos Zorrozúa

Supervisors

Iñaki Inza Cano
Department of Computer Science and Artificial Intelligence
Computer Science Faculty

Jose A. Lozano Alonso
Department of Computer Science and Artificial Intelligence
Computer Science Faculty

informatika fakultatea
facultad de informática

KZAA /CCIA

Septiembre 2012

Índice

1. INTRODUCCIÓN

1.1. Planteamiento del problema

El consumo de tabaco es la principal causa, aislada y evitable, de muerte e invalidez en el mundo desarrollado. La razón es que la nicotina contenida en los cigarros es una sustancia psicoactiva con un poder adictivo muy alto. Además su consumo mediante la inhalación tiene una gran toxicidad.

El tabaquismo se definió en 1984 por la OMS como una forma de drogodependencia y más tarde en 1987, la Sociedad Americana de Psiquiatría determinó que la nicotina era una sustancia psicoactiva que es capaz de producir dependencia sin abuso. Al ser considerado como tal, al tabaco se le reconocen características comunes de las demás adicciones:

a) Conducta de autoadministración con dificultad en el control de la ingesta.

b) Aparición de deseos intensos y urgentes de consumir asociados a estímulos previamente neutros.

c) Asociado con la cesación de su consumo puede aparecer, en mayor o menor medida, un síndrome de abstinencia.

La dependencia tiene muchos factores influyentes, como por ejemplo: adaptación cerebral, alteraciones electrofisiológicas, anormalidades funcionales y estructurales en el cerebro y cambios neurocognitivos. La vulnerabilidad a la dependencia tiene un importante componente genético. Además los factores genéticos también pueden asociarse a la vulnerabilidad a la dependencia a la nicotina, la intensidad a la abstinencia y las dificultades para poder dejar de consumir tabaco.

Actualmente, la prevalencia de la adicción al tabaco se estima que es de un 22 % en mayores de 15 años en el mundo. El consumo es mayor entre los hombres con un 41.1 % que entre las mujeres donde la prevalencia es del 8.9 % [1]. La cifra de muertes anuales como consecuencia del tabaco asciende a alrededor de 5 millones de personas y si no se es capaz de reducir la prevalencia en un futuro próximo, durante los siguientes 50 años las muertes atribuibles al tabaco ascenderán a 500 millones [2].

La prevalencia en España entre la población mayor de 16 años es de un 26.2 %. Al igual que en el resto del mundo, el consumo es más elevado entre los hombres con un 31.2 % que entre las mujeres con un 21.3 %, aunque en comparación con los datos mundiales el consumo entre ellas es superior a la proporción mundial [3]. Anualmente en España las muertes causadas por el consumo de tabaco ascienden a unas 50000.

Concretamente, en Euskadi, según la ESCAV-02, el 26 % de la población mayor de 16 años fuma habitualmente. Siguiendo la tendencia mostrada anteriormente, la población fumadora masculina es mayor que la femenina con un 31 % frente a un 21 %. Además, este estudio sugiere que las diferencias socioeconómicas en el consumo de tabaco tenderán a aumentar, ya que será mayor entre las clases más desfavorecidas y especialmente entre las mujeres.

Con lo cual, el tabaquismo es un problema de salud pública y queda patente la complejidad existente a la hora de tratar a los fumadores.

Hoy en día, el tratamiento del tabaquismo se realiza dentro de la atención primaria y especializada, con una serie de intervenciones que varían desde una intervención mínima hasta una intensiva. Esta última requiere de personal especializado y asocia tratamientos farmacológicos con tratamientos psicológicos. La eficacia se estima entre un 30-40 % de abstinencia al año y el objetivo del tratamiento farmacológico es mejorar el pronóstico a medio y largo plazo.

Aunque como se ha visto antes, la dependencia tiene asociada factores genéticos, los conocimientos de farmacogenética no se han aplicado demasiado. El motivo de la aplicación de dichos conocimientos podría tener un interés alto ya que se podría establecer una relación entre las variantes genéticas y las respuestas a diferentes tratamientos de deshabituación. En el caso de la vareniclina, debido a su reciente comercialización, no se han publicado estudios de asociación con variante genéticas.

La eficacia de los tratamientos farmacológicos, asociados con un programa de deshabituación, está alrededor de un 30 %. Esto determina que la selección entre las opciones terapéuticas se realiza principalmente en base al criterio individual de cada médico y no sobre bases racionales.

La estructura de este trabajo es la siguiente. Después de la introducción del tema y las hipótesis del mismo, la sección 2 habla de la importancia de los estudios de asociación y de la aplicación de la minería de datos como técnica de análisis. La sección 3 describe las variables utilizadas en el estudio así como los diferentes métodos de selección de variables y concluye con la descripción de las técnicas de minería de datos empleadas en el análisis de los datos. La sección 4 expone los resultados obtenidos y finalmente se discuten en la última sección, donde además se incluyen las limitaciones y posibles estudios futuros.

1.2. Objetivos del trabajo

Establecer si mediante la información genética y el uso de escalas médicas relacionadas con el tabaquismo, es posible predecir la abstinencia tras 12 meses de tratamiento con vareniclina. Para ello se utilizarán técnicas de minería de datos para comprobar si su capacidad de predicción es superior a la obtenida mediante un clasificador que utiliza la regresión logística.

El protocolo de análisis que se ha seguido consta de dos partes:

Parte 1: Preprocesado.

Parte 2: Aplicación de los clasificadores.

En la primera parte se imputarán los valores perdidos utilizando la moda y la media, según sea la variable discreta o continua. Después se realizará un análisis de las escalas clínicas para poder utilizar la putuación total de manera correcta, una vez hecho esto discretizaremos las variables continuas para facilitar el funcionamiento de los clasificadores y finalmente utilizaremos algoritmos de selección de variables para maximizar el poder de clasificación, eliminando las variables poco informativas.

La segunda fase se centra en el análisis de los datos por parte de los dos clasificadores. Para ello se llevará a cabo el mismo proceso tanto para el Naive-Bayes como para el modelo que utiliza regresión logística. De esta manera la comparación será en igualdad de condiciones. El proceso de análisis será k-fold cross validation, con $k = 10$.

Una vez obtenidos los datos, en función de ellos, se seleccionará el modelo que mejor capacidad de clasificación tenga.

2. ESTADO DEL ARTE

2.1. Estudios genéticos de asociación

Tras una búsqueda en la literatura, concretamente a través de PUBMED, se han encontrado estudios de asociación entre polimorfismos genéticos y cesación en el consumo de nicotina [4] y [5].

El primero de ellos es una revisión de más de 30 estudios en los que se buscan variables genéticas que estén relacionadas con la cesación cuando se siguen diferentes tratamientos de sustitución.

El segundo de los artículos encontrados también busca Single Nucleotide Polymorphisms (SNPs) que estén relacionados con la cesación de consumo de tabaco bajo diversos tratamientos, entre los que se encuentra la vareniclina.

Debido a que la vareniclina es un medicamento nuevo, no se han encontrado más estudios farmacogenéticos relacionados con este tratamiento.

La aplicación de la farmacogenética a la hora de buscar tratamientos para dejar de fumar está ganando cada vez más peso, ya que la utilización de inforamción genética y su relación con el éxito de los tratamientos puede permitir el establecimiento de tratamientos a medida de cada paciente. Esto puede ser muy positivo ya que permitiría seleccionar en función de datos genéticos, el mejor tratamiento para la deshaituación tabáquica, garantizando de este modo que el paciente vaya a seguir el tratamiento con el que tenga las mayores posibilidades de éxito.

3. MATERIAL Y MÉTODOS

3.1. Datos

La base de datos que se va a utilizar consta de un total de 472 sujetos para los que se tiene información sobre un total de 329 variables divididas en:

i) Variables genéticas (SNP's): 325.

ii) Variables clínicas (escalas de valoración clínica): 8.

iii) Variable de estudio (variable dependiente): 1.

3.1.1. Recogida de datos.

Los pacientes incluidos en el estudio proceden de diferentes unidades de tabaquismo: Hospital de Cruces (Barakaldo), Hospital de Galdakao, Centro de Salud Mental de Santurtzi, Unidad de Tabaquismo Servicio Cántabro de Salud, Hospital de Requena (Valencia), Hospital 9 de Octubre (Clínica Galiana, Valencia), Hospital Clínico Universitario Lozano Bielsa (Zaragoza), Centro de Salud Universitario "Delicias Sur" (Zaragoza), Unidad de Tabaquismo de Ceuta, Hospital General Yagüe (Burgos), Hospital de Móstoles, Hospital de Córdoba, Hospital Virgen de la Arrixaca (Murcia), Hospital Son Dureta (Palma de Mallorca), Centro de Salud Iturrama (Pamplona), Hospital Río Hortega (Valladolid), Hospital Alta Resolución El Toyo (Almería), Hospital Univesitari de Girona y Unidade de Tabaquismo (Facultade de Medicina da Beira Interior UBI, Potugal).

Todos los pacientes firmaron un consentimiento informado para el genotipado y el tratamiento con vareniclina del tabaquismo. En la primera sesión, los pacientes entregaron el consentimiento firmado y se les extrajo una muestra de sangre de 5 ml, además fueron seleccionados como aptos para formar parte de este estudio tras la realización de un cuestionario básico de tabaquismo, donde están descritos los criterios de inclusión y exclusión. Los pacientes que superaron esta fase fueron sometidos a una valoración médica para descartar criterios médicos que estuvieran incluidos dentro de los criterios de exclusión.

La información genética fue obtenida al analizar el ADN extraído de la muestra de sangre, para lo que se utilizó el procedimiento estándar. El genotipado de los diferentes SNPs se llevó a cabo mediante la tecnología SNPlex [6].

La cronología del estudio es la siguiente:

- Selección de pacientes, en base a los criterios de inclusión y exclusión, firma del consentimiento informado y extracción de sangre. Octubre 2008 - Enero 2009.

- Búsqueda de genes a incluir en el estudio, análisis bibliográfico. Diciembre 2008 - Enero 2009.

- Selección de SNPs, en cada gen se seleccionaron aquellos que pudieran tener implicación funcional. Enero 2009 - Febrero 2009.

- Genotipado, puesta a punto. Febrero 2009 - Abril 2009.

- Genotipado de los SNPs seleccionados en la fase previa. Junio 2009 - Julio 2009.

- Establecimiento de los grupos de tratamiento. Julio 2009 - Septiembre 2009.

- Contraste de eficacia del tratamiento, seguimiento de pacientes y determinación de eficacia y seguridad con asignación a tratamiento aleatoria frente a grupo población con asignación en función del perfil genotípico. Septiembre 2009 - Octubre 2009.

- Determinación de asociaciones. Octubre 2009 - Diciembre 2009.

- Establecimiento de conclusiones y difusión. Noviembre 2009 - Diciembre 2010.

Originalmente, el estudio tambíen analizaba otros tratamientos (TSN y bupropion), pero solamente se han tenido en cuenta en este estudio los pacientes que recibieron vareniclina, debido al tamaño muestral.

En el siguiente cuadro (ver cuadro 1) se recogen las características relacionadas con la edad y el sexo de los pacientes en la rama de vareniclina.

	Edad ($\overline{x} \pm SD$)	Sexo
	$46{,}9 \pm 11{,}51$	Varón 237
	$45{,}52 \pm 10{,}06$	Mujer 241
Total	$46{,}20 \pm 10{,}81$	478

Cuadro 1: Características de los pacientes antes del filtrado.

3.1.2. Descripción de las variables.

Como se ha recogido con anterioridad, las variables que vamos a manejar son de tres tipos: las variables genéticas, las variables clínicas y la variable dependiente.

Variables genéticas:

Single Nucleotide Polymorphism (SNPs).

Es una variación en la secuencia del ADN que afecta a una sola base (adenina (A), timina (T), citosina (C) o guanina (G)) de una secuencia del genoma.

Los SNPs constituyen hasta el 90 % de todas las variaciones que se producen en el genoma humano y pueden afectar a la respuesta de los individuos a: enfermedades, bacterias, virus, productos químicos, fármacos, etc... [7]. Por lo que esta información será vital para realizar una Medicina personalizada en cuanto a prevención y tratamiento [8].

La selección de los SNPs se llevó a cabo en el Departamento de Farmacología de la UPV/EHU siguiendo las referencias publicadas en la literatura acerca de las vías asociadas a la dependencia y a la respuesta a fármacos, en este caso vareniclina. Para conocer la lista donde vienen recogidos los SNPs incluidos en el estudio basta con ir al anexo 1.

Variables clínicas:

Test de Fagerström para la dependencia de la nicotina (FTND).

Esta escala tiene su origen en la escala FTQ construida por Fagerstrom en 1978. En un principio la FTQ constaba de 8 ítems y se diseñó como una herramienta que permitiera mediante autoadministración, medir la dependencia a la nicotina. Estudios posteriores a su publicación encontraron que la FTQ tenía varias deficiencias psicométricas: inaceptable consistencia interna ($\alpha < 0{,}6$) y una estructura multifactorial donde únicamente el primer factor explicaba una parte importante de variablidad [9]. Es por esta razón por la que se decidió excluir dos ítems (contenido de nicotina en cigarrillo y frecuencia de inhalación) y ampliar la puntuación de otros dos (tiempo hasta el primer cigarro y número de cigarrillos al día) [13].

Finalmente la escala utilizada es la FTND, consta de 6 ítems de los cuales:

- Dos de ellos son dicotómicos (respuesta si o no).
- Otro es ordinal con dos posibles respuestas.
- Los restantes tienen un formato Likert con 4 categorías.

En cuanto a sus cualidades psicométricas cabe decir que tiene una validez interna moderada (α de $0{,}58, 0{,}57, 0{,}56, 0{,}61, 0{,}74$) [11] y [12]. Su estructura factorial estudiada mediante análisis factorial exploratorio (EFA) y confirmatorio (CFA), es principalmente de dos factores, uno que hace referencia a la urgencia con la que se restablece el nivel de nicotina tras el sueño y otro que se asocia al mantenimiento del nivel de nicotina a lo largo del día. La diversidad de muestras donde

se ha encontrado esta estructura bifactorial es muy amplia [11], [10], [13], [12], [14] y [15], aunque existen publicaciones en la literatura que abogan por un único factor [13] y [16].

Finalmente cabe decir que es una de las escalas que más se utiliza en el campo del tabaquismo y ha sido validada en multitud de idiomas [10], [15] y [16], entre los cuales está el español.

Escala de estrés percibido (PSS-10).

Tiene sus orígenes en la escala de 14 ítems desarrollada por Cohen, Kamarck y Mermelstein [17]. La PSS se diseñó para poder medir el grado de estrés que un individuo puede experimentar en situaciones de su vida. La estructura de la PSS está formada por dos factores, uno relacionado con los ítems redactados en sentido negativo y el otro con los que lo están en sentido positivo. Posteriormente se descubrió que eliminando los cuatro ítems con menor carga factorial, la escala resultante tenía mejores propiedades en cuanto a validez interna $\alpha = 0{,}78$ frente a $\alpha = 0{,}75$ y además mantenía la misma estructura factorial [24].

Hay que decir que la PSS-10, una escala autoadministrada, es una de las más utilizadas en el mundo para medir el estrés percibido y queda justificado su uso en el ámbito del tabaquismo porque las personas que tienen puntuaciones más altas son más reacias a dejar de fumar que las que obtienen puntuaciones más bajas.

Ha sido validada a multitud de lenguas, entre ellas el castellano [18], y se han obtenido valores de $\alpha = 0{,}82, 0{,}74, 0{,}83, 0{,}86, 0{,}85$ [18], [19], [20], [21], [22] y [23] indicativo de que la validez interna es más que aceptable. Esto junto con la diversidad de muestras utilizadas y los buenos resultados al estudiar su validez externa, hace posible que sea una escala que se pueda utilizar en varios tipos de población, lo cual es muy interesante.

Finalmente en cuanto a su estructura factorial, cabe decir que se ha analizado mediante EFA [20], CFA [19] y [21] y CFA para corroborar los resultados obtenidos con previamente EFA [22] y [23]. La estructura bifactorial está ampliamente aceptada, así como la composición y el porcentaje de varianza explicada por los mismos.

Criterio DSM-IV para evaluar la dependencia de la nicotina.

Es una escala formada por 7 ítems que hacen referencia a 7 síntomas asociados a la dependencia de la nicotina. Cada uno de los ítems tiene una respuesta dicotómica en el sentido de si el paciente presenta el síntoma correspondiente o no. Una vez pasada la escala, mediante la suma de las puntuaciones obtenidas en cada ítem se obtiene una puntuación total y si ésta es superior a 3 [25] el individuo se clasifica como dependiente.

En cuanto al estudio de su estructura, se encuentran en la literatura publicaciones que la analizan desde dos puntos de vista: análisis clásicos y mediante

teoría de respuesta a los ítems (IRT). Según la primera técnica, el α de Chronbach que se tiene es de $\alpha = 0{,}588, 0{,}66$ [25] y [29] lo que indica que el conjunto de ítems mide el mismo constructo, lo cual queda corroborado con EFA y CFA [26], [27], [28] y [29]. Los resultados de los análisis basados en IRT también apoyan la unidimensionalidad [26], [27], [28] y [29] y además establecen un orden en los ítems de la escala en función de su gravedad [27] y [28].

Por lo tanto, la escala DSM-IV es una herramienta válida para poder evaluar la dependencia y como tal, ha sido utilizada en diversos tipos de muestras. Así, ha sido empleada en este estudio.

Escala de ansiedad y depresión de Goldberg (EDAG).

Esta escala fue desarrollada por Goldberg [30], para evaluar el estado de ansiedad y depresión. Está formado por 18 ítems, 9 haciendo referencia a ansiedad y los restantes a depresión. La estructura de respuesta es dicotómica (si o no) en cada ítem y el proceso de administración es diferente al de las escalas anteriores. El autor, utilizando una técnica llamada análisis de rasgo latente, estableció en cada escala una serie de ítems que simplemente sirvieran para discriminar y en función de las respuestas obtenidas, extraer información acerca del resto [31].

El análisis de la estructura de dicha escala se ha hecho desde diferentes puntos de vista: teoría clásica de test [33] y [34], teoría de respuesta a los ítems [31], [33] y [34] y mediante el cálculo de la sensibilidad y especificidad al compararla con otras escalas [32]. En cuanto a la consistencia interna los valores de α obtenidos fueron $\alpha = 0{,}81, 0{,}82$ [33] y [34] que son valores buenos. En cuanto a la estructura factorial cabe decir que verdaderamente es bifactorial con un factor haciendo referencia a la ansiedad y el otro a la depresión, corroborando las hipótesis de su autor [31], [33] y [34]. Finalmente, la sensibilidad y especificidad es del $83{,}1\,\%$ y $81{,}8\,\%$ respectivamente, obtenidos en la validación al castellano de la escala [32].

Debido a sus propiedades, esta escala es utilizada ampliamente y en muestras tan diferentes unas de otras como puede ser la que aquí nos ocupa.

Escala de predicción.

Desarrollada por el hospital Henri Mondor de París, está compuesta por 15 ítems que evalúan las posibilidades de éxito que tiene un paciente a la hora de dejar de fumar [35].

La puntuación total se obtiene sumando las puntuaciones obtenidas en cada uno de los ítems, 5 de ellos tienen una puntuación de 2 puntos si están presentes y el resto 1. De acuerdo con la puntuación total, se pueden clasificar las posibilidades del individuo como: $punt \geq 16$ muchas, $12 \leq punt < 16$ bastantes, $6 \leq punt < 12$ reales y $punt < 6$ muy reducidas [35] y [36].

Por último, queda hablar de su estructura factorial, la cual es unifactorial, lo que permite que la puntuación total obtenida sea una buena herramienta para la

evaluación del paciente.

Escala de confianza.

Es una reducción de la escala de confianza en situaciones de fumar de Condoite y Litchenstein [37]. Está formada por 14 ítems con 10 alternativas de respuesta en una escala de 100 puntos con intervalos de 10 [38]. La obtención de una puntuación de 100 significa que existe una total confianza en que el individuo no fumaría y si es de 0, todo lo contrario.

Es una medida de autoeficacia, es decir, de la convicción del individuo para ejecutar las actitudes necesarias para no fumar [39]. En varios estudios ha sido propuesta como un predictor de las conductas de abstinencia [41]. Por tanto, esta escala es útil para que el paciente conozca qué estímulos están asociados a su comportamiento como fumador [40].

Escala de locus de control (HLC).

Esta escala fue desarrollada por Wallston [42] para poder medir hasta qué punto el individuo puede controlar los eventos que le afectan.

Está compuesta por 11 ítems que tienen un formato de escala de Likert de 6 puntos, 6 de ellos son de dirección externa (percepción sobre eventos que ocurren al azar) y los 5 restantes tienen una dirección interna (percepción de los eventos que ocurren por las propias acciones). La puntuación total de la escala se calcula sumando las puntuaciones obtenidas en cada ítem, teniendo en cuenta que la puntuación de los ítems de dirección interna tiene que ser revertida. El rango de puntuación va de 11 a 66 puntos y a mayor puntuación mayor percepción [42].

Los estudios psicométricos realizados arrojan un coeficiente $\alpha = 0{,}49, 0{,}58, 0{,}72$ [42], [43] y [44]. En cuanto a la estructura factorial de la escala, los resultados encontrados no apoyan la unidimensionalidad de la misma, sino que han encontrado que es bifactorial y los factores están compuestos por los ítems con dirección interna y externa respectivamente [43] y [44].

Test de Russel para la evaluación de los motivos para fumar (RAM test).

Esta escala fue desarrollada por Russell en 1988 y se utiliza para evaluar los motivos que tiene un individuo para fumar. Consta de 24 ítems con un formato Likert de 4 puntos. La puntuación total se calcula sumando las respuestas a cada uno de los ítems y su rango va desde 0 hasta 72. Una mayor puntuación indica una mayor motivación para fumar.

Variable dependiente

Abstinencia a los 12 meses (abst12). Es la variable de interés en el estudio y recoge la información acerca de la abstinencia a los 12 meses de empezar el tratamiento, en este caso con vareniclina. Es una variable binaria que toma el valor 1 cuando el individuo permanece sin fumar a los 12 meses y 0 en otro caso.

Variables genéticas	Variables clínicas	Variable clase
SNPs	FTND PSS-10 DSM-IV EDAG predicción confianza HLC RAM	abst12
Nominales	Continuas (puntuaciones totales)	Binaria

Cuadro 2: Variables originales y sus tipos.

3.1.3. Tratamiento de la base de datos.

Antes de comenzar con los análisis se procedió a analizar cada una de las variables por separado para comprobar si los datos recogidos tenían algún error, lo que puede distorsionar los resultados y en consecuencia hacer que las conclusiones extraídas fueran erróneas.

En cuanto a las variables genéticas (SNPs), las razones para eliminarlas fueron:

i) Falta de variablidad en los datos recogidos.

ii) Presencia de valores perdidos.

Ya que **i)** no aporta ningún tipo de información y **ii)** está relacionado con errores de genotipado. Los SNPs suprimidos fueron: rs1801272, rs28399433, rs6090378 y rs2735917.

El proceso de filtrado de las variables que recogen las respuestas a los ítems de cada una de las escalas (variables clínicas) fue diferente. Los pasos que se siguieron fueron los siguientes:

i) Cálculo del patrón de valores perdidos para la eliminación de los individuos que no hubieran respondido ningún ítem o tuviesen un número elevado de valores perdidos.

ii) Imputación de los valores perdidos con la puntuación más conservadora según fuera la escala analizada. De este modo tratábamos de evitar una sobrestimación en la puntuación de la escala.

El número mínimo de valores perdidos (ítems no respondidos) por individuos según la escala es de:

- 5 para la escala de estrés percibido (PSS-10).

- 5 para la escala de ansiedad y depresión de Goldberg (EDAG).

- 5 para la escala de predicción.

- 5 para la escala de confianza.

- 5 para la escala de locus de control (HLC).

- 5 para la escala de evaluación de los motivos para fumar (RAM).

En cuanto a los valores perdidos presentes en la variable edad, cabe decir que fueron imputados utilizando la edad media.

Finalmente también fueron eliminados los sujetos que tuvieran la variable principal del estudio (abst12) no recogida, 6 sujetos. El siguiente cuadro (ver cuadro 3) recoge las características de la muestra filtrada

	Edad ($\overline{x} \pm SD$)	Sexo	abst12
	$47{,}01 \pm 11{,}51$	Varón 234	Si 222
	$45{,}52 \pm 10{,}02$	Mujer 238	No 250
Total	$46{,}3 \pm 10{,}8$	$n = 472$	

Cuadro 3: Características de los pacientes tras el filtrado.

3.2. Modelos aprendidos: preprocesado de los datos, selección de variables y algoritmos de clasificación.

3.2.1. Introducción a la minería de datos.

La minería de datos surgió a mediados de los años 90. Se puede definir como "el proceso de extraer conocimiento útil y comprensible desde grandes cantidades de datos almacenados en distintos formatos" [45].

Está compuesta por una colección de algoritmos y técnicas procedentes de varias disciplinas como: estadística clásica, inteligencia artificial, visualización de datos, obtención de información,... [46] que posibilitan un análisis moderno de los datos, para lo cual hacen un uso intensivo de las capacidades de memoria y cálculo de los ordenadores.

Las tareas que se llevan a cabo con la minería de datos se pueden clasificar como predictivas y descriptivas. En las tareas predictivas cada elemento de la base de datos se caracteriza por tener unos parámetros de entrada y un parámetro de salida. El objetivo es intentar predecir el valor del parámetro de salida utilizando

la información proporcionada por los parámetros de entrada. Dentro de las tareas predictivas existen también dos tipos:

i) Clasificación. Cada elemento de la base de datos tiene asociado un valor discreto (clase) y el objetivo que se persigue es maximizar el poder de predicción de la clasificación de nuevos elementos para los cuales la clase es desconocida.

ii) Regresión. En este caso el valor asociado a cada elemento es un número real. El objetivo es el mismo, maximizar la capacidad predicción de este valor para un elemento nuevo a través de una función real que se debe aprender.

En cuanto a las tareas descriptivas, los elementos de la base de datos sólo tienen atributos de entrada y el objetivo es el de agruparlos maximizando la similitud entre los elementos de un mismo grupo y minimizando dicha similitud entre los diferentes grupos. Un ejemplo claro de esto es el clustering.

Los campos de aplicación de la minería de datos son variopintos: telefonía móvil, banca, marketing,... Pero dos de los campos donde la aplicación de la minería de datos ha tenido un mayor impacto han sido la medicina y la bioinformática. A raíz del desarrollo de numerosas técnicas que permiten la recogida de datos biológicos, la cantidad de información ha sufrido un grandísimo aumento, lo que ha obligado a utilizar estas técnicas de análisis de datos.

3.2.2. Preprocesado.

Una vez filtrada la base de datos siguiendo el procedimiento mencionado en el apartado anterior, imputados los valores perdidos mediante la media o moda según sea la variable continua o discreta y discretizadas las variables continuas en cuatro intervalos de igual frecuencia, se comenzó el análisis de la misma de acuerdo a la hipótesis del trabajo: comparar la capacidad de predicción de la abstinencia cuando se usan modelos de regresión logística y de minería de datos.

El primer paso fue analizar la estructura factorial de las escalas, ya que en la literatura se ha visto que muchas de ellas no eran unidimensionales. Para ello utilizamos técnicas de teoría de respuesta a los ítems no paramétricas, puesto que son más adecuadas que las utilizadas en los artículos citados en la descripción de cada una de las escalas. Las técnicas utilizadas fueron el Mokken Scale Analysis y el algoritmo AISP.

El Mokken Scale Analysis es una técnica no paramétrica de teoría de respuesta a los ítems que se basa en tres hipótesis: unidimensionalidad, independencia local y monotonía [47]. Permite que los ítems sean tanto dicotómicos como politómicos [48], como es nuestro caso. Además, permite conocer la estructura latente de cada una de las escalas que conforman las variables clínicas, es decir, permite

distinguir los diferentes conceptos que miden (en caso de que las escalas no sean unidimensionales) y proporciona los subconjuntos unidimensionales (subescalas) asociados con cada uno de ellos [49]. Esto es muy interesante, ya que aunque aumente el número de variables clínicas, permite utilizar la suma de las puntuaciones de los ítems de cada una de las subescalas de manera correcta, es decir, sin que esté contaminada, ya que la puntuación no se verá alterada por la presencia de ítems que midan un aspecto diferente al de la subescala.

La reducción de las escalas se llevó de acuerdo a esta técnica, calculando los índices de escalabilidad, H_{ij}, H_i y H [50], indicativos de la calidad de los ítems y de cada subescala. Mediante el algoritmo AISP [50] se construyeron conjuntos de ítems que verificaran las tres hipótesis anteriores y además tales que el índice H de cada uno de ellos fuera $H_i \geq c$ con $c = 0{,}3$, recogido en la literatura como el valor mínimo que se debe tener. Los análisis fueron realizados con el software estadístico R [52] mediante la librería mokken [50].

Para facilitar el proceso de estimación cuando se aplicaron los algoritmos de clasificación (en particular Naive-Bayes), las variables fueron discretizadas en cuatro intervalos de igual frecuencia [51].

3.2.3. Selección de variables.

Previo a la utilización de los algoritmos de clasificación, se llevó a cabo un proceso de selección de variables, para poder eliminar las que no aportan información y así poder mejorar la capacidad predictiva de los clasificadores [62] y [67]. Los algoritmos de selección de variables se pueden clasificar principalmente en dos grupos diferentes: filters y wrappers. Los métodos de selección pertenecientes al primer grupo utilizan heurísticos de búsqueda basados en las características de los datos para seleccionar las variables. En cambio, los pertenecientes al segundo grupo emplean un algoritmo de clasificación para estimar la precisión de cada uno de los subconjuntos que se van obteniendo (ver figura 1), de tal modo que aquel subconjunto de variables con una mayor precisión de acuerdo al algoritmo empleado es el que se utilizará en el problema de clasificación [60].

Por tanto, la diferencia principal entre estos dos tipos de métodos de selección de variables es la utilización de un algoritmo de clasificación, lo que hace que, en general los filters sean computacionalmente más rápidos, pero son los wrappers los que ofrecen mejores subconjuntos, entre otras diferencias [53]

i) Relief.

ii) Probabilistic feature selection (PFS).

iii) Fast Correlation-Based Filter (FCBF).

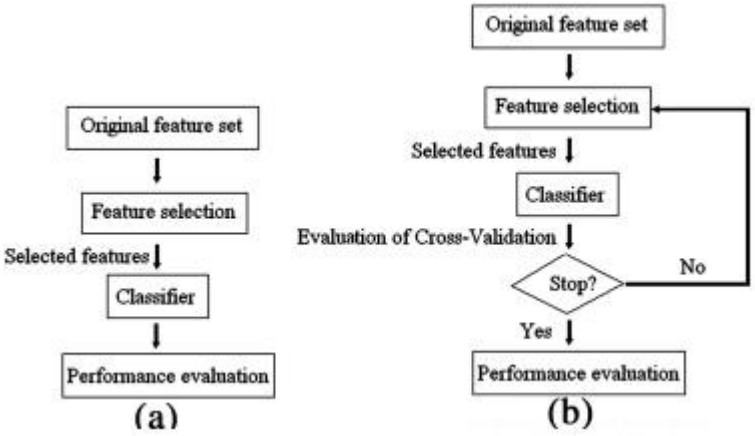

Figura 1: a) método filter; b) método wrapper

iv) Correlated Feature Selection (CFS).

v) Wrapper.

Antes de comenzar con la explicación de los diferentes algoritmos de selección y los métodos de clasificación, vamos a introducir la notación necesaria para dichas explicaciones:

- $\mathbf{X} = \langle X_1, X_2, \ldots, X_k \rangle$ es un vector de variables aleatorias que denotan los valores observados en las variables de cada instancia.
- $\mathbf{x} = \langle x_1, x_2, \ldots, x_k \rangle$ denota a una instancia.
- $\mathbf{X} = \mathbf{x}$ es una abreviatura de $X_1 = x_1 \wedge X_2 = x_2 \wedge \ldots \wedge X_k = x_k$.
- C es una variable aleatoria que denota la clase a la que pertenece cada instancia y c denota un valor determinado.
- n denota el número total de instancias.

i) Relief:

Es un algoritmo que pondera las variables de la base de datos, originalmente fue diseñado para problemas que involucraban dos clases [54], pero más tarde fue generalizado para poder ser utilizado en otros tipos de problemas [56]. Dados un conjunto de datos D de tamaño n y un umbral de relevancia $\tau, 0 \leq \tau \leq 1$, el algoritmo detecta aquellas variables que sean estadísticamente significativas con respecto a la variable de estudio utilizando un método estadístico.

Dadas dos instancias **X** e **Y** la diferencia entre ellas se calcula [55]:

$$diff(x_i, y_i) = \begin{cases} 0 & \text{si } x_i = y_i \\ 1 & \text{si } x_i \neq y_i \end{cases}$$

con $i = 1, \ldots, k$ en el caso de que sean nominales. En caso de que sean numéricas:

$$diff(x_i, y_i) = \frac{x_i - y_i}{nu_i}$$

con $i = 1, \ldots, k$ siendo nu_k un normalizador para que $0 \leq diff(x_i, y_i) \leq 1$.

Relief escoge un conjunto de m ternas: X, near-hit (instancia con la misma clase más próxima) y near-miss (instancia con distinta clase más próxima). Mediante un proceso iterativo actualiza el peso de la i-ésima variable mediante la fórmula

$$W_i = W_i - diff(x_i, near - hit_i)^2 + diff(x_i, near - miss_i)^2, i = 1, \ldots, k$$

seleccionando aquellas variables cuyo $W_i \geq \tau$ una vez finalizado el proceso iterativo.

Una vez ordenadas las variables según su relevancia, las variables seleccionadas finalmente fueron aquellas con una relevancia positiva [54].

ii) Probabilistic feature selection (PFS):

Este algoritmo de selección de variables utiliza elecciones aleatorias para guiar de una manera más rápida la búsqueda de una solución correcta [57]. Además utiliza la aleatoriedad para garantizar que, aunque las decisiones que se hayan tomado no hayan sido correctas, la solución sea válida. Si esto sucediese, el algoritmo sólo necesitaría más tiempo [58].

El algoritmo genera en cada paso, de manera aleatoria, subconjuntos de variables. Cuando crea un subconjunto (S_{new}) de menor cardinalidad que el mejor subconjunto obtenido hasta el momento, es decir $|S_{new}| \leq |S_{best}|$, lo que hace es tomar las variables que forman S_{new} y utiliza los datos para chequear el criterio de inconsistencia para S_{new}. Si la tasa de inconsistencia para S_{new} es menor que un valor determinado γ, entonces se hace $S_{new} = S_{best}$.

La clave del funcionamiento del algoritmo es el criterio de inconsistencia que especifica hasta qué punto se puede reducir el número de variables. El índice de inconsistencia se calcula de la siguiente manera: dos instancias se consideran inconsistentes si coinciden en todos los valores de las variables salvo en la clase. Después se calcula para todas las instancias, sin tener en cuenta la clase, el inconsistency count de la siguiente manera:

$$IC = n - max_{i=1,\ldots,k}(n_{c_i})$$

siendo n_{c_i} el número de instancias pertenecientes a la clase i-ésima y n el número total de instancias. Finalmente la tasa de inconsistencia se calcula:

$$IR = \frac{\sum_{i=1}^{k}(IC_i)}{n}$$

El método de búsqueda utilizado fue Greedy Stepwise. Este heurístico construye un subconjunto de variables eligiendo en cada paso la mejor opción. La búsqueda puede ser hacia adelante, añadiendo variables, empezando con un subconjunto de variables vacío o hacia atrás, eliminando variables, comenzando con un subconjunto que tiene a todas las variables, de tal manera que mejore la calidad en cada decisión [59]. Pero hay que destacar que la solución final que proporciona (el subconjunto de variables), es subóptima ya que es un óptimo local. La búsqueda finaliza cuando al añadir o eliminar variables no mejora la calidad del subconjunto.

iii) Fast Correlation-Based Filter (FCBF): Este algoritmo de selección de variables se basa en la correlación. Antes de nada se deben definir varios conceptos necesarios para poder explicar el funcionamiento del algoritmo: relevancia y redundancia.

Relevancia: Se dice que una variable X_i es relevante [60] respecto a la clase C si y sólo si existe x_i y c con $P(X_i = x_i) > 0$ tal que:

$$P(C = c|X_i = x_i) \neq P(C = c)$$

Redundancia: Se dice que X es redundante [60] si existe al menos otra variable altamente correlada con ella. Estas dos definiciones nos llevan a la siguiente hipótesis [60]: "Un subconjunto de variables es bueno si contiene variables altamente correladas con la clase pero a la vez independientes entre si".

Dicho esto, lo que debemos hacer es establecer una medida de correlación. En un primer momento podríamos utilizar la correlación de Pearson, pero tiene varias limitaciones [61], por ejemplo las variables no tienen que ser todas continuas. Es por esto que la medida de correlación que se va a usar está basada en la entropía, que mide la incertidumbre de un sistema. Dada una variable X, se define su entropía de la siguiente manera:

$$H(X) = -\sum_{i=1} P(x_i)log_2(P(x_i))$$

La entropía de una variable X dada otra variable Y se calcula mediante:

$$H(X|Y) = \sum_j P(y_j) \sum_i P(x_i|y_j)log_2(P(x_i|y_j))$$

Definidos estos conceptos podemos calcular la información sobre X conocida Y como:

$$IG(X,Y) = H(X) - H(X|Y),$$

que es una medida simétrica [61]. Finalmente, como esta medida está sesgada a favor de las variables con muchos valores, construimos la medida que vamos a utilizar para la selección de las variables. Se define la symmetrical uncertainty como:

$$SU(X,Y) = 2[\frac{IG(X,Y)}{H(X)+H(Y)}] \in [0,1]$$

$SU(X,Y) = 1$ indica que el conocimiento de cualquiera de las dos variables predice completamente la otra, mientras que $SU(X,Y) = 0$ quiere decir que son independientes.

Una vez definida la medida de correlación que se va a utilizar, vamos a introducir dos conceptos más que se necesitan para comprender el funcionamiento del algoritmo FCBF: correlación predominante y variable predominante.

Correlación predominante: Dado un subconjunto de variables se dice que la correlación existente entre una variable $X_i \in S$ y la variable clase C es predominante si y sólo si $SU(X_i,C) \geq \delta$ y para cualquier $X_j \in S(i \neq j)$, no existe $X_j|SU(X_j,X_i) \geq SU(X_i,C)$. Si existe tal X_j para X_i, se dice que es un par redundante de X_i y se define S_{P_i} como el conjunto de todos los pares redundantes de X_i que se puede dividir en dos partes: $S^+_{P_i} = \{X_j|X_j \in S_{P_i}, SU(X_j,c) > SU(X_i,C)\}$ y
$S^-_{P_i} = \{X_j|X_j \in S_{P_i}, SU(X_j,c) \leq SU(X_i,C)\}$.

Variable predominante: Una variable X_i es predominante para la variable clase C si y sólo si $SU(X_i,C)$ es predominante o puede llegar a serlo al eliminar todas aquellas variables que sean pares redundantes de X_i.

El algoritmo FCBF está compuesto de dos partes. En la primera de ellas, el algoritmo calcula la SU para cada variable, selecciona las variables relevantes tomando un δ predeterminado y las ordena en orden descendente de acuerdo a SU en un conjunto S'_{list}. En la segunda parte el algoritmo se encarga de eliminar todas las variables redundantes de S'_{list} hasta que quedan únicamente aquellas que son predominantes, para lo que utiliza tres heurísticos [61].

Una de las principales características de este algoritmo es que reduce ampliamente el número de variables en un tiempo de computación asumible [61].

iv) Correlated Feature Selection (CFS):

Este algoritmo de selección de variables tiene en cuenta la utilidad de las variables para predecir la clase y también el nivel de correlación entre ellas [67]. La hipótesis que plantea la hemos visto antes, un subconjunto de variables será "bueno" si las variables que lo forman están altamente correladas con la clase

mientras que muy poco entre si. La clave del algoritmo reside en esta fórmula:

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Esta fórmula no es más que la correlación de Pearson cuando las variables han sido estandarizadas [68]. El numerador recoge la correlación que existe entre las variables del subconjunto S con la variable clase (dependiente) mientras que en el denominador se calcula el nivel de correlación que existe entre todos los pares de variables de S [60]. Por lo tanto un subconjunto S será bueno si $Merit_S$ está próximo a 1, es decir, si $\overline{r_{ff}}$ es inferior a $\overline{r_{cf}}$ lo que no es más que la hipótesis planteada anteriormente.

La medida de correlación utilizada vuelve a ser la symmetrical uncertainty debido a las limitaciones de la correlación de Pearson explicadas antes. El método de búsqueda de las variables que forman los subconjuntos es el BestFirst [66]. Este heurístico es una variante del greedy forward; BestFirst mantiene una lista ordenada con las mejores soluciones, en nuestro caso los subconjuntos de variables más prometedores, encontradas hasta el momento. El heurístico finaliza la búsqueda cuando la expansión de la lista con un número determinado de soluciones (en WEKA [65] por defecto son 5) no aporta una solución que mejore a la mejor solución hallada hasta ese momento, que es la que devuelve.

v) Wrapper: Este método de selección de variables es diferente a los que se han utilizado previamente. La diferencia principal radica en que utiliza un clasificador para evaluar cada uno de los diferentes subconjuntos que se van seleccionando [62], mientras que los anteriores únicamente tienen en cuenta las características de las variables (filters) [63].

Como tenemos dos clasificadores, Naive-Bayes y el basado en regresión logística, vamos a utilizarlos como evaluadores de los subconjuntos. Es decir, cuando seleccionemos las variables para el clasificador Naive-Bayes, lo utilizaremos como evaluador en el proceso de selección. Lo mismo haremos para el clasificador que utiliza la regresión logística, también lo emplearemos para poder seleccionar el mejor subconjunto. Como criterio de selección utilizaremos el % de bien clasificados, por lo que se utilizará el subconjunto que mayor % de bien clasificados obtenga cuando sea testado durante el proceso.

Para mantener la objetividad durante todo el proceso vamos a seleccionar las variables mediante 10-fold cross validation. En otras palabras, realizaremos la búsqueda de subconjuntos sobre 9 de las 10 fold y emplearemos la restante para calcular el % de bien clasificados de cada subconjunto de variables. Al final, el subconjunto seleccionado será el que mejor % tenga.

El método de búsqueda en el espacio de subconjuntos de variables es nuevamente el Best-First [66].

3.2.4. Algoritmos de clasificación.

A continuación vamos a describir los algoritmos de clasificación utilizados. Son los siguientes:

i) Naive-Bayes.

ii) Regresión logística.

Estos algoritmos requieren una partición de la base de datos en dos conjuntos independientes. El primero de ellos denominado *training set* sirve para construir el modelo y el segundo conjunto, *test set*, sirve para predecir las clases de las instancias (desconocidas para el modelo). Es decir, el *test set* se emplea como si fuera un conjunto de datos "nuevo" para el que el modelo construido con el *training set* debe predecir la clase [60].

i) Naive-Bayes: Este clasificador se basa en el teorema de Bayes para predecir la clase de una nueva instancia. El teorema es el siguiente:

$$P(C = c|\mathbf{X}{=}\mathbf{x}) = \frac{P(C = c)P(\mathbf{X}{=}\mathbf{x}|C = c)}{P(\mathbf{X}{=}\mathbf{x})}$$

El denominador es invariante respecto de la clase a la que pertenezcan las instancias por lo que su eliminación en la expresión anterior no afecta la clasificación con lo que:

$$P(C = c|\mathbf{X}{=}\mathbf{x}) \propto P(C = c)P(\mathbf{X}{=}\mathbf{x}|C = c)$$

Donde $P(C = c)$ y $P(\mathbf{X}{=}\mathbf{x}|C = c)$ se estiman utilizando los datos del *training set*.

En el proceso de estimación de $P(\mathbf{X}{=}\mathbf{x}|C = c)$ nos encontramos con un problema, en general no se puede calcular directamente de los datos porque no siempre se tienen en el *training set* todas las posibles instancias **x**. Para poder solucionarlo tenemos que asumir una fuerte hipótesis sobre los datos, esta hipótesis es que las variables $X_1, \ldots, X_k$ son condicionalmente independientes dada la clase. Es decir, que una vez proporcionado el valor de la clase (conocido en el *training set*), las variables son independientes, lo que hace posible que:

$$P(\mathbf{X}{=}\mathbf{x}|C = c) = P(\bigwedge_{i=i}^{k} X_i = x_i|C = c) = \prod_{i=1}^{k} P(X_i = x_i|C = c),$$

lo que nos lleva a la ecuación que estima la probabilidad de la clase dados los valores de las variables:

$$P(C = c|\mathbf{X}{=}\mathbf{x}) \propto P(C = c) = \prod_{i=1}^{k} P(X_i = x_i|C = c).$$

De este modo, una vez calculadas estas probabilidades a partir del *training set*, las clases de las instancias pertenecientes al *test set* se estimarán de la siguiente manera:

$$c^* = argmax_c(P(C = c|\mathbf{X}=\mathbf{x}))$$

para cada **x**.

Cuando se enfrenta a una variable X_i discreta utiliza las tablas de frecuencia en el proceso de estimación de las probabilidades, mientras que si es continua asume una cierta distribución de probabilidad (debemos de conocer los parámetros). Es por esto que si la variable es continua [72] y aunque sea discreta pero tome muchos valores, para favorecer la estimación las variables conviene que se discreticen en un número no muy grande de intervalos. De este modo se consigue que tomen un número reducido de valores y así se evite la presencia de valores con una frecuencia demasiado pequeña [51].

El proceso de aprendizaje del modelo es muy sencillo [60], además en contra de lo que se puede pensar, la presencia de variables correladas entre si no afecta tanto a las estimaciones del modelo [69]. Además proporciona una alta precisión y velocidad cuando se aplica a grandes bases de datos [70] y es especialmente popular en el ámbito médico [71].

Como se puede apreciar en el dibujo (ver figura 2), es un caso particular de red bayesiana. Basta con no permitir relaciones entre las variables, salvo que sea la variable clase, que es ancestro de todas.

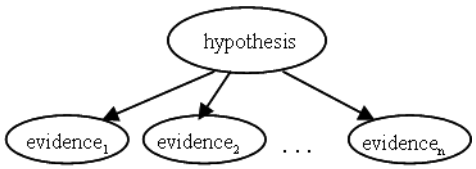

Figura 2: Ejemplo de Naive-Bayes

ii) Regresión logística:

El modelo de regresión logística permite establecer una relación entre un conjunto de variables independientes, continuas o discretas, y una variable dependiente dicotómica [73]. La estructura general del modelo es la siguiente:

$$logit(p(\mathbf{x})) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

Donde la función logit (ver figura 3) que es $logit : [0, 1] \longrightarrow \mathbb{R}$ tal que $logit(p) = ln\frac{p}{1-p}$ siendo $p = E(c_i|\mathbf{x})$, se emplea como función de enlace.

En el caso de que tengamos variables discretas como es el nuestro, la fórmula varía un poco puesto que tenemos que utilizar *dummy variables* para recoger la

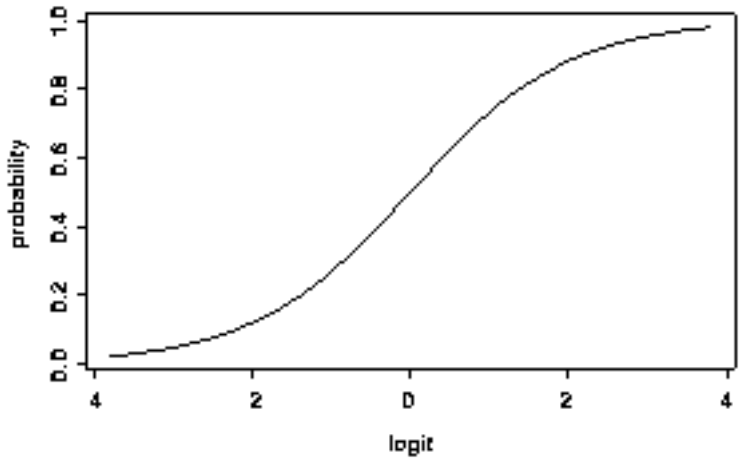

Figura 3: Función logit

información de cada uno de los valores que toma la variable. Suponiendo que la variable X_i toma l_i valores, el número de *dummy variables* necesarias es $l_i - 1$ [73]. Por lo tanto la expresión del modelo en nuestro caso es el siguiente:

$$logit(p(\mathbf{x})) = \beta_0 + \sum_{h=1}^{l_1-1} \beta_{1h} D_{1h} + \ldots + \sum_{h=1}^{l_k-1} \beta_{kh} D_{kh}$$

En la construcción del modelo se estiman los parámetros β a partir de los datos. Como ya hemos dicho, los algoritmos de clasificación requieren una partición de la base de datos en *training test* y *test set*. El bloque de entrenamiento (en el que conocemos la clase a la que pertenece cada instancia) se va a emplear para calcular los estimadores $\widehat{\beta}$ de los parámetros β del modelo. Una vez conocidos los estimadores emplearemos la fórmula:

$$\widehat{p} = \frac{e^{\mathbf{x}\widehat{\beta}}}{1 + e^{\mathbf{x}\widehat{\beta}}},$$

para predecir la clase a la que pertenecen las instancias del *test set* y así validar el modelo. Si $\widehat{p} > 0{,}5$ a la instancia correspondiente se le asociará $c = 1$ y en caso contrario se le asignará $c = 0$.

El método de estimación de los parámetros del modelo de regresión logística en WEKA es el que proporciona *ridge estimators* [74].

3.3. Estimación de la capacidad de predicción de los modelos aprendidos y selección del modelo final.

Cuando se construye un modelo, lo que interesa es que su capacidad de predicción sea alta, es decir, que su error de predicción, la probabilidad de clasificar mal a una nueva instancia, sea bajo. Cada clasificador tiene asociado dicho error pero en general no es conocido, esto se debe a que la distribución de la variable clase es desconocida [75]. Por lo tanto el objetivo es tener un estimador de este error lo más preciso posible, es decir, con sesgo y varianza muy pequeños [76]. El sesgo y su varianza se definen como:

$$Bias = \epsilon - E(\widehat{\epsilon})$$

$$Var = E(\epsilon - E(\widehat{\epsilon}))$$

Donde ϵ es el error real del clasificador y $\widehat{\epsilon}$ es la estimación del error [75].

Entre los diferentes métodos de estimación que existen, el que se ha seleccionado, debido a sus propiedades [76], ha sido el k-fold cross validation [77] (ver figura 4) con $k = 10$ [75]. Este método divide los datos en k subconjuntos mutuamente excluyentes aproximadamente del mismo tamaño. El clasificador se va a entrenar y testar k veces de tal manera que en cada paso i con $i \in \{1, \ldots, k\}$, se utilizan todos los subconjuntos excepto el i-ésimo como *training test* para construir los clasificadores (NB y RL) y el i-ésimo como *test set* para evaluarlos. En cada paso se calcula el error de predicción del clasificador $\widehat{p_i}$ y como estimador del error del modelo se proporciona:

$$\widehat{p_M} = \frac{1}{k} \sum_{i=1}^{k} \widehat{p_i}.$$

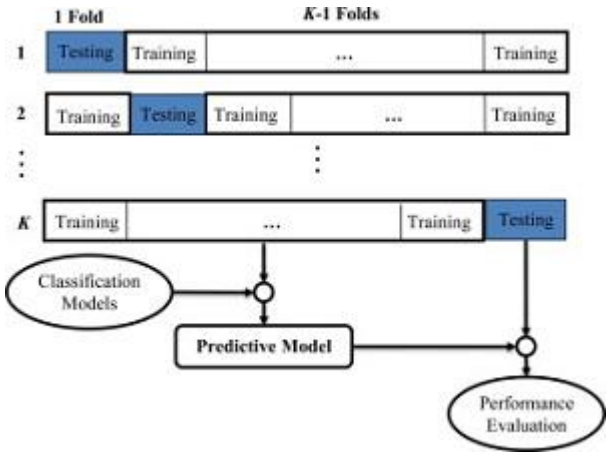

Figura 4: k-fold cross validation

Finalmente, para comparar los dos modelos construidos, vamos a utilizar los estimadores proporcionados por el k-fold cross validation del error del modelo y el del área bajo la curva ROC.

Esta curva es una representación gráfica de la sensibilidad frente a (1 - especificidad) para un sistema clasificador binario según se varía el umbral de discriminación. El análisis del área bajo esta curva permite seleccionar aquellos modelos de clasificación óptimos. Dependiendo del área bajo la curva, los modelos se clasifican como: $[0{,}5, 6)$ test malo, $[0{,}6, 0{,}75)$ test regular, $[0{,}75, 0{,}9)$ test bueno, $[0{,}9, 0{,}97)$ test muy bueno y $[0{,}97, 1)$ test excelente.

3.4. Software

El software utilizado en el estudio de las propiedades psicométricas y de la estructuras factoriales de las escalas fue R en su versión 2.15.0 [52]. En el proceso de discretización, selección de variables, clasificación, evaluación y selección de los modelos, utilizando Naive-Bayes y regresión logística, se empleo el paquete de minería de datos WEKA en su versión 3.7 con los módulos Explorer, KnowledgeFlow y Experimenter.

4. RESULTADOS

Lo primero que se analizó fueron las estructuras de las variables clínicas (escalas) mediante Mokken Scale Analysis [47] empleando para ello el algoritmo AISP [50] disponible en el paquete de R [52] mokken [50]. En las siguientes tablas vamos a ver en cuántas subescalas unidimensionales han quedado divididas dichas variables, ya que de este modo podremos utilizar la puntuación total de cada una de ellas como variables explicativas en los modelos. Como se explicó anteriormente, el número de variables clínicas puede aumentar ya que no todas las escalas originales tienen porque ser unidimensionales y es lo que ha ocurrido. Para que quede explicado, una vez presentados los resultados obtenidos del análisis de Mokken, se resumirá en una tabla cuántas subescalas se han obtenido y qué aspecto miden. De este modo quedará justificado el porqué del aumento del número de variables clínicas.

Subescala 1	Subescala 2
8	3
7	1
5	2
4	10
	6
	9
$H_1 = 0{,}46$	$H_2 = 0{,}42$

Cuadro 4: Escala PSS-10

Los ítems que forman la escala de estrés percibido (PSS-10) [17], se distribuyen en dos subescalas (ver cuadro 4) cuyos coeficientes de escalabilidad indican que es media [47]. Además, no ha dejado ningún ítem sin clasificar. La primera subescala contiene ítems que hacen referencia a la capacidad del sujeto de controlar la situación durante el último mes, mientras que las preguntas que forman la segunda subescala están relacionadas con el descontrol de la situación a lo largo del último mes.

De los ítems que forman la escala de ansiedad y depresión de Goldberg (EDAG) (ver cuadro 5) [30], únicamente el a6 y el d6 han quedado fuera de las dos subescalas. Los índices de escalabilidad de las mismas indican que su escalabilidad es media, a pesar de que los aspectos que miden las dos subescalas son muy similares, ya que tienen ítems que miden ansiedad y depresión. Para que la puntuación total tomando todos los íems a la vez no se vea contaminada por estos 2 factores, vamos a tomar las puntuaciones totales de las dos subescalas por separado.

Subescala 1	Subescala 2
d8, a4	a9
d1, a1	a5
d2, a3	a7
d3, a2	d7
d4, d9	
d5, a8	
$H_1 = 0{,}41$	$H_2 = 0{,}4$

Cuadro 5: Escala EDAG

Subescala 1	Subescala 2
7	4
5	1
13	8
14	3
6	
11	
2	
10	
9	
$H_1 = 0{,}51$	$H_2 = 0{,}68$

Cuadro 6: Escala de confianza

Los ítems que forman la escala de predicción [35] salvo 1b, 1c, 2 y 8 forman una escala unidimensional con índice $H = 0{,}4$, lo que significa que además de medir el mismo constructo, predicción del éxito a la hora de dejar de fumar, lo hace con una escalabilidad media [47].

Los ítems de la escala de confianza [37] se han clasificado en dos subescalas (ver cuadro 6) cuyos índices H son superiores a $0{,}5$ lo que indica que tienen una escalabilidad fuerte [47]. La primera de ellas trata de medir la asociación del tabaco con otros estímulos y con la imagen que se tiene de uno mismo. La segunda subescala, recoge información relacionada con ansiedad.

Los ítems que forman la escala de locus de control (HLC) [42] no tienen la suficiente calidad como para que se tengan en cuenta en el análisis posterior. No sólo las subescalas están formadas por dos ítems (ver cuadro 7), lo que no tiene sentido, sino que además los ítems 1, 4 y 5 no han sido asignados a ninguna subescala.

Subescala 1	Subescala 2	Subescala 3	Subescala 4
9	11	7	8
3	10	6	2
$H_1 = 0{,}46$	$H_2 = 0{,}43$	$H_3 = 0{,}35$	$H_4 = 0{,}34$

Cuadro 7: Escala HLC

Subescala 1	Subescala 2	Subescala 3	Subescala 4
22, 23	16	17, 12	15, 14
10, 19	3	8, 24	2, 21
5	18	13	9, 1
6		7	11
$H_1 = 0{,}4$	$H_2 = 0{,}39$	$H_3 = 0{,}36$	$H_4 = 0{,}36$

Cuadro 8: Escala RAM

Los ítems que forman la escala RAM están distribuidos en cuatro subescalas (ver cuadro 8) cuyas escalabilidades no son excesivamente buenas, pero como tienen un número suficiente de ítems se van a utilizar en el análisis. La primera de las subescalas contiene ítems que están relacionados con la imagen que proyecta el fumador, mientras que las cuestiones que forman la segunda, recogen información sobre el momento del día en el que más se disfruta fumando. La tercera de las subescalas se centra en la sensación de concentración y activación en el momento de fumar, y la cuarta y última hace referencia al automatismo y la dificultad a la hora de dejar de fumar.

Finalmente decir que tanto la escala DSM-IV [25] y el test de Fagerstrom (FTND) [9] se asumieron ambas unidimensionales puesto que se utilizan para medir un único constructo, la dependencia a la nicotina.

En la siguiente tabla (ver cuadro 9) se recoge la información de cada una de las subescalas que se han obtenido mediante el Mokken analysis y la aplicación del algoritmo AISP.

Para conocer los ítems de cada una de las escalas basta con ir al anexo 2.

Con respecto al proceso de selección de variables, cabe decir que se aplicaron diferentes métodos, desde los que utilizaban las propiedades de las variables únicamente: Relief [54], PFS [57], basado en la medida de correlación symmetrical uncertainty [61] y CFS [68], hasta los que mediante la utilización de un algoritmo de aprendizaje evalúan a los subconjuntos que se van obteniendo y finalmente seleccionar el que menor tasa de error proporciona: wrapper con Naive-Bayes o Regresión Logística como evaluador de los subconjuntos.

En el cuadro 11 se puede observar el número de variables seleccionadas por

Escala	Subescalas	Aspecto de medición
FTND	1	Dependencia
PSS-10	1	Control situación
	2	Descontrol situación
DSM-IV	1	Dependencia
EDAG	1	Ansiedad y depresión
	2	Ansiedad y depresión
Predicción	1	Éxito dejar de fumar
Confianza	1	Autoimagen
	2	Ansiedad
RAM	1	Imagen exterior
	2	Disfrute
	3	Estimulación
	4	Dependencia

Cuadro 9: Subescalas obtenidas y aspecto que miden.

Variables genéticas	Variables clínicas ($\overline{x} \pm SD$)	Variable clase
SNPs	FTND ($6{,}1 \pm 2{,}3$)	abst12
	PSS-10 subescala 1 ($10{,}6 \pm 3{,}5$)	
	PSS-10 subescala 2 ($11{,}8 \pm 6{,}1$)	
	DSM-IV ($3{,}9 \pm 1{,}7$)	
	EDAG subescala 1 ($18{,}2 \pm 6{,}1$)	
	EDAG subescala 2 ($6 \pm 2{,}1$)	
	predicción ($8{,}8 \pm 3{,}3$)	
	confianza subescala 1 ($431{,}4 \pm 315{,}3$)	
	confianza subescala 2 ($186{,}6 \pm 129{,}8$)	
	RAM subescala 1 ($5{,}4 \pm 5$)	
	RAM subescala 2 ($6{,}5 \pm 2{,}2$)	
	RAM subescala 3 ($11{,}6 \pm 4{,}2$)	
	RAM subescala 4 ($12{,}4 \pm 5$)	
Nominales	Continuas (puntuaciones totales)	Binaria

Cuadro 10: Variables finales y sus tipos.

Método	Vars. seleccionadas	reducción (%)
Relief + Ranker	172	49.26
PFS + Greedy Stepwise	8	97.64
FCBF + Symmetrical Uncertainty	8	97.64
CFS + BestFirst	59	82.6
Wrapper (Naive-Bayes)+ Best First	6	98.23
Wrapper (Reg. logística)+ Best First	7	97.94
Nºde variables		339

Cuadro 11: Selección por filtro y heurístico de búsqueda.

Método	% bien	AUC (ROC)
Relief + Ranker	$56{,}22 \pm 7{,}09$	$0{,}57 \pm 0{,}07$
Probabilístico + Greedy Stepwise	$44{,}11 \pm 6{,}66$	$0{,}60 \pm 0{,}08$
FCBF + Correlación	$62{,}55 \pm 6{,}71$	$0{,}67 \pm 0{,}07$
CFS + BestFirst	$67{,}29 \pm 5{,}90$	$0{,}73 \pm 0{,}07$
Wrapper (Naive-Bayes) + Best First	$61{,}77 \pm 7{,}49$	$0{,}58 \pm 0{,}08$
Completa	$50{,}80 \pm 7{,}52$	$0{,}51 \pm 0{,}08$

Cuadro 12: Comportamiento del clasificador Naive-Bayes.

cada método, así como el porcentaje de variables descartadas por los mismos con respecto al número total de variables presentes en la base de datos.

En los dos siguientes cuadros (ver cuadros 12 y 13) se puede ver el comportamiento de cada uno de los clasificadores utilizados, Naive-Bayes (ver cuadro 12) y el basado en la regresión logística (ver cuadro 13). Los parámetros recogidos para ver dicho comportamiento son: porcentaje de bien clasificados y área bajo la curva ROC. Los tres parámetros están acompañados del error estándar originado al realizar la 10-fold cross validation.

Para poder comparar la información presente en ambas tablas de una manera gráfica, se han construido las siguientes figuras (ver figuras 5 y 6) siendo los métodos de selección de variables los siguientes: 1 Totalidad de variables; 2 Relief; 3 PFS; 4 FCBF; 5 CFS; 6 Wrapper:

Método	% acierto	AUC (ROC)
Relief + Ranker	$49{,}40 \pm 6{,}15$	$0{,}49 \pm 0{,}07$
PFS + Greedy Stepwise	$57{,}18 \pm 7{,}07$	$0{,}60 \pm 0{,}08$
FCBF + Symmetrical Uncertainty	$61{,}76 \pm 6{,}55$	$0{,}67 \pm 0{,}08$
CFS + BestFirst	$64{,}25 \pm 6{,}86$	$0{,}69 \pm 0{,}07$
Wrapper (Reg. logística) + Best First	$60{,}63 \pm 7{,}05$	$0{,}60 \pm 0{,}08$
Completa	$50{,}38 \pm 6{,}68$	$0{,}50 \pm 0{,}08$

Cuadro 13: Comportamiento del clasificador basado en regresión logística.

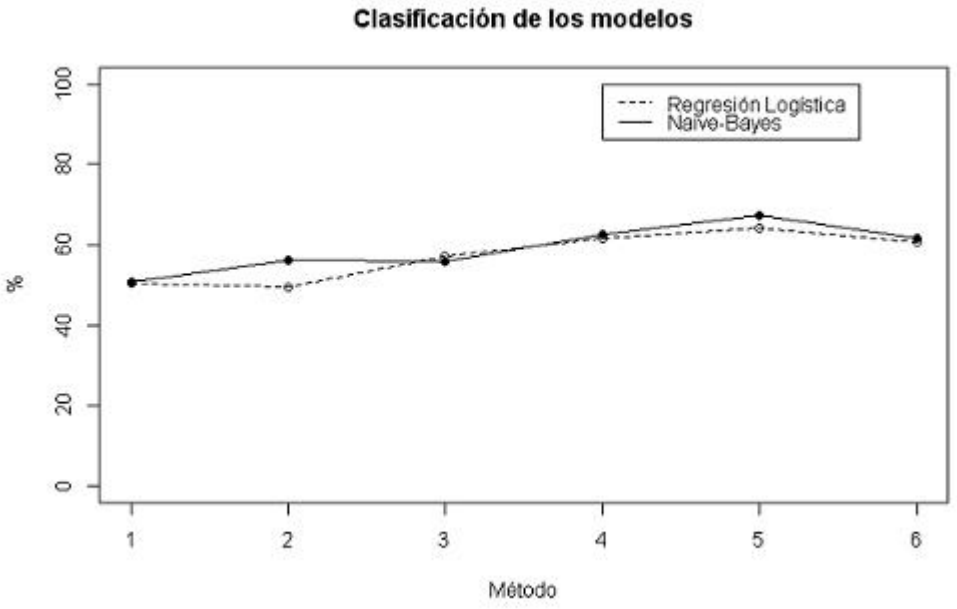

Figura 5: Precisión de los modelos, porcentaje de acierto en la clasificación. (discontinua RL; continua NB)

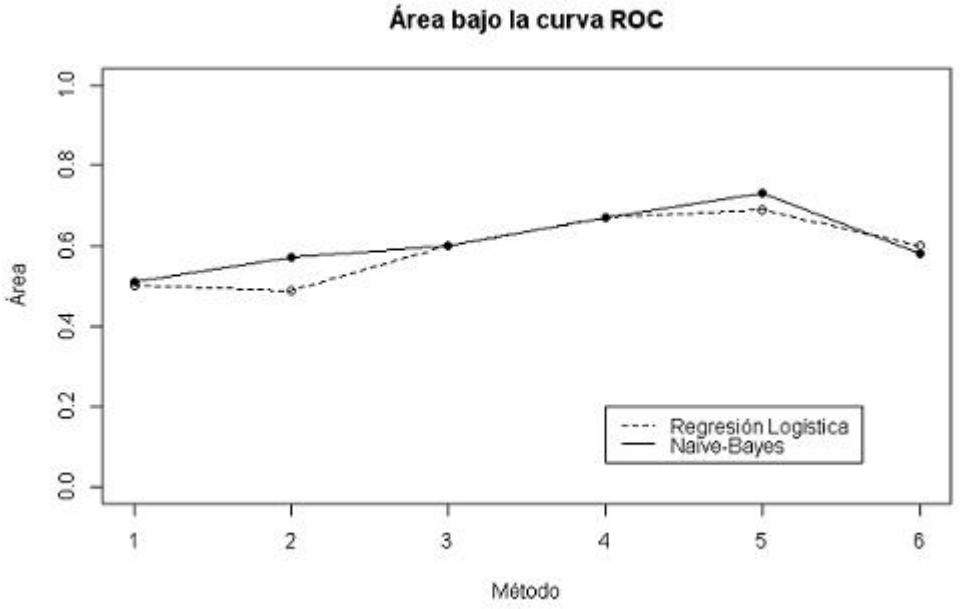

Figura 6: Área bajo la curva ROC. (discontinua RL; continua NB)

5. DISCUSIÓN

Hoy en día el tabaquismo está definido por la OMS desde 1984 como una forma de drogodependencia y la Sociedad Americana de Psiquiatría en 1987 clasificó a la nicotina como una sustancia psicoactiva, que produce dependencia sin abuso. Se estima que la prevalencia de la adicción al tabaco en mayores de 15 años en todo el mundo es del 22 %. Actualmente, se estima que el 41.1 % de los hombres y el 8.9 % de las mujeres en todo el mundo son fumadores [1]. Anualmente es el responsable directo de la muerte de 5 millones de personas [1] y si la prevalencia se mantiene como hasta ahora, será la causa de la muerte de 500 millones de personas durante los próximos 50 años [2].

En España, el porcentaje de fumadores mayores de 16 años es el 26.2 %, 31.2 % hombres y 21.3 % mujeres [3]. Es responsable de multitud de enfermedades y responsable de alrededor de 50000 muertes al año. Concretamente en Euskadi, el porcentaje de población mayor de 16 años que fuma habitualmente es del 26 % y entre los hombres asciende a un 31.2 % nuevamente superior al 21.3 % entre mujeres.

Debido a la magnitud del problema, existen diferentes tratamientos para la deshabituación tabáquica. En este trabajo se ha querido evaluar la deshabituación

a 12 meses en una muestra de pacientes que reciben un determinado tratamiento (vareniclina). En el estudio se pretendía encontrar un perfil genético que estuviese asociado al éxito del tratamiento, por esta razón la mayor parte de las variables son genéticas (SNPs). Estas variables están acompañadas de varias escalas clínicas asociadas con rasgos relacionados con la adicción, por lo que también formaban parte del conjunto de variables del estudio.

Nuestro objetivo es construir un modelo capaz de predecir en función de las variables, tanto genéticas como clínicas, si un individuo va a ser capaz de estar sin fumar 12 meses después de iniciar el tratamiento. La medida a los 12 meses se cree que es la más adecuada ya que desaparece cualquier factor subjetivo y por lo tanto es la medida que mejor recoge la capacidad del fármaco.

El procesado de los datos comemzó con un filtrado de los mismos. En esta fase se imputaron los valores perdidos mediante la moda en las variables discretas y mediante la media en caso de que fueran continuas. Una vez filtrados, se llevó a cabo un estudio de la estructura factorial de las escalas clínicas para encontrar subescalas unidimensionales de las mismas y poder utilizar sus puntuaciones totales correctamente.

Nºde sujetos	Variables clínicas	Variables genéticas	Variable interés
472	13	325	1 (abst12)

Cuadro 14: Composición de la base de datos.

El siguiente paso fue la discretización de las variables continuas para mejorar así las prestaciones de los clasificadores. El método empleado fue la discretización en intervalos de igual frecuencia [51] y el número de intervalos creados, cuatro.

Una vez discrtetizadas las variables continuas en cuatro intervalos, se pasó a la selección de variables [62]. Se utilizaron diferentes métodos: Relief, PFS con greedy stepwise, FCBF con best first, CFS con best first y wrapper con best first.

Puesto que el tamaño de la muestra lo permite, se emplearon técnicas de minería de datos. Los clasificadores que se han empleado han sido:

i) Clasificador basado en regresión logística.

ii) Clasificador Naive-Bayes.

La utilización de diferentes métodos de selección de variables propició la obtención de diferentes subconjuntos de variables recogidos en el cuadro 5. De entre todos los métodos, el que proporciona el subconjunto con menor porcentaje de reducción fue Relief. Se puede deber a que este método de selección es subóptimo en el sentido de que no proporciona siempre el menor subconjunto de variables

[54] y [55]. El CFS, que utiliza la medida $Merit_S$ además de SU para seleccionar las variables junto con el heurístico BestFirst, selecciona un subconjunto de 59 variables, lo que supone una reducción del 82.6 %. Los subconjuntos con menor número de variables fueron los obtenidos por: PFS, FCBF y wrapper(NB) y wrapper(RL) con 8, 8, 6 y 7 variables respectivamente, lo que representa una reducción del 97.64 %, 97.64 %, 98.23 % y 97.24 %.

La gran diferencia entre PFS, FCBF, wrapper(NB), wrapper(RL) y Relief puede radicar en que al basarse en el criterio de inconsistencia, SU, y entropía sean mucho más estrictos y además los heurísticos que utilizan tienen unos criterios de parada que dejan de seleccionar variables muy pronto, sin olvidar que Relief no tiene porque seleccionar el mejor subconjunto de variables [54] y [55]. Por otro lado, se puede ver en los experimentos publicados, que al utilizar estas técnicas el número de variables seleccionadas es bastante reducido en comparación con el número inicial de las mismas [57] y [61].

Las variables seleccionadas por cada método se pueden observar en el anexo 3.

Los dos clasificadores se aplicaron tanto con los subconjuntos seleccionados como con la base de datos completa mediante 10-fold cross validation. La calidad de cada uno de ellos aplicado a cada conjunto de variables se midió tanto con el porcentaje de bien clasificados, como con el área bajo la curva ROC. Para calcularlos se utilizó el módulo Experimenter de WEKA [65], los datos recogidos son los valores medios y las desviaciones estándar tras la validación cruzada.

Los cuadros 12 y 13 recogen la información sobre la calidad de los análisis. Éstos fueron realizados siguiendo el siguiente proceso: imputación de los valores perdidos, discretización de las variables continuas, selección de variables mediante diferentes métodos y por último la aplicación de clasificadores supervisados. En ellas se puede ver como en el caso del Naive-Bayes (cuadro 11) la reducción de variables tiene el resultado esperado. La calidad del modelo tanto medido con el % de bien clasificados como mediante el área bajo la curva ROC es mejor cuando utilizamos los subconjuntos proporcionados por los algoritmos de reducción. Se ve cómo con la totalidad de variables se alcanza el 50.80 % de bien clasificados y un área de 0.51, bastante inferiores a los valores obtenidos con los subconjuntos de variables. En el caso del modelo basado en la regresión logística (cuadro 12), salvo con el subconjunto seleccionado por Relief, los resultados son los mismos. Hay mejoría al utilizar los subconjuntos de variables. Esto apoya lo recogido en la literatura acerca de la utilización de algoritmos de selección, que mejoran la calidad de los clasificadores [62] y [67].

El % de bien clasificados obtenido por el modelo que utiliza la regresión logística (cuadro 12) alcanza el máximo con el algoritmo de selección CFS, 64.25 %. El peor porcentaje, 49.40 % se consigue utilizando Relief y si aplicamos el clasificador con la totalidad de las variables el porcentaje que se tiene es del 50.38 %

lo que no aporta ninugún tipo de información. Con el resto de subconjuntos, los porcentajes alcanzados son muy similares y están en torno al 60 %, un poco por debajo que el porcentaje alcanzado con CFS.

En cuanto al porcentaje de bien clasificados utilizando el Naive-Bayes (cuadro 13), las conclusiones son similares. Nuevamente el mayor porcentaje se alcanza con el CFS y roza el 70 % (67.29 %), lo que supone la mejor clasificación de todos los modelos evaluados. El peor porcentaje de acierto se tiene al utilizar el PFS y no llega al 50 %. La clasificación utilizando la totalidad de variables, al igual que en el caso del otro clasificador, tampoco aporta información ya que no pasa del 50 %. Finalmente, con el resto de algoritmos de selección, el porcentaje de bien clasificados se encuentra en torno al 60 %.

El modelo basado en la regresión logística (tabla 12) en cuanto al área bajo la curva ROC se refiere, se comporta de igual manera que con el % de bien clasificados. Las mayores áreas se obtienen con los algoritmos que utilizan la SU, 0.69 y 0.67 con CFS y FCBF respectivamente indicando que la capacidad de predicción es regular llegando casi a buena. Las áreas más bajas, como era de esperar, se tienen con Relief y con la totalidad de las variables, 0.49 y 0.50 respectivamente indicando que la capacidad de predicción con estas variables es mala. Con el resto de algoritmos las áreas se encuentran alrededor de 0.60, indicando una capacidad de predicción regular.

Lo mismo ocurre con el clasificador Naive-Bayes (cuadro 13). La mayor área, 0.73, se alcanza con el subconjunto seleccionado por CFS y al igual que con el % de bien clasificados, es el mejor valor que se obtiene entre todos los modelos ajustados. Los peores valores se obtienen con Relief y la base de datos completa 0.57 y 0.51. Con wrapper(NB) y PFS tenemos áreas en torno al 0.60, lo que supone una capacidad regular de predicción.

Comparando ambos clasificadores se ve como la mejor calidad se alcanza con los algoritmos de selección que utilizan la SU en los que la cantidad de variables seleccionadas varía mucho (59 con CFS y 8 con FCBF), lo que demuestra que son más adecuados que el resto y por supuesto que la utilización de todas las variables. Entre el modelo Naive-Bayes y el que utiliza regresión logística, el modelo con mayor capacidad de predicción es el primero, tanto si atendemos al % como si lo hacemos al área bajo la curva. Todo esto demuestra que el mejor modelo es el Naive-Bayes con CFS.

Al comparar nuestros resultados con los presentes en la literatura, cabe destacar que en [5] se presentan diferentes SNPs asociados con abstinencia. Los SNPs utilizados en este trabajo no coinciden con los meniconados en la publicación, la razón principal de que no se encuentren coincidencias es que en [5] el estudio se hace a corto plazo (9-12 semanas) mientras que en este trabajo asciende a 52 semanas.

Con respecto a la capacidad predictiva de nuestros mejores modelos, Naive-

Bayes con CFS (67.29 % de bien clasificados y área de 0.73 bajo la curva ROC) y FCBF (62.55 % y área bajo la curva ROC de 0.67), no se han encontrado en la literatura estudios similares en los que se construyan modelos predictivos para evaluar el éxito de la abstinencia a los 12 meses en pacientes tratados con vareniclina. Esto hace imposible una comparación directa de la capacidad predictiva de los modelos aquí construidos.

La utilización de modelos predictivos en otras áreas de la medicina, concretamente los modelos APACHE, que utiliza regresión logística, es bastante común. Por lo tanto, con carácter meramente informativo, podemos comparar nuestro modelo frente a ellos para hacernos una idea de su capacidad de predicción. Los campos en los que se utilizan para predecir mortalidad son: pacientes con cirrosis con un área bajo la curva de 0.759 y superior [78], coma no traumático donde el % de acierto es de 79.9 % y el área es 0.86 [79], infarto agudo de miocardio con área de 0.94 [80] y cirugía cardiaca con área de 0.86 y superior. Hay que decir que en la mayoría de esos estudios, el tamaño de la muestra es muy superior al nuestro.

Las limitaciones del estudio son dos. La primera de ellas es el tamaño muestral, ya que con la cantidad de variables medidas inicialmente, la n de que disponemos es un poco reducida, ya que el criterio que se utiliza es el de 10 observaciones por variable. Por lo tanto sería positivo poder llevar a cabo un estudio con una mayor cantidad de muestra para ver si los resultados siguen la tendencia de los aquí mostrados. La segunda limitación es que sólo se ha utilizado un tratamiento, vareniclina, lo que hace que la posible generalización de estos resultados no pueda extenderse a aquellos estudios que utilicen otro tratamiento.

Para concluir, como se decía en el planteamiento de las hipótesis del trabajo, la minería de datos es una técnica válida en este campo al permitir la construcción de modelos capaces de predecir la abstinencia después de 12 meses en pacientes tratados con vareniclina. De los dos clasificadores empleados, el que mejores resultados ha dado ha sido el Naive-Bayes, sobre todo con los métodos de selección de variables basados en medidas de correlación, lo cual indica qué métodos de selección de variables hay que utilizar en un futuro, además de quedar demostrado que la regresión logística no tiene porque ser el mejor modo de analizar los datos, en la que se basa los modelos predictivos APACHE.

La realización de futuros estudios de asociación para detectar variables genéticas relacionadas con la abstinencia sería muy positiva. Esto permitiría llevar a cabo estudios de este tipo con una mayor cantidad de información. Esto junto con una mayor muestra (como los que se emplean en la mayoría de los estudios que utilizan APACHE) y un mayor número de fármacos, permitiría la construcción de modelos predictivos con una mayor cantidad de datos que proporcionarían un mejor aprendizaje. Este mejor aprendizaje quedaría reflejado en un aumento de la capacidad de predicción de nuevos casos, por lo que sería posible a partir de la información genética de un individuo establecer un tratamiento a medida selec-

cionando el que tuviera la mayor probabilidad de eficacia.

6. BIBLIOGRAFÍA

Referencias

[1] WHO Report on the Global Tobacco Epidemic. Geneva, Switzerland: WHO, 2009.

[2] The Tobacco Atlas. 1st Edition. Geneva, Switzerland: WHO, 2002.

[3] WHO Report on the Global Tobacco Epidemic, 2011.

[4] KORTMANN, G. L, DOBLER, C. J., BIZARRO, L. & BAU, C. H. D. (2009) Pharmacogenetics of Smoking Cessation Therapy, *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 153B (1), pp. 17-28.

[5] KING, D. P., PACIGA, S., PICKERING, E., BENOWITZ, N. L., BEIRUT, L. J., CONTI, D. V., KAPRIO, J., LERMAN, C. & PARK, P. W. (2012) Smoking Cessation Pharmacogenetics: Analysis of Varenicline and Bupropion in Placebo-Controlled Clinical Trials, *Neuropsychopharmacology*, 37, pp. 641.650

[6] SNPlex^{TM}Applied Biosystems, Inc.

[7] http://es.wikipedia.org/wiki/Polimorfismo_de_nucleótido_simple.

[8] http://www.medmol.es/glosario/66/.

[9] HEATHERTON, T. F.,KOZLOWSKI, L. T., FRECKER, R. C. & FAGERSTRÖM, K.-O. (1991) The Fagerström Test for Nicotine Dependence: a revision of the Fagerström Tolerance, *British Journal of Addiction*, 86, pp. 1119-1127.

[10] MENESES-GAYA, C., ZUARDI, A. W., MAZZONCINI DE AZEVEDO MARQUES, J., SOUZA, R. M., LOUREIRO, S. R. & CRIPPA, JA. S. (2009) Psychometric qualities of the Brazilian versions of the Fagerström for Nicotine Dependencie and the Heaviness of Smoking Index, *Nicotine & Tobacco Research*, 11 (10), pp. 1160-1165.

[11] RICHARDSON, C. G. & RATNER, P. A. (2005) A confirmatory fator analysis of the Fagerström Test for Nicotine Dependence, *Addictive Behaviors*, 30, pp. 697-709

[12] JHANJE, S. & SETHI H. (2010) The Fagerström Test For Nicotine Dependence in An Indian Sample of Daily smokers With Poly Drug Use, *Nicotine & Tobacco Research*, 12 (11), pp. 1162-1166.

[13] MENESES-GAYA, I. C., ZUARDI, A. W., LOUREIRO S. R. & CRIPPA, JA. S. (2009) Psychometric properties of the Fagerström Test for Nicotine Dependence, *The Jornal Brasileiro de Pneumologia*, 35 (1), pp. 73-82.

[14] RADZIUS, A., GALLO, J. J., EPSTEIN, D. H., GORELICK D. A., CADET, J. L., UHL G. E. & MOOLCHAN, E. T. (2003) A factor analysis of the Fagerström Test for Nicotine Dependence (FTND), *Nicotine & Tobacco Research*, 5, pp. 255-260.

[15] UYSAL, M. A., KADAKAL, F., KARSIDAG, C., BAYRAM, N. G., UYSAL, O. & YILMAZ, V. (2004) Fagerstrom test for nicotine dependence: Reliability in a Turkish sample and factor analysis, *Tüberküloz ve Toraks Dergisi*, 52 (2), pp. 115-121.

[16] ETTER J.-F., VU DUC, T. & PERENEGER T. V. (1999) Validity of the Fagerström test for nicotine dependence and of the Heaviness of Smoking Index among relatively light smokers, *Addiction*, 94 (2), pp. 269-281.

[17] COHEN, S., KAMARCK, T. & MERMELSTEIN, R. (1983) A Global Measure of Percieved Stress, *Journal of Health and Social Behavior*, 24 (4), pp. 385-396.

[18] REMOR, E. (2006) Psychometric Properties of a European Spanish Version of the Percieved Stress Scale (PSS), *The Spanish Journal of Psychology*, 9 (1), pp. 86-93.

[19] ANDREOU, E., ALEXOPOULUS, E. C., LIONIS, C., VARVOGLI, L., GNARDELLIS, C., CHOROUSOS, G. P. & DARVIRI, C. (2011), *International Journal of Environmental Research and Public Health*, 8, pp. 3287-3298.

[20] CHAAYA, M., OSMAN, H., NAASSAN, G. & MAHFOUD, Z. (2010) Validation of the Arabic version of the Cohen percieved stress scale (PSS-10) among pregnant and postpartum women, *BMC Psychiatry*, 10:111.

[21] LEUNG, D. YP., LAM, T. & CHAN, S. SC. (2010) Three versions of Percieved Stress Scale: validation in a sample of Chinese cardiac patients who smoke, *BMC Public Health*, 10:513.

[22] WANG, Z., CHEN, J., BOYD, J. E., ZHANG, H., JIA, X., QIU, J. & XIAO, Z. (2011) Psychometric Properties of the Chinese Version of the Percieved Stress Scale in Policewomen, *PLos ONE*, 6 (12).

[23] WONGPAKARAN, N. & WONGPAKARAN, T. (2010) The Thai version of the PSS-10: An Investigation of its psychometric properties, *BioPsychoSocial Medicine*, 4:6.

[24] COHEN, S. & WILLIAMSON, G. (1988) Percieved stress in a probability sample of the United States, *The Social Psychology of Health: Claremont Symposium on Applied Social Psuchology*, Edited by: SPACAPAN, S. & OSKAMP, S. Newubury Park CA: Sage, pp. 31-67

[25] CLEMENTE JIMÉNEZ, M. L., PÉREZ TRULLÉN, A., RUBIO ARANDA, E., MARRÓN TUNDIDOR, R., RODRÍGUEZ IBÁÑEZ, M. L. & HERRERO LABARGA, I. (2003) A version of DSM-IV criteria adapted for adolescents and applied to young smokers, *Archivos de Bronconeumología*, 39 (7), pp. 303-309.

[26] ROSE, J. S. & DIEKER, L. C. (2010) DSM-IV nicotine dependence symptom characteristics for recent-onset smokers, *Nicotine & Tobacco Research*, 12 (3), pp. 278-286.

[27] STRONG, D. R., KAHLER, C. W., RAMSEY, S. E. & BROWN, R. A. (2003) Finding order in the DSM-IV nicotine dependence syndrome: a Rasch analysis, *Drug and Alcohol Dependence*, 72, pp. 151-162.

[28] McBRIDE, O., STRONG, D. R. & KAHLER, C. W. (2010) Exploring the role of a nicotine quantity-frecuency use criterion in the classification of nicotine dependence and the stability of a nicotine dependence continuum over time, *Nicotine & Tobacco Research*, 12 (3), pp. 207-216.

[29] SHMULEWITZ, D., KEYES, K. M., WALL, M. M., AHARONOVICH, E., AIVADYAN, C., GREENSTEIN, E., SPIVAK, B., WEIZMAN, A., FRISCH, A., CRANT, B. F., & HASIN, D. (2011) Nicotine dependence, abuse and craving: dimensionality in an Israeli sample, *Addiction*, 106, pp. 1675-1686.

[30] GOLDBERG, D., BRIDGES, K., DUNCAN-JONES, P. & GRAYSON, D. (1988) Detecting anxiety and depression in general medical settings, *British Medical Journal*, 297, pp. 897-899.

[31] KOLOSKI, N. A., SMITH, N. & PACHANA, N. A. (2008) Performance of the Goldberg Anxiety and Depression Scale in older women.

[32] MONTÓN, C., PÉREZ ECHEVARRÍA, MJ., CAMPOS, R., GARCÍA CAMPAYO, J. & LOBO, A. (1993) Escalas de ansiedad y depresión de Goldberg: una guía de entrevista eficaz para la detección del malestar psíquico, *Atención primaria*, 12, pp. 345-349.

[33] MACKINNON, A., CHRISTENSEN, H., JORM, A. F., HENDERSON, A. S., SCOTT, R. & KORTEN, A. E. (1994) A latent trait analysis of an inventory designed to detect symptoms of anxiety and depression using an elderly community sample, *Psychological Medicine*, 24, pp. 977-986.

[34] CHRISTENSEN, H., JORM, A. F., MACKINNON, A. J., KORTEN, A. E., JACOMB, P. A., HENDERSON, A. S. & RODGERS, B. (1999) Age differences in depression and anxiety symptoms: a structural equation modelling analysis of data from a general population sample, *Psychological Medicine*, 29, pp. 325-339.

[35] PÉREZ-TRULLÉN, A., BARTOLOMÉ, C. B. & BANEGAS, J. R. (2006), Nuevas perspectivas en el diagnóstico y la evolución del consumo de tabaco: marcadores de susceptibilidad y lesión, *Medicina Clínica*, 126 (16), pp. 628-631.

[36] http://www.navarra.es/NR/rdonlyres/A7EBFCE1-89F0-4406-81A7-7786919A25F7/193854/EducarenSaludI.pdf

[37] CONDIOTTE, M. M. & LICHTENSTEIN, E. (1981) Self-efficacy and relapse in smoking cessation programs, *Journal of Consulting and Clinical Pschology*, 49, pp. 648-658.

[38] BECOÑA, E. & LORENZO, M^{a}. C. (2004) Evaluación de la conducta de fumar, *Adicciones*, 16, supl. 2.

[39] SECADES, R. (1997) Evaluación conductual en prevención de recaídas en la adicción a las drogas: estado actual y aplicaciones clínicas, *Psicothema*, 9 (2), pp. 259-270.

[40] AYESTA, F. J. & OTERO, M. (2004) El tabaquismo como una enfermedad adictiva crónica. En: Jiménez-Ruiz CA, Fagerström KO, editores. Manual de Tabaquismo

[41] BAER. J. S., HOLT, C. S. & LICHTENSTEIN, E. (1986) Self-Efficacy and Smoking Reexaminated: Construct Validity and Clinical Utility, *Journal of Consulting and Clinical Psychology*, 54 (6), pp. 846-852. (en prensa). Madrid: Aula Médica Editorial.

[42] WALLSTON, B. S., WALLSTON, K. A., KAPLAN, G. D. & MAIDES, S. A. (1976) Development and Validation of the Health Locus of Control (HLC) Scale, *Journal of Consulting and Clinical Psychology*, 44 (4), pp. 580-585.

[43] MEYERS, R., DONHAM, G. W. & LUDENIA, K. (1982) The Psychometric Properties of the Health Locus of Control Scale with Medical and Surgical Patients, *Journal of Consulting and Clinical Psychology*, 38 (4), pp. 783-787.

[44] BOYLE, E. S. & HARRISON, B. E. (1981) Factor Structure of the Health Locus of Control Scale, *Journal of Consulting and Clinical Psychology*, 37 (4), pp. 819-824.

[45] WITTEN, I. H. & FRANK, E. (2000) Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, 2000.

[46] MITCHELL, T. M. (1997) Machine Learning. Series in Computer Science. McGraw-Hill, 1997.

[47] SIJTSMA, K.& VERWEIJ, A. C. (1992) Mokken Scale Analysis: Theoretical Considerations and an Aplication to Transitivity Tasks, *Applied Measurement in Education*, 5 (4), pp. 355-373.

[48] SIJTSMA, K., DEBETS, P. & MOLENAAR, I. W. (1990) Mokken scale analysis for polychotomous items: theory, a computer program and an empirical application, *Quality & Quantity*, 24, pp. 173-188.

[49] GILLESPIE, M., TENVERGERT, E. M. & KINGMA, J. (1987) Using Mokken scale analysis to develop unidimensional scales: Do the six abortion items in the NORC GSS form one or two scales?, *Quality & Quantity*, 21, pp. 393-408.

[50] ANDRIES van der ARK, L. (2007) Mokken Scale Analysis in **R**, *Journal of Statistical Software*, 20 (11), 1-19. URL http://www.jstatsoft.org/v20/i11/.

[51] YANG, Y. & WEBB, G. I. (2002) A Comparative Study of Discretization Methods for Naive-Bayes Classifiers, *Proceedings of PKAW 2002: The 2002 Pacific Rim Knowledge Acquisition Workshop*, pp. 159-173.

[52] R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[53] SAEYS, Y., INZA, I. & LARAÑAGA, P. (2007) A review of feature selection techniques in bioinformatics, *Bioinformatics*, 23 (19), pp. 2507-2517.

[54] KIRA, K. & RENDELL, L. A. (1992) A Practical Approach to Feature Selection, *Proceedings of the ninth international workshop on Machine learning*, pp. 249-256.

[55] KIRA, K. & RENDELL, L. A. (1992) The Feature Selection Problem: Traditional Methods and a New Algorithm, *AAAI-92 Proceedings*, pp. 129-134.

[56] KONONENKO, I. (1994) Estimating Attributes: Analysis and Extensions of RELIEF, *European Conference on Machine Learning*, pp. 171-182.

[57] LIU, H. & SEITONO, R (1996) A Probabilistic Approach to Feature Selection - A Filter Solution, *13th International Conference on Machine Learning*, pp. 319-327.

[58] BRASSARD, G. & BRATLEY, P. (1996) Fundamentals of Algorithms, *Prentice Hall*, New Jersey.

[59] CORMEN, T. H., LEISERSON, C. E. & RIVEST, R. L. (1990) Introduction to Algorithms, *MIT Press and McGraw-Hill*, ISBN 0-262-03141-8.

[60] HALL, M.A. (1998) Correlation-based feature selection machine learning, Ph.D. Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

[61] LEI, Y. & HUAN, L. (2003) Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, *Proceedings of the 20th International Conference on Machine Learning (ICML-2003)*, pp. 856-863.

[62] GUYON, I., & ELISSEEFF, A. (2003) An Introduction to Variable and Feature Selection, *Journal of Machine Laerning Research*, 3, pp. 1157-1182.

[63] HUAN, L. & LEI, Y. (2005) Toward Integrating Feature Selection Algorithms for Classification and Clustering, *IEEE Transactions on Knowledge and Data Engineering*, 17 (4), pp. 491-502.

[64] QUINLAN, J. R. (1993) C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

[65] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER B., REUTERMANN,P. & WITTEN, I. H. (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[66] RUSSELL, S. J. & NORVIG, P. (2009) Aritificial Intelligent: A Modern Approach (3 ed.), *Prentice Hall*, ISBN: 0136042597.

[67] HALL, M. A. & SMITH, L. A. (1999) Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper, *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, pp. 235-239.

[68] HALL, M. A. & SMITH, L. A. (1997) Feature Suset Selection: A Correlation Based Filter Approach, *In 1997 International Conference on Neural Information Processing and Intelligent Information Systems*, pp. 855-858.

[69] DOMINGOS, P. & PAZZANI, M (1997) On the optimality of the simple Bayesian classifier under zero-one loss, *Machine Learning*, 29, pp. 103-130.

[70] ZANG, S., TJORTIS, C., ZENG, X., QIAO, H., BUCHAN, I. & KEANE, J. (2009) Comparing Data Mining MOdels with Logistic Regression in Childhood Obesity Prediction, *Information Systems Frontiers*, 11 (4), pp. 449-460.

[71] KOKONENKO, I. (1993) Inductive and Bayesian Learning in Medical Diagnosis, *Applied Artificial Intelligence*, 7, pp. 317-337.

[72] DOUGHERTY, J., KOHAVI, R. & SAHAMI, M. (1995) Supervised and Unsupervised Discretization of Continuous Features, *Machine Learning: Proceedings of the Twelfth International Conference*, pp. 194-202.

[73] HOSMER, D. W. & LEMESHOW, S. (2000) Applied Logistic Regression, *Wiley series in probability and statistics*, John Wiley & Sons, 2000. Second Edition.

[74] LE CESSIE, S. & VAN HOUWELINGEN, J. C. (1992) Ridge Estimators in Logistic Regression, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 41, pp. 191-201.

[75] RODRÍGUEZ, J. D., PÉREZ, A. & LOZANO, J. A. (2010) Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32 (3), pp. 569-575.

[76] KOHAVI, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI, 1995)*, Vol 2, pp. 1137-1143.

[77] STONE, M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal Statistical Society. Series B (Methodological)*, 36 (2), pp. 111-147.

[78] CHATZICOSTAS, C., ROUSSOMOUSTAKAKI, M, NOTAS, G., VLACHONIKOLIS, I. G., SAMONAKIS, D., ROMANOS J., VARDAS E. & KOUROUMALIS E. A. (2003) A comparison oh Child-Plugh, APACHE II and Apache III scoring systems in predicting hospital mortality of patients with liver cirrhosis, *BMC Gastroenterology*, 3:7.

[79] GRMEC, S., & GASPAROVIC, V. (2000) Comparison of APACHEII, MEES and Glasgow Coma Scale in patients with nontraumatic coma for prediction of mortality, *Critical Care*, 5, pp. 19-23.

[80] MERCADO-MARTÍNEZ, J., RIVERA-FERNÁNDEZ, R., AGUILAR-ALONSO, E., GARCÍA-ALCÁNTARA, A., ESTIVILL-TORRULL, A., ARANDA-LEÓN A., GUIA-RAMBLA, M.C. & FUSET-CABANES, M. P. (2010) APACHE-II score and Killip class for patients with acute myocardial infarction, *Intensive Care Medicine*, 36, pp. 1579-1586.

[81] HEKMAT, K., DOERR, F., KROENER A., HELDWEIN, M., BOSSERT, T., BADRELDIN, A. M. A. & LICHTENBERG, A. (2010) Prediction of mortality in intensive care unit cardiac surgical patients, *European Journal of Cardio-thoracic Surgery*, 38, pp. 104-109.

7. ANEXOS

7.1. Anexo 1. Variables genéticas (SNPs genotipados.)

rs6660775	rs2072660	rs9427092	rs1127326
rs9616	rs1127311	rs3935384	rs723105
rs12416004	rs1125524	rs1644391	rs1757070
rs1660622	rs629039	rs678188	rs1778871
rs10825324	rs4919652	rs10786684	rs6585674
rs7920207	rs2292692	rs3842748	rs2070762
rs6356	rs4929966	rs11564710	rs11564710
rs11564709	rs4758287	rs11041740	rs1055233
rs4343012	rs7116230	rs11023059	rs10766758
rs10734298	rs6589354	rs2011505	rs10891510
rs4938012	rs2734849	rs1800497	rs12422191
rs2234689	rs2242592	rs6277	rs1076560
rs2283265	rs2075654	rs1079727	rs2734833
rs1076562	rs1079597	rs17529477	rs4245147
rs4936270	rs7131056	rs4648317	rs18091556
rs6589377	rs7930792	rs4935872	rs11604309
rs12423809	rs4760813	rs17110459	rs4448731
rs4570625	rs11178997	rs11178998	rs4565946
rs11179000	rs1843809	rs1386494	rs1386493
rs7978482	rs7305115	rs1386497	rs1352251
rs1487278	rs1473437	rs1487275	rs4290270
rs17110747	rs1872824	rs4766674	rs9658498
rs733334	rs4262815	rs7995907	rs11069434
rs4262815	rs7995907	rs11069434	rs7139689
rs9550153	rs883473	rs8024987	rs1913456
rs2175886	rs982574	rs6494211	rs6494212
rs7175581	rs12438848	rs6494223	rs9788679
rs1500948	rs1039394	rs12915265	rs1909884
rs2651417	rs2611604	rs2611605	rs7178176
rs2337980	rs12906868	rs8034191	rs2036527
rs16969968	rs10517130	rs3743078	rs1317286
rs2869546	rs6495308	rs1948	rs7178270
rs37824	rs9923657	rs4782031	rs7223372
rs3794808	rs9923657	rs4782031	rs7223372
rs3794808	rs140701	rs4583306	rs140700
rs2020942	rs12150214	rs4251417	rs16965628
rs1050565	rs981577	rs7215201	rs2240152
rs2240154	rs2285907	rs7146	rs739884

rs17761012	rs4803380	rs8192726	rs28399435
rs1496402	rs3844443	rs7251950	rs110835985
rs2099361	rs100458	rs16974799	rs3745274
rs8192719	rs1042389	rs11666982	rs7255146
rs17726861	rs2160695	rs12468478	rs11883614
rs9677968	rs921573	rs11675607	rs1712905
rs10176971	rs1426707	rs1434064	rs4849074
rs11689769	rs1607373	rs10177758	rs6431541
rs4522666	rs3787138	rs1044397	rs1044393
rs2273502	rs2273504	rs2273505	rs755203
rs3746372	rs9605030	rs6518591	rs2020917
rs737865	rs933271	rs1544325	rs4633
rs4818	rs4680	rs4646316	rs165774
rs174696	rs9332377	rs165599	rs887199
rs4821566	rs5751222	rs5758589	rs764481
rs11129660	rs1965458	rs1237403	rs17397636
rs17399554	rs4683831	rs6439919	rs13151552
rs7671397	rs1435480	rs11932367	rs4696168
rs462761	rs12516758	rs27072	rs1042098
rs40184	rs3776511	rs6347	rs37022
rs2042449	rs2975292	rs464049	rs2975292
rs464049	rs460700	rs403636	rs6350
rs2937639	rs2652511	rs6413429	rs3756450
rs2078247	rs2915438	rs7731574	rs13160445
rs2032863	rs10485171	rs806365	rs7766029
rs806366	rs806368	rs12720071	rs1049353
rs806369	rs806374	rs806375	rs806377
rs2023239	rs6454672	rs6454674	rs10485170
rs9398910	rs763132	rs1799971	rs477292
rs3823010	rs495491	rs1381376	rs3778151
rs9479757	rs2075572	rs562859	rs609148
rs648893	rs132041	rs13194785	rs12200296
rs223659	rs1918760	rs2281617	rs4318891
rs1040822	rs2293537	rs6935927	rs9479797
rs1554817	rs9383697	rs7759388	rs6455600
rs12672248	rs4392794	rs12700327	rs1174692
rs900047	rs7037246	rs177040	rs10820820
rs7030238	rs16920504	rs12353519	rs9299345
rs4743473	rs7858998	rs10119861	rs10121600
rs4743474	rs4742820	rs7856074	rs1323416
rs2485531	rs10989591	rs10989598	rs1337696
rs1337684	rs1415644	rs4474069	rs11146020

rs2301364 rs4074426 rs5905512	rs10870198 rs1799836 rs1183035	rs6293 rs10521432	rs10747050 rs3027450

7.2. Anexo 2. Variables clínicas (escalas clínicas)

7.2.1. Escala de Fagerström para la dependencia de la nicotina (FTND)

Ítems que forman la escala:

1. ¿Cuántos cigarrillos fuma al día? 1- 10 ó menos; 2- 11 a 20; 3- 21 a 30; 4- 31 ó más.

2. ¿Cuánto tiempo transcurre desde que se levanta hasta que fuma el primer cigarro? 1- menos de 5 min ;2- de 6 a 30 min; 3- de 31 a 60 min; 4- más de 60 min.

3. ¿Fuma más en las primeras horas del día? 1- Si; 2- No.

4. ¿Tiene alguna dificultad para estar sin fumar en lugares donde está prohibido? 1- Si; 2- No.

5. ¿Fuma cuando no se encuentra bien o cuando está enfermo en la cama? 1- Si; 2- No.

6. ¿A qué cigarrillo le gustaría más renunciar? 1- Al primero del día; 2- A otros.

7.2.2. Escala de estrés percibido (PSS-10)

Estos son los ítems que forman esta escala: ¿Con que frecuencia en el último mes...

1. ... ha estado afectado por algo que le ha ocurrido inesperadamente? 0; 1; 2; 3; 4.
2. ... se ha sentido incapaz de controlar las cosas importantes en su vida? 0; 1; 2; 3; 4.
3. ... se ha sentido nervioso y estresado? 0; 1; 2; 3; 4.
4. ... ha estado seguro sobre su capacidad para manejar sus problemas personales? 0; 1; 2; 3; 4.
5. ... ha sentido que las cosas le van bien? 0; 1; 2; 3; 4.
6. ... ha sentido que puede afrontar todas las cosas que tenía que hacer? 0; 1; 2; 3; 4.
7. ... ha podido controlar las dificultades de su vida? 0; 1; 2; 3; 4.
8. ... se ha sentido con el control de todo? 0; 1; 2; 3; 4.
9. ... se ha enfadado porque las cosas que le han ocurrido estaban fuera de su control? 0; 1; 2; 3; 4.
10. ... ha sentido que las dificultades se acumulan tanto que no puede superarlas? 0; 1; 2; 3; 4.

0- Nunca; 1- Casi nunca; 2- A veces; 3- A menudo; 4- Muy a menudo.

7.2.3. Criterio DSM-IV para evaluar la dependencia de la nicotina

Los ítems que forman la escala son:

1. ¿Suele sentir nauseas o mareos cuando fuma varios cigarrillos? Si; No.
2. ¿Se ha encontrado físicamente mal cuando ha estado bastantes horas o un par de días sin fumar? Si; No.
3. ¿Fuma más de lo que desaría? Si; No.
4. ¿Ha intentado sin éxito disminuir su consumo de cigarrillos? Si; No.
5. ¿Dedica mucho tiempo del día a fumar? Si; No.
6. ¿Se ha ido antes de alguna reunión o actividad o incluso no ha asistido a ella porque no se podía fumar? Si; No.
7. ¿Cree que el tabaco le está ocasionando algún trastorno en su salud? Si; No.

7.2.4. Escala de ansiedad y depresión de Goldberg (EDAG)

Los ítems de las dos subescalas son: ¿Ha tenido en las 2 últimas semanas alguno de estos síntomas?

Subescala A.

1. ¿Se ha sentido muy excitado, nervioso o en tensión? 1- Si; 2- No.
2. ¿Ha estado muy preocupado por algo? 1- Si; 2- No.
3. ¿Se ha sentido muy irritable? 1- Si; 2- No.
4. ¿Ha tenido dificultad para relajarse? 1- Si; 2- No.
5. ¿Ha dormido mal? 1- Si; 2- No.
6. ¿Ha tenido dolores de cabeza o nuca? 1- Si; 2- No.
7. ¿Ha tenido alguno de los siguientes síntomas: temblores, hormigueos, mareos, sudores, diarrea? 1- Si; 2- No.
8. ¿Ha estado preocupado por su salud? 1- Si; 2- No.
9. ¿Ha tenido alguna dificultad para conciliar el sueño? 1- Si; 2- No.

Subescala B.

1. ¿Se ha sentido con poca energía? 1- Si; 2- No.
2. ¿Ha perdido usted su interés por las cosas? 1- Si; 2- No.
3. ¿Ha perdido la confianza en sí mismo? 1- Si; 2- No.
4. ¿Se ha sentido usted desesperanzado, sin esperanzas? 1- Si; 2- No.
5. ¿Ha tenido dificultades para concentrarse? 1- Si; 2- No.
6. ¿Ha perdido peso? (A causa de falta de apetito) 1- Si; 2- No.
7. ¿Se ha estado despertando demasiado temprano? 1- Si; 2- No.
8. ¿Se ha sentido usted enlentecido? 1- Si; 2- No.
9. ¿Cree que en general se ha encontrado peor por las mañanas? 1- Si; 2- No.

7.2.5. Escala de predicción

Señalar si se considera verdaderas o falsas cada una de las siguientes afirmaciones:

1.1 Vengo a la consulta espontáneamente V; F.

1.2 Vengo a la consulta por indicación médica V; F.

1.3 Vengo a la consulta por presión familiar V; F.

2. Anteriormente ya he dejado de fumar durante más de un mes V; F.

3. Actualmente no tengo demasiados problemas en el trabajo V; F.

4. Actualmente no tengo demasiados problemas en el plano familiar V; F.

5. Hago deporte o tengo intención de hacerlo V; F.

6. Voy a estar en mejor forma física V; F.

7. Voy a cuidar mi aspecto físico V; F.

8. Estoy embarazada (o mi pareja lo está) V; F.

9. Tengo costumbre de lograr lo que emprendo V; F.

10. Soy más bien de temperamento tranquilo V; F.

11. Mi peso es habitualmente estable V; F.

12. Estoy con buena moral habitualmente V; F.

13. Quiero liberarme de esta dependencia V; F.

14. Tengo hijos de corta edad V; F.

15. Voy a acceder a una mayor calidad de vida V; F.

7.2.6. Escala de confianza

Las cuestiones que forman la escala son: En la siguiente escala conteste el grado de resistencia que cree que podría oponer al deseo de fumar si se dieran las siguientes situaciones:

1. Si me sintiera ansioso 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

2. Si quisiera sentarme cómodamente y disfrutar de un cigarrillo 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

3. Si termino una comida o un tentempié 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

4. Si me sintiera nervioso 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

5. Si quisiera sentirme más atractivo 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

6. Si quisiera relajarme 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

7. Si pensara que fumar es parte de mi imagen 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

8. Si me sintiera tenso 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

9. Si estuviera bebiendo una bebida alcohólica 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

10. Si viera a otros fumando 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

11. Si alguien me ofreciera un cigarrillo 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

12. Si quisiera comer algo dulce 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

13. Si quisiera sentirme más maduro y sofisticado 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

14. Si pensara que esto me ayudaría a mantener el peso 0 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %.

Un 100 % significa resistencia absoluta (NO fumaría), en cambio un 0 % significa ninguna resistencia (SI fumaría).

7.2.7. Escala de locus de control (HLC)

Los ítems son: Señale su grado de acuerdo o desacuerdo con las siguientes afirmaciones:

1. Si me cuido puedo prevenir las enfermedades 1; 2; 3; 4; 5; 6.
2. Cuando me pongo enfermo es por algo que he hecho o que he dejado de hacer 1; 2; 3; 4; 5; 6.
3. Tener buena salud es en gran parte cuestión de suerte 1; 2; 3; 4; 5; 6.
4. Independientemente de lo que hagas te vas a poner enfermo 1; 2; 3; 4; 5; 6.
5. La mayor parte de la gente no es consciente de hasta qué grado sus enfermedades dependen de hecho casuales 1; 2; 3; 4; 5; 6.
6. Sólo puedo hacer lo que el médico me dice que haga 1; 2; 3; 4; 5; 6.
7. Hay tantas enfermedades extrañas por ahí que uno nunca sabe cuando puede coger una 1; 2; 3; 4; 5; 6.
8. Cuando me encuentro mal es porque no he hecho el suficiente ejercicio o no he comido bien 1; 2; 3; 4; 5; 6.
9. La gente que nunca enferma es simplemente porque tiene muy buena suerte 1; 2; 3; 4; 5; 6.
10. La mala salud de la gente suele resultar de su falta de cuidado 1; 2; 3; 4; 5; 6.
11. Soy directamente responsable de mi salud 1; 2; 3; 4; 5; 6.

1- Muy en desacuerdo; 6- Muy de acuerdo.

7.2.8. Test de Russell para la evaluación de los motivos para fumar

Los ítems son los siguientes: Indique en qué medida cada afimración se corresponde con lo que le ocurre a usted:

1. Siento un gran deseo de fumar cuando tengo que parar cualquier actividad por un momento 0; 1; 2; 3.
2. Enciendo un cigarrillo sin darme cuenta de que tengo otro encendido en el cenicero 0; 1; 2; 3.
3. Me gusta fumar sobre todo cuando estoy descansando tranquilamente 0; 1; 2; 3.
4. Obtengo un gran placer fumando sea cuando sea 0; 1; 2; 3.
5. Tener un cigarrillo entre los dedos es parte del placer que da fumar 0; 1; 2; 3.
6. Pienso que mejora mi aspecto con un cigarro entre los dedos 0; 1; 2; 3.
7. Fumo más cuando estoy preocupado por algo 0; 1; 2; 3.
8. Me siento más estimulado y alerta cuando fumo 0; 1; 2; 3.
9. Fumo automáticamente a pesar de estar atento 0; 1; 2; 3.
10. Fumo por tener algo que hacer con las manos 0; 1; 2; 3.
11. Cuando me quedo sin cigarrillos me es casi insoportable hasta que puedo volver a tenerlos 0; 1; 2; 3.
12. Cuando me siento infeliz fumo más 0; 1; 2; 3.
13. Fumar me ayuda a aguantar cuando estoy cansado 0; 1; 2; 3.
14. Me resulta difícil estar sin fumar 0; 1; 2; 3.
15. Me encuentro a mi mismo fumando sin recordar haber encendido el cigarrillo 0; 1; 2; 3.
16. Cuando estoy cómodo y relajado es cuando más deseo fumar 0; 1; 2; 3.
17. Fumar me ayuda a pensar y a concentrarme 0; 1; 2; 3.
18. Tengo muchas ganas de fumar cuando no he fumado durante un rato 0; 1; 2; 3.

19. Me siento más maduro y sofisticado cuando fumo 0; 1; 2; 3.
20. Cuando no estoy fumando soy muy consciente de ello 0; 1; 2; 3.
21. Me resultaría muy difícil estar una semana sin fumar 0; 1; 2; 3.
22. Fumo para tener algo que ponerme en la boca 0; 1; 2; 3.
23. Me siento más atractivo frente a personas del sexo contrario cuando fumo 0; 1; 2; 3.
24. Enciendo un cigarrillo cuando estoy enfadado 0; 1; 2; 3.

0- no le sucede en absoluto; 1- le sucede un poco; 2- le sucede bastante; 3- le sucede mucho.

7.3. Anexo 3. Variables extraídas por cada algoritmo de selección de variables

7.3.1. Relief

- Evaluator: weka.attributeSelection.ReliefFAttributeEval -W -M -1 -D 1 -K 472 -A 2

- Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 172

Selected attributes (172): 1 2 4 5 6 7 12 13 14 15 16 17 18 26 29 38 40 42 43 44 49 59 61 62 64 67 68 69 76 77 78 79 81 84 88 89 90 92 96 97 98 99 100 102 107 108 109 110 113 114 118 119 120 121 123 126 127 128 129 131 137 138 139 140 141 142 143 147 152 153 154 155 156 158 159 160 162 163 164 165 166 167 168 169 171 173 174 175 176 179 180 183 185 188 189 193 194 196 198 200 201 203 206 209 216 218 219 220 222 223 225 228 235 238 239 240 244 245 246 247 248 249 250 251 253 255 256 258 261 262 267 272 273 277 279 280 281 282 283 286 288 289 294 297 298 299 301 302 303 304 306 307 308 309 310 311 312 313 314 316 319 321 322 324 325 327 328 329 331 334 336 338.

Relief	Variable	Nombre
0.0594515	98	rs8024987
0.056036	162	rs11666982
0.0496767	168	rs9677968
0.0490376	140	rs1050565
0.045717	225	rs40184
0.0413366	127	rs7178270
0.0356386	244	rs10485171
0.0353639	97	rs883473
0.0352576	14	rs1644391
0.034865	38	rs10766758
0.0342301	245	rs806365
0.0342029	16	rs1660622
0.0338775	99	rs1913456
0.0331508	248	rs806368
0.0322708	286	rs6455600
0.0303764	159	rs3745274
0.0297442	1	edad
0.0289479	250	rs1049353
0.0282844	129	rs9923657
0.0276408	194	rs1544325
0.0272485	301	rs10119861
0.0271843	143	rs2240152
0.0271733	174	rs1434064
0.026968	40	rs6589354
0.0268721	302	rs10121600
0.026416	81	rs1386497
0.025892	251	rs806369
0.0255171	173	rs1426707
0.0254015	228	rs37022
0.0243584	100	rs2175886
0.0243566	142	rs7215201
0.0243199	246	rs7766029
0.02407	334	punt.ram.sub2
0.0239801	68	rs4760813
0.0232342	294	rs10820820
0.022913	158	rs16974799
0.0225063	280	rs2293537
0.0223235	89	rs4766674
0.0220393	255	rs2023239
0.0214484	176	rs11689769
0.021234	119	rs2036527

Relief	Variable	Nombre
0.0209108	309	rs10989598
0.0208873	175	rs4849074
0.0207908	256	rs6454672
0.019773	322	rs10521432
0.019454	249	rs12720071
0.019266	49	rs6277
0.0185916	328	punt.estres.sub1
0.0181757	118	rs8034191
0.0181704	123	rs1317286
0.0180662	78	rs1386493
0.0178453	7	rs1127326
0.0176205	160	rs8192719
0.0173194	188	rs3746372
0.0171279	171	rs1712905
0.0169971	108	rs1500948
0.016847	324	rs5905512
0.0167936	102	rs6494211
0.0164376	26	rs3842748
0.0163983	299	rs4743473
0.0163917	76	rs1843809
0.0163894	155	rs11083595
0.0162302	304	rs4742820
0.0159955	29	rs4929966
0.0159463	165	rs2160695
0.0157971	222	rs12516758
0.0157047	297	rs12353519
0.0152086	282	rs9479797
0.0151692	67	rs12423809
0.0151372	201	rs9332377
0.0150986	126	rs1948
0.0147747	240	rs2915438
0.0144987	107	rs9788679
0.0144983	258	rs10485170
0.0143764	325	rs1183035
0.0141803	220	rs4696168
0.0134702	261	rs1799971
0.0129632	44	rs2734849
0.0128891	209	rs1965458
0.0128478	120	rs16969968
0.0123076	64	rs7930792
0.01179	128	rs37824

Relief	Variable	Nombre
0.0115484	18	rs678188
0.0111416	154	rs7251950
0.0111167	121	rs1051730
0.0110933	77	rs1386494
0.0110532	267	rs9479757
0.0107743	218	rs11932367
0.0107341	5	rs2072661
0.0105003	308	rs10989591
0.0103187	185	rs2273504
0.010094	15	rs1757070
0.009746	69	rs17110459
0.0096671	311	rs1337684
0.0094444	289	rs12700327
0.0092517	109	rs1039394
0.009149	273	rs13194785
0.009095	316	rs2301364
0.0089717	92	rs4262815
0.0088486	338	fag.total
0.0087017	219	rs4696397
0.0085633	169	rs921573
0.0083212	156	rs2099361
0.0080863	313	rs4743485
0.0079911	196	rs4818
0.0079866	279	rs1040822
0.0076991	12	rs12416004
0.0076332	167	rs11883614
0.007619	306	rs1323416
0.0076157	239	rs2078247
0.007599	223	rs27072
0.0074346	17	rs629039
0.0072961	277	rs2281617
0.0072057	138	rs4251417
0.0071872	153	rs3844443
0.0068566	329	punt.estres.sub2
0.0067396	113	rs2611604
0.0067363	203	rs887199
0.0067135	110	rs12915265
0.0063621	281	rs6935927
0.0060635	235	rs2937639
0.0060038	310	rs1337696
0.0058314	139	rs16965628

Relief	Variable	Nombre
0.0058149	206	rs5758589
0.0056586	90	rs9658498
0.0055226	131	rs7223372
0.0050985	59	rs4936270
0.0049826	84	rs1473473
0.0049142	147	rs739884
0.0049046	62	rs10891556
0.0049026	253	rs806375
0.0048996	189	rs9605030
0.0047634	307	rs2485531
0.0046045	164	rs17726861
0.0045954	114	rs2611605
0.0043896	4	rs2072660
0.004373	166	rs12468478
0.0043184	314	rs4474069
0.0042133	6	rs9427092
0.0041874	238	rs3756450
0.0038376	61	rs4648317
0.0038246	288	rs4392794
0.0028867	331	punt.ad.sub1
0.0027192	2	sexo
0.0025902	336	punt.ram.sub4
0.0022038	43	rs4938012
0.0021801	327	punt.conf.sub2
0.002163	321	rs1799836
0.0021275	79	rs7978482
0.0019191	88	rs1872824
0.0018573	163	rs7255146
0.0017107	312	rs1415644
0.0016614	183	rs1044393
0.0016144	272	rs1323041
0.0015491	303	rs4743474
0.0014698	13	rs11255241
0.001443	180	rs4522666
0.0013674	42	rs10891510
0.0013523	262	rs477292
0.0009484	200	rs174696
0.0008532	137	rs12150214
0.0006753	198	rs4646316
0.0005997	283	rs1554817
0.0003998	216	rs7671397

Relief	Variable	Nombre
0.0002753	319	rs10747050
0.0002532	179	rs6431541
0.0001756	141	rs981577
0.0001668	152	rs1496402
0.0001655	298	rs9299345
0.0001041	247	rs806366
0.0000879	96	rs9550153
0.0000112	193	rs933271

7.3.2. Probabilistic feature selection

- Evaluator: weka.attributeSelection.ConsistencySubsetEval

- Search: weka.attributeSelection.GreedyStepwise -T -1.7976931348623157E308 -N -1

Selected attributes (8):9 70 244 247 326 327 329 335.

Variable	Nombre
9	rs1127311
70	rs4448731
244	rs10485171
247	rs806366
326	punt.conf.sub1
327	punt.conf.sub2
329	punt.estres.sub2
335	punt.ram.sub3

7.3.3. Fast Correlation-Based Filter (FCBF)

- Evaluator: weka.attributeSelection.SymmetricalUncertAttributeSetEval
- Search: weka.attributeSelection.FCBFSearch -D false -T -1.7976931348623157E308 -N -1

Selected attributes (8):42 110 137 166 204 218 255 329.

Medida Correlación	Variable	Nombre
0.02246	329	punt.estres.sub2
0.01272	255	rs2023239
0.01122	166	rs12468478
0.01087	42	rs10891510
0.01063	204	rs4821566
0.00958	218	rs11932367
0.0095	137	rs12150214
0.00649	110	rs12915265

7.3.4. Correlated Feature Selection (CFS)

- Evaluator: weka.attributeSelection.CfsSubsetEval
- Search: weka.attributeSelection.BestFirst -D 1 -N 5

Selected attributes (58):6 9 31 40 42 50 59 69 72 89 91 99 100 102 108 110 112 129 134 137 139 162 164 166 167 168 176 198 200 201 204 207 216 218 220 227 229 233 247 249 251 255 257 283 298 313 316 319 320 328 329 330 332 334 335 336 337 338.

Variable	Nombre
6	rs9427092
9	rs1127311
31	rs11564709
40	rs6589354
42	rs10891510
50	rs1076560
59	rs4936270
69	rs17110459
72	rs11178997
89	rs4766674
91	rs733334
99	rs1913456
100	rs2175886
102	rs6494211
108	rs1500948
110	rs12915265
112	rs2651417
129	rs9923657
134	rs4583306
137	rs12150214
139	rs16965628
162	rs11666982
164	rs17726861
166	rs12468478
167	rs11883614
168	rs9677968
176	rs11689769
198	rs4646316
200	rs174696
201	rs9332377
204	rs4821566
207	rs764481
216	rs7671397
218	rs11932367
220	rs4696168
227	rs6347
229	rs2042449
233	rs403636
247	rs806366
249	rs12720071
251	rs806369

Variable	Nombre
255	rs2023239
257	rs6454674
283	rs1554817
298	rs9299345
313	rs4743485
316	rs2301364
319	rs10747050
320	rs4074426
328	punt.estres.sub1
329	punt.estres.sub2
330	punt.pred
332	punt.ad.sub2
334	punt.ram.sub2
335	punt.ram.sub3
336	punt.ram.sub4
337	dsm4.total
338	fag.total

7.3.5. Wrapper (Naive-Bayes)

- Evaluator: weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.bayes.NaiveBayes -F 10 -T 0.01 -R 1

- Search: weka.attributeSelection.BestFirst -D 1 -N 5

Selected attributes (6):21 223 249 289 298 328.

Variable	Nombre
21	rs4919652
223	rs27072
249	rs12720071
289	rs12700327
298	rs9299345
328	punt.estres.sub1

7.3.6. Wrapper (Regresión logística)

- Evaluator: weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.functions.Logistic -F 10 -T 0.01 -R 1 – -R 1.0E-8 -M -1

- Search: weka.attributeSelection.BestFirst -D 1 -N 5

Selected attributes (7):25 51 166 201 223 298 328

Variable	Nombre
25	rs2292692
51	rs2283265
166	rs12468478
201	rs9332377
223	rs27072
298	rs9299345
328	punt.estres.sub1