GRADO EN INGENIERÍA EN SISTEMAS DE TELECOMUNICACIÓN

Trabajo Fin de Grado

APRENDIZAJE AUTOMÁTICO PARA LA CLASIFICACIÓN DE PACIENTES CRÓNICOS CON COMORBILIDADES

Autora : Paula Vecino Rodríguez
Tutora : Inmaculada Mora Jiménez
Co-tutora : Cristina Soguero Ruíz

Curso Académico 2019/2020

Agradecimientos

A mi madre y a mi padre que SIEMPRE han estado dándome ánimos y nunca me han fallado. Sois mis pilares.

A mi hermana y mejor amiga, ¿qué haría yo si no estuvieras en mi vida?.

A mis sobrinas, os amo más que a nadie y lo sabéis.

A mis amigas de y para siempre, lo que une el baloncesto que no lo separe el tiempo.

A todos los profesores que desde el colegio me han apoyado y ayudado a superar etapa tras etapa.

A todos mis amigos de la universidad, muchas gracias por la compañía en esas horas desesperadas en los laboratorios o en la biblioteca, pero sobre todo gracias por las cervezas de después.

A mi gente del Erasmus, que siempre *abbiamo Sant'Erasmo dentro al nostro cuore*. ¡¡*ESN a Milano siamo noi*!!

VECINO RODRIGUEZ PAULA - 70252704D

Firmado digitalmente por VECINO RODRIGUEZ PAULA - 70252704D
Fecha: 2019.11.04 08:30:00 +01'00'

Este trabajo ha sido parcialmente financiado por el Ministerio de Ecnonomía y Competitividad, como parte del proyecto TEC2016-75361-R, y por el Instituto de Salud Carlos III, como parte del proyecto DTS17/00158. Agradecimiento también al personal del Hospital Universitario de Fuenlabrada por el apoyo prestado para que este proyecto saliera adelante.

Resumen

El volumen de datos que se almacena sobre las personas crece exponencialmente en el tiempo y no tiene previsto parar. El hecho de que las bases de datos estén en muchas ocasiones saturadas de datos, trae consigo que aumente la brecha entre el almacenamiento de esos datos y su utilización real. Para extraer información útil de los datos almacenados es posible usar técnicas de *Data Mining*, pudiendo construir modelos a partir de los cuales realizar inferencia sobre nuevos casos. El área de *Data Mining* denominada aprendizaje automático permite abordar, entre otras, tareas de clasificación si se dispone de un conjunto de casos previamente etiquetados y suficiente representativo de la tarea a resolver.

Las enfermedades crónicas suponen una gran demanda de recursos. En España, más del $43\,\%$ de la población adulta es hipertensa, y casi el $20\,\%$ diabética. Las cronicidades nunca aparecen solas, y con los años aparecen comorbilidades en los pacientes crónicos. Los pacientes hipertensos duplican el gasto sanitario de los pacientes con tensión normal, y un paciente diabético cuesta al año un $67\,\%$ más que un paciente sin diabetes. A esas cifras hay que sumarle los tratamientos y gastos asociados a las comorbilidades. Normalmente, para clasificar a los pacientes crónicos se considera un estado de salud por categoría (enfoque multi-clase). El problema de tener categorías excluyentes entre sí, es cuando aparecen estados de salud que encajan en más de una categoría (enfoque multi-etiqueta). Ese enfoque multi-etiqueta es la manera natural de relacionar estados de salud con categorías, debido a que cada paciente es distinto y no siempre el estado de salud "encaja" en una sola categoría. La realización de este trabajo encuentra su motivación en el diseño de sistemas automáticos predictivos del estado de salud, abordando el diseño de dichos sistemas bajo dos enfoques: multi-clase y multi-etiqueta.

Este trabajo tiene como objetivo diseñar modelos predictivos del estado de salud considerando pacientes sanos y crónicos (de manera conjunta), y únicamente pacientes crónicos. Las cronicidades consideradas son hipertensión, diabetes y sus comorbilidades. Disponemos de 65201 casos asignados al Hospital Universitario de Fuenlabrada., habiendo considerado para este trabajo 2735 pacientes para representar cada estado de salud con el mismo número de casos. El vector de características de un paciente está formado por características demográficas, diagnósticos y dispensación farmacéutica (2265 campos). Las características clínicas pueden estar codificadas bien como la presencia/ausencia de un diagnóstico o fármaco (características binarias), o bien como el número de veces que ese diagnóstico/fármaco aparece en un periodo de tiempo (ocurrencia). Dado el número elevado de características frente al número de casos disponibles (2265 *versus* 2735), se realiza un proceso de selección de características previo al diseño de modelos de clasificación, tanto para multi-clase como para multi-etiqueta. Los esquemas de clasificación evaluados en este trabajo son: Máquinas de Vectores Soporte, Regresión Logística Nominal, Árboles de decisión, *Random Forest* y Perceptrón Multicapa. Las prestaciones de los modelos se evalúan en base a la tasa de acierto (enfoque multi-clase y multi-label), obteniendo varias conclusiones. Por un lado, es preferible considerar características clínicas binarias en modelos no lineales, y basadas en ocurrencia en modelos lineales. La mejor tasa de acierto se obtiene con características clínicas binarias. Por otro lado, se observa un ligero incremento en las prestaciones de los modelos no lineales multi-etiqueta frente a los esquemas multi-clase.

Tras la finalización de este trabajo, se concluye que aunque las tasas de acierto para los enfoques multi-clase y multi-etiqueta son muy prometedoras, el número de pacientes considerado para cada categoría limita el aprendizaje de los modelos propuestos. En esa línea, sería muy interesante analizar un número de pacientes más elevado de cada categoría.

Índice general

Acrónimos y Siglas

ADAM *Adaptative Moment Estimation*. 42

ANOVA ANalysis Of VAriance. 34

ATC Sistema de Clasificación Anatómica, Terapéutica, Química. 10, 13, 14, 17–28, 49, 58–60, 63, 64, 76–78

BBDD Bases de Datos. 1, 3, 17, 33, 47–50, 52–60, 63, 69, 72–76, 79–90

BP *BackPropagation*. 41

CIE Sistema de Clasificación Internacional de Enfermedades. 12, 13, 17–28, 49, 58–60, 63, 64, 76–78

CPU *Central Processing Unit*. 47, 48

CRG *Clinical Risk Groups*. 14, 15, 17–28, 35, 48, 49, 55, 57, 58, 63, 64

CV *Cross Validation*. 34, 35, 52

DT *Decision Tree*. 39, 40, 42, 53, 79–90

EM *Embedded Methods*. 40

ENT Enfermedades No Transmisibles. 7, 8

ETSIT Escuela Técnica Superior de Ingeniería de Telecomunicación. 47

HT *Hyperparameter tuning*. 34

HUF Hospital Universitario de Fuenlabrada. 3, 12, 17, 63, 64, 69

KDD *Proceso Knowledge Discovery in Database*. 31

***k*NN** Esquema de los *k*vecinos más cercanos.. 39, 42, 53, 79–90

MLP *Multi-Layer Perceptron*. 41, 43, 53, 55, 57–60, 64, 79–90

OMS Organización Mundial de la Salud. 7, 8, 12

OOB *Out Of Bag*. 40

RAM *Random Access Memory*. 47

RBF *Radial Basis Function*. 38

RLM Regresión Logística Multinomial. 37, 38, 42, 52–54, 64, 79–83

SGD *Stochastic Gradient Descent*. 42

SVM *Support Vector Machines*. 38, 42, 52–59, 64, 79–90

URJC Universidad Rey Juan Carlos. 47

Índice de figuras

Índice de tablas

Capítulo 1

Introducción y objetivo

La Sección 1.1 justifica y pone en contexto el tema tratado en este trabajo: "Aprendizaje automático para la clasificación de pacientes crónicos y sus comorbilidades". La Sección 1.2 presenta el objetivo que se trata de alcanzar en este trabajo. En la Sección 1.3 se expondrá la metodología seguida para alcanzar el objetivo. Por último en la Sección 1.4 se realiza un breve resumen del contenido de cada capítulo que compone este trabajo.

1.1. Motivación

El volumen de información que se genera supera los $2{,}5$ *quintillones* de bytes cada día [12], y seguirá creciendo exponencialmente sin previsión de detenerse. Esto se debe a que las nuevas tecnologías tienen mucha facilidad de generar datos digitalmente. Es un hecho que las bases de datos (BBDD) están saturadas de información. Hay mucho conocimiento útil, desaprovechado, y que simplemente está ahí, encerrado en esta gran cantidad de datos.

La minería de datos o *Data Mining* trata de analizar datos almacenados, clasificarlos y hacer predicciones para nuevos datos. Las tareas de predicción las lleva a cabo un área del *Data Mining* llamada aprendizaje automático que se basa en la clasificación, esto implica la búsqueda de una función clasificadora que permita identificar si un caso pertenece a una u otra clase (codificada a través de una etiqueta). En su origen, la tarea de clasificación se definió como la asignación de una clase única para cada caso. Si se consideran sólo dos clases mutuamente excluyentes, se habla de clasificación binaria; si son más de dos clases las posibles alternativas, la tarea se suele denominar clasificación multi-clase. Hay muchos problemas en la vida real donde la restricción de una única clase por caso, no se cumple. La clasificación multi-etiqueta cubre la necesidad de que un caso pueda estar asociado simultáneamente a más de una clase. Un ejemplo del enfoque multi-etiqueta es la asignación de géneros de una película en *Internet Movie Database* (IMDb), con varios géneros posibles, por ejemplo la película *The Dark Knight*(2008) entra en las clases de acción, crimen y drama. La clasificación multi-etiqueta se ha aplicado para resolver problemas en múltiples campos de conocimiento, tales como categorización de documentos [13], bioinformática [14], clasificación de imágenes [15] y sonidos [16], o medicina entre otras.

Este trabajo se va a centrar en la aplicación del aprendizaje automático en el campo de la medicina, específicamente en la predicción del estado de salud de los pacientes crónicos. El enfoque multi-etiqueta se adapta a la necesitad de clasificar patologías distintas que comparten el mismo estado de salud (síntomas). Por ejemplo, los síntomas de la esquizofrenia son muy parecidos a los de la paranoia. Este trabajo se centrará en las enfermedades crónicas de hipertensión, diabetes y comorbilidades crónicas . Cuando se padece una enfermedad crónica hay un riesgo muy alto de

padecer otra enfermedad crónica asociada a la primera (comorbilidad), es decir, tendríamos a un paciente con un estado de salud concreto asociado a más de dos enfermedades crónicas, enfoque multi-etiqueta.

La utilización del aprendizaje automático en el ámbito de la diagnosis no es novedoso, son numerosos los trabajos que exploran ese camino. El trabajo presentado en [17] se apoya en el aprendizaje automático y la multi-clase para identificar y diagnosticar dos enfermedades graves de retina. Los autores utilizan un esquema de redes neuronales artificiales que se entrena con casi un millón de imágenes de retina, córnea y nervio óptico, en lugar de considerar imágenes completas del ojo. Este enfoque han permitido que el modelo diseñado pueda reconocer más fácilmente los daños en las distintas partes del ojo. Con los resultados obtenidos, los autores observaron que el esquema de clasificación propuesto diagnosticaba igual de bien que el oftalmólogo, siendo capaz de tomar una decisión acerca del tipo de enfermedad en "*3 segundos con una eficacia de 95 %*" [17].

Kang Zhang, director del Instituto de medicina genómica de la Escuela de Medicina de la Universidad de California San Diego (EE.UU.), sobre la aplicación en la medicina del aprendizaje automático, comenta: "*la inteligencia artificial puede ayudar a los médicos a realizar un diagnóstico al instante. Y eso es incluso más importante en áreas rurales o países en vías de desarrollo, donde además no hay especialistas*" [18]. Por tanto, el uso del aprendizaje automático para ayudar al personal clínico a realizar diagnósticos, es vista de manera positiva. Sólo queda averiguar los modelos que mejor se adapten a los estados de salud. Este trabajo introduce varios modelos de clasificación bajo los enfoques multi-clase y multi-etiqueta.

En el artículo [19], el doctor Luis Eduardo Juarez-Orozco, autor del estudio perteneciente al Centro de PET (*Positron Emission Tomography*) de Turku (Finlandia), indica que "*el algoritmo aprende progresivamente de los datos y, después de numerosas rondas de análisis, determina los patrones de alta dimensión que deben usarse para identificar de manera eficiente a los pacientes que tienen el evento. El resultado es una puntuación de riesgo individual*". Con la repetición y el ajuste en el algoritmo *LogitBoost*, han conseguido que con sólo 85 variables se pueda predecir un ataque cardiaco. El entrenamiento se llevó a cabo con datos de pacientes con enfermedades cardiacas, monitorizados en un periodo de seis años. El clasificador *LogitBoost* es un esquema de regresión logística aditiva. Otras aplicaciones clásicas del aprendizaje automático se pueden extrapolar a otros campos de la medicina, tales como la categorización de historiales clínicos [20] o de radiografías [21].

Tratando el problema de la diagnosis, hay varios trabajos que aplican técnicas multi-etiqueta. Por ejemplo, la aplicación del algoritmo *RandomForest* para seleccionar las mejores características y diagnosticar el nivel de gravedad para la gastritis crónica [22]. En este artículo, las características seleccionadas con *Random Forest*, se asemejan a las seleccionadas por expertos clínicos.

Otro problema que se ha modelado mediante técnicas multi-etiqueta es el de la entre medicamentos. Una posible aplicación es la que se presenta en el artículo [23], que trata de diagnosticar problemas cardiovasculares con un vector de características compuesto por los medicamentos que han podido originar la enfermedad cardiovascular asociada.

1.2. Objetivo

El objetivo de este trabajo es evaluar, en términos de prestaciones, métodos que permitan predecir el estado de salud de pacientes que presenten o no algún tipo de cronicidad considerando dos escenarios: pacientes sanos y crónicos (de manera conjunta); y únicamente pacientes crónicos. Esta evaluación se analizará bajo los enfoques de multi-clase y multi-etiqueta. Este trabajo se centrará en las siguientes cronicidades: hipertensión, diabetes y sus comorbilidades.

Desde un punto de vista teórico, se definen los elementos básicos relacionados con el aprendizaje automático, tales como validación cruzada o selección de características, entre otros. Los esquemas de clasificación usados para este trabajo son: Máquinas de Vectores Soporte (lineal y no lineal), Regresión Logística Nominal, Árboles de decisión, *Random Forest* y Perceptrón Multicapa. Para su evaluación definiremos medidas de prestaciones para los enfoques multi-clase y multi-etiqueta.

A nivel práctico, se realizan experimentos con los esquemas indicados anteriormente utilizando en el lenguaje de programación *Python*. Compararemos los resultados obtenidos para los enfoques multi-clase y multi-etiqueta.

1.3. Metodología

La metodología seguida en este trabajo se resume en los siguientes puntos:

- Definición y situación de las enfermedades crónicas en España, haciendo hincapié en la hipertensión, la diabetes y sus comorbilidades.
- Definición de los sistemas de codificación de datos clínicos y los sistemas de clasificación de pacientes.
- Análisis descriptivo de las BBDD. Analizaremos la distribución de los pacientes con respecto a la edad y el género. Además, haremos un análisis poblacional de las características clínicas y farmacológicas.
- Definición y explicación de los métodos utilizados en el trabajo para clasificar las cronicidades: selección de características, algoritmos y evaluación.
- Definición de modelos lineales y no lineales que ayuden a predecir el estado de salud de los pacientes. Los modelos usados son: regresión logística multinominal, máquinas de vectores soporte (lineal y no lineal), *k*-NN, Árboles de decisión, *Random Forest* y *Multi-layer Perceptron*.
- Discusión sobre los resultados obtenidos con los diferentes modelos, tanto desde el punto de vista de clasificación multi-clase como multi-etiqueta.

1.4. Estructura de la memoria

A continuación se muestra un breve resumen del contenido de cada capítulo.

Capítulo 1: Introducción y objetivos. Es un capítulo introductorio que se divide en 4 partes. La primera parte justifica y pone en contexto el tema que se trata en el trabajo. La segunda parte presenta los objetivos del proyecto, mientras que la tercera describe la metodología usada para su consecución. La última parte resume la estructura del proyecto por capítulos.

Capítulo 2: Conceptos previos. Se describen las enfermedades crónicas analizadas en este trabajo: hipertensión, diabetes y sus comorbilidades, los sistemas de codificación de datos clínicos y los sistemas clínicos de clasificación de pacientes.

Capítulo 3: Base de datos y análisis descriptivo. Se realiza el pre-procesamiento de los datos proporcionados por el HUF. Se analiza la distribución poblacional en base a la edad y al género de los pacientes según su estado de salud. Para cada grupo de salud se estudia la presencia de los códigos de diagnósticos y farmacéuticos, en porcentaje para cada género.

Capítulo 4: Métodos. Se detallan las fases de los esquemas de aprendizaje máquina y los requisitos para obtener un buen modelo. Además, se explican los métodos de selección de características, los modelos de clasificación y las medidas de prestaciones para la evaluación de los modelos.

Capítulo 5: Experimentos y Resultados. Se explica el proceso seguido para predecir el estado de salud de los pacientes. Se muestran los resultados obtenidos en función de distintas medidas de prestación para los distintos algoritmos definidos en el Capítulo 4 para los dos tipos de clasificación: multi-clase y multi-etiqueta.

Capítulo 6: Conclusiones y líneas futuras. Se desarrollan las conclusiones que se derivan del trabajo y se analizan posibles líneas futuras de trabajo.

Anexo A: Código *Python*. Se presenta el código usado en el lenguaje de programación *Python*.

Anexo B: Características escogidas aplicando distintos métodos de selección de características. Se listan las diferentes características obtenidas al aplicar diferentes métodos de selección de características.

Anexo C: Tablas de resultados obtenidos en los experimentos multi-clase y multi-etiqueta. Tablas compuestas por los valores de las medidas de evaluación para los experimentos del Capítulo 5.

Anexo D: Diagrama de Gantt y presupuesto del trabajo. Se detalla la evolución temporal del trabajo y el coste económico asociado.

Bibliografía. Listado de referencias consultadas en la realización de este trabajo.

Capítulo 2

Conceptos Previos

En este capítulo se exponen los conceptos básicos en los que se sustenta este trabajo. Primero se realiza una definición de las enfermedades crónicas y su situación actual en España. Consecutivamente se describen las enfermedades crónicas consideradas en este trabajo. expondrán los sistemas de clasificación para la codificación de los datos cínicos utilizados en este trabajo. Además de los sistemas de clasificación de pacientes poblacionales.

2.1. Enfermedades crónicas

Tal y como indica la Organización Mundial de la Salud (OMS), las enfermedades crónicas *son aquellas enfermedades de larga duración, por lo general de evolución lenta, y cuya curación no se puede prever* [24]. También se refieren a ellas como enfermedades no transmisibles (ENT) [25]. Para combatir las enfermedades crónicas. Hay dos tipos de tratamientos, por un lado los tratamientos que las cure de manera definitiva. Por otro lado, los tratamientos para retrasar su evolución.

Las ENT afectan a todos los grupos de edad, todas las regiones y países. Estas enfermedades se suelen asociar a los grupos de edad más avanzada. Los datos a nivel mundial muestran que 15 millones de todas las defunciones atribuidas a las ENT se producen entre los 30 y los 69 años de edad. Más del $85\,\%$ de estas defunciones *"prematuras"* ocurren en países de ingresos bajos y medianos. Niños, adultos y ancianos son todos ellos vulnerables a padecer, al menos, una enfermedad crónica. Esto implica que los costes de la atención sanitaria pueden mermar los recursos económicos de las familias.

2.1.1. Situación actual de las enfermedades crónicas en España

Según recoge el informe de la OMS del perfil poblacional de España en 2016 [1], la población de 46 millones se redujo en $400{,}000$ defunciones debido a las enfermedades crónicas. En la Figura 2.1 se muestra el porcentaje de defunciones prematuras de las principales cronicidades, tales como, cardiovasculares (cardiopatía isquémica, insuficiencia cardíaca, enfermedad cerebro vascular), respiratorias crónicas (enfermedad pulmonar obstructiva crónica y asma crónico) y la diabetes, entre otras. La OMS también ha elaborado un informe [2] que describe las *"Mejores inversiones"* para luchar contra las enfermedades crónicas. En este informe se detallan:

Principales factores de riesgo.

- Inactividad física.

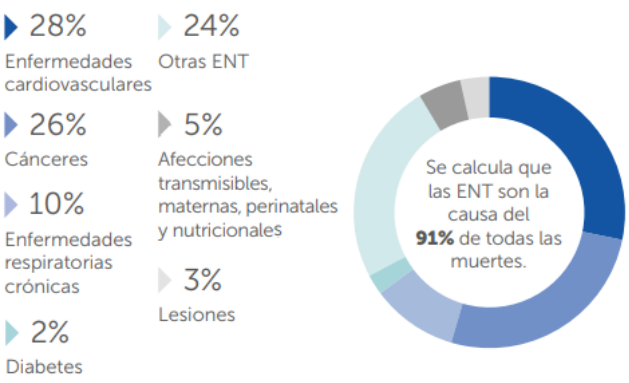


Figura 2.1: Porcentaje de defunciones prematuras con respecto a las ENT. Tomada de [1].

- Consumo de tabaco y alcohol.
- Alimentación poco sana.

Reducción de los factores de riesgo. Promoviendo la creación de entornos que fomenten una vida saludable. Por ejemplo, aumentando los impuestos especiales al tabaco y al alcohol o reduciendo de la cadena alimenticia las grasas trans industriales, la sal, el azúcar, etc.; poner en marcha educación y concienciación pública para fomentar la actividad física.

Gestionar las enfermedades crónicas: entre otras medidas destacan:

- Explorar mecanismos viables de financiación sanitaria e instrumentos económicos innovadores basados en los datos recogidos en la Historia Clínica Electrónica.
- Capacitar al personal sanitario y fortalecer la capacidad del sistema de salud, sobre todo en la atención primaria, con el objetivo de abordar la prevención y el control de las ENT.
- Ampliar el uso de las tecnologías digitales para incrementar el acceso a los servicios de salud y su eficacia para la prevención de las ENT.

Con esas *"mejores inversiones"*, la OMS desea reducir la probabilidad de defunción prematura en las ENT, tal y como se muestra en la Figura 2.2.

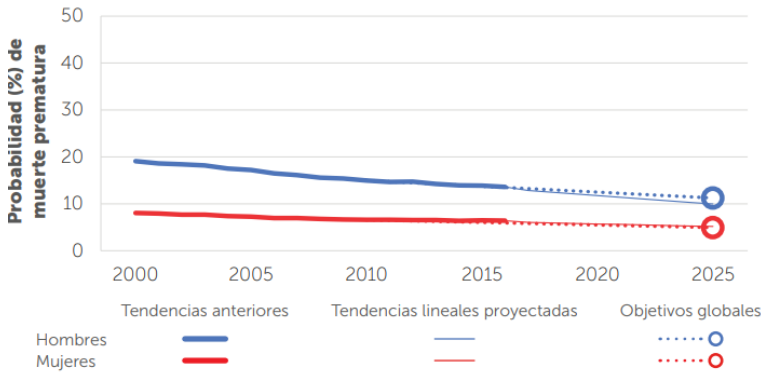


Figura 2.2: Riesgo de mortalidad prematura debido a ENT (%). Tomada de [2].

En cifras de presencia de enfermedades crónicas [3], España se sitúa el decimocuarto país en porcentaje de pacientes crónicos de la Unión Europea. Alrededor del $42\,\%$ de la población española padece al menos una patología crónica. Estas cifras aumentan si se tiene en cuenta el envejecimiento

de la población, alcanzando hasta el $70\,\%$ de los mayores de 65 años, con una media de cuatro patologías crónicas por persona. Las cronicidades más comunes y sus porcentajes se muestran en la Figura 2.3.

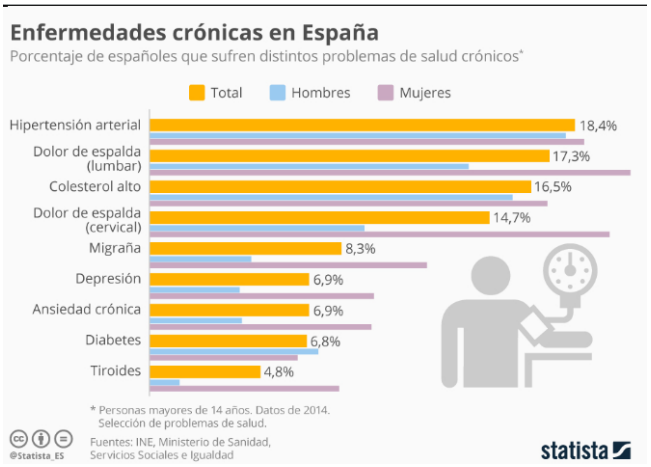


Figura 2.3: Porcentaje de enfermedades crónicas en España. Tomada de [3].

En las siguientes subsecciones se explican las cronicidades con las que se va a trabajar. Se trata de la hipertensión, la diabtetes y sus comorbilidades. Este tipo de enfermedades supone el $80\,\%$ de las consultas de Atención Primaria, el 60 % de ingresos hospitalarios y el $85\,\%$ de los pacientes ingresados en Medicina Interna, siendo la primera causa de gasto sanitario en nuestro país [3].

El barómetro realizado por *EsCronicos* [26] en 2017 da una nota de $6,4$ a la calidad en la asistencia sanitaria en España. La nota es buena, pero casi todos los encuestados afirman que, para un mejor tratamiento de sus enfermedades, las Comunidades Autónomas deberían eliminar sus diferencias asistenciales para lograr una continuidad asistencial independiente de la Comunidad Autónoma en la que el paciente resida. Como novedad en este 2019, se ha introducido la receta electrónica interoperable, entre Comunidades Autónomas. Para los pacientes crónicos supone la posibilidad de moverse por todo el territorio nacional y obtener los medicamentos prescritos en cualquier farmacia, independientemente de la comunidad autónoma en la que hayan sido recetados.

2.1.2. Hipertensión

La hipertensión *es la elevación de los niveles de presión arterial. Los niveles de presión arterial máxima se producen cuando el corazón se contrae y ejerce presión sobre las arterias para que éstas conduzcan la sangre hacia los diferentes órganos del cuerpo humano* [27]. Este efecto hace que el corazón se sobre esfuerce y aumente su masa muscular, y las arterias se vuelven rígidas y estrechas. Las principales causas de la hipertensión son, entre otras, la obesidad, niveles altos y continuados de ansiedad y estrés o consumo de tabaco, alcohol y grandes cantidades de sal.

Para medir la presión arterial se usa un brazalete conectado a una pequeña bomba y a un medidor, que muestra dos valores numéricos sistólico y diastólico. El valor sistólico es *la presión sanguínea máxima cuando su corazón está expulsando la sangre*. El valor diastólico indica *la presión cuando su corazón se llena de sangre* [27].

Hay dos tipos de hipertensión:

Hipertensión primaria o esencial. Este es el tipo más común de hipertensión arterial y, por lo general, tarda muchos años en aparecer. Probablemente sea resultado del estilo de vida del

paciente, su entorno o cómo su cuerpo cambia a medida que envejece.

Hipertensión secundaria. Este tipo se produce cuando un problema de salud o un medicamento provoca la hipertensión arterial. Las causas más comunes son las enfermedades que afectan los riñones, las arterias, el corazón o el sistema endocrino.

Los medicamentos más comunes para tratar la hipertensión son:

- Diuréticos: *son medicamentos que actúan sobre los riñones para ayudar al cuerpo a eliminar el sodio y el agua y, de este modo, reducir el volumen de sangre* [27]. El código correspondiente al subgrupo terapéutico en el sistema de clasificación anatómica, terapéutica y química (ATC) es C03.

- Agentes activos sobre el sistema renina-angiotensina. *Ayudan a relajar los vasos sanguíneos al bloquear la formación de una sustancia química natural que los estrecha* [27]. El código ATC perteneciente al subgrupo terapéutico es C09.

2.1.3. Diabetes

La Diabetes es una enfermedad que *se origina porque el páncreas no sintetiza la cantidad de insulina que el cuerpo humano necesita, la elabora de una calidad inferior o no es capaz de utilizarla con eficacia* [28]. La ínsula es una hormona que regula la glucosa en sangre. Esta hormona consigue que la glucosa entre en el organismo y sea transportada al interior de las células. Donde se transforma en energía para que funcionen los músculos y los tejidos. Altos niveles de glucosa en sangre pueden ser perjudicialaes para el corazón, el riñón y las arterias, esto implica que las personas que tienen diabetes y no lo saben o no siguen el tratamiento, tienen más riesgo de problemas renales, infartos, pérdida de visión y amputaciones de miembros inferiores. Las principales causas de la diabetes son, entre otras, la obesidad, el envejecimiento, el embarazo o el consumo de alcohol o tabaco.

Hay varios tipos de diabetes [28]:

Tipo 1. Suele aparecer en la infancia y ocurre cuando el páncreas ataca a sus propias células como si fueran extrañas.

Tipo 2. Aparece en la edad adulta y su principal causa es la obesidad, ya que el páncreas no produce suficiente cantidad de insulina o las células del cuerpo ignoran o reconocen mal la insulina por culpa de la grasa corporal.

Gestacional. Durante el embarazo, la insulina aumenta para incrementar las reservas de energía. Suele desaparecer tras el parto, pero las mujeres con diabetes gestacional tienen alto riesgo de desarrollar diabetes Tipo 2 a lo largo de su vida.

Otros tipos. Pueden derivarse por una lesión del páncreas, por causas genéticas o por el consumo de ciertos fármacos.

La insulina se puede regular inyectándola diariamente, utilizando una aguja/jeringa o un autoinyector. También se pueden usar medicamentos de tipo oral, se inhalan en polvo por la boca, llegan a los pulmones y pasan rápidamente a la sangre. Los medicamentos asociados a la insulina tienen el código ATC A10 (subgrupo terapéutico).

2.1.4. Comorbilidades

Las comorbilidades son las patologías crónicas que aparecen debido a los tratamiento o complicaciones de otras patologías. En este apartado se presentan las patologias crónicas más comunes asociadas a la hipertensión (Sección 2.1.2) y a la diabetes (Sección 2.1.3).

Las comorbilidades que un paciente con hipertensión puede desarrollar están relacionadas con los tres grandes sistemas vasculares: corazón, cerebro y riñon.

- Enfermedad cardíaca: problemas cardíacos que ocurren debido a la hipertensión arterial, como la insuficiencia cardiaca.
- Insuficiencia renal: si los vasos sanguíneos de los riñones se dañan, es posible que dejen de eliminar los desechos y el exceso de líquido del cuerpo [29]. En este caso, puede que el exceso de líquido en los vasos sanguíneos aumente aún más la presión arterial. Es un ciclo peligroso.
- Dislipidemia diabética: *elevación anormal de concentración de grasas en la sangre (colesterol o triglicéridos), suele causar dolor abdominal, pancreatitis, fatiga, zumbido de oídos y dolor de ardor en miembros inferiores* [29].

Con respecto a las comorbilidades de la diabetes, la mayoría están asociadas a la diabetes del Tipo 2, como indica la Figura 2.4 perteneciente al estudio realizado en [4]. En el supuesto caso de que la diabetes Tipo 2 se haya desarrollado a raíz del sobrepeso o la obesidad, es muy probable que el paciente también tenga hipertensión, ya que al aumentar de peso se eleva la presión arterial. En el caso de padecer diabetes y no tratarla, la glucosa en la sangre se eleva produciendo graves problemas de salud, entre otros enfermedades cardíacas y daños en los nervios y riñones. Las siguientes patologías crónicas suelen ser las más comunes según [4].

- Neuropatía diabética. *Daño producido en los nervios de todo el cuerpo, en muchos casos en piernas y en los pies* [4]. Esto hace perder sensibilidad en algunas partes del cuerpo o sentir dolor, hormigueo o ardor en las extremidades.
- Dislipidemia diabética y enfermedad cardiaca. Mismas patologías crónicas que para la hipertensión.

Comorbilidad	Frecuencia	%
Hipertension arterial	245	64,14
Neuropatiadiabetica	103	26,96
Dislipidemia	61	15,97
Hipotiroidismo	41	10,73
Retinopatiadiabetica	20	5,24
Pie diabetico	20	5,24
Nefropatiadiabetica	8	2,09
Cardiopatia	7	1,83
Depresion	6	1,57
Otros	43	11,25

Figura 2.4: Distribución de comorbilidades pacientes con diabetes Tipo 2. Tomada de [4].

2.2. Sistemas de codificación de datos clínicos

En esta sección se revisa la definición de los sistemas de clasificación de diagnósticos y medicamentos utilizados en los sistemas de codificación de datos clínicos.

2.2.1. Sistema de Clasificación Internacional de Enfermedades

El Sistema de Clasificación Internacional de Enfermedades (CIE), redactado y publicado por la OMS, *es un índice de clasificación y codificación de las enfermedades y una amplia variedad de signos, síntomas, circunstancias sociales y causas externas de enfermedades* [11]. El principal objetivo de la CIE es clasificar las enfermedades, afecciones y causas externas de enfermedades y traumatismos, con objeto de recopilar información sanitaria útil relacionada con defunciones, enfermedades y traumatismos (mortalidad y morbilidad). Cabe destacar que los códigos CIE-9 pueden clasificarse dentro de una única categoría, pues éstas son excluyentes. La codificación CIE-9 es utilizado por las aseguradoras médicas cuyos reembolsos dependen de la codificación de la CIE; por los administradores de los programas nacionales de salud; por los especialistas en recopilación de datos; y por otras personas que hacen un seguimiento de los progresos en la salud mundial y determinan la asignación de los recursos sanitarios.

Aunque el sistema CIE ya se encuentre en la onceava versión [30], este trabajo usa la novena edición publicada en 1975 [31], debido a que los datos de pacientes proporcionados por el HUF, pertenecen al año 2012, el último año que se utilizó esta versión. A partir del 2012 se utiliza CIE versión 10 . La clasificación de códigos CIE-9 se muestra en la Tabla 2.1, donde la categoría y la subclasificación indican la gravedad del diagnóstico. Su codificación se muestra a continuación y está sacada de [11].

1. Enfermedades y lesiones A cada código de tres dígitos se le asocia una enfermedad o lesión. Cada código se divide en subcategorías y subclasificaciones. Las enfermedades y lesiones se agrupan de la siguiente manera:

001-139 Enfermedades infecciosas y parasitarias

140-239 Neoplasias

240-279 Enfermedades endocrinas, nutricionales y metabólicas, y trastornos de la inmunidad

280-289 Enfermedades de la sangre y órganos formadores de sangre

290-319 Desórdenes mentales

320-289 Enfermedades del sistema nervioso

290-459 Enfermedades de los órganos sensoriales

460-519 Enfermedades del sistema respiratorio

580-629 Enfermedades del sistema genitourinario

630-679 Complicaciones del embarazo, parto y puerperio

680-709 Enfermedades de la piel y tejido subcutáneo

710-739 Enfermedades del sistema musculoesquelético y tejido conectivo

740-759 Anomalías congénitas

760-779 Ciertas condiciones originadas en el período perinatal

[**780-799** Síntomas, signos y condiciones mal definida

800-999 Lesiones y envenenamientos

A continuación se muestra un ejemplo de cómo se estructura el código de un diagnóstico:
Categoría: 250 - Diabetes.

Subcategoría: 250.7 - Diabetes con trastornos circulatorios periféricos.
Subclasificación: 250.73 - Tipo I [tipo juvenil], incontrolada.

2. Códigos E
Los códigos E indican las causas externas de la lesión e intoxicación, es decir, que clasifican la naturaleza de la enfermedad registrada. Al igual que las enfermedades y lesiones, cada código tiene asociado una subcategoría y una subclasificación.

3. Códigos V
Con los códigos V tienen como función la clasificación los factores que influyen en el estado de salud y el contacto con los servicios sanitarios, es decir, que cubren ciertas circunstancias que no sean ni enfermedades ni lesiones. Sólo tienen asociado una subcategoría.

4. Códigos M
Indican la morfología de las neoplastias/tumores, como se indica en [11] cada código MXXXX es un tipo de tumor asociado a un código entre $140 - 239$ de la categoría Neoplasias .

Tipo de Clasificación	Categoría	Subcategoría	Subclasificación
Enfermedades: 000-999	XXX	XXX.X	XXX.XX
Códigos V: V01-V89	VXX	VXX.X	VXX.XX
Códigos E: E800-E999	EXXX	EXXX.X	—
Códigos M: M8000-M9970	MXXXX	—	—

Tabla 2.1: Codificación de los códigos CIE. Fuente [11].

2.2.2. Sistema de Clasificación Anatómica, Terapéutica, Química

El Sistema de Clasificación Anatómica, Terapéutica, Química Europeo (ATC) *codifica las sustancias farmacéuticas y medicamentos en cinco niveles con arreglo al sistema u órgano afectado y al efecto farmacológico, las indicaciones terapéuticas y la estructura química de un fármaco* [32] . A cada fármaco le corresponde un código ATC, y éste se especifica en prospecto del medicamento. Un ejemplo de codificación con códigos ATC se muestra en la Tabla 2.2.

Los códigos ATC se estructuran de la siguiente manera:

- Nivel 1. Grupo anatómico principal: la letra del alfabeto el órgano o sistema sobre el que actúa el fármaco.

A Sistema digestivo y metabolismo

B Sangre y órganos hematopoyéticos

C Sistema cardiovascular

D Medicamentos dermatológicos

G Aparato genitourinario y hormonas sexuales

H Preparados hormonales sistémicos

J Antiinfecciosos en general para uso sistémico

L Agentes antineoplásicos e imunomoduladores

M Sistema musculoesquelético

N Sistema nervioso

P Productos antiparasitarios, insecticidas y repelentes

R Sistema respiratorio

S Órganos de los sentidos

V Varios

- Nivel 2. Subgrupo terapéutico.
- Nivel 3. Subgrupos terapéuticos-farmacológicos.
- Nivel 4. Subgrupos químico-terapéuticos
- Nivel 5. Principios activos o asociaciones farmacológicas

Nivel	ATC	Descripción
1: Grupo anatómico principal	M	Sistema musculoesquelético
2: Subgrupo terapéutico	M01	Antiinflamatorios y antirreumáticos
3: Subgrupo terapéutico farmacológico	M01A	Antiinflamatorios y antirreumáticos no esteroiditos
4: Subgrupo químico-terapéutico	M01AE	Antiinflamatorios: derivados del ácido propiónico
5: Principio activo ATC	M01AE01	Ibuprofeno

Tabla 2.2: Ejemplo de codificación ATC.

2.3. Sistemas de clasificación de poblacional

Los sistemas de clasificación de pacientes poblacional se basan en *modelos predictivos de ajuste de riesgo y permiten normalizar, medir, evaluar y gestionar la actividad asistencial*. Una de las muchas finalidades de estos sistemas de clasificación es la estratificación poblacional: *es una estrategia de análisis orientada a identificar subgrupos con diferentes niveles de necesidades asistenciales*. Ambas definiciones están tomadas de [33].

Un sistema de estratificación poblacional son los *Clinical Risk Groups* (CRG) [5], que clasifican a las personas en categorías clínicas mutuamente excluyentes a partir de los contactos en cualquier ámbito asistencial (Atención Primaria, Especializada, Dispensación Farmacéutica, Hora/Día, etc.), durante un período determinado. En este trabajo vamos a analizar el periodo de 2012.

Los CRG permiten al personal clínico:

- Planificar y evaluar los sistemas de salud.

- Asignar los recursos asistenciales en base a la carga real de trabajo definida por el estado de salud de una población.
- Analizar los patrones de frecuentación y consumo de servicios.
- Realizar el seguimiento de las tasas de prevalencia de enfermedades crónicas.
- Monitorizar los estados de salud de la población.

Los CRG se clasifican en las categorías mostradas en la Tabla 2.5. En este trabajo vamos a tratar con las categorías: 1 - *Healthy* (pacientes sanos); 5 - *Single dominant or moderate chronic disease* (pacientes con una enfermedad crónica); 6 - *Significant chronic disease in multiple organ systems (pairs)* (pacientes con dos enfermedades crónicas); y 7 - *Dominant chronic disease in 3 or more organ systems (triplets)* (pacientes con tres enfermedades crónicas). Pero para nosotros los CRG base interesantes son los pertenecientes a los pacientes sanos, hipertensos, diabéticos y sus comorbilidades. Dichos CRG base se muestran Tabla 2.3.

3M CRG core health status groups (1-9)	Base 3M CRGs (Total = 330)	Description/Example of base 3M CRG	Severity levels	Number of 3M CRGs (Total = 1,408)
9 - Catastrophic condition status	10	History of major organ transplant	4	40
8 - Dominant and metastatic malignancies	30	Colon malignancy - under active treatment	4	120
7 - Dominant chronic disease in 3 or more organ systems (triplets)	28	Diabetes mellitus, congestive heart failure (CHF) and chronic obstructive pulmonary disease (COPD)	6	168
6 - Significant chronic disease in multiple organ systems (pairs)	78	Diabetes mellitus and CHF	6	468
5 - Single dominant or moderate chronic disease	125	Diabetes mellitus	4	500
4 - Minor chronic disease in multiple organ systems	1	Migraine and benign prostatic hyperplasia (BPH)	4	4
3 - Single minor chronic disease	50	Migraine	2	100
2 - History of significant acute disease	6	Chest pains	None	6
1 - Healthy/Non-Users	2	Healthy (no chronic health problems)	None	2

Figura 2.5: Categorías básicas en el CRG y en ejemplo de cada. Tomada de [5].

CRG base	Descripción del CRG
1000	Sanos
5192	Hipertensión. Niveles: 1,2,3,4
5424	Diabetes. Niveles: 1,2,3,4
6144	Diabetes - Hipertensión. Niveles: 1,2,3,4,5,6
7071	Diabetes - Hipertensión - Otra Enfermedad Crónica Dominante. Niveles: 1,2,3,4,5,6

Tabla 2.3: CRG base considerados en el trabajo. Cada CRG base agrupa varios niveles de gravedad.

Capítulo 3

Base de datos y análisis descriptivo

ste capítulo se divide en dos partes, la primera, Sección 3.1, donde se describe la composición de la base de datos de pacientes entregada por el HUF y los trabajos previos en relación a ella. La segunda parte de este capítulo, Sección 3.2, se analiza la distribución de las características CIE-9 y ATC para los pacientes del HUF asociados a los CRG indicados en la Tabla 2.3, de la Sección 2.3.

3.1. Base de datos

El Hospital Universitario de Fuenlabrada (HUF) es un hospital público que se encuentra en Fuenlabrada (Madrid) y cubre a una población de $220{,}000$ habitantes. En ese número de habitantes están incluidas las poblaciones situadas a menos de 5 km de distancia de Fuenlabrada, que son Moraleja de Enmedio y Humanes de Madrid, así como los habitantes no empadronados en dichas poblaciones pero que residen en ellas. El HUF facilitó una BBDD con información demográfica, clínica y de dispensación farmacológica de sus pacientes. Todos los datos proporcionados estaban debidamente anonimizados, teniendo asociado cada paciente un identificador único. Por cada paciente, se dispone del número de contactos que éste ha tenido con el Sistema Nacional de Salud en el periodo de 2012. Cada contacto contiene: (1) la información demográfica (edad y género); (2) fecha en la que tuvo lugar el contacto; (3) especialidad médica a la que corresponde el contacto ; (4) diagnósticos y procedimientos (codificados en CIE-9); y (5) información de dispensación farmacéutica (codificada en ATC). Siguiendo las recomendaciones de nuestros colaboradores en el HUF, en este trabajo no se tendrán en cuenta los procedimientos, por aportar escasa información para identificar cronicidades.

Como se indica en la Tabla 2.3, de la Sección 2.3, en este trabajo consideramos 5 grupos de CRG base (véase la Tabla 3.2): (1) CRG1000, con 46835 pacientes sanos; (2) CRG5192, con 12447 pacientes con hipertensión; (3) CRG5424, con 2166 pacientes con diabetes; (4) CRG6144, con 3179 pacientes con hipertensión y diabetes; y (5) CRG7071, con 547 pacientes con hipertensión, diabetes y otra cronicidad.

El primer paso en el pre-procesado de datos fue la eliminación de pacientes itinerantes, es decir, pacientes que no están asociados a ningún centro de salud del HUF y su contacto con dichos centros es esporádico. El siguiente paso fue la limpieza de los códigos proporcionados. Así, con ayuda de personal del HUF, sustituimos los códigos que contenían letras y exclamaciones, por ejemplo $\tilde{A}¡\tilde{A}¡$, por los códigos CIE-9 correctos. Con la ayuda del personal del HUF, se redujo el número de códigos CIE-9 en base a los más relacionados con las cronicidades de hipertensión, diabetes y sus

comorbilidades. En definitiva, hemos utilizado 1517 códigos CIE-9 considerando únicamente los 3 primeros dígitos de los códigos de Enfermedades y códigos V; para los códigos E y M, se consideran códigos de una longitud de 4 dígitos.

En cuanto a los códigos ATC, están formados por 7 caracteres: 3 letras y 4 números. Se procedió a eliminar los dígitos que hacen referencia a *"Principios activos o asociaciones farmacológicas"*. Tras la reagrupación, se obtuvo un conjunto de 746 códigos ATC compuesto por 4 caracteres: 2 letras y 2 dígitos. En total, el vector de características posee 2265 campos, mostrados en la Tabla 3.1.

Edad	**Género**	**001-999 Enfermedades**	**V01-V89 Códigos V**	**E000-E999 Códigos E**	**M888-M997 Códigos M**	**Letra - Número - Número - Letra = ATC**

Tabla 3.1: Características en los CRGs a analizar en este trabajo.

Considerando el vector de características indicado en la Tabla 3.1 y los distintos CRG de la Tabla 2.3, se han realizado otros trabajos fin de grado/máster:

- Francisco Javier Gutiérrez Expósito [34] realizó parte del pre-procesado y limpieza de códigos ATC, estudiando a la vez su distribución estadística con respecto a los pacientes crónicos.
- Javier Fernández Sánchez [35], buscó asociaciones lineales que pudieran guardar las características con cada CRG, haciendo uso la técnica de análisis de correspondencias.
- Ana Alberca Díaz-Plaza [36] realizó un análisis predictivo del estado de salud de los paciente utilizando árboles de decisión.

En este trabajo se completará el estudio realizado por Ana Alberca con un punto de vista nuevo respecto a la forma de construir modelos, usando para ello la clasificación multi-etiqueta.

3.2. Análisis de los grupos poblacionales

En este apartado se analiza la distribución de características CIE-9-CM y ATC para los pacientes asociados a cada CRG, como indica la Tabla 3.2.

CRG base	**Número de pacientes**
CRG 1000	46835
CRG 5192	12447
CRG 5424	2166
CRG 6144	3179
CRG 7071	547

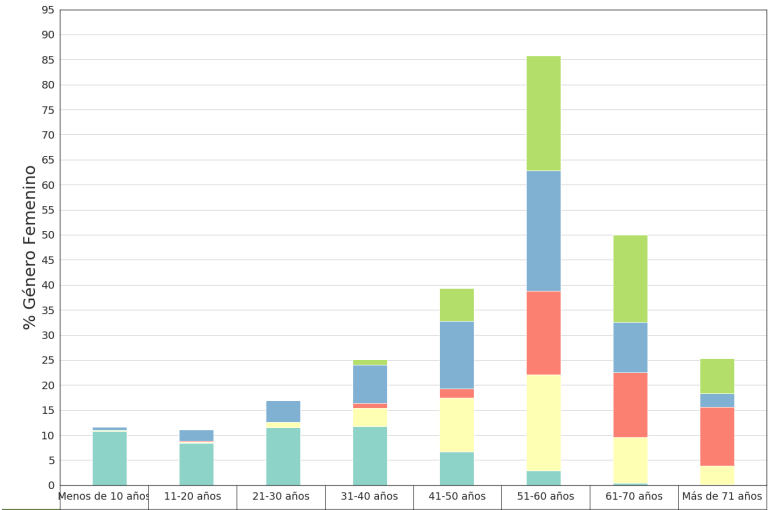
Tabla 3.2: Pacientes en los CRGs a analizar.

Las Figuras 3.1 y 3.2 muestran, en porcentajes, para cada CRG el porcentaje de mujeres y hombres por rangos de edad. Se observa que para ambos géneros se sigue la misma tendencia. Hasta los 23 años, predominan los pacientes sanos, es decir, ninguna cronicidad asociada. Cuando se superan los 40 años ya empiezan a aparecer las cronicidades. Esto es normal, ya que el envejecimiento del cuerpo es un factor de riesgo para la aparición de cronicidades.

Casi todos los valores que componen la Tabla 3.3, se han obtenido de operar con la presencia de los códigos CIE-9 y ATC referentes a cada CRGs. El único valor que se ha obtenido de la ocurrencia de los códigos CIE-9 y ATC, ha sido *Valor medio de códigos* ATC paciente del grupo. Esa tasa indica la

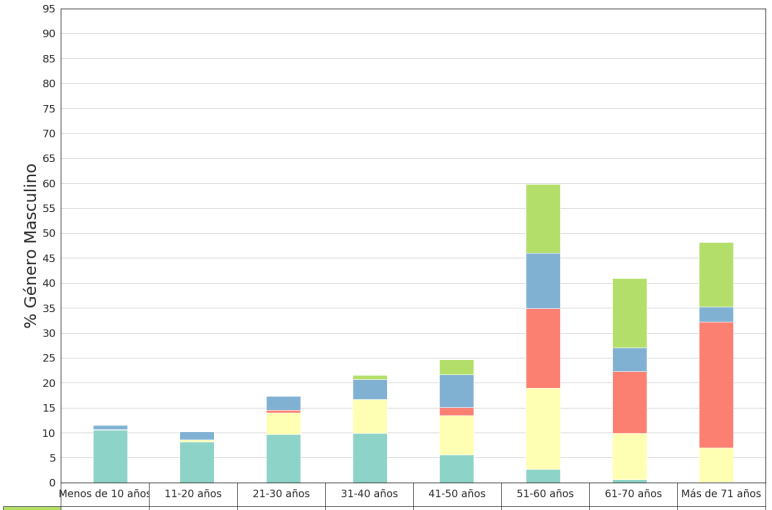
ocurrencia media de los códigos CIE-9 y ATC para cada CRG. Observando los valores que aporta la Tabla 3.3 vemos que cuando los pacientes tienen varias cronicidades asociadas, más códigos CIE-9 tiene asociados (Filas: *Códigos CIE-9 diferentes por paciente en cada grupo* y *Valor medio de códigos CIE-9 por paciente del grupo*). Lo mismo pasa con los códigos ATC, cuantas más patologías crónicas tienen los pacientes, más medicamentos se les han dispensado (Filas: *Códigos ATC diferentes por paciente en cada grupo* y *Valor medio de códigos ATC por paciente del grupo*).

En las subsecciones siguientes se presenta, por género y estado de salud, el número medio de códigos CIE-9 y ATC diferentes por paciente. También se proporciona el porcentaje de presencia de algunos códigos CIE-9 y ATC considerando género y estado de salud.



	Menos de 10 años	11-20 años	21-30 años	31-40 años	41-50 años	51-60 años	61-70 años	Más de 71 años
CRG6144	0.03	0.0	0.09	1.13	6.54	22.93	17.46	7.01
CRG5424	0.65	2.26	4.39	7.66	13.53	24.1	9.97	2.68
CRG7071	0.0	0.18	0.0	0.91	1.83	16.64	12.98	11.7
CRG5192	0.14	0.24	1.01	3.68	10.75	19.25	9.1	3.74
CRG1000	10.81	8.4	11.54	11.74	6.66	2.88	0.45	0.14

Figura 3.1: Para cada CRG considerado, porcentaje de mujeres por rangos de edad



	Menos de 10 años	11-20 años	21-30 años	31-40 años	41-50 años	51-60 años	61-70 años	Más de 71 años
CRG6144	0.03	0.0	0.16	0.85	2.99	13.84	13.9	13.02
CRG5424	0.92	1.62	2.77	4.02	6.51	11.13	4.71	3.0
CRG7071	0.0	0.0	0.55	0.0	1.65	15.9	12.43	25.23
CRG5192	0.06	0.43	4.3	6.8	7.89	16.24	9.25	6.82
CRG1000	10.55	8.13	9.71	9.9	5.58	2.72	0.63	0.14

Figura 3.2: Para cada CRG considerado, porcentaje de hombres por rangos de edad.

	CRG 1000	CRG 5192	CRG 5424	CRG 6144	CRG 7071
Pacientes en cada grupo	46835	12447	2166	3179	547
Pacientes con al menos un código CIE-9	45043 (96 %)	11903 (96 %)	2101 (97 %)	3090 (97 %)	533 (97 %)
Número posible de códigos CIE-9	1517	1517	1517	1517	1517
Códigos CIE-9 diferentes utilizados	614	588	404	445	387
Códigos Diagnósticos diferentes utilizados	548	541	380	417	360
Códigos V diferentes utilizados	42	47	24	28	27
Códigos E diferentes utilizados	23	0	0	0	0
Códigos M diferentes utilizados	1	0	0	0	0
Códigos CIE-9 diferentes por paciente en cada grupo	2,7	4,84	4,31	6,01	9,52
Valor medio de códigos CIE-9 por paciente del grupo	4,27	12,96	13,28	18,74	32,15
Pacientes con al menos un código ATC	35054	12243	2079	3163	545
Número posible de códigos ATC	746	746	746	746	746
Códigos ATC diferentes utilizados	252	298	227	258	277
Códigos ATC diferentes por paciente en cada grupo	2,12	5,71	5,3	9,11	14,94
Valor medio de códigos ATC por paciente del grupo	2,95	22,01	21,81	50,07	86,44

Tabla 3.3: Análisis descriptivo de los CRG utilizados en este trabajo.

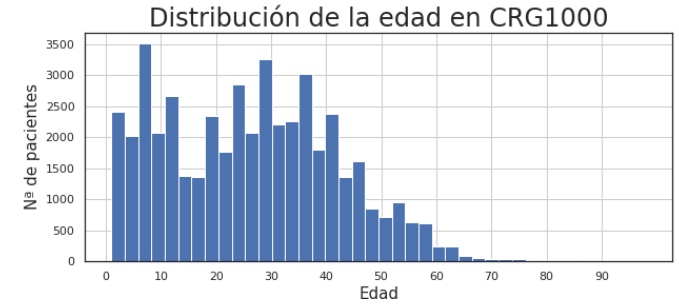
3.2.1. Pacientes sanos

Este grupo poblacional se corresponde con el CRG 1000, al que pertenecen 46835 pacientes, siendo 24649 mujeres. La distribución de edad en base al género de los pacientes sanos se muestra en la Figura 3.3.

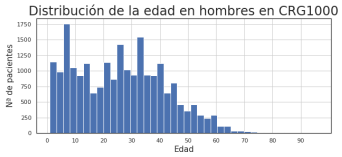
El número medio de códigos CIE-9 diferentes por paciente es 2,90 para mujeres y 3,02 para hombres. Los códigos referentes a enfermedades y lesiones más frecuentes son los problemas dentales [526] (6,60 % hombres, 5,50 % mujeres), y resfriado común [460] (5,01 % hombres y 4,60 % mujeres). Los códigos V más frecuentes son la asistencia anticonceptiva [V25] (59,88 % mujeres y 53,02 % hombres), embarazo [V22] (20,65 % mujeres) y observación / evaluación por sospecha de enfermedades, pero no encontradas [V71] (11,65 % hombres). Cabe destacar que hay presencia de códigos E, que indican la clasificación de los acontecimientos, circunstancias y condiciones ambientales. Los más frecuentes por género son el lugar donde ocurrió la lesión o envenenamiento [E849] (16,00 % mujeres), actividad que se realizaba asociada a la causa externa [E000] (16,32 % hombres) y causas ambientales y accidentales [E928] (15,33 % hombres y 12,00 % mujeres).

El número medio de códigos ATC diferentes por paciente es de 2,16 en mujeres y 1,44 en hombres. Siendo los códigos ATC más frecuentes, por género los antiinflamatorios y antirreumáticos sin esteroides, derivados del ácido propiónico [M01AE] (15,08 % hombres, 12,90 % mujeres), otros analgésicos y antipiréticos: paracetamol [N02BE] (9,66 % hombres 8,87 % mujeres), anticonceptivos hormonales sistémicos [G03AA] (8,05 % mujeres) y otros antihistamínicos para uso sistémico [R06AX] (6,67 % hombres).

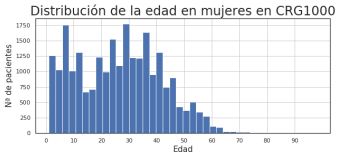
Los perfiles de pacientes sanos (CRG 1000) según las codificaciones CIE-9 y ATC, se muestran en la Figura 3.4.



(a)

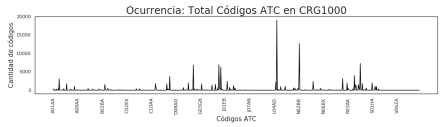


(b)

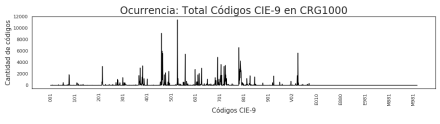


(c)

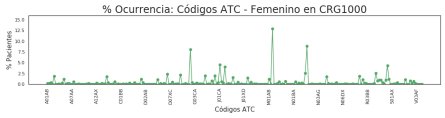
Figura 3.3: Distribución de edad en base al género para el CRG 1000. La gráfica (a) considera ambos géneros. Las gráficas (b) y (c) consideran sólo el género masculino y el femenino, respectivamente.



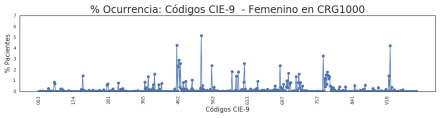
(a)



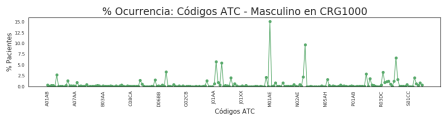
(b)



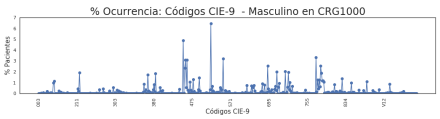
(c)



(d)



(e)



(f)

Figura 3.4: Perfiles de códigos ATC (gráficas (a), (c) y (e)) y CIE-9 (gráficas (b), (d) y (f)) en base al género de los pacientes que componen el CRG 1000. Las gráficas (a) y (b) consideran ambos géneros. Las gráficas (c) y (d) pertenecen al género femenino y las gráficas (e) y (f) al género masculino.

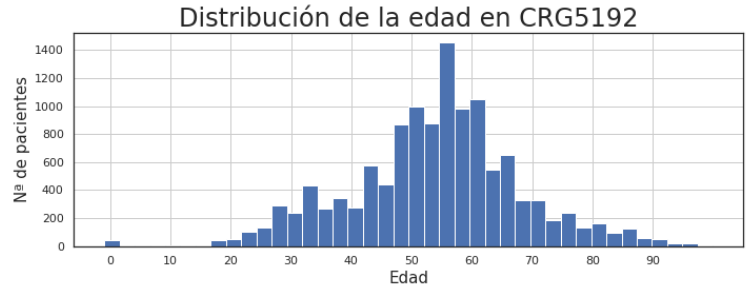
3.2.2. Pacientes hipertensos

Cuando consideramos los pacientes clasificados en el CRG 5192, pacientes con hipertensión, resultan ser un total de 12447 pacientes, de los cuales 6464 son mujeres y 5983 son hombres. La distribución de edad en base al género de los pacientes con hipertensión se muestra en la Figura 3.5.

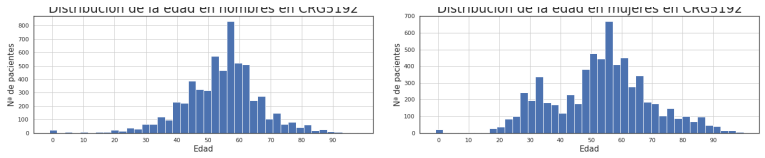
Cada género tiene un número medio de códigos CIE-9 diferentes por paciente, siendo este valor $5{,}57$ en las mujeres y $4{,}03$ los hombres. Los códigos referentes a enfermedades y lesiones más frecuentes por género, son: hipertensión [401] ($36{,}50\,\%$ hombres y $33{,}90\,\%$ mujeres), dolores en articulaciones [272] ($3{,}10\,\%$ hombres y $2{,}90\,\%$ mujeres), trastornos en la descomposición del alimento en azucares y ácidos [719] ($3{,}34\,\%$ hombres) y trastornos relacionados con los dolores de espalda [724] ($3{,}20\,\%$ mujeres). Los códigos V más frecuentes por género, son el embarazo [V22] ($54{,}68\,\%$) y consulta post-parto [V27] ($13{,}93\,\%$) en mujeres; e ingresos/admisión para cuidados posteriores y uso prolongado de medicamentos [V58] ($25{,}61\,\%$) y asistencia anticonceptiva [V25] ($17{,}87\,\%$) en hombres.

El número medio de códigos ATC diferentes por paciente para mujeres es de $6{,}11$ y $7{,}34$ para hombres. Los códigos ATC más frecuentes son los medicamentos para la terapia cardíaca - glucósidos digitálicos [C10AA] ($12{,}49\,\%$ hombres y $8{,}93\,\%$ mujeres), medicamentos para el control de la insuficiencia cardíaca [C09AA] ($13{,}26\,\%$ hombres y $7{,}82\,\%$ mujeres), y medicamentos para reducir la cantidad de ácido gástrico [A02BC] ($7{,}48\,\%$ hombres y $7{,}37\,\%$ mujeres).

Los perfiles de pacientes con hipertensión (CRG5192) según las codificaciones CIE-9 y ATC, se muestran en la Figura 3.6.



(a) [



(b) (c)

Figura 3.5: Distribución de edad en base al género para el CRG 5192. La gráfica (a) considera ambos géneros. Las gráficas (b) y (c) consideran sólo el género masculino y el femenino, respectivamente.

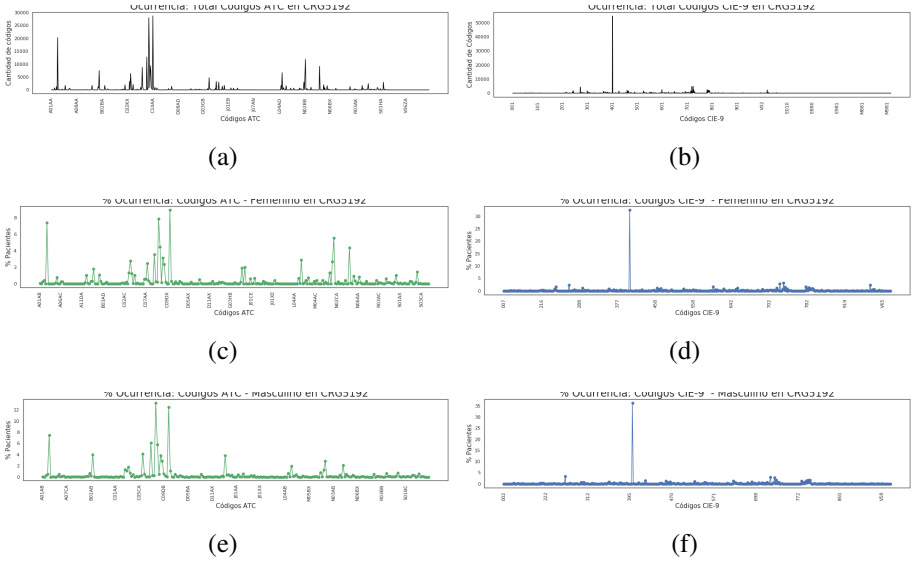


Figura 3.6: Perfiles de códigos ATC (gráficas (a), (c) y (e)) y CIE-9 (gráficas (b), (d) y (f)) en base al género de los pacientes que componen el CRG 5192. Las gráficas (a) y (b) consideran ambos géneros. Las gráficas (c) y (d) pertenecen al género femenino y las gráficas (e) y (f) al género masculino.

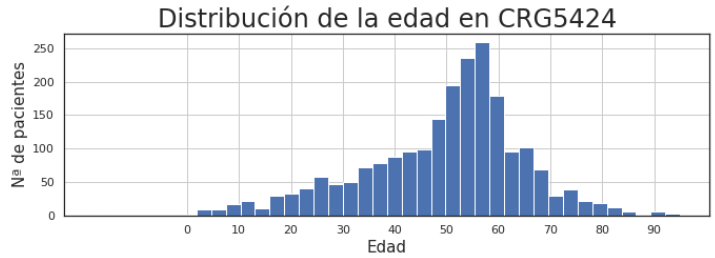
3.2.3. Pacientes diabéticos

El CRG 5424 tiene a un total de 2166 pacientes clasificados con la cronicidad diabetes, de los cuales el 1413 son hombres y el 743 son mujeres. La distribución de edad en base al género de los pacientes con diabetes se muestra en la Figura 3.7.

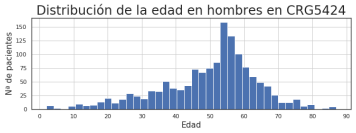
El número medio de códigos CIE-9 diferentes por paciente con diabetes es de 4,97 en las mujeres y 3,95 los hombres. Los códigos referentes a enfermedades y lesiones más frecuentes por género son la diabetes [250] (49,42 % hombres y 41,83 % mujeres), dolores en articulaciones [272] (3,70 % mujeres y 2,65 % hombres) y trastornos en la descomposición del alimento en azucares y ácidos [719] (2,74 % hombres y 2,02 % mujeres). Los códigos V más frecuentes por género son el embarazo [V22] (25,82 % mujeres), consulta/detección en el historial familiar de enfermedades crónicas [V25] (17,17 % hombres), e ingresos/admisión para cuidados posteriores y uso prolongado de medicamentos [V58] (28,12 % hombres y 16,48 % mujeres).

El número medio de códigos ATC diferentes por paciente es de 5,65 en mujeres y 5, 11 en hombres. Los códigos ATC más frecuentes por género son los antidiabeticos orales [A10BA] (17,04 % hombres y 13,89 % mujeres), medicamentos para reducir la cantidad de ácido gástrico [C10AA] (16,32 % hombres y 13,41 % mujeres), insulina por inyección [A10AE] (6,90 % mujeres) y fármacos antitrombóticos sin heparina [B01AC] (7,20 % hombres).

En la Figura 3.8 se muestran los perfiles de pacientes con diabetes (CRG5424) según las codificaciones CIE-9 y ATC.



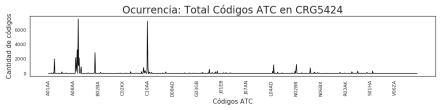
(a)



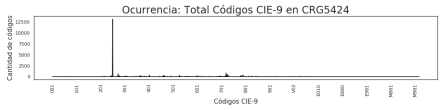
(b)

(c)

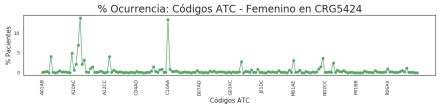
Figura 3.7: Distribución de edad en base al género para el CRG 5424. La gráfica (a) considera ambos géneros. Las gráficas (b) y (c) consideran sólo el género masculino y el femenino, respectivamente.



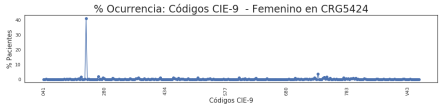
(a)



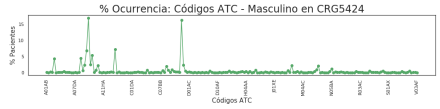
(b)



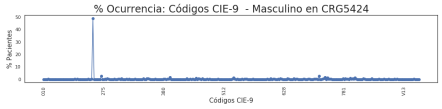
(c)



(d)



(e)



(f)

Figura 3.8: Perfiles de los códigos ATC (izquierda) y CIE-9 (derecha) en base al género para el CRG 5424. Las gráficas (a) y (b) consideran ambos géneros. Las gráficas (c) y (d) pertenecen al género femenino y las gráficas (e) y (f) al género masculino.

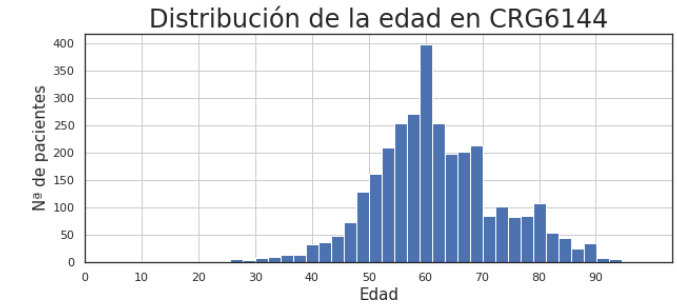
3.2.4. Pacientes diabéticos e hipertensos

Hay 3179 pacientes con hipertensión y diabetes (CRG 6144) de los cuales 1424 son mujeres y 1755 son hombres. La distribución de edad en base al género de los pacientes se muestra en la Figura 3.9.

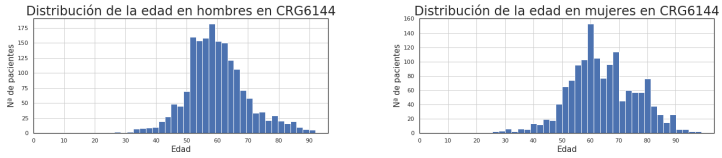
El número medio de códigos CIE-9 diferentes por paciente es de 6,98 en las mujeres y 5,22 los hombres. Los códigos referentes a enfermedades y lesiones más frecuentes son la diabetes [250] (30,32 % hombres y 24,72 % mujeres), hipertensión [401] (16,27 % mujeres y 17,70 % hombres) y dolores en articulaciones [719] (3,15 % mujeres y 2,62 % hombres). Los códigos V más frecuentes son los ingresos/admisión para cuidados posteriores y/o uso prolongado de medicamentos [V58] (46,12 % mujeres y 42,02 % hombres), estado problemático en los miembros superiores e inferiores [V49](14,01 % hombres) y embarazo [V22] (9,48 % mujeres).

El número medio de códigos ATC diferentes por paciente es de 10,42 en mujeres y 8, 01 en hombres. Los códigos ATC más frecuentes son el fármaco que bloquea la enzima necesaria que se produzca colesterol [C10AA] (12,97 % hombres y 10,58 % mujeres), antidiabéticos orales [A10BA] (11,31 % hombres y 8,85 % mujeres), medicamentos para reducir la cantidad de ácido gástrico [A02BC] (7,7 % mujeres) y medicamentos para el control de la insuficiencia cardíaca [C09AA] (8,13 % hombres).

En la Figura 3.10 se pueden observar los perfiles de pacientes con hipertensión y diabetes (CRG6144).



(a)



(b) (c)

Figura 3.9: Distribución de edad en base al género para el CRG 6144. La gráfica (a) considera ambos géneros. Las gráficas (b) y (c) consideran sólo el género masculino y el femenino, respectivamente.

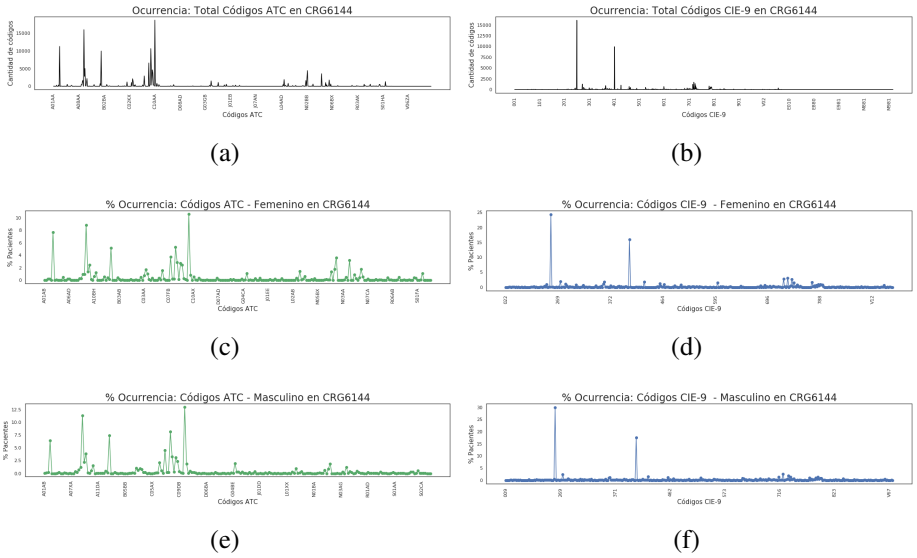


Figura 3.10: Perfiles de los datos de pacientes de los códigos CIE-9 y ATC en CRG 6144. Las gráficas (a) y (b) consideran ambos géneros. Las gráficas (c) y (d) pertenecen al género femenino y las gráficas (e) y (f) al género masculino.

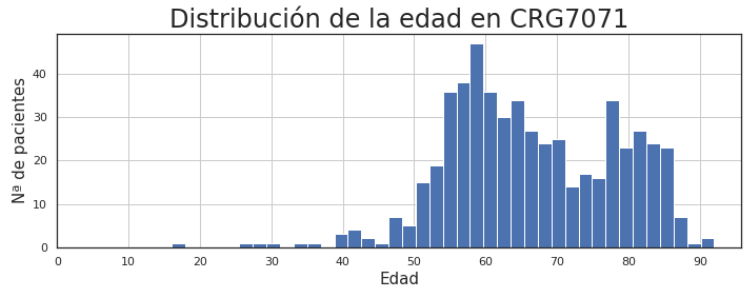
3.2.5. Pacientes diabéticos, hipertensos y otras comorbilidades

La distribución en base a la edad de los 547 pacientes con diabetes, hipertensión y otras enfermedades crónicas (CRG 7071) se muestra en la Figura 3.9. De esos pacientes, 305 son mujeres y 242 son hombres.

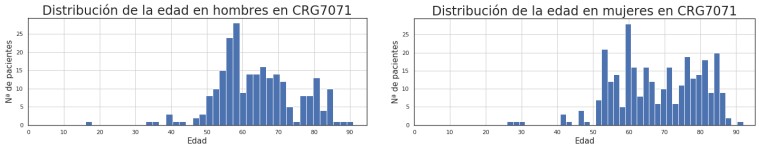
El número medio de códigos CIE-9 diferentes por paciente es de $9{,}93$ en las mujeres y $8{,}98$ los hombres. Los códigos referentes a enfermedades y lesiones más frecuentes por género son la diabetes [250] ($19{,}38\,\%$ hombres y $17{,}52\,\%$ mujeres), la hipertensión [401] ($12{,}23\,\%$ mujeres y $9{,}45\,\%$ hombres), síntomas generales [780] ($2{,}98\,\%$ mujeres), dolores de espalda [724] ($2{,}8\,\%$ mujeres), obstrucción crónica de vías respiratorias [496] ($6{,}13\,\%$ hombres) y enfermedades hepáticas y cirrosis crónicas [571] ($2{,}30\,\%$ hombres). Los códigos V más frecuentes son los ingresos/admisión para cuidados posteriores y/o uso prolongado de medicamentos [V58] ($35{,}38\,\%$ mujeres y $26{,}23\,\%$ hombres), sustitución de órgano o tejido [V43] ($20{,}62\,\%$ mujeres) e historial personal que presenta riesgos para la salud: alergias, cirugía, traumas, lesión o envenenamiento [V15] ($12{,}87\,\%$ hombres).

El número medio de códigos ATC diferentes por paciente es de $15,65$ en mujeres y $14{,}04$ en hombres. Los códigos ATC más frecuentemente dispensados son los medicamentos que reducen la cantidad de ácido gástrico [A02BC] ($7{,}89\,\%$ hombres y $7{,}76\,\%$ mujeres), el fármaco que bloquea la enzima necesaria que se produzca colesterol [C10AA] ($6{,}75\,\%$ hombres y $6,49\,\%$ mujeres), antidiabéticos orales [A10BA] ($5{,}3\,\%$ mujeres y $5{,}01\,\%$ hombres), analgésicos y antipiréticos con ácido salicílico [N02BE] ($4{,}55\,\%$ mujeres) y medicamentos para el control de la insuficiencia cardíaca [C09AA] ($4{,}37\,\%$ hombres).

En la Figura 3.12 se muestran los perfiles de pacientes con hipertensión, diabetes y otra enfermedad crónica (CRG7071), según las codificaciones CIE-9 y ATC.

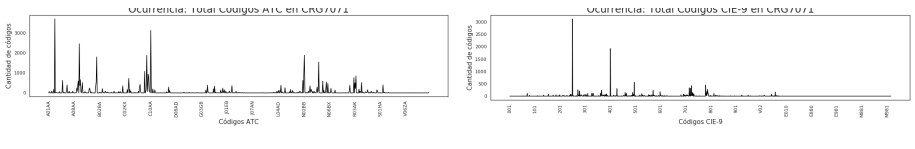


(a)

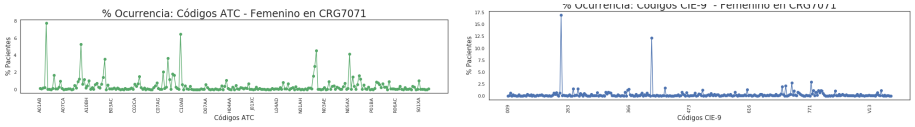


(b) (c)

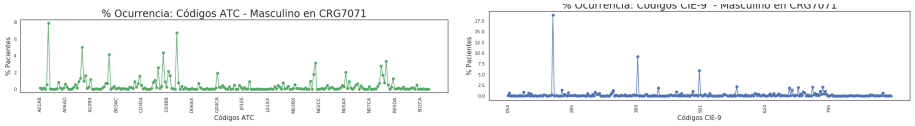
Figura 3.11: Distribución de edad en base al género para el CRG7071. La gráfica (a) considera ambos géneros. Las gráficas (b) y (c) consideran sólo el género masculino y el femenino, respectivamente.



(a) (b)



(c) (d)



(e) (f)

Figura 3.12: Perfiles de los códigos ATC (izquierda) y CIE-9 (derecha) en base al género de los pacientes que componen el CRG 7071. Las gráficas (a) y (b) consideran ambos géneros. Las gráficas (c) y (d) pertenecen al género femenino y las gráficas (e) y (f) al género masculino.

3.2.6. Conclusiones

Teniendo en cuenta el análisis presentado en el principio de la Sección 3.2 y posteriores subsecciones, podemos extraer las siguientes conclusiones:

1. El análisis de la edad por CRG indica que, a medida que aumenta la edad y las enfermedades crónicas, los valores medios de códigos CIE-9 y ATC por paciente, aumentan en la misma media.

2. El análisis por CRG del número medio de códigos CIE-9 diferentes por paciente muestra que las mujeres usan más la atención sanitaria que los hombres. Esta misma tendencia es la seguida al considerar el número medio de códigos ATC diferentes por paciente.

3. Sobre los códigos CIE-9 y ATC que más frecuentemente aparecen para cada CRG, se sostiene que la literatura descrita en las Subsecciones 2.1.2, 2.1.3 y 2.1.4 coincide con el análisis mostrado en los apartados de cada CRG. Por ejemplo, los medicamentos más comunes para tratar la diabetes son los códigos ATC que empiezan por A10, y tras el análisis de los pacientes del CRG5424, se observa que el código ATC A10BA posee un porcentaje para los hombres de un $17{,}04\,\%$ y $13{,}89\,\%$ para las mujeres. Cuando en cualquier CRG aparece la diabetes, el código V con más presencia es embarazo [V22], siendo un indicador de diabetes gestacional.

Capítulo 4

Métodos de aprendizaje automático para clasificación

Este capítulo se compone de tres partes. La primera, Sección 4.1, se definen e introducen los conceptos de aprendizaje automático, tipos de clasificación, validación cruzada, balanceo de observaciones y métodos de selección de características. En la Sección 4.2 se explican detalladamente esquemas de clasificación y su correspondencia en *Python*. Para terminar, Sección 4.3 se hará un repaso de las medidas de prestaciones para los enfoques multi-clase y multi-etiqueta.

4.1. Aprendizaje automático

Definimos el aprendizaje automático como *el algoritmo encargado de extraer/descubrir patrones a partir de una base de datos* [9]. El proceso seguido en el aprendizaje automático se ilustra en la Figura 4.1.

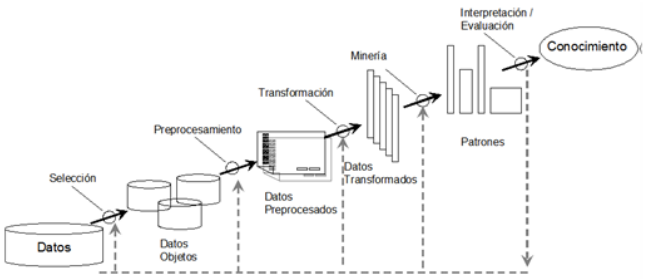


Figura 4.1: *Knowledge Discovery in Databases* (KDD). Tomada de [6]

El aprendizaje automático trata de abordar dos tareas: descripción y predicción. La tarea de descripción de datos o aprendizaje no supervisado, es aquella que trata de obtener información acerca de la estructura de los datos, sin asignarles ninguna etiqueta a los casos. *Clustering* y reglas de asociación son métodos de aprendizaje no supervisado. El *clustering* trata de agrupar los casos de acuerdo a un criterio de proximidad haciendo uso de alguna función de la distancia. Las reglas de asociación tratan de descubrir relaciones o asociaciones entre las distintas características que definen a los casos. El aprendizaje supervisado trata de determinar, para cada caso, una etiqueta, haciendo para ello uso de casos ya etiquetados (e.g. tareas de regresión y clasificación). Tanto la regresión como la clasificación tienen como objetivo aprender una función que represente la correspondencia

que existe entre casos y etiquetas asociadas. La principal diferencia entre las dos es el formato de la salida: en regresión se consideran valores reales, mientras que en clasificación se consideran valores discretos o clases [9]. Abordamos este trabajo fin de grado desde el punto de vista de la clasificación, ya que el objetivo es determinar el estado de salud de los pacientes.

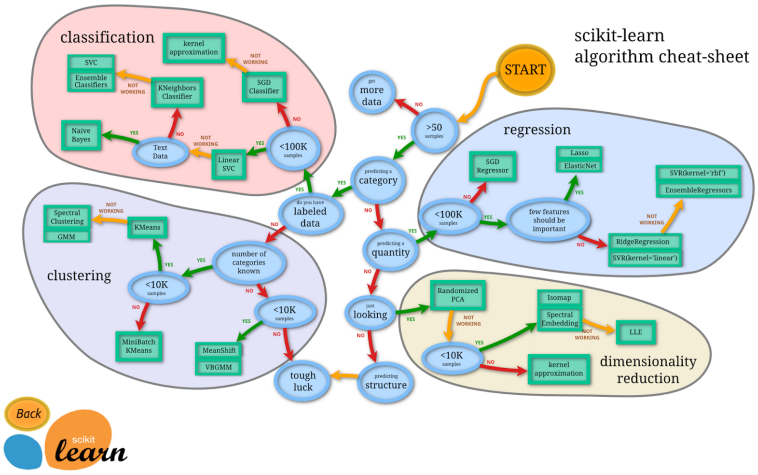


Figura 4.2: Resumen de métodos de aprendizaje automático ofrecidos por el paquete Scikit-learn. Tomada de [7].

El principal objetivo del aprendizaje automático es que el modelo construido tenga capacidad de generalización, es decir, pueda ofrecer una respuesta adecuada a casos nunca vistos [9]. Para ello, el conjunto disponible de casos ya clasificados se suele dividir en dos subconjuntos, uno para entrenar el modelo de clasificación (*train*) y otro para evaluar el modelo (*test*). Existen factores que dificultan el éxito de la generalización, como pueden ser el desbalanceo de clases o la alta dimensionalidad de las características. El desbalanceo de clases consiste en que el número de casos de cada clase no está equilibrado, por lo que unas clases están sobrerrepresentadas respecto a otras [9]. La alta dimensionalidad hace alusión a la presencia de un número muy alto de características en comparación al número de casos que componen el conjunto de *train* [9]. Un problema asociado a la alta dimensionalidad es el sobreajuste, el cual provoca que modelos que presenten errores de entrenamiento bajos dan lugar a errores de *test* altos, es decir, se generaliza mal. La selección de características consiste en seleccionar las características más importantes y/o relevantes para una tarea, con el objetivo de mejorar el rendimiento en la predicción de nuevos casos [9]. Al reducir el número de características hay que tener especial cuidado en los casos duplicados, especialmente al considerar características binarias, ya que al reducir muchas características es posible que un mismo vector de características tenga dos etiquetas diferentes.

Los algoritmos de aprendizaje automático tienen mejor rendimiento cuando tratan con características que están en la misma escala. Una forma de conseguirlo es aplicando un proceso de estandarización, que consiste en centrar cada característica para que tenga media 0 y desviación estándar 1. De esta forma, es mucho más fácil para los algoritmos de clasificación "aprender" los pesos de los parámetros.

4.1.1. Clasificación multi-clase y multi-etiqueta

De forma clásica, el problema de clasificación se ha formulado asumiendo la restricción de que cada caso ha de estar asociado a solamente una clase de las definidas en el problema. En esta formulación clásica se habla de clasificación binaria cuando el número de posibles etiquetas es dos, y de clasificación multi-clase cuando el número de etiquetas es mayor que dos [37]. No obstante, hay muchos problemas donde la restricción de una clase por caso no es la más natural. Entre

estos problemas, merece especial atención el paradigma en el que un caso puede estar asociado simultáneamente a más de una clase, lo que se denomina clasificación multi-etiqueta [37].

4.1.2. Balanceo de las observaciones

Desbalanceo de las observaciones se refiere a *una situación en la que el número de observaciones asociadas a las disitntas clases no es igual* [38]. Los modelos de clasificación descritos en la Sección 4.2 son sensibles a las proporciones de las diferentes clases. Como consecuencia, estos modelos tienden a favorecer el aprendizaje de la clase mayoritaria.

Hay dos estrategias de muestreo que permiten tratar las BBDD con escenarios desbalanceados: submuestreo y sobremuestreo. Ambas estrategias modifican la proporción de casos en cada clase y el tamaño de las base de datos originales. Los métodos de submuestreo eliminan observaciones de la clase mayoritaria con el fin de igualar el número de casos en ambas clases. Por el contrario, los métodos de sobremuestreo "crean" nuevas observaciones de la clase minoritaria. En este trabajo vamos a usar técnicas de submuestreo para balancear las clases.

4.1.3. Selección de características

Dentro del aprendizaje automático, la parte de selección de características tiene como objetivo identificar las características más relevantes y reducir el tiempo computacional asociado al aprendizaje del modelo. Vamos a tratar dos tipos de métodos de selección: filtrado y envoltura.

Los métodos de filtrado o *filter* son procedimientos de selección de características donde se evalúan las características en base a cómo estén correlacionadas con las clases. A este grupo pertenecen los métodos de información mutua o entropía, correlación de Pearson y prueba F de Fisher, entre otros, que ofrece el paquete `sklearn` de *Python*. Suelen ser métodos computacionalmente rápidos, pero no tiene en cuenta la posible relación entre las características [9].

Los métodos de envoltura o *wrapper* combinan la búsqueda de las características más relevantes con la clasificación, evaluando los subconjuntos de características en base al comportamiento en el proceso de clasificación. La mayor desventaja de estos métodos es el problema del sobreajuste, especialmente si el conjunto de *train* es pequeño [9].

Los procedimientos de los métodos de filtrado y de envoltura se ilustran en las Figuras 4.3 y 4.4, respectivamente.

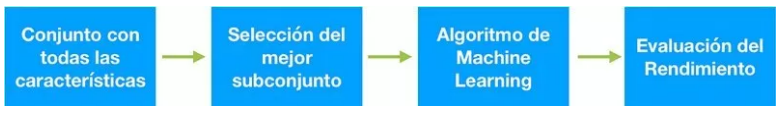


Figura 4.3: Procedimiento de los métodos de filtrado.

Figura 4.4: Procedimiento de los métodos de envoltura.

En este trabajo vamos a usar tres métodos para la selección de características. Los tres métodos se realizarán en un entorno de clases balanceadas, para que la selección no se vea condicionada por la clase mayoritaria.

1. Basado en la frecuencia de las características. Método que selecciona las características más frecuentes. Este método ha obtenido buenos resultados en el trabajo de Ana Alberca [36]. La mayor desventaja es que no tiene en cuenta la relación entre las características y las etiquetas.

2. Basado en la prueba F-Fisher. Método de filtrado. Consiste en el análisis de las varianzas (ANalysis Of VAriance, ANOVA) de cada característica en relación a una etiqueta o clase (prueba F-Fisher). La prueba F-Fisher calcula, para una etiqueta o clase, cómo de diferente es la varianza inter-características (s_x) con respecto a la varianza intra-características (s_w):

$$Prueba\ F = \frac{s_x^2}{s_w^2} = \frac{ns_{\overline{x}}^2}{(s_1^2 + s_2^2 + s_3^2 ... s_k^2)/k} \tag{4.1}$$

   En la Ecuación 4.1, n es el número de observaciones, k es el número de las características. La varianza inter-características o varianza entre características (s_x), mide la variabilidad entre la media de cada característica respecto a la media total de las observaciones asociadas una clase [9]. La varianza intra-características o varianza dentro de las características (s_w), mide la variabilidad de cada observación asociada a una clase respecto a la media de las características [9].

   En *Python*, la función es *f classif* y sólo es válida para el enfoque multi-clase, por lo que para el enfoque multi-etiqueta debemos evaluar cada clase por separado [39] [40].

3. Basado en la clasificación *Random Forest*. Método de envoltura que está explicado con más detalle en la Sección 4.2.2.5. Como la selección se hace en función del rendimiento del clasificador, suele obtener buenos resultados [41].

4.1.4. Prevención del sobreajuste y validación cruzada

Para prevenir el sobreajuste en un modelo de clasificación, debemos buscar los mejores parámetros o *hyperparameter tuning* (HT) que ayuden al modelo a obtener la mejor tasa de acierto. Para dicha búsqueda se suele hacer uso de una estrategia de validación cruzada (*cross-validation*, CV) [9]. El procedimiento de CV consiste en: (1) se divide toda la base de datos en K subconjuntos, (2) se eligen unos valores de parámetros y se entrena el algoritmo de clasificación con $K-1$ subconjuntos (conjunto de *train*) (3) con el subconjunto restante (conjunto validación) se evalúa su rendimiento en predicción con la tasa de acierto (u otra medida de prestación). Este procedimiento se ejecuta K veces. Una vez ejecutado en su totalidad, se realiza la media entre las K tasas de acierto. El procedimiento se repite con todas las combinaciones posibles de parámetros. Gráficamente se puede ver en la Figura 4.5.

Hay dos tipos de CV [42]:

- *K-Fold*. Sigue el procedimiento indicado anteriormente,donde la elección del valor de K tiene que ser representativo de la base de datos, es decir, si la base de datos se compone de 3 clases, el valor de K recomendado sería igual o menor a 3.

- *Leave One Out*. Sigue el procedimiento indicado anteriormente, pero K es el número total de observaciones que se tenga en la base de datos.

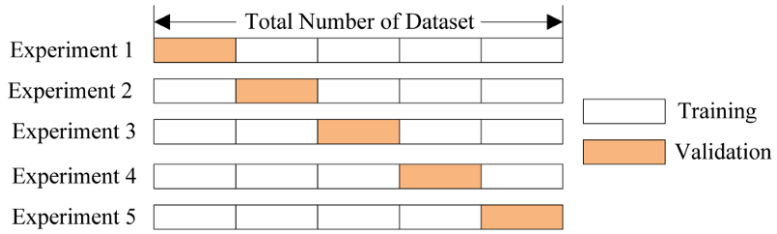


Figura 4.5: Ejemplo de 5 *K-Fold*. Fuente [8].

Python proporciona dos métodos de búsqueda de parámetros libres que hacen uso de CV. A la vez que se busca el mejor parámetro para un modelo, CV permite validar la solidez del modelo frente al sobreajuste.

- Búsqueda en cuadrícula. Viene implementada en *Python* con la función `GridSearchCV` y consiste en la búsqueda de los parámetros que optimizan el modelo en función de una medida estadística de evaluación, normalmente, la tasa de acierto. El mayor beneficio es que garantiza la combinación de parámetros que mejor tasa de acierto ofrece. El inconveniente es que invierte mucho tiempo en la búsqueda y es computacionalmente muy costoso.

```
sklearn.model_selection.GridSearchCV(estimator,
                                      param_grid,
                                      scoring=None,cv)
```

- Búsqueda aleatoria. En *Python* se corresponde con la función `RandomizedSearchCV`, disminuye el tiempo de ejecución y reduce el cómputo a expensas de no probar todas las combinaciones posibles de los valores de los parámetros, sino solo un cierto número de ellos. Por lo tanto, no encontrará los valores de paramteros con la mejor tasa de acierto, sino que encontrará unos valores de parámetros que sean próximos a la mejor tasa de acierto.

```
sklearn.model_selection.RandomizedSearchCV(estimator,
                                           param_grid,
                                           scoring=None,
                                           cv)
```

En este trabajo se ha elegido `GridSearchCV` ya que, aunque se sacrifique tiempo y recursos en su ejecución, sabemos que tendremos mejor combinación de parámetros. Para implementar *K-Fold* hemos elegido $K = 5$, dicho valor sale de que tenemos hasta 5 clases oCRG. (Tabla 3.2).

4.2. Métodos de clasificación

En esta sección se exponen las principales técnicas que se han utilizado en este trabajo. Se ha tratado de resolver los problemas de clasificación mediante dos métodos diferentes: transformación de problemas y adaptación de algoritmos.

4.2.1. Métodos de transformación de problemas

Estos métodos sólo se usan cuando la clasificación es multi-etiqueta. Consisten en la transformación del problema multi-etiqueta a un enfoque multi-clase. La ventaja de este método es que permite utilizar algoritmos de clasificación clásicos, más ampliamente probados y aceptados que los adaptados a multi-etiqueta. La mayor desventaja de estos métodos es la pérdida de información de la correlación entre clases.

A continuación vamos a explicar los utilizados en este trabajo, en base a la Figura 4.6.

Example	Attributes	Label set
1	$\mathbf{x_1}$	$\{\lambda_1, \lambda_4\}$
2	$\mathbf{x_2}$	$\{\lambda_3, \lambda_4\}$
3	$\mathbf{x_3}$	$\{\lambda_1\}$
4	$\mathbf{x_4}$	$\{\lambda_2, \lambda_3, \lambda_4\}$

Figura 4.6: Ejemplo de conjunto multi-etiqueta [9].

4.2.1.1. Relevancia binaria

El método de relevancia binaria o *Binary Relevance* [9] consiste en la generación de n clasificadores binarios, un clasificador por cada etiqueta del conjunto original. Como se ve en la Figura 4.7, se consideran las decisiones positivas de esa etiqueta si en el conjunto original contiene esa etiqueta, considerando la decisión negativa en caso contrario. Para la clasificación de un nuevo dato, evalúo la tasa de acierto para cada modelo de clasificación generado, decido sus clases en función de si la salida es decisión positiva o negativa. Este método tiene como ventaja que es una transformación sencilla, además de ser reversible. La desventaja es que asume que las etiquetas son independientes y al aplicar la trasformación se pierde la información de dependencia entre etiquetas.

Ex.	Label
1	λ_1
2	$\neg\lambda_1$
3	λ_1
4	$\neg\lambda_1$

Ex.	Label
1	$\neg\lambda_2$
2	$\neg\lambda_2$
3	$\neg\lambda_2$
4	λ_2

Ex.	Label
1	$\neg\lambda_3$
2	λ_3
3	$\neg\lambda_3$
4	λ_3

Ex.	Label
1	λ_4
2	λ_4
3	$\neg\lambda_4$
4	λ_4

Figura 4.7: Ejemplo de transformación de relevancia binaria, aplicado a la Figura 4.6.

En *Python* se utiliza la función de uno-contra-todos (*One-Vs-Rest*) [43]. En esta función no hay búsqueda de parámetros libres, ya que dados por el clasificador binario que se elija. En este trabajo hemos probado con los algoritmos de Máquinas de Vectores Soporte, definidos en la Sección 4.2.2.2.

```
sklearn.multiclass.OneVsRestClassifier(estimator)
```

4.2.1.2. Agrupación de etiquetas

El método de agrupación de etiquetas [9] usa la técnica de transformación de los conjuntos multi-etiqueta agrupando las etiquetas asignadas a cada observación. Cuando ya se han detectado todas las posibles combinaciones de etiquetas, se genera una base de datos multi-clase, con tantas clases como combinaciones tenga el conjunto original. Para clasificar un nuevo dato dependerá del

clasificador multi-clase que se use. La mayor ventaja de este método es que es simple y tiene en cuenta la relación entre etiquetas. Pero los problemas que introduce son que para las combinaciones de etiquetas menos frecuentes habrá pocas observaciones asociadas, y sólo tendrá en cuenta las combinaciones presentes en el conjunto de *train*, haciendo imposible predecir nuevas combinaciones de etiquetas.

Ex.	Label
1	$\lambda_{1,4}$
2	$\lambda_{3,4}$
3	λ_{1}
4	$\lambda_{2,3,4}$

Figura 4.8: Ejemplo de transformación de agrupación de etiquetas, aplicado a la Figura 4.6.

En *Python*, la función utilizada es *Label Powerset* [44] y se ejecuta igual que la función *One-Vs-Rest*.

```
skmultilearn.problem_transform.LabelPowerset(classifier=None)
```

4.2.2. Métodos de adaptación de algoritmos

Los métodos de adaptación se refieren a los algoritmos de clasificación que aceptan los enfoques multi-clase y multi-etiqueta sin necesidad de usar una función, como en la Sección 4.2.1. A continuación se hace un breve repaso de los principales algoritmos, indicando las funciones *Python* utilizadas en este trabajo y sus parámetros libres.

4.2.2.1. Regresión logística multinomial

La Regresión Logística Multinominal (RLM) generaliza el método de regresión logística para problemas multi-clase [45]. La RLM aplica a un nuevo dato un modelo de regresión logística y a los valores resultantes se les calcula su probabilidad de aparecer en una clase (*Softmax function*). Se aplica la función de entropía cruzada para calcular la distancia entre los valores y sus probabilidades. La clasificación resultante se mostrará en una matriz binaria de $n \times m$, siendo n el número de clases y m el número de observaciones.

La función en *Python* se corresponde a `LogisticRegression`:

```
sklearn.linear_model.LogisticRegression(C, solver,
                                        multi_class = 'multinominal')
```

- `C`: es el parámetro que define al método de regularización L2 (*Ridge Regression*) que es la penalización por clasificar erróneamente una observación. Cuando `C` es pequeño, el clasificador tiene alto sesgo y baja varianza, es decir, las fronteras de decisión no son muy severas. Y viceversa, cuando C es grande, el clasificador tiene bajo sesgo y alta varianza, es decir, sobreajusta las fronteras de decisión a los datos.

- `solver`: es la función que se usa para optimizar la función de coste con el término de penalización por clasificar erróneamente. Buscaremos entre dos algoritmo de optimización: `newton-cg` y `lbfgs`. Son dos algoritmos que buscan valores mínimos, `newton-cg` los

busca en la matriz de las segundas derivadas parciales (Hessiana), y `lbfgs` en el gradiente de la función.

4.2.2.2. Máquinas de Vectores Soporte

Las máquinas de vectores soporte o SVM) [9], del inglés *Support Vector Machines*, constituyen una técnica de clasificación que busca tantos hiperplanos como clases tengas en el espacio de características, que maximicen la separación entre clases. Los hiperplanos se pueden construir en base a diferentes *kernels* o funciones núcleo, a saber, lineal, polinómico, función de base radial (*Radial basis function*, RBF) o gaussiano.

El algoritmo SVM, en principio se concibió para clasificación binaria, aunque posteriormente se ha adaptado para abordar tareas de clasificación multi-clase. Para aplicar estos algoritmos a los problemas multi-etiqueta debemos utilizar los métodos propuestos en el apartado anterior (Sección 4.2.1).

Este trabajo se basa en el modelo propuesto por Wan et al. [46], que aborda la localización subcelular de proteínas multi-etiqueta con un algoritmo que utiliza conjuntamente las SVM con un enfoque *One-Vs-Rest*. Los autores hacen que cada SVM de *One-Vs-Rest*, uno por cada etiqueta, tenga un umbral adaptativo. Ese umbral es el valor máximo de la función que prevé la localización subcelular, definida en su trabajo.

En *Python* que se corresponde a la función `SVC`.

```
sklearn.svm.SVC(C, kernel,gamma,
                class_weight=None,
                decision_function_shape='ovr')
```

En este trabajo vamos a considerar para el parámetro `kernel` lineal (Ecuación 4.2) y RBF (Ecuación 4.3).

$$K(X) = W * X + c \tag{4.2}$$

La Ecuación 4.2 representa el hiperplano en el espacio de características y donde X es el conjunto de *train* y W son los pesos que vienen determinados por el hiperplano [47].

$$K(x, x') = exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right) \tag{4.3}$$

En la Ecuación 4.3 donde $||x - x'||^2$ es la distancia Euclídea al cuadrado entre dos observaciones x y x' [48].

Los parámetros con los que podemos personalizar la función de *Python* son los siguientes:

- `C`: es la penalización por clasificar erróneamente una observación. Definida anteriormente en RLM.
- `gamma`: es un parámetro de propagación, es decir, cuando el valor de `gamma` es cercano a 0, la rontera de decisión entre clases es muy baja y, por lo tanto, la región de decisión es muy amplia. Cuando `gamma` es un valor alto, las curvas de decisión se sobreajustan a los datos. Sólo se usa para el *kernel* RBF.

Otros parámetros que utiliza la función son `decision_function_shape` indica cómo se va a resolver el problema de multi-clase; `class_weight` por defecto asume que las clases están balanceadas, que es nuestro caso.

4.2.2.3. *k Nearest Neighbour*

El algoritmo de clasificación de los *k* vecinos más cercanos (distancia euclídea) por votación (*k Nearest Neighbour*, *k*NN) [9], es uno de los modelos más ampliamente utilizados en clasificación. Permite determinar si un caso pertenece a una clase eligiendo los *k* casos más cercanos del conjunto de *train* por votación y examinando la clase a la que pertenecen.

En [49] Zhang y Zhou proponen una modificación del algoritmo *k*NN, denominada ML-*k*NN, que adapta este modelo a la utilización de clasificación multi-etiqueta. Para un nuevo dato se buscan los *k* vecinos más próximos, por distancia euclídea, y se crea un vector en el que se contabiliza la clase de los *k* vecinos. Se elije la clasificación en base a las clases mayoritarias. En caso de empate se revolverá mirando el orden de la matriz de distancias euclídeas de los *k* vecinos más próximos.

La función en *Python* se corresponde a `KNeighborsClassifier`:

```
sklearn.neighbors.KNeighborsClassifier(n_neighbors,
                                       weights, p=2,
                                       metric='minkowski')
```

El parámetro que utilizamos para generalizar el modelo es `k` que indica el número de vecinos más próximos para decidir por votación la etiqueta de una observación desconocida. Para un valor bajo de `k`, la clasificación es muy sensible al ruido y a las características irrelevantes. Al contrario, si `k` es mayor se consideran más vecinos y el ruido pierde influencia.

Otros parámetros de esta función son `weights`, que indica el peso de cada observación en la elección de una nueva etiqueta: `uniform` (por votación) y `distance` (inverso de la distancia); `metric` y `p` indican la medida de distancia elegida, por defecto es la distancia Minkowski (Ecuación 4.4), que es la generalización de las distancias Manhattan (`p=1`) y Euclídea (`p=2`) para dos puntos X : $(x_1, x_2, ..., x_n)$ y Y : $(y_1, y_2, ..., y_n)$.

$$D(X,Y) = \left(\sum_{i=1}^{k} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{4.4}$$

En la Ecuación 4.4

4.2.2.4. Árboles de Decisión

Los árboles de decisión (*Decision Tree*, DT) [9] es una estructura muy utilizada en clasificación, ya que dada una base de datos, construye reglas lógicas que sirven para categorizar los casos.

Los árboles están formados por nodos internos, nodos terminal y ramas. Los nodos internos contienen la regla lógica sobre el valor de las características, los nodos terminal indican la salida propiamente dicha del clasificador, y las ramas comunican los nodos entre sí.

El trabajo de Clare y King [50] presenta una adaptación del algoritmo C4.5 que modifica la forma de construir el árbol para que éste permita la utilización de datos multi-etiqueta. En concreto, se

presenta una definición de la entropía multi-etiqueta, que permite distingue qué atributos se utilizarán en la construcción del árbol, además de permitir que haya múltiples etiquetas en los nodos internos y terminales.

La función `DecisionTreeClassifier` es la encargada del algoritmo C4.5 en *Python*:

```
sklearn.tree.DecisionTreeClassifier(criterion='entropy',
                                    max_depth,
                                    min_samples_split,
                                    class_weight=None)
```

Los parámetros con los que vamos a definir nuestros árboles son:

- `max_depth`. Máxima profundidad del árbol, es decir, cuanto más profundo es el árbol, más divisiones tiene y captura información y ruido, es decir, sobreajusta.
- `min_samples_split`. Mínimo número de observaciones que se requieren en un nodo interno para ser considerado para ramificación. Valores más altos previenen que el modelo aprenda relaciones muy específicas.

Otros parámetros interesantes son el `criterion`, que es el criterio que toma el árbol como referencia para ramificarse, en nuestro caso se ha elegido `'entropy'`; y `class_weight`, definido anteriormente.

Los DT en *Python* se representan gráficamente con el paquete `Graphviz`, pero por extensión no vamos a representar los árboles resultantes sobre nuestra base de datos. Además, *Python* no soporta los procesos de poda y ha de hacerse manualmente. La poda de un árbol de decisón consite en eliminar los nodos internos de un árbol y convirtiendolos en hojas.

4.2.2.5. Multiclasificadores: Random Forests

Los métodos multiclasificadores (*Embedded Methods*, EM) [9] son técnicas en las que se combinan las respuestas de varios clasificadores que presentan una exactitud (*recall*) no muy elevada, para formar uno más fuerte.

El método que vamos a usar en este trabajo es *bagging*. Sirve para reducir la varianza de las predicciones a través de la combinación de los resultados de varios clasificadores, cada uno de ellos diseñado con diferentes subconjuntos tomados de la misma base de datos.

A este método pertenecen los *Random Forests* [22]. Es la combinación de n Árboles de decisión.Se construye de la siguiente manera:

– Se selecciona, del conjunto de *train*, aleatoriamente con reemplazamiento (*Bootstrap*) dos subconjuntos, uno para entrenar el árbol y otro servirá para validarlo (*Out of bag*, OOB).

– En cada nodo, al seleccionar la partición óptima, tenemos en cuenta sólo una porción de las características, elegidas al azar en cada ocasión.

– Cada árbol creado se evalúa de forma independiente, y la predicción del bosque será la predicción media de los N árboles. La proporción de árboles que toman una misma respuesta se interpreta como la probabilidad de la misma.

La parte más notable que introducen los *Random Forest* es la importancia de las características de la base de datos. Para cada árbol del bosque se establece una tasa de acierto de referencia usando todas las características. En una segunda vuelta ,se quita una característica al azar y se vuelve a calcular la tasa de acierto. La importancia de la característica es la diferencia entre la tasa de referencia y la de la base de datos modificada. Este proceso se repetirá tantas veces como características tengamos. Es independiente de la optimización de los parámetros libres.

En *Python* usaremos la función `RandomForestClassifier`:

```
sklearn.ensemble.RandomForestClassifier(n_estimators,
                                        criterion='entropy',
                                        max_depth,
                                        min_samples_split,
                                        class_weight=None)
```

Los parámetros que debemos modificar para mejorar la capacidad predictiva son los mismos que en la función `DecisionTreeClassifier` (Sección 4.2.2.4). La novedad es el parámetro `n_estimators`, que indica el número de árboles en el bosque. A medida que aumenta el número de árboles, la tasa de acierto aumenta hasta cierto punto, pero no hay mucho beneficio en aumentar el número de árboles más allá del punto óptimo, sólo se ve afectado el tiempo de entrenamiento.

4.2.2.6. *Multi-Layer Perceptron*

El perceptrón multicapa (*Multi-Layer Perceptron*, MLP) [9] es una red neuronal artificial, formada por múltiples niveles de neuronas que permiten resolver problemas de forma no linear. Se suele entrenar por medio de un algoritmo de retropropagación de errores (*BackPropagation*, BP) [9]. A partir de un conjunto de *train* etiquetado, se determina la ecuación de un hiperplano que separa las diferentes clases del problema pasando por capas de neuronas de procesamiento. Siendo el objetivo final minimizar una error cometido por la red al clasificar. Al utilizar la función introducida por *Python*, el error cometido se define con la función entropía cruzada o *log-loss* (Ecuación 4.5). Esta ecuación compara la etiqueta estimada con la de test correspondiente. Si el valor de la entropía cruzada es cercano a 0, la predicción es casi perfecta.

$$L_{log}(Y,P) = -log\ \rho(Y|P) = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k} log\ p_{i,k} \tag{4.5}$$

donde $Y : (y_1, y_2, ..., y_n)$ representa a las etiquetas del conjunto de test, $P : (p_1, p_2, ..., p_n)$ es el conjunto de etiquetas estimadas por el clasificador, N es el número de etiquetas del conjunto y K los posibles valores de dichas etiquetas.

En el artículo [51], Zhang y Zhou, utilizan un MLP con un conjunto multi-etiqueta con múltiples valores en la salida.

En *Python* la función correspondiente es `MLPClassifier`:

```
sklearn.neural_network.MLPClassifier(hidden_layer_sizes,
                                     activation,
                                     solver)
```

Los parámetros libres de dicha función son:

- `hidden_layer_sizes`. Número de neuronas por capa oculta.
- `activation`. Función que hay dentro de las neuronas que activa el tipo de etiquetado para las neuronas de la capa oculta. Se muestran en la Figura 4.9.

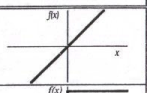
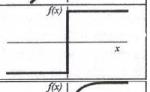
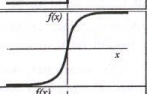
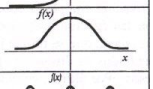
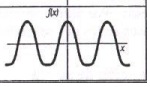
	Función	Rango	Gráfica
Identidad	$y = x$	$[-\infty, +\infty]$	
Escalón	$y = sign(x)$ $y = H(x)$	$\{-1, +1\}$ $\{0, +1\}$	
Sigmoidea	$y = \frac{1}{1+e^{-x}}$ $y = tgh(x)$	$[0, +1]$ $[-1, +1]$	
Gaussiana	$y = Ae^{-Bx^2}$	$[0,+1]$	
Sinusoidal	$y = A\,\mathrm{sen}(\omega x + \varphi)$	$[-1,+1]$	

Figura 4.9: Funciones de activación. Fuente [10].

- `solver`. Función que se usa para optimizar la función de coste. Buscaremos entre dos algoritmo de optimización:
  - descenso de gradiente estocástico (*Stochastic Gradient Descent*, SGD): es la aproximación estocástica del método de gradiente descendiente.
  - Estimación del momento adaptativo (*Adaptive Moment Estimation*, ADAM): es una optimización del SGD que utiliza el gradiente de primer orden de funciones estocásticas [52]. Este método es eficiente computacionalmente hablando, necesita poca capacidad de memoria y es muy bueno resolviendo problemas con gran cantidad de características y etiquetas.

4.2.3. Conclusión

Los motivos para elegir los modelos descritos han sido los siguientes:

- **RLM**. Se ha elegido por ser un método simple y validado para multi-clase, aunque sólo esté adaptado para la clasificación multi-etiqueta.
- **SVM**. Tiene una buena generalización con nuevos datos cuando el modelo está bien parametrizado. Para el enfoque multi-etiqueta debemos usar las funciones descritas en la Sección 4.2.1.
- ***k*NN** es un método sencillo, pero sensible a la presencia de atributos irrelevantes. Por ello, resulta de interés para evaluar subconjuntos de atributos calculados con las técnicas de selección presentadas en la Sección 4.1.3. La mayor desventaja con respecto a los otros métodos es que no realiza la optimización de una función en entrenamiento, sólo es posible considerar varios valores de `k` para mejorar la tasa de acierto.
- **DT** es una técnica muy potente, basada en una estructura de árbol y que hace uso de la entropía. Presenta tolerancia a la presencia de atributos irrelevantes y/o redundantes. Al contrario que el resto de modelos, este algoritmo sí se puede considerar para problemas multi-etiqueta, pero por extensión en este trabajo no vamos a profundizar.

- ***Random Forest*** aporta las ventajas de los árboles de decisión, unidas a una mayor robusted al ruido.
- **MLP** es considerado un aproximador universal, siendo capaz de aprovechar relaciones no-lineales entre los atributos.

4.3. Evaluación de prestaciones

En esta sección se va a definir los tipos de evaluación para la clasificación multi-etiqueta: basados en etiquetas y basados en observaciones [9]. Además, se va a mostrar sus correspondientes funciones en *Python*.

4.3.1. Medidas de prestaciones basadas en etiquetas

Las medidas de prestaciones basadas en etiquetas [9] están definidas para la clasificación clásica y se calculan para cada etiqueta distinta, siendo posteriormente promediadas para obtener un valor único. Todas las medidas de prestaciones definidas en esta sección toman como base la matriz de confusión (véase la Figura 4.1).

		Predicho	
		Si	No
Real	Si	tp	fn
	No	fp	tn

Tabla 4.1: Matriz de confusión.

- tp indica las observaciones positivos clasificados correctamente.
- fp indica las observaciones incorrectamente clasificados como positivos o falsos positivos.
- tn indica las observaciones negativos correctamente asignados.
- fn indica las observacioness no asignados o falsos negativos.

Para usar estas medidas de prestaciones en la clasificación multi-clase y multi-etiqueta, se proponen dos enfoques, *macro* y *micro* [9]. El planteamiento *macro* calcula cualquier métrica (M) para cada etqueta (n) y hace la media por etiquetas (Ecuación 4.6). En el enfoque *micro*, primero ha de calcularse la matriz de confusión para cada n etiqueta, posteriormente se calcula la métrica M con esos valores (Ecuación 4.7).

$$M_{macro} = \frac{1}{n}\sum_{i=1}^{n} M(tp_i, fp_i, tn_i, fn_i) \tag{4.6}$$

$$M_{micro} = M(\sum_{i=1}^{n} tp_i, \sum_{i=1}^{n} fp_i, \sum_{i=1}^{n} tn_i, \sum_{i=1}^{n} fn_i)) \tag{4.7}$$

La métrica más usada es *accuracy* o acierto, se define como *el cociente del número de etiquetas correctamente clasificadas* (Ecuación 4.8).

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{4.8}$$

La *precision* o precisión de un clasificador se define como *el cociente del número de etiquetas correctamente clasificadas entre todas las etiquetas asignadas por el clasificador* [9] (Ecuación 4.9).

$$precision = \frac{tp}{tp + fp} \tag{4.9}$$

El *recall* o exhaustividad de un clasificador es *el cociente del número de etiquetas correctamente clasificadas entre todas las etiquetas realmente correctas* [9] (Ecuación 4.10)

$$recall = \frac{tp}{tp + fn} \tag{4.10}$$

Uno de los objetivos del aprendizaje automático es maximizar *precision* y *recall*, para ello se utiliza la métrica *F-score* o *F1-score*. Es *la media armónica entre precision y recall* (Ecuación 4.11)

$$f - score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4.11}$$

Para la clasificación multi-clase, en *Python* vamos a implementar las ecuaciones antes descritas. La función de tasa de acierto, en *Python*, soporta multi-clase y multi-etiqueta (se explicará más adelante):

```
accuracy_score(y_true, y_pred)
```

Las funciones en *Python* para clasificación multi-etiqueta son:

```
precision_score(y_true, y_pred,
                average=['macro', 'micro'])
recall_score(y_true, y_pred,
             average=['macro', 'micro'])
f1_score(y_true, y_pred,
         average=['macro', 'micro'])
```

La correspondencia de los parámetros es la siguiente:

- `y_true` es el conjunto de etiquetas reales Y.
- `y_pred` es el conjunto de etiquetas estimadas P.
- `average` la opción de elegir una métrica bajo el enfoque *macro* o *micro*.

4.3.2. Medidas de prestaciones basadas en observaciones

Las medidas de prestaciones basadas en observaciones [9] están específicamente concebidas para clasificación multi-etiqueta, y por esta razón se calculan para cada observación teniendo en cuenta

tanto la etiqueta real (etiqueta perteneciente al conjunto *test*) como la etiqueta estimada (etiqueta estimada de aplicar un algoritmo clasificador sobre una observación).

Para la definición de estas medidas de prestaciones consideraremos un conjunto de *train* (x_i, Y_i), $i = 1 \ldots m$, donde $Y_i \subseteq L$, siendo L el conjunto de etiquetas $L = \{\lambda_j : j = 1 \ldots n\}$ y n indica el número máximo de etiquetas que puede tener asociadas un caso. La función clasificador se puede denominar como: $H : X \rightarrow P \subseteq L$ que, para cada caso i, realice una medida estadística que minimice la diferencia entre el conjunto de etiquetas estimadas P_i y el conjunto de etiquetas reales Y_i.

La medida más usada para evaluar clasificadores multi-etiqueta es la distancia Hamming (*Hamming loss*) que evalúa la diferencia simétrica entre los conjuntos de etiquetas estimadas y reales, y los promedia (Ecuación 4.12). Al considerar tanto los errores de clasificación (el hecho de que una etiqueta incorrecta sea estimada) como los errores por omisión (cuando una etiqueta que debería estar presente en el clasificador no lo está) estará acotada entre 0 y 1, siendo 0 su mejor resultado y 1 su peor.

$$Hamming\ loss = \frac{1}{n}\sum_{i=1}^{n}\frac{|P_i \triangle Y_i|}{n} \tag{4.12}$$

La *accuracy* se ha adaptado para que pueda abarcar la clasificación multi-etiqueta . Se define como el promedio del cociente de aciertos del clasificador frente a la unión de etiquetas reales y estimadas (Ecuación 4.13).

$$accuracy = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i \bigcap P_i}{Y_i \bigcup P_i} \tag{4.13}$$

Para las medidas de prestaciones descritas en este apartado, las funciones correspondientes en *Python* son:

```
hamming_loss(y_true, y_pred)

accuracy_score(y_true, y_pred)
```

Siendo `y_true` e `y_pred` los conjuntos de etiquetas Y y P, respectivamente.

Capítulo 5

Experimentos y resultados

En este capítulo se empieza exponiendo el entorno de ejecución (Sección 5.1) en el que se ha realizado el procedimiento seguido para el ajuste de los modelos, definidos en la Sección 4.2.2, a las distintas BBDD de pacientes, para su posterior evaluación. Los primeros pasos son realizar el balanceo de las bases de datos y la división de las BBDD en los conjuntos de *train* y *test*(Sección 5.2). Acontinuación se realiza la selección de características más relevantes según los criterios de: frecuencia, prueba F-Fisher y el algoritmo *Random Forest* (Sección 5.3) . Para finalizar, se exponen los resultados obtenidos para los enfoques multi-clase y multi-etiqueta (Sección 5.4).

5.1. Entorno de ejecución

Uno de las dificultades encontradas en este trabajo ha sido en la ejecución de algoritmos computacionalmente costosos en portátiles de uso personal. Para ello se ha trasladado toda la carga de cómputo a los ordenadores reservados para alumnos de la Escuela Técnica Superior de Ingeniería de Telecomunicación (ETSIT) de la Universidad Rey Juan Carlos (URJC). Gracias al administrador de los laboratorios, Antonio Gutiérrez Mayoral, se ha podido disponer de ejecuciones ininterrumpidas y configuraciones especiales.

El *hardware* que se ha utilizado en este trabajo corresponde al de los ordenadores de los Laboratorios 3 de la URJC del Campus de Fuenlabrada:

- *Central Processing Unit* (CPU): Intel Core i5-8500 3.00GHz.
- *Random Access Memory* (RAM): 8Gb DDR3.
- Sistema Operativo: Ubuntu Linux 18.04.3, Version bionic.

El *software* utilizado al principio de este trabajo fue *Jupyter Notebook* [53]. Es un entorno de trabajo interactivo que permite programar el código en *Python*, actualmente en la *release* 3 [54]. La manera de desarrollar es dinámica, ya que se puede integrar en un mismo documento tantos bloques de código como texto, gráficas o imágenes, haciendo que su compresión, a ojos externos al programador, sea más sencilla.

Una vez que se empezó a entrenar los algoritmos se notó un gran incremento en tiempos de ejecución. Para ejecutar *Jupyter Notebook* se necesita un navegador como por ejemplo *FireFox*, esto supone que la CPU esté compartida entre el navegador y *Jupyter Notebook*. Encontrada la causa, se procedió

a trasladar el código a un archivo ejecutable de *Python*. Con ese archivo ejecutable el tiempo de ejecución se redujo, ya que la CPU se dedica exclusivamente a ejecutar el código. Para que un archivo de *Python* se ejecutase sin interrupciones, recurrimos al comando `screen` [55]. Este comando permite abrir múltiples instancias de terminal dentro de una sesión de usuario, que se ejecutan en segundo plano. Esto quiere decir que si salimos de la sesión y cerramos el terminal, el proceso ejecutándose en las terminales `screen` no se interrumpirá.

5.2. Creación de las BBDD, balanceo y conjuntos de *train* y *test*

Para crear las BBDD siguiendo las definiciones de clasificación multi-clase y multi-etiqueta explicadas en la Sección 4.1.1, los escenarios a estudiar serán distintos en función de los CRG:

1. 5 CRGs: CRG 1000 (pacientes sanos), CRG 5192 (pacientes hipertensos), CRG 5424 (pacientes diabéticos), CRG 6144 (pacientes hipertensos y diabéticos), CRG 7071 (pacientes hipertensos, diabéticos y comorbilidades). De aquí en adelante nos referiremos a este escenario como base de datosde pacientes sanos y crónicos.

2. 4 CRGs que incluyen sólo pacientes crónicos. De aquí en adelante nos referiremos a este escenario como base de datos de pacientes crónicos.

Estos escenarios de análisis se escogen debido a que con ellos se produce un cambio de criterios entre los pacientes sanos y crónicos.

Para el caso multi-clase no hace falta idear una solución de clasificación, ya que se dispone de los pacientes ya clasificados en base a su CRG. Se tradujo los CRG a valores nominales para que los algoritmos puedan ejecutarse. En la Tabla 5.1 se muestran los valores asociados a las clases para cada CRG. El caso multi-etiqueta se resolvió mediante las enfermedades crónicas que componen a los CRG, explicada en el Capítulo 3. Utilizamos una clasificación binaria en base a la presencia de las enfermedades crónicas. La clasificación multi-etiqueta se muestra en la Tabla 5.2.

Multi-clase	
CRG	Clase
CRG1000	0
CRG5192	1
CRG5424	2
CRG6144	3
CRG7071	4

Tabla 5.1: Clasificación multi-clase de los CRG estudiados.

Multi-clase			
CRG	Hipertensión	Diabetes	Comorbilidades
CRG1000	0	0	0
CRG5192	1	0	0
CRG5424	0	1	0
CRG6144	1	1	0
CRG7071	1	1	1

Tabla 5.2: Clasificación multi-etiqueta de los CRG estudiados.

Una vez creadas las BBDD etiquetadas, debemos balancearlas. Para solucionarlo usamos la técnica de submuestreo. Limitamos el número de pacientes de cada clase en base a el CRG con menor número de pacientes, es decir, la clase minoritaria. En nuestro caso, la clase minoritaria corresponde al CRG 7071 con 547 pacientes. En la base de datos de pacientes sanos y crónicos consideraremos 2735 pacientes, y en las BBDD de pacientes crónicos 2188. Con ese submuestreo, disponemos de varios

subconjuntos de los CRG no minoritarios. Construiremos 50 subconjuntos para pacientes sanos y crónicos y 50 subconjuntos para pacientes crónicos.

Como usamos las técnicas de aprendizaje supervisado, dividiremos nuestros subconjuntos balanceados en dos conjuntos: *train* y *test*. Para ello usaremos de cada subconjuntos un $80\,\%$ para *train* y un $20\,\%$ para *test*.

5.3. Selección de características

Para este trabajo nos hemos basado en los mejores resultados proporcionados por el trabajo de Ana Alberca Díaz-Plaza [36]: 100 mejores características del conjunto formado por ATC y CIE-9, añadiendo las características demográficas. Con este escenario arriesgamos a que las características seleccionadas de ATC y CIE-9 no estén balanceadas. En resumen, vamos a usar 102 características para entrenar los distintos algoritmos de clasificación. Para una mejor selección, hemos ejecutado cada método con 50 subconjuntos de pacientes sanos y crónicos, y 50 subconjuntos de pacientes crónicos. Las características clínicas las seleccionaremos considerando la presencia (características binarias) y basadas en la ocurrencia.

5.3.1. Basada en frecuencia.

Para cada subconjunto balanceado se calcula la frecuencia de cada característica. Estos valores se ordenan de mayor a menor y se seleccionan los 100 primeros, pues en principio son los que más información aportan al problema. Este tipo de selección no depende de la posible relación entre las características y las clases o etiquetas. No condiciona que el tipo de clasificación sea multi-clase o multi-etiqueta. Vamos a tener 4 listados diferentes dependiendo de las BBDD consideradas y el tipo de características clínicas usadas, binarias y basadas en la ocurrencia. En el Anexo B se pueden ver los diferentes listados obtenidos (Tabla B1) y las diferencias entre listados (Tabla B6). En la Figura 5.1, se comparan las 25 características de los dos escenarios de BBDD propuestos. Para los subconjuntos de pacientes sanos y crónicos salen mayores tasas de frecuencia en las características con respecto a los subconjuntos de pacientes crónicos. Esto se debe al número de pacientes de cada subconjunto: en los subconjuntos de pacientes sanos y crónicos, son 2735 pacientes y en los subconjuntos de pacientes crónicos 2188.

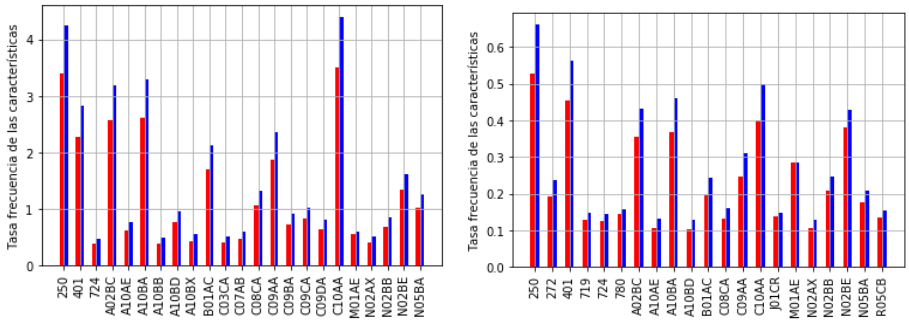
5.3.2. Basada en la prueba F-Fisher

Con este método se evalúa la relación entre varianzas de las características con respecto a las clases/etiquetas. Las características seleccionadas son las 100 características que tienen la correlación más alta con respecto a las clases/etiquetas consideradas en cada tipo de clasificación. Por tanto, salen 8 listados de características. Tenemos 4 listados más que cuando seleccionamos en base a la frecuencia, debido a que debemos de tener en cuenta el tipo de clasificación. Las Tablas B2 y B3 del Anexo B corresponden a la clasificación multi-clase y multi-etiqueta, respectivamente. Comparando los listados para los dos escenarios de BBDD propuestos, se obtienen los códigos CIE-9 y ATC que se muestran en las Tablas B7 y B8. Los códigos diferentes son mayores para este método que para el resto de métodos, por tanto, podemos asumir que el análisis de varianzas se adapta mejor a los dos escenarios con respecto a las BBDD de pacientes. En la Figura 5.2 se observan 25 características de

las 100 características seleccionadas para cada tipo de clasificación y para las características clínicas binarias o basadas en la ocurrencia.

5.3.3. Basada en la importancia de las características de *Random Forest*

Para cada subconjunto balanceado se entrena el algoritmo *Random Forest* y se almacena en un vector el valor de importancia que cada característica aporta en los árboles. Tras ejecutar los 50 subconjuntos, estos valores para cada característica se suman, se ordenan de mayor a menor y se seleccionan las 100 primeras características. Los listados que se obtienen son 8 en base al tipo de clasificación y la configuración de clas caracterisitcas clínicas binarias u ocurrencia. Las Tablas B4 y B5 indican esos listados. Comparando las diferencias entre los listados anteriores se obtienen las tablas B9 y B10. Podemos observar que el número de características diferentes no es tan alto como las seleccionadas en base a la prueba F-Fisher, pero aún así son bastantes para esperar una buena adaptación a los distintos escenarios de este trabajo, En la Figura 5.3 se observa la misma tendencia que en la selección basada en la frecuencia: mejores prestaciones en las BBDD de pacientes sanos y crónicos, tanto para características clínicas binarias como para las basadas en ocurrencia .



(a) (b)

Figura 5.1: Muestra de las primeras 25 características seleccionadas en base a la frecuencia. La gráfica (a) evalúa las características clínicas basadas en ocurrencia. La gráfica (b) se evalúa con las características clínicas binarias. En las gráficas las barras azules simbolizan los valores para los subconjuntos pacientes sanos y crónicos y las barras rojas para los sobconjuntos pacientes crónicos. En este tipo de selección de características no hace falta diferenciar los escenarios de multi-clase y multi-etiqueta.

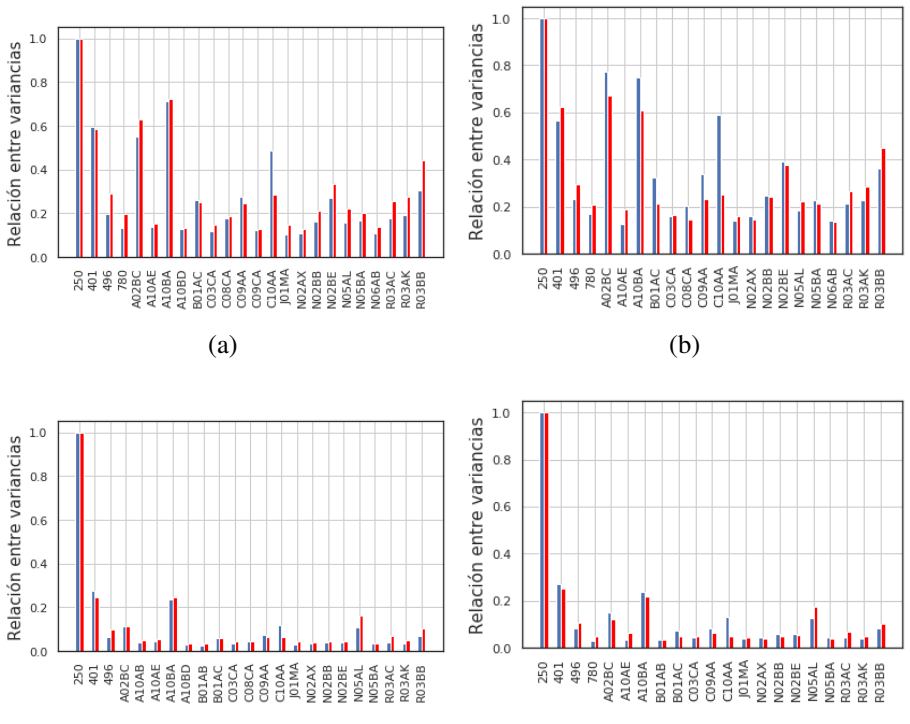


Figura 5.2: Primeras 25 características seleccionadas en base a la prueba F-Fisher. Las gráficas (a) y (b) evalúan las características clínicas basadas en ocurrencia para los enfoques multi-clase y multi-etiqueta, respectivamente. Las gráficas (c) y (d) evalúan las características en forma binaria para los enfoques multi-clase y multi-etiqueta. Las barras azules hacen referencia a los subconjuntos de pacientes sanos y crónicos, y las barras rojas a los subconjuntos de pacientes crónicos.

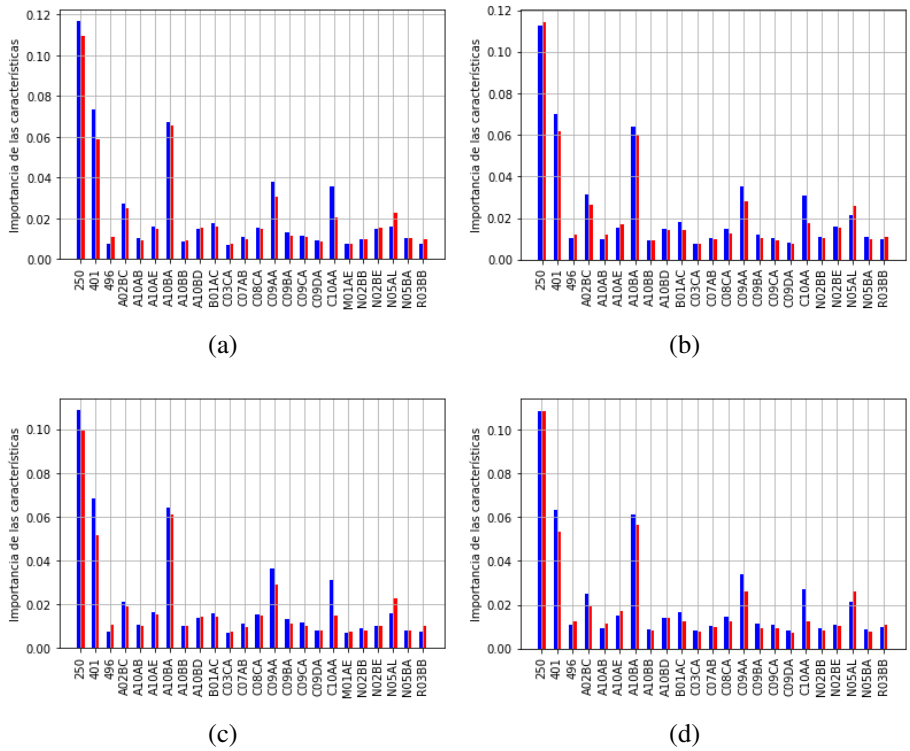


Figura 5.3: Muestra de las primeras 25 características seleccionadas con el algoritmo *Random Forest*. Las gráficas (a) y (b) evalúan las características clínicas basadas en ocurrencia para los enfoques multi-clase y multi-etiqueta, respectivamente. Las gráficas (c) y (d) evalúan las características en forma binaria para los enfoques multi-clase y multi-etiqueta. Las barras azules hacen referencia a los subconjuntos de pacientes sanos y crónicos, y las barras rojas a los subconjuntos de pacientes crónicos.

5.4. Experimentos

El procedimiento seguido para cada ejecución de los algoritmos es el mismo, cambiando para cada algoritmo el tipo de clasificación (multi-clase o multi-etiqueta) y sus medidas de prestaciones asociadas; los dos escenarios con respecto a las BBDD de los pacientes; y distintos métodos de selección de características, considerando las características clínicas en ocurrencia o binarias. Como se ha indicado anteriormente, se han utilizado 50 subconjuntos de cada escenario para entrenar y ejecutar los algoritmos de clasificación. Elegiremos el mejor modelo en base a la tasa de acierto media obtenida por el clasificador con el conjunto de *train*.

El procedimiento seguido para cada algoritmo es el siguiente:

1. Carga de los 50 subconjuntos de *train* y de *test* y el listado de características seleccionadas perteneciente a esa configuración de BBDD.
2. Filtrado de datos duplicados en los subconjuntos con un listado de características clínicas seleccionadas.
3. Normalización del conjunto de *train*.
4. Ejecución de la función `GridSearch` con CV para la búsqueda de parámetros libres en base al algoritmo de clasificación.
5. Terminada la búsqueda de parámetros libres, se evalúan los parámetros que ofrecen mejor tasa de acierto y se guardan.
6. Los pasos 2-4 se repiten para los 50 subconjuntos, y los parámetros libres que se seleccionan son los que mejor tasa de acierto aporten.
7. Con los mejores parámetros procedemos a evaluar el conjunto de *test*.
8. Normalizamos el conjunto de *test* con la media y la desviación típica del conjunto de *train*.
9. Predecidos las etiquetas para *test* con el algoritmo de clasificación parametrizado.
10. Dependiendo si el problema es multi-clase o multi-etiqueta usamos medidas de prestaciones diferentes para evaluar los resultados que aportan las etiquetas estimadas y las etiquetas reales.
11. Los pasos 7-10 se ejecutan para los 50 subconjuntos y realizamos la media de las medidas de prestaciones obtenidas.

Una nota sobre el duplicado, para ningún escenario propuesto se ha producido eliminación de pacientes iguales asignados a distintas clases o conjunto de etiquetas.

Los parámetros libres y sus valores seleccionados para cada método son:

RLM:
- `C`: [0,001, 0,0025, 0,005, 0,0075, 0,01, 0,025, 0,05, 0,075, 0,1, 0,25, 0,5, 0,75 , 1,0, 2,5, 5,0, 7,5, 10,0]
- `solver`: $['newton-cg', 'lbfgs']$

SVM:
- `kernel`: $['linear', 'rfb']$
- `C`:[0,001, 0,0025, 0,005, 0,0075, 0,01, 0,025, 0,05, 0,075, 0,1, 0,25, 0,5, 0,75

, 1,0, 2,5, 5,0, 7,5, 10,0, 15,0, 20,0]

- gamma:[$1.e-09$, $1.e-08$, $1.e-07, 1.e-06, 1.e-05, 1.e-04, 1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02, 1.e+03$] . Sólo para SVM no linear.

***k*NN:**
- k: [$i\ for\ i\ in\ range(10, neigh)$]. Siendo neigh 50 el número máximo de vecinos.

DT:
- max_depth: [$i\ for\ i\ in\ range(10, depth)$]. Siendo depth = número máximo de características por 2, es decir, 204.
- min_samples_split: [$i\ for\ i\ in\ range(1, samples\ split)$]. Siendo samples split = número máximo de observaciones en las BBDD de *train*.

***Random Forest*:** hemos aprovechado los parámetros max depth y min_samples_split de los DT clasificados para ahorrarnos tiempo y recursos.
- estimator: [$i\ for\ i\ in\ range(1, max\ estimator)$]. Siendo max estimator = número máximo de características por 2, es decir, 204.

MLP:
- activation: [$'logistic'$, $'tanh'$, $'relu'$].
- solver: [$'sgd'$, $'adam'$].
- hidden_layer_sizes: [$(i,)\ for\ i\ in\ range(5, max\ layers)$]. Siendo max layers = número máximo de características por 2, es decir, 204.

Cómo se puede ver en los parámetros se barren muchos valores. Al tratar con diferentes escenarios de BBDD y tipos de clasificación queríamos asegurarnos de que encontrábamos los mejores parámetros para cada algoritmo. En las subsecciones siguientes se resumen los resultados en base a la selección de características utilizada. Los resultados de todos los experimentos realizados se muestran en el Anexo C.

5.4.1. Clasificación multi-clase

En las Tablas 5.3 y 5.4 se muestran los resultados resumidos de los experimentos con multi-clase realizados mediante el procedimiento descrito al principio de la Sección 5.4. El sombreado de las Tablas indica el mejor resultado. Podemos ver que el mejor método de selección de características es el basado en la prueba F-Fisher, tampoco es una novedad ya que las características seleccionadas por ese método se escogen por la alta correlación con las etiquetas. Aunque esperábamos mejores resultados para la selección de características en base a la clasificación *Random Forest*, ya que es la única que explora el comportamiento entre características y etiquetas, no están tan alejado de los resultados de la prueba de F-Fisher. Con respecto a los métodos, no es ninguna sorpresa que cuando se consideran las características clínicas binarias despuntan las tasas de acierto para los algoritmos no lineales, en este caso SVM; y para ocurrencia algoritmos lineales, ya sean RLM o SVM lineal. Si nos fijamos en las tasas de acierto, tanto si consideramos las características clínicas binarias o en ocurrencia, vemos que los subconjuntos de pacientes sanos y crónicos tienen la tasa de acierto más elevada que los subconjuntos de pacientes crónicos. La respuesta es muy simple, los primeros tiene más muestras que los segundos, de nuevo resultados esperados.

Selección de Características	Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
BBDD sanos y crónicos - Características clínicas basadas en ocurrencia						
Frecuencia	Random Forest	Estimators: 85	0,853 ± 0,014	0,854 ± 0,014	0,853 ± 0,014	0,852 ± 0,014
F-Fisher	RLM	C: 10,0 Solver: $newton - cg$	0,888 ± 0,011	0,888 ± 0,010	0,888 ± 0,011	0,887 ± 0,011
Random Forest	Random Forest	Estimators: 81	0,860 ± 0,014	0,861 ± 0,015	0,860 ± 0,014	0,859 ± 0,014
BBDD crónicos - Características clínicas basadas en ocurrencia						
Frecuencia	Random Forest	Estimators: 70	0,832 ± 0,016	0,833 ± 0,016	0,832 ± 0,016	0,831 ± 0,016
F-Fisher	Linear SVM	C: 1,0	0,886 ± 0,014	0,888 ± 0,014	0,886 ± 0,014	0,886 ± 0,015
Random Forest	RLM	C: 10,0 Solver: $newton - cg$	0,849 ± 0,017	0,849 ± 0,017	0,849 ± 0,017	0,848 ± 0,017

Tabla 5.3: Mejores algoritmos de clasificación en función de las medidas estadísticas (media y desviación típica) obtenidas de los distintos métodos de selección de características. Considerando el enfoque multi-clase y la Características clínicas basadas en ocurrencia de características. Se muestran los resultados para los dos escenarios de BBDD conjuntamente.

Selección de Características	Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
BBDD sanos y crónicos - Características clínicas binarias						
Frecuencia	Linear SVM	C: 0,025	0,858 ± 0,013	0,860 ± 0,014	0,858 ± 0,013	0,857 ± 0,013
F-Fisher	No Linear SVM	C: 10,0 Gamma: 0,001	0,916 ± 0,012	0,919 ± 0,012	0,916 ± 0,012	0,916 ± 0,012
Random Forest	No Linear SVM	C: 10,0 Gamma: 0,001	0,889 ± 0,013	0,892 ± 0,013	0,889 ± 0,013	0,889 ± 0,013
BBDD crónicos - Características clínicas binarias						
Frecuencia	No Linear SVM	C: 7,5 Gamma: 0,001	0,837 ± 0,016	0,840 ± 0,017	0,837 ± 0,016	0,835 ± 0,016
F-Fisher	No Linear SVM	C: 10,0 Gamma: 0,001	0,909 ± 0,012	0,911 ± 0,012	0,909 ± 0,012	0,909 ± 0,012
Random Forest	No Linear SVM	C: 10,0 Gamma: 0,001	0,877 ± 0,014	0,881 ± 0,014	0,877 ± 0,014	0,877 ± 0,014

Tabla 5.4: Mejores algoritmos de clasificación en función de las medidas estadísticas (media y desviación típica) obtenidas de los distintos métodos de selección de características. Considerando el enfoque multi-clase y la Características clínicas binarias de características. Se muestran los resultados para los dos escenarios de BBDD conjuntamente.

5.4.2. Clasificación multi-etiqueta

El problema multi-etiqueta lo hemos dividido en dos, dependiendo del método: transformado o adaptado.

Empezamos con el método de transformación de algoritmos, en las Tablas 5.5 y 5.6 podemos observar los resultados resumidos de los experimentos. Los resultados completos están en el Anexo B. En ellos se puede observar que el algoritmo de transformación elegido es indiferente, ya que, las tasas de acierto para *Label Powerset* y *One-vs-Rest* son muy similares. Cuando evaluamos las Tablas 5.7 y 5.8, vemos que mejores tasas de acierto tienen valores por debajo de los resultados multi-clase. Hay que tener en cuenta que la clase minoritaria, CRG7071, condiciona todo el procedimiento haciendo que tengamos como máximo 1750 pacientes en los subconjuntos de *train*. Aunque intentamos paliar el efecto con los 50 subconjuntos, los pacientes que aporta CRG7071 siempre serán los mismos. Los mejores algoritmos en este enfoque (*Random Forest* y MLP) destacan del aprendizaje automático cuando hay una gran cantidad de datos a clasificar. Como en la clasificación multi-clase, la selección de características elegida es la prueba F-Fisher. Tanto los resultados como los algoritmos elegidos se asemejan a la clasificación multi-clase.

El mejor algoritmo cuando se considera la ocurrencia de las características clínicas es *Random Forest*, y cuando las características clínicas son binarias es el MLP. Destaca que el valor del parámetro `hidden_layer_sizes` sea en todos los experimentos mayor que el número de características seleccionada por cáda método (102), ya que queríamos "barrer" distintos valores para asegurar el mejor valor del parámetro. Viendo los resultados que aporta el paso 6 del procedimiento inicialmente descrito, hemos observado que lo que diferencian el escoger el valor `hidden_layer_sizes` = 155 y `hidden_layer_sizes` = 45 son milésimas de unidad, por tanto despreciable.

Selección de Características	Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Métodos de transformación de problemas - BBDD sanos y crónicos - Características clínicas basadas en ocurrencia										
Frecuencia	Linear SVM OneVsRest	C: 5,0	0,852 ± 0,015	0,053 ± 0,006	0,937 ± 0,014	0,963 ± 0,008	0,862 ± 0,018	0,921 ± 0,01	0,895 ± 0,013	0,941 ± 0,007
F-Fisher	Linear SVM OneVsRest	C: 0,5	0,895 ± 0,013	0,038 ± 0,005	0,974 ± 0,007	0,979 ± 0,005	0,917 ± 0,014	0,940 ± 0,009	0,944 ± 0,008	0,959 ± 0,005
Random Forest	Linear SVM OneVsRest	C: 2,5	0,877 ± 0,014	0,044 ± 0,005	0,957 ± 0,011	0,971 ± 0,006	0,891 ± 0,014	0,933 ± 0,009	0,921 ± 0,009	0,952 ± 0,006
Métodos de transformación de problemas - BBDD crónicos - Características clínicas basadas en ocurrencia										
Frecuencia	Linear SVM Label Powerset	C: 0,25	0,827 ± 0,019	0,070 ± 0,008	0,925 ± 0,016	0,956 ± 0,009	0,874 ± 0,017	0,923 ± 0,011	0,897 ± 0,012	0,939 ± 0,007
F-Fisher	Linear SVM Label Powerset	C: 1,0	0,886 ± 0,015	0,045 ± 0,007	0,964 ± 0,013	0,973 ± 0,008	0,925 ± 0,015	0,949 ± 0,009	0,943 ± 0,010	0,961 ± 0,006
Random Forest	Linear SVM Label Powerset	C: 0,25	0,864 ± 0,019	0,056 ± 0,008	0,952 ± 0,014	0,966 ± 0,009	0,906 ± 0,015	0,937 ± 0,010	0,927 ± 0,012	0,952 ± 0,007

Tabla 5.5: Mejores algoritmos de clasificación con métodos de transformación en función de las medidas estadísticas (media y desviación típica) obtenidas de los distintos métodos de selección de características. Considerando el enfoque multi-etiqueta con Características clínicas basadas en ocurrencia de características. Se muestran los resultados para los dos escenarios de BBDD conjuntamente.

Selección de Características	Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Métodos de transformación de problemas - BBDD sanos y crónicos - Características clínicas binarias										
Frecuencia	Linear SVM OneVsRest	C: 0,025	0,854 ± 0,012	0,053 ± 0,005	0,930 ± 0,012	0,959 ± 0,007	0,865 ± 0,015	0,927 ± 0,008	0,893 ± 0,012	0,943 ± 0,005
F-Fisher	Linear SVM OneVsRest	C: 0,1	0,912 ± 0,010	0,031 ± 0,003	0,964 ± 0,011	0,971 ± 0,007	0,935 ± 0,009	0,963 ± 0,005	0,948 ± 0,006	0,967 ± 0,004
Random Forest	Linear SVM OneVsRest	C: 0,075	0,891 ± 0,011	0,038 ± 0,004	0,947 ± 0,012	0,966 ± 0,006	0,905 ± 0,014	0,951 ± 0,007	0,923 ± 0,01	0,959 ± 0,005
Métodos de transformación de problemas - BBDD crónicos - Características clínicas binarias										
Frecuencia	No Linear SVM Label Powerset	C: 7,5 Gamma: 0,001	0,84 ± 0,014	0,062 ± 0,006	0,928 ± 0,013	0,960 ± 0,007	0,881 ± 0,015	0,932 ± 0,009	0,903 ± 0,011	0,946 ± 0,005
F-Fisher	No Linear SVM Label Powerset	C: 10,0 Gamma: 0,001	0,903 ± 0,012	0,036 ± 0,005	0,974 ± 0,007	0,978 ± 0,005	0,935 ± 0,011	0,959 ± 0,006	0,953 ± 0,007	0,968 ± 0,004
Random Forest	No Linear SVM Label Powerset	C: 10,0 Gamma: 0,001	0,886 ± 0,010	0,043 ± 0,004	0,964 ± 0,010	0,976 ± 0,006	0,913 ± 0,013	0,950 ± 0,007	0,937 ± 0,007	0,963 ± 0,004

Tabla 5.6: Mejores algoritmos de clasificación con métodos de transformación en función de las medidas estadísticas (media y desviación típica) obtenidas de los distintos métodos de selección de características. Considerando el enfoque multi-etiqueta con Características clínicas binarias de características. Se muestran los resultados para los dos escenarios de BBDD conjuntamente.

Selección de Características	Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Métodos de adaptación de algoritmos - BBDD sanos y crónicos - Características clínicas basadas en ocurrencia										
Frecuencia	Random Forest	Estimators: 50	0,847 ± 0,014	0,054 ± 0,005	0,945 ± 0,013	0,960 ± 0,007	0,852 ± 0,016	0,924 ± 0,008	0,888 ± 0,012	0,942 ± 0,005
F-Fisher	Random Forest	Estimators: 100	0,857 ± 0,012	0,050 ± 0,004	0,955 ± 0,012	0,965 ± 0,006	0,860 ± 0,014	0,928 ± 0,007	0,897 ± 0,010	0,946 ± 0,005
Random Forest	Random Forest	Estimators: 37	0,855 ± 0,013	0,051 ± 0,005	0,948 ± 0,012	0,963 ± 0,006	0,860 ± 0,015	0,925 ± 0,008	0,895 ± 0,011	0,944 ± 0,005
Métodos de adaptación de algoritmos - BBDD crónicos - Características clínicas basadas en ocurrencia										
Frecuencia	Random Forest	Estimators: 80	0,825 ± 0,018	0,062 ± 0,007	0,948 ± 0,013	0,960 ± 0,008	0,862 ± 0,017	0,933 ± 0,008	0,895 ± 0,013	0,946 ± 0,006
F-Fisher	Random Forest	Estimators: 95	0,842 ± 0,017	0,056 ± 0,006	0,958 ± 0,013	0,968 ± 0,008	0,869 ± 0,014	0,936 ± 0,007	0,904 ± 0,011	0,951 ± 0,005
Random Forest	Random Forest	Estimators: 66	0,827 ± 0,017	0,061 ± 0,006	0,955 ± 0,012	0,961 ± 0,007	0,863 ± 0,015	0,934 ± 0,007	0,897 ± 0,012	0,947 ± 0,005

Tabla 5.7: Mejores algoritmos de clasificación adaptados a multi-etiqueta en función de las medidas estadísticas (media y desviación típica) obtenidas de los distintos métodos de selección de características. Considerando la Características clínicas basadas en ocurrencia de características. Se muestran los resultados para los dos escenarios de BBDD conjuntamente.

Selección de Características	Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Métodos de adaptación de algoritmos - BBDD sanos y crónicos - Características clínicas binarias										
Frecuencia	MLP	Activation: $relu$ Solver: sgd Hidden Layers: 155	0,844 ± 0,014	0,057 ± 0,005	0,915 ± 0,012	0,95 ± 0,007	0,873 ± 0,017	0,927 ± 0,009	0,892 ± 0,012	0,938 ± 0,006
F-Fisher	MLP	Activation: $relu$ Solver: sgd Hidden Layers: 155	0,903 ± 0,011	0,035 ± 0,004	0,955 ± 0,012	0,968 ± 0,007	0,934 ± 0,011	0,958 ± 0,006	0,944 ± 0,008	0,963 ± 0,004
Random Forest	MLP	Activation: $relu$ Solver: sgd Hidden Layers: 170	0,877 ± 0,015	0,044 ± 0,005	0,935 ± 0,014	0,96 ± 0,008	0,906 ± 0,014	0,945 ± 0,008	0,919 ± 0,011	0,952 ± 0,006
Métodos de adaptación de algoritmos - BBDD crónicos - Características clínicas binarias										
Frecuencia	Random Forest	Estimators: 95	0,817 ± 0,017	0,064 ± 0,006	0,947 ± 0,013	0,958 ± 0,006	0,855 ± 0,015	0,930 ± 0,008	0,889 ± 0,013	0,944 ± 0,005
F-Fisher	MLP	Activation: $relu$ Solver: sgd Hidden Layers: 145	0,883 ± 0,015	0,042 ± 0,005	0,958 ± 0,008	0,969 ± 0,005	0,938 ± 0,012	0,96 ± 0,007	0,947 ± 0,008	0,964 ± 0,005
Random Forest	MLP	Activation: $relu$ Solver: sgd Hidden Layers: 190	0,868 ± 0,012	0,047 ± 0,005	0,947 ± 0,011	0,966 ± 0,007	0,921 ± 0,012	0,953 ± 0,007	0,933 ± 0,008	0,959 ± 0,004

Tabla 5.8: Mejores algoritmos de clasificación adaptados a multi-etiqueta en función de las medidas estadísticas (media y desviación típica) obtenidas de los distintos métodos de selección de características. Considerando la Características clínicas binarias de características. Se muestran los resultados para los dos escenarios de BBDD conjuntamente.

5.4.3. Comparación de resultados entre conjuntos multi-clase y multi-etiqueta

La comparación entre las tasas de acierto para los enfoques multi-etiqueta y multi-clase se muestra en las Tablas 5.9 y 5.10, pudiendo extraer las siguientes conclusiones:

- Las tasas de acierto para los subconjuntos de pacientes sanos y crónicos siempre tiene un valor más elevado que las tasas para los subconjuntos de pacientes crónicos. Esto es debido a que en los pacientes sanos (CRG1000) es más fácil diferenciar sus características principales que los que sufren alguna cronicidad.

- La mejor configuración para tratar nuestro problema es con la selección de características basada en la prueba F-Fisher y las características en forma binaria. Con respecto las características clínicas binarias o basadas en ocurrencias, las tasas no muestran mucha desviación la una con respecto a la otra, pero elegimos la forma binaria por la rapidez en la ejecución de los algoritmos.

- Los algoritmos que mejor funcionan con esa configuración son: SVM no lineal para multi-clase, y MLP para multi-label.

- En multi-etiqueta, cuando usamos algoritmos de transformación a multi-clase, las tasas de acierto se parecen mucho a las de multi-clase, pero cuando evaluamos los algoritmos adaptados, vemos cómo baja la tasa de acierto más o menos un 5 % . Esto se debe a que los algoritmos adaptados necesitan un gran volumen de datos para ser eficientes y son muy sensibles a los datos, ya sea por la composición de los subconjuntos (pacientes sanos y crónicos; y crónicos) o por la forma de las características clínicas binarias u ocurrencia. presencia u ocurrencia.

- Tener una clase minoritaria de sólo 547 pacientes (CRG7071) e utilizarla en los 50 subconjuntos ha condicionado todos los resultados y ha fijado un límite a la tasa de acierto.

- Para clasificar los subconjuntos con las características clínicas basadas en ocurrencia el mejor modelo es el SVM lineal y para los subconjuntos con las características clínicas binarias el mejor modelo es el SVM no lineal. Independientemente de la elección del enfoque multi-clase o multi-etiqueta.

- Para todos los métodos de selección de características no hemos encontrado pacientes con las mismas características pero distinta clase (datos duplicados). Esto indica que las 102 características son suficientes para la clasificación sin duplicados.

- Hay ligero aumento en la tasa de acierto para la clasificación multi-etiqueta con respecto a la multi-clase, más concretamente en el caso de MLP y *Random Forest*. Esto es un claro indicativo de que vamos por el camino correcto, pero hace falta una clase minoritaria con más pacientes para entrenar mejor el modelo.

Comparativa entre los mejores resultados de las clasificaciones multi-clase y multi-etiqueta - BBDD pacientes sanos y crónicos					
Selección de Características	BBDD	Clasificación	Algoritmo	Parámetros libres	Tasa de acierto
F-Fisher	Ocurrencia	Multi-clase	Regresión Logistica Multinominal	C: 10,0 Solver: $newton-cg$	$0{,}888 \pm 0{,}011$
		Multi-etiqueta	Linear SVM OneVsRest	C: 0,5	$0{,}895 \pm 0{,}013$
		Multi-etiqueta	Random Forest	Estimators: 100	$0{,}857 \pm 0{,}012$
F-Fisher	Presencia	Multi-clase	No Linear SVM	C: 10,0 Gamma: 0,001	$0{,}916 \pm 0{,}012$
		Multi-etiqueta	Linear SVM Label Powerset	C: 1,0	$0{,}886 \pm 0{,}015$
		Multi-etiqueta	MLP	Activation: $relu$ Solver: sgd Hidden Layers: 155	$0{,}903 \pm 0{,}011$

Tabla 5.9: Comparativa entre las mejores tasas de acierto (media y desviación típica) de los modelos sobre le conjunto de *test* en los escenarios multi-clase y multi-etiqueta con las BBDD pacientes sanos y crónicos.

Comparativa entre los mejores resultados de las clasificaciones multi-clase y multi-etiqueta - BBDD pacientes crónicos					
Selección de Características	BBDD	Clasificación	Algoritmo	Parámetros libres	Tasa de acierto
F-Fisher	Ocurrencia	Multi-clase	Linear SVM	C: 1,0	$0{,}886 \pm 0{,}014$
		Multi-etiqueta	Linear SVM Label Powerset	C: 1,0	$0{,}886 \pm 0{,}015$
		Multi-etiqueta	Random Forest	Estimators: 95	$0{,}842 \pm 0{,}017$
F-Fisher	Presencia	Multi-clase	No Linear SVM	C: 10,0 Gamma: 0,001	$0{,}909 \pm 0{,}012$
		Multi-etiqueta	No Linear SVM Label Powerset	C: 10,0 Gamma: 0,001	$0{,}903 \pm 0{,}012$
		Multi-etiqueta	MLP	Activation: $relu$ Solver: sgd Hidden Layers: 145	$0{,}883 \pm 0{,}015$

Tabla 5.10: Comparativa entre las mejores tasas de acierto (media y desviación típica) de los modelos sobre le conjunto de *test* en los escenarios multi-clase y multi-etiqueta con las BBDD pacientes crónicos.

El siguiente experimento llevado a cabo ha sido comprobar la relevancia de los diferentes tipos de características que componen nuestro vector de características. Dichos tipos son: demográficos (Edad y Género), clínicos (códigos CIE-9 - Sección 2.2.1) y farmacológicos (códigos ATC - Sección 2.2.2). Además, sirve para ver hasta dónde es posible reducir el número de características y seguir obteniendo buenos resultados en predicción. Este experimento se ha realizado para los dos tipos de clasificación y sólo teniendo en cuenta las características clínicas binarias y seleccionas con la prueba F-Fisher. El algoritmo lineal elegido ha sido el SVM lineal (resultados en la Tabla 5.12) y para como algoritmo no

lineal el MLP (resultados en la Tabla 5.13). Cabe destacar que en la realización de este experimento se han encontrado datos duplicados según reducimos el número de características (Tabla 5.11).

Los resultados mostrados en las Tablas 5.12 y 5.13 muestran la misma tendencia: mejor o parecida tasa de acierto para la selección de características de CIE-9 y ATC conjuntamente comparado con el vector de características original; el mejor tipo de características son las ATC o farmacológicas, lo indica la tasa de acierto, y el hecho de ser más del 70 % del vector de características, indica que son las más correlacionadas con las etiquetas; y por último sólo utilizando las características CIE-9 o las demográficas no son suficientes para predecir correctamente nuestros conjuntos de *test*.

Características	Multi-clase		Multi-etiqueta	
	BBDD sanos y crónicos	BBDD crónicos	BBDD sanos y crónicos	BBDD crónicos
Género + Edad + CIE-9 + ATC	2188 (102)	1750 (102)	2188 (102)	1750 (102)
CIE-9 + ATC	1657 (100)	1316 (100)	1690 (100)	1516 (100)
CIE-9	449(33)	308 (34)	453 (33)	310 (34)
ATC	1419 (69)	1316 (66)	1444 (67)	1353 (66)
Edad + Género	9 (2)	22 (2)	13 (2)	25 (2)

Tabla 5.11: Número de pacientes duplicados en cada escenario para multi-clase y multi-etiqueta. Los resultados se muestran Número de pacientes (Número de características)

Linear SVM			
Clasificador	Configuración elegida	Características	Tasa de acierto
BBDD de pacientes sanos y crónicos			
Multi-clase	C: 0,075	Género + Edad + CIE-9 + ATC (102)	$0{,}915 \pm 0{,}012$
	C: 0,075	CIE-9 + ATC (100)	$0{,}914 \pm 0{,}012$
	C: 0,1	CIE-9 (31)	$0{,}622 \pm 0{,}063$
	C: 0,1	ATC (69)	$0{,}849 \pm 0{,}015$
	C: 2,5	Edad + Género (2)	$0{,}310 \pm 0{,}010$
Multi-etiqueta	C: 0,01	Género + Edad + CIE-9 + ATC (102)	$0{,}912 \pm 0{,}010$
	C: 0,075	CIE-9 + ATC (100)	$0{,}912 \pm 0{,}009$
	C: 0,025	CIE-9 (33)	$0{,}597 \pm 0{,}093$
	C: 0,05	ATC (67)	$0{,}839 \pm 0{,}013$
	C: 0,075	Edad + Género (2)	$0{,}218 \pm 0{,}038$
BBDD de pacientes crónicos			
Multi-clase	C: 0,025	Género + Edad + CIE-9 + ATC (102)	$0{,}907 \pm 0{,}013$
	C: 0,05	CIE-9 + ATC (100)	$0{,}907 \pm 0{,}013$
	C: 0,075	CIE-9 (30)	$0{,}467 \pm 0{,}012$
	C: 0,075	ATC (70)	$0{,}850 \pm 0{,}017$
	C: 2,5	Edad + Género (2)	$0{,}211 \pm 0{,}020$
Multi-etiqueta	C: 0,025	Género + Edad + CIE-9 + ATC (102)	$0{,}907 \pm 0{,}013$
	C: 0,05	CIE-9 + ATC (100)	$0{,}888 \pm 0{,}014$
	C: 0,1	CIE-9 (34)	$0{,}579 \pm 0{,}038$
	C: 0,5	ATC (66)	$0{,}802 \pm 0{,}015$
	C: 0,001	Edad + Género (2)	$0{,}250 \pm 0{,}001$

Tabla 5.12: Tasa de acierto (media y desviación típica) sobre el conjunto de *test* al utilizar el clasificador Lineal SVM en los tipos de clasificación: multi-clase y multi-etiqueta. La configuración se ha elegido teniendo en cuenta la presencia de las características seleccionadas con la prueba F-Fisher para los dos escenarios: pacientes sanos y crónicos; y pacientes crónicos.

MLP			
Clasificador	Configuración elegida	Características	Tasa de acierto
BBDD de pacientes sanos y crónicos			
Multi-clase	Activation: $relu$ Solver: sgd Hidden Layers: 125	Género + Edad + CIE-9 + ATC (102)	$0{,}903 \pm 0{,}01$
	Activation: $relu$ Solver: sgd Hidden Layers: 36	CIE-9 + ATC (100)	$0{,}728 \pm 0{,}075$
	Activation: $rtanh$ Solver: sgd Hidden Layers: 46	CIE-9 (31)	$0{,}540 \pm 0{,}094$
	Activation: $relu$ Solver: sgd Hidden Layers: 58	ATC (69)	$0{,}742 \pm 0{,}088$
	Activation: $tanh$ Solver: sgd Hidden Layers: 12	Edad + Género (2)	$0{,}318 \pm 0{,}020$
Multi-etiqueta	Activation: $relu$ Solver: sgd Hidden Layers: 155	Género + Edad + CIE-9 + ATC (102)	$0{,}903 \pm 0{,}011$
	Activation: $relu$ Solver: sgd Hidden Layers: 44	CIE-9 + ATC (100)	$0{,}895 \pm 0{,}012$
	Activation: $rule$ Solver: sgd Hidden Layers: 48	CIE-9 (33)	$0{,}455 \pm 0{,}196$
	Activation: $relu$ Solver: sgd Hidden Layers: 52	ATC (67)	$0{,}820 \pm 0{,}035$
	Activation: $tanh$ Solver: sgd Hidden Layers: 10	Edad + Género (2)	$0{,}251 \pm 0{,}053$
BBDD de pacientes crónicos			
Multi-clase	Activation: $logistic$ Solver: sgd Hidden Layers: 100	Género + Edad + CIE-9 + ATC (102)	$0{,}880 \pm 0{,}015$
	Activation: $relu$ Solver: sgd Hidden Layers: 22	CIE-9 + ATC (100)	$0{,}890 \pm 0{,}014$
	Activation: $relu$ Solver: sgd Hidden Layers: 52	CIE-9 (30)	$0{,}328 \pm 0{,}112$
	Activation: $relu$ Solver: sgd Hidden Layers: 22	ATC (70)	$0{,}728 \pm 0{,}075$
	Activation: $tanh$ Solver: sgd Hidden Layers: 10	Edad + Género (2)	$0{,}221 \pm 0{,}020$
Multi-etiqueta	Activation: $relu$ Solver: sgd Hidden Layers: 145	Género + Edad + CIE-9 + ATC (102)	$0{,}883 \pm 0{,}015$
	Activation: $relu$ Solver: sgd Hidden Layers: 54	CIE-9 + ATC (100)	$0{,}881 \pm 0{,}015$
	Activation: $tanh$ Solver: sgd Hidden Layers: 26	CIE-9 (34)	$0{,}551 \pm 0{,}063$
	Activation: $relu$ Solver: sgd Hidden Layers: 48	ATC (66)	$0{,}823 \pm 0{,}015$
	Activation: $tanh$ Solver: sgd Hidden Layers: 14	Edad + Género (2)	$0{,}304 \pm 0{,}025$

Tabla 5.13: Tasa de acierto (media y desviación típica) sobre el conjunto de *test* al utilizar el clasificador MLP en los tipos de clasificación: multi-clase y multi-etiqueta. La configuración se ha elegido teniendo en cuenta la presencia de las características seleccionadas con la prueba F-Fisher para los dos escenarios: pacientes sanos y crónicos; y pacientes crónicos.

Capítulo 6

Conclusiones y líneas futuras

En este capítulo se presentan las conclusiones (Sección 6.1) del trabajo y algunas líneas futuras (Sección 6.2) de trabajo.

6.1. Conclusiones

España es el decimocuarto país en porcentaje de pacientes crónicos de la Unión Europea, con un $42\,\%$ de la población que padece, al menos, una enfermedad crónica. Esta cifra aumenta con la edad, alcanzando hasta el $70\,\%$ de los mayores de 65 años, con una media de cuatro patologías crónicas por persona. Es evidente que el tratamiento de dichas cronicidades supone un gran gasto económico, tanto para el sistema sanitario como para la economía familiar. En este contexto, el objetivo de este trabajo ha sido evaluar la utilización de esquemas de aprendizaje automático para ayudar a predecir el estado de salud de los pacientes con alguna (una o más) de las dos patologías crónicas más presentes en nuestra sociedad, i.e., hipertensión y diabetes. Los modelos diseñados podrían utilizarse para mejorar la eficiencia del sistema sanitario, porque se podría utilizar que introduce el aprendizaje automático es la rapidez (en tiempo) de predicción del estado de salud de un paciente con alguna cronicidad. Esto influye a la hora de realizar un diagnóstico, el tratamiento seguido y la asignación de recursos económicos.

Los análisis sobre la población asociada al HUF indican que el valor medio de códigos CIE-9 y ATC diferentes por paciente es más alto para mujeres que para hombres. A medida que aumenta la edad de la población, valores medios son equivalentes para mujeres y para hombres. Esto se debe a que, debido al envejecimiento de las personas, las patologías crónicas tienden a aparecer y no suelen aparecer solas. Por otro lado, también se confirma que a mayor número de comorbilidades, los pacientes tienen más contactos con el sistema sanitario y, por tanto, más códigos CIE-9 y ATC.

En el trabajo hemos considerado dos escenarios de clasificación: (1) BBDD de pacientes sanos y crónicos, compuesta por CRG1000 (sanos, es decir, sin enfermedades crónicas), CRG5192 (hipertensión), CRG5424 (diabetes), CRG6144 (hipertensión y diabetes) y CRG5192 (hipertensión, diabetes y otra cronicidad); (2) BBDD pacientes crónicos, excluyendo al CRG1000. Para dichos escenarios hemos considerado las características clínicas binarias y basadas en la ocurrencia.

Para todos los escenarios hemos aplicado la selección de características y la que mejor resultados ha obtenido es la basada en la prueba F-Fisher. Una característica altamente correlacionada con la clase recibe una puntuación más alta. Otra característica menos correlacionada obtienen una puntuación

más baja, dándonos una relación lineal entre características y clases. Estos vectores de características están formados por 100 códigos desbalanceados de CIE-9 y ATC más las características demográficas.

En este trabajo hemos considerado los modelos RLM, SVM lineal y no lineal, árboles de decisión, *Random Forest* y MLP para predecir el estado de salud de los pacientes del HUF . Las tasas de acierto, tanto para multi-clase como para multi-etiqueta, son siempre mayores cuando se consideran pacientes sanos y crónicos de manera conjunta. Este resultado es razonable, ya que los diagnósticos y fármacos considerados por los pacientes crónicos son, en general, muy diferentes a los de los pacientes sanos. Centrándonos en los resultados para multi-etiqueta, dividimos su análisis considerando de manera independiente métodos de transformación y modelos adaptados. Para los métodos de transformación sobresalen en prestaciones los siguientes modelos: SVM lineal considerando la ocurrencia de las características, y SVM no lineal cuando consideramos características binarias. Estos modelos son los que también ofrecen mejores resultados en clasificación multi-clase, algo lógico porque cuando se usan métodos de transformación, la clasificación multi-etiqueta se transforma en multi-clase.

Del análisis de los resultados obtenidos por los modelos multi-etiqueta adaptados, se concluye que éstos ofrecen prestaciones inferiores con respecto a los métodos de transformación. No obstante, para un mismo tipo de clasificador, se observa que un modelo multi-etiqueta adaptado ofrece, en general, prestaciones ligeramente superiores a las del modelo multi-clase. Por ejemplo, analizando las tasas de acierto para MLP, hay un incremento de $0{,}3\,\%$ en la tasa de acierto del esquema multi-etiqueta frente al esquema multi-clase. El modelo MLP es uno de los mejores para detectar si se está generalizando bien. Nuestras observaciones del conjunto de *train*, tanto para multi-clase como para multi-etiqueta, nunca es superior a 2735, es decir, 547 pacientes de cada clase, insuficientes para realizar la mejor generalización.

Al evaluar la composición del vector de características (prueba F-Fisher) podemos extraer varias conclusiones. Por un lado, si sólo utilizamos los 100 códigos CIE-9 y ATC, se obtienen tasas de acierto similares a cuando utilizamos el vector de características original. Si sólo tenemos en cuenta los códigos CIE-9, los códigos ATC y las características demográficas por separado, las tasas de acierto disminuyen mucho. Esta disminución en las tasas, indica que las características por separado, son insuficientes para una buena predicción del estado de salud de los pacientes, debido a todos los pacientes duplicados que se encuentran bajo cada selección.

6.2. Líneas Futuras

En este trabajo hemos contribuido a la investigación de la aplicación del aprendizaje automático para predecir el estado de salud en los pacientes crónicos. Los resultados obtenidos son buenos, pero se hace evidente que queda bastante trabajo por delante. Por ello se proponen varias líneas futuras:

- Aumentar la población de estudio, para que a su vez aumente la clase minoritaria por la cual balanceamos los pacientes. Necesitaríamos más pacientes para representar la clase minoritaria de nuestro problema, es decir, CRG7071 con 547 pacientes.

- En el caso de multi-etiqueta, aumentar el número de etiquetas considerando los niveles de gravedad de cada CRG. Esto proporcionaría resultados más específicos desde el punto de vista clínico, por otro lado nos llevaría a reducir el tamaño del conjunto de *train* usado para entrenar los modelos.

- Con respecto al balanceo de pacientes, sería interesante encontrar algoritmos de clasificación que puedan entrenarse sin que el conjunto de *train* esté balanceado, es decir, algoritmos que no

están sesgados a la clase mayoritaria.

- Mejorar el control sobre los datos duplicados que salen al reducir las características. Programar una función para sólo descartar los pacientes duplicados pertenecientes a distinta clase.
- Trabajar en la parte de visualización, ya que estamos tratando con médicos y no les sirve que una "caja negra" les dé una predicción.
- Para futuros trabajos, sería interesante utilizar *TensorFlow* o MEKA (*A Multi-label/Multi-target Extension to WEKA*). *TensorFlow* porque es una librería de *Python* y se pueden ejecutar en *Amazon Web Services* (AWS), y MEKA porque, al ser una extensión WEKA, sería muy interesante ejecutar los modelos que ofrece.
- Entrenar los modelos más pesados (*Random Forest* en tiempos de ejecución y cómputo en un *cluster* de servidores sería muy beneficioso para reducir el tiempo de entrenamiento del modelo.

Anexos

Anexo A: Código *Python*

Todo el código programado en este trabajo está disponible en el siguiente repositorio de GitHub: https://github.com/pvecino/trabajo_fin_de_grado_urjc

Las BBDD utilizadas al ser pertenecientes al HUF no podemos compartirlas, pero en dicho repositorio de GitHub hay un ejemplo sintético de cómo están formadas las BBDD para clasificación multi-clase y multi-etiqueta.

Anexo B: Características escogidas aplicando distintos métodos de selección de características

En este anexo se muestran los listados de las características seleccionadas por los diferentes métodos de selección descritos en la Secciónn 4.1.3 y usados para entrenar los algoritmos de clasificación, cuyos resultados se muestran en la Sección 5.4 y en más detalladamente en el Anexo C.

Selección de características en Frecuencia

Ocurrencia de las características				Presenciade las características			
BBDD de pacientes sanos y crónicos		BBDD de pacientes crónicos		BBDD de pacientes sanos y crónicos		BBDD de pacientes crónicos	
244	C03BA	244	C03CA	244	B01AB	244	A12AX
250	C03CA	250	C03EA	250	B01AC	250	B01AB
272	C03EA	272	C07AA	272	B03AA	272	B01AC
278	C07AA	278	C07AB	278	C02CA	278	B03AA
300	C07AB	300	C07AG	300	C03BA	300	C02CA
365	C07AG	365	C08CA	305	C03CA	305	C03BA
366	C08CA	366	C08DB	311	C05CA	311	C03CA
401	C09AA	401	C09AA	362	C07AB	362	C05CA
427	C09BA	427	C09BA	366	C08CA	365	C07AB
460	C09CA	460	C09CA	372	C09AA	366	C08CA
465	C09DA	465	C09DA	380	C09BA	372	C09AA
490	C09DB	490	C09DB	382	C09CA	380	C09BA
496	C10AB	496	C10AB	401	C09DA	386	C09CA
526	C10AX	526	C10AX	460	C10AA	401	C09DA
571	D01AC	571	D01AC	462	C10AB	460	C10AA
599	D07AA	599	D07AA	463	D01AC	462	C10AB
692	D07AC	692	D07AC	465	D06AX	465	D01AC
715	G04CA	715	G04CA	490	D07AC	490	D06AX
719	H02AB	719	H02AB	496	G04CA	496	D07AC
724	H03AA	724	H03AA	526	H02AB	526	G04CA
726	J01CA	726	J01CA	558	H03AA	558	H02AB
729	J01CR	729	J01CR	599	J01CA	599	H03AA
780	J01MA	780	J01MA	692	J01CR	692	J01CA
782	M01AB	782	M01AB	715	J01FA	715	J01CR
784	M01AE	784	M01AE	719	J01MA	719	J01FA
786	M02AA	786	M02AA	723	J01XX	723	J01MA
787	N02AX	788	N02AX	724	M01AB	724	J01XX
788	N02BB	789	N02BB	726	M01AE	726	M01AB
789	N02BE	790	N02BE	729	M02AA	729	M01AE
790	N03AX	A02BA	N03AX	780	M03BX	780	M02AA
A02BA	N05AL	A02BC	N05AL	782	N02AA	782	M03BX
A02BC	N05BA	A03FA	N05BA	784	N02AX	784	N02AX
A03FA	N05CD	A06AD	N05CD	786	N02BB	786	N02BB
A06AD	N05CF	A10AB	N05CF	787	N02BE	787	N02BE
A10AB	N06AA	A10AC	N06AA	788	N03AX	788	N03AX
A10AC	N06AB	A10AD	N06AB	789	N05AL	789	N05AL
A10AD	N06AX	A10AE	N06AX	790	N05BA	790	N05BA
A10AE	R01AD	A10BA	R01AD	A02AD	N05CD	A02AD	N05CD
A10BA	R01AX	A10BB	R01AX	A02BA	N06AB	A02BA	N06AB
A10BB	R03AC	A10BD	R03AC	A02BC	N06AX	A02BC	N06AX
A10BD	R03AK	A10BH	R03AK	A03FA	R01AD	A03FA	R01AD
A10BH	R03BA	A10BX	R03BA	A06AD	R03AC	A06AD	R03AC
A10BX	R03BB	A12AX	R03BB	A10AB	R03AK	A10AB	R03AK
A12AX	R05CB	B01AA	R05CB	A10AD	R03BB	A10AD	R03BB
B01AA	R06AX	B01AB	R06AX	A10AE	R05CB	A10AE	R05CB
B01AB	S01ED	B01AC	S01ED	A10BA	R05DA	A10BA	R05DA
B01AC	S01EE	B03AA	S01EE	A10BB	R05DB	A10BB	R05DB
B03AA	S01XA	C02CA	S01XA	A10BD	R06AX	A10BD	R06AX
C02CA	V58	C03AA	V58	A10BX	S01XA	A10BH	S01XA
C03AA	C10AA	C03BA	C10AA	A12AX	V58	A10BX	V58

Tabla B1: Listado de las características seleccionadas en base a la frecuencia, considerando la presencia y la ocurrencia de las características. Se diferencian las BBDD en los dos escenarios propuestos.Estos listados son independiente del tipo de clasificación.

Selección de características con F-Fisher y clasificación multi-clase

Ocurrencia de las características				Presenciade las características			
BBDD de pacientes sanos y crónicos		BBDD de pacientes crónicos		BBDD de pacientes sanos y crónicos		BBDD de pacientes crónicos	
199	C08CA	199	C08DB	250	C09AA	199	C09BA
250	C08DB	250	C09AA	272	C09BA	250	C09CA
272	C09AA	294	C09BA	278	C09CA	294	C09DA
294	C09BA	311	C09CA	294	C09DA	311	C10AA
311	C09CA	327	C09DA	311	C10AA	331	D07AA
327	C09DA	331	C10AA	331	C10AB	332	G04CA
331	C10AA	332	D07AA	332	D07AA	333	H02AB
332	C10AB	333	G04CA	333	G03AA	366	H03CA
333	D07AA	386	H02AB	362	G04CA	386	H04AA
365	G03AA	401	H03CA	366	H02AB	401	J01CR
386	G04CA	427	H04AA	386	H03AA	427	J01DD
401	H02AB	490	J01CR	401	H03CA	492	J01MA
427	H03AA	491	J01DD	427	H04AA	496	L04AA
463	H03CA	492	J01MA	463	J01CR	518	L04AD
490	H04AA	496	L04AA	492	J01DD	519	M02AA
492	J01CR	518	L04AD	496	J01MA	571	M04AA
496	J01DD	519	M02AA	518	L04AA	650	N02AX
518	J01MA	571	M04AA	519	L04AD	715	N02BB
571	L04AA	650	N02AX	571	M02AA	733	N02BE
650	L04AD	664	N02BB	650	M04AA	780	N03AE
724	M02AA	733	N02BE	715	N02AX	781	N03AX
733	M04AA	780	N03AE	724	N02BB	786	N04BA
780	N02AX	781	N03AX	780	N02BE	788	N04BC
781	N02BB	786	N04BA	781	N03AE	A01AB	N05AA
786	N02BE	788	N04BC	786	N03AX	A02BA	N05AD
788	N03AX	789	N05AA	788	N04BA	A02BC	N05AH
789	N04BA	A01AB	N05AH	A01AB	N04BC	A03FA	N05AL
A01AB	N04BC	A02BC	N05AL	A02BC	N05AA	A06AD	N05AX
A02BC	N05AH	A03FA	N05AX	A03FA	N05AH	A07EC	N05BA
A03FA	N05AL	A06AD	N05BA	A06AD	N05AL	A10AB	N05CD
A06AD	N05AX	A07EC	N05CD	A10AB	N05AX	A10AD	N06AB
A10AB	N05BA	A10AB	N06AA	A10AD	N05BA	A10AE	N06AX
A10AD	N05CD	A10AD	N06AB	A10AE	N05CD	A10BA	N06DA
A10AE	N06AB	A10AE	N06AX	A10BA	N06AB	A10BB	N07CA
A10BA	N06AX	A10BA	N06DA	A10BB	N06AX	A10BD	R01AX
A10BB	N06DA	A10BB	R01AX	A10BD	N06DA	A10BH	R03AC
A10BD	R01AX	A10BD	R03AC	A10BH	N07CA	A10BX	R03AK
A10BX	R03AC	A10BX	R03AK	A10BX	R01AX	A12AX	R03BA
A12AX	R03AK	A12AX	R03BA	A12AX	R03AC	A12BA	R03BB
A12BA	R03BA	A12BA	R03BB	A12BA	R03AK	B01AA	R03DA
B01AA	R03BB	B01AA	R03DA	B01AA	R03BA	B01AB	R03DC
B01AB	R05CB	B01AB	R03DC	B01AB	R03BB	B01AC	R05CB
B01AC	S01XA	B01AC	R05CB	B01AC	R03DA	C02CA	S01XA
B03AA	V10	C02CA	S01XA	C02CA	R05CB	C03BA	V10
C02CA	V12	C03BA	V10	C03BA	S01XA	C03CA	V12
C03BA	V15	C03CA	V12	C03CA	V12	C03DA	V15
C03CA	V22	C03DA	V15	C03DA	V15	C07AA	V22
C03DA	V25	C07AA	V22	C07AA	V25	C07AB	V27
C07AA	V27	C07AB	V27	C07AB	V27	C08CA	V45
C07AB	V58	C08CA	V58	C08CA	V58	C09AA	V58

Tabla B2: Listado de las características seleccionadas en base a la prueba F-Fisher para la clasificación multi-clase. Se diferencian para la ocurrencia y la presencia de las características y para los dos escenarios de BBDD propuestos.

Selección de características con F-Fisher y clasificación multi-etiqueta

Ocurrencia de las características				Presenciade las características			
BBDD de pacientes sanos y crónicos		BBDD de pacientes crónicos		BBDD de pacientes sanos y crónicos		BBDD de pacientes crónicos	
199	C07AB	199	C08DB	199	C03CA	199	C08CA
250	C08CA	250	C09AA	250	C03DA	250	C09AA
272	C08DB	294	C09BA	272	C07AA	294	C09BA
278	C09AA	311	C09CA	278	C07AB	311	C09CA
311	C09BA	327	C09DA	294	C08CA	327	C09DA
327	C09CA	331	C10AA	311	C09AA	331	C10AA
331	C09DA	332	D07AA	327	C09BA	332	D07AA
332	C10AA	333	G04CA	331	C09CA	333	H02AB
333	C10AB	386	H02AB	332	C09DA	366	H03CA
362	C10AX	401	H03CA	333	C10AA	386	H04AA
365	D07AA	427	H04AA	362	C10AB	401	J01CR
366	G04CA	490	J01CR	366	D07AA	427	J01DD
386	H02AB	491	J01DD	386	G04CA	486	J01MA
401	H03AA	492	J01MA	401	H02AB	492	L04AA
427	J01CR	496	L04AA	427	H03AA	493	L04AD
490	J01DD	518	L04AD	463	J01CR	496	M02AA
492	J01MA	519	M02AA	492	J01DD	518	M04AA
496	L04AA	571	M04AA	496	J01MA	519	N02AX
518	L04AD	650	N02AX	518	L04AA	571	N02BB
519	M02AA	724	N02BB	519	L04AD	585	N02BE
571	M04AA	733	N02BE	571	M02AA	650	N03AE
715	N02AX	780	N03AE	585	M04AA	715	N03AX
724	N02BB	781	N03AX	715	N02AX	733	N04BA
733	N02BE	786	N04BA	724	N02BB	780	N04BC
780	N03AE	788	N04BC	733	N02BE	781	N05AA
781	N03AX	789	N05AA	780	N03AE	786	N05AH
786	N04BA	A01AB	N05AH	781	N03AX	788	N05AL
788	N04BC	A02BC	N05AL	786	N04BA	A01AB	N05AX
789	N05AH	A03FA	N05AX	788	N04BC	A02BA	N05BA
A02BC	N05AL	A06AD	N05BA	A01AB	N05AH	A02BC	N05CD
A03FA	N05AX	A07EC	N05CD	A02AD	N05AL	A03FA	N06AB
A06AD	N05BA	A10AB	N06AA	A02BA	N05AX	A06AD	N06AX
A10AB	N05CD	A10AD	N06AB	A02BC	N05BA	A10AB	N06DA
A10AD	N06AA	A10AE	N06AX	A03FA	N05CD	A10AD	N07CA
A10AE	N06AB	A10BA	N06DA	A06AD	N06AB	A10AE	R01AX
A10BA	N06AX	A10BB	N07CA	A10AB	N06AX	A10BA	R03AC
A10BB	N06DA	A10BD	R01AX	A10AD	N06DA	A10BB	R03AK
A10BD	N07CA	A10BX	R03AC	A10AE	N07CA	A10BD	R03BA
A10BX	R01AX	A12AX	R03AK	A10BA	R01AX	A10BX	R03BB
A12AX	R03AC	A12BA	R03BA	A10BB	R03AC	A12AX	R03DA
A12BA	R03AK	B01AA	R03BB	A10BD	R03AK	A12BA	R03DC
B01AA	R03BA	B01AB	R03DA	A10BH	R03BA	B01AA	R05CB
B01AB	R03BB	B01AC	R03DC	A10BX	R03BB	B01AB	S01XA
B01AC	R05CB	C02CA	R05CB	A12AX	R03DA	B01AC	V10
B03AA	S01XA	C03BA	S01XA	A12BA	R05CB	C02CA	V12
C02CA	V10	C03CA	V10	B01AA	S01XA	C03BA	V15
C03BA	V12	C03DA	V12	B01AB	V12	C03CA	V22
C03CA	V15	C07AA	V15	B01AC	V15	C03DA	V27
C03DA	V25	C07AB	V27	C02CA	V25	C07AA	V45
C07AA	V58	C08CA	V58	C03BA	V58	C07AB	V58

Tabla B3: Listado de las características seleccionadas en base a la prueba F-Fisher para la clasificación multi-etiqueta. Se diferencian para la ocurrencia y la presencia de las características y para los dos escenarios de BBDD propuestos.

Selección de características con Random Forest y clasificación multi-clase

Ocurrencia de las características				Presenciade las características			
BBDD de pacientes sanos y crónicos		BBDD de pacientes crónicos		BBDD de pacientes sanos y crónicos		BBDD de pacientes crónicos	
244	B01AA	250	B01AC	244	B01AA	244	B01AC
250	B01AB	272	B03AA	250	B01AB	250	B03AA
272	B01AC	278	C02CA	272	B01AC	272	C02CA
278	B03AA	300	C03AA	278	B03AA	278	C03AA
300	C02CA	305	C03BA	300	C02CA	300	C03BA
305	C03AA	311	C03CA	305	C03AA	305	C03CA
311	C03BA	331	C03DA	311	C03BA	311	C03DA
331	C03CA	362	C03EA	331	C03CA	331	C03EA
362	C03EA	365	C07AA	362	C03EA	362	C07AA
365	C07AA	366	C07AB	366	C07AA	366	C07AB
366	C07AB	380	C08CA	380	C07AB	380	C08CA
380	C08CA	386	C09AA	386	C08CA	386	C09AA
386	C09AA	401	C09BA	401	C09AA	401	C09BA
401	C09BA	460	C09CA	460	C09BA	427	C09CA
460	C09CA	462	C09DA	462	C09CA	460	C09DA
462	C09DA	465	C10AA	463	C09DA	462	C10AA
463	C10AA	496	C10AB	465	C10AA	465	C10AB
465	C10AB	526	D01AC	496	C10AB	496	D01AC
496	D01AC	571	D07AC	526	D01AC	526	D07AC
526	D07AC	599	G04CA	558	D07AC	571	G04CA
558	G04CA	692	H02AB	571	G04CA	599	H02AB
571	H02AB	715	H03AA	599	H02AB	692	H03AA
599	H03AA	719	H03CA	692	H03AA	715	H03CA
692	H03CA	724	J01CA	715	H03CA	719	J01CA
715	J01CA	726	J01CR	719	J01CA	724	J01CR
719	J01CR	729	J01FA	724	J01CR	726	J01FA
724	J01FA	780	J01MA	726	J01FA	729	J01MA
726	J01MA	782	J01XX	729	J01MA	780	J01XX
729	J01XX	784	M01AB	780	J01XX	782	M01AB
780	M01AB	786	M01AE	782	M01AB	784	M01AE
782	M01AE	787	M02AA	784	M01AE	786	M02AA
784	M02AA	788	N02AX	786	M02AA	787	N02AX
786	N02AX	789	N02BB	787	N02AX	788	N02BB
787	N02BB	790	N02BE	788	N02BB	789	N02BE
788	N02BE	A02AD	N03AX	789	N02BE	790	N03AX
789	N03AX	A02BC	N05AH	790	N03AX	A02BC	N05AH
790	N05AL	A03FA	N05AL	A02AD	N05AL	A03FA	N05AL
A02BC	N05BA	A06AD	N05BA	A02BC	N05BA	A06AD	N05BA
A03FA	N05CD	A10AB	N05CD	A03FA	N05CD	A10AB	N05CD
A06AD	N06AB	A10AC	N06AB	A06AD	N06AB	A10AC	N06AB
A10AB	N06AX	A10AD	N06AX	A10AB	N06AX	A10AD	N06AX
A10AC	N06DA	A10AE	N06DA	A10AC	N06DA	A10AE	N06DA
A10AD	R01AD	A10BA	R01AD	A10AD	R01AD	A10BA	R01AD
A10AE	R03AC	A10BB	R03AC	A10AE	R03AC	A10BB	R03AC
A10BA	R03AK	A10BD	R03AK	A10BA	R03AK	A10BD	R03AK
A10BB	R03BB	A10BH	R03BB	A10BB	R03BB	A10BH	R03BB
A10BD	R05CB	A10BX	R05CB	A10BD	R05CB	A10BX	R05CB
A10BH	R06AX	A12AX	R06AX	A10BH	R06AX	A12AX	R06AX
A10BX	S01XA	B01AA	S01XA	A10BX	S01XA	B01AA	S01XA
A12AX	V58	B01AB	V58	A12AX	V58	B01AB	V58

Tabla B4: Listado de las características seleccionadas aplicando *Random Forest* para la clasificación multi-clase. Se diferencian para la ocurrencia y la presencia de las características y para los dos escenarios de BBDD propuestos.

Ocurrencia de las características				Presenciade las características			
BBDD de pacientes sanos y crónicos		BBDD de pacientes crónicos		BBDD de pacientes sanos y crónicos		BBDD de pacientes crónicos	
244	B01AA	250	C03AA	244	B01AB	250	C02CA
250	B01AB	272	C03BA	250	B01AC	272	C03AA
272	B01AC	278	C03CA	272	B03AA	278	C03BA
278	B03AA	300	C03DA	278	C02CA	300	C03CA
300	C02CA	311	C03EA	300	C03AA	305	C03DA
305	C03AA	331	C07AA	305	C03BA	311	C03EA
311	C03BA	362	C07AB	311	C03CA	331	C07AA
331	C03CA	365	C08CA	331	C03DA	362	C07AB
362	C03DA	366	C09AA	362	C03EA	365	C08CA
365	C03EA	386	C09BA	365	C07AA	366	C09AA
366	C07AA	401	C09CA	366	C07AB	386	C09BA
386	C07AB	427	C09DA	386	C08CA	401	C09CA
401	C08CA	460	C10AA	401	C09AA	460	C09DA
427	C09AA	462	C10AB	460	C09BA	462	C10AA
460	C09BA	465	D01AC	462	C09CA	465	C10AB
462	C09CA	496	D07AC	463	C09DA	496	D01AC
463	C09DA	526	G04CA	465	C10AA	526	D07AC
465	C10AA	571	H02AB	496	C10AB	571	G04CA
496	C10AB	599	H03AA	526	D01AC	599	H02AB
526	D01AC	692	H03CA	558	D07AC	692	H03AA
571	D07AC	715	J01CA	571	G04CA	715	H03CA
599	G04CA	719	J01CR	599	H02AB	719	J01CA
692	H02AB	724	J01MA	692	H03AA	724	J01CR
715	H03AA	726	J01XX	715	J01CA	726	J01FA
719	J01CA	729	M01AB	719	J01CR	729	J01MA
724	J01CR	780	M01AE	724	J01FA	780	J01XX
726	J01MA	782	M02AA	726	J01MA	782	M01AB
729	M01AB	784	N02AX	729	J01XX	784	M01AE
780	M01AE	786	N02BB	780	M01AB	786	M02AA
782	M02AA	788	N02BE	782	M01AE	787	N02AX
784	N02AX	789	N03AX	784	M02AA	788	N02BB
786	N02BB	790	N04BA	786	N02AX	789	N02BE
787	N02BE	A02BC	N04BC	787	N02BB	790	N03AX
788	N03AX	A03FA	N05AH	788	N02BE	A02BC	N04BC
789	N05AH	A06AD	N05AL	789	N03AX	A03FA	N05AH
790	N05AL	A10AB	N05AX	790	N05AH	A06AD	N05AL
A02AD	N05BA	A10AC	N05BA	A02BC	N05AL	A10AB	N05AX
A02BC	N05CD	A10AD	N05CD	A03FA	N05BA	A10AC	N05BA
A03FA	N06AB	A10AE	N06AB	A06AD	N05CD	A10AD	N05CD
A06AD	N06AX	A10BA	N06AX	A10AB	N06AB	A10AE	N06AB
A10AB	N06DA	A10BB	N06DA	A10AC	N06AX	A10BA	N06AX
A10AC	R01AD	A10BD	R01AD	A10AD	N06DA	A10BB	N06DA
A10AD	R03AC	A10BH	R03AC	A10AE	R01AD	A10BD	R01AD
A10AE	R03AK	A10BX	R03AK	A10BA	R03AC	A10BH	R03AC
A10BA	R03BA	A12AX	R03BA	A10BB	R03AK	A10BX	R03AK
A10BB	R03BB	B01AA	R03BB	A10BD	R03BB	A12AX	R03BB
A10BD	R05CB	B01AB	R05CB	A10BH	R05CB	B01AA	R05CB
A10BH	R06AX	B01AC	R06AX	A10BX	R06AX	B01AB	R06AX
A10BX	S01XA	B03AA	S01XA	A12AX	S01XA	B01AC	S01XA
A12AX	V58	C02CA	V58	B01AA	V58	B03AA	V58

Tabla B5: Listado de las características seleccionadas aplicando *Random Forest* para la clasificación multi-etiqueta. Se diferencian para la ocurrencia y la presencia de las características y para los dos escenarios de BBDD propuestos.

Ocurrencia	Presencia de las características		
787(Síntoma Aparato Digestivo)	463 (Amigdalitis Aguda)	386 (Trastorno Aparato Vesicultar)	N02AA (Alcaloides del opio: morfina, codeina, ...)
C08DB (Derivados de la benzotiazepina)	382 (Otitis media supurativa)	365 (Glaucoma)	A10BH (Inhibidores de DPP-4)

Tabla B6: Listado de los códigos CIE-9 y ATC diferentes comparando los listados obtenidos aplicar la selección de características en base a la frecuencia: Tabla B1. Diferenciamos entre la ocurrencia y la presencia de las características.

Ocurrencia de las características			Presencia de las características		
463(Amigdalitis aguda)	C10AB (Fibratos)	V25 (Asistencia anticonceptiva)	362 (Enfermedad Retina)	463 (Amigdalitis Aguda)	C10AB (Fibratos)
365(Glaucoma)	724 (Enfermedad de espalda)	G03AA (Progestágenos y estrógenos)	V25 (Asistencia Anticonceptiva)	724 (Enfermedad de espalda)	278 (Obesidad y otra hiperalimentación)
B03AA (Hierro bivalente)	272 (Trastorno metabolismo lípido)	H03AA (Hormonas tiroideas)	G03AA (Progestágenos y estrógenos)	272 (Trastorno metabolismo lípido)	H03AA (Hormonas tiroideas)
N03AE (Derivados de benzodiazepina)	A07EC (Aminosalicílico ácido)	R03DC (Antagonistas del receptor de leucotrienos)	V10 (H.P. Neoplasia maligna)	199 (Neoplasia maligna)	N05AD (Antipsicóticos derivados de Butirofenona)
N05AA (Fenotiazinas con cadena lateral alifática)	664 (Traumatismo perineo y vulva en parto)	N06AA (Inhibidores no selectivos de recaptación de monoaminas)	A07EC (Aminosalicílico ácido)	R03DC (Antagonistas del receptor de leucotrienos)	V22 (Supervisión embarazo)
491 (Bronquitis crónica)	R03DA (Xantinas)	519 (Enfermedad aparato respiratorio)	733 (Enfermedad de huesos y cartigalos)	A02BA (Antagonistas del receptor H2)	V45 (Estado postquirúrgico)

Tabla B7: Listado de los códigos CIE-9 y ATC diferentes comparando los listados obtenidos de aplicar la selección de características con la prueba F-Fisher en multi-clase (Tabla B2). Diferenciamos entre la ocurrencia y la presencia de las características.

Ocurrencia de las características			Presencia de las características		
362 (Enfermedad Retina)	715 (Osteoartrosis)	C10AB (Fibratos)	362 (Enfermedad Retina)	463 (Amigdalitis aguda)	G04CA (Antagonistas de receptores alfa - adrenérgicos)
V25 (Asistencia Anticonceptiva)	365 (Galucoma)	C10AX (Otros agentes modificadores de los lípidos)	C10AB (Fibratos)	V25 (Asistencia Anticonceptiva)	724 (Enfermedad de espalda)
278 (Obesidad y otra hiperalimentación)	366 (Catarata)	B03AA (Hierro bivalente)	278 (Obesidad y otra hiperalimentación)	A02AD (Compuestos de aluminio, calcio y magnesio)	A10BH (Inhibidores de DPP-4)
272 (Trastorno metabolismo lípido)	H03AA (Hormonas tiroideas)	H04AA (Hormonas glucogenolíticas)	272 (Trastorno metabolismo lípido)	H03AA (Hormonas tiroideas)	493 (Asma)
V27 (Parto)	650 (Parto normal)	A07EC (Aminosalicílico ácido)	H04AA (Hormonas glucogenolíticas)	V10 (H.P. Neoplasia maligna)	V27 (Parto
R03DC (Antagonistas del receptor de leucotrienos)	N05AA (Fenotiazinas con cadena lateral alifática)	H03CA (Terapia de iodo)	650 (Parto normal)	R03DC (Antagonistas del receptor de leucotrienos)	V22 (Supervisión embarazo)
294 (Estado psicotico organoico crónico)	A01AB (Antiinfecciosos y antisépticos)	491 (Bronquitis crónica)	H03CA (Terapia de iodo)	N05AA (Fenotiazinas con cadena lateral alifática)	486 (Neumonía)
R03DA (Xantinas)			V45 (Estado postquirúrgico)		

Tabla B8: Listado de los códigos CIE-9 y ATC diferentes comparando los listados obtenidos de aplicar la selección de características con la prueba F-Fisher en multi-etiqueta (Tabla B3). Diferenciamos entre la ocurrencia y la presencia de las características.

Ocurrencia de las características		Presencia de las características		
558 (Gastritis,enteritis y colitis no infecciosa)	463 (Amigdalitis Aguda)	A02AD (Compuestos de aluminio, calcio y magnesio)	558 (Gastritis,enteritis y colitis no infecciosa)	463 (Amigdalitis Aguda)
N05AH (Diazepinas, oxazepinas, tiazepinas y oxepinas)	C03DA (Antagonistas de la aldosterona)	N05AH (Diazepinas, oxazepinas, tiazepinas y oxepinas)	427 (Disritmia)	C03DA (Antagonistas de la aldosterona

Tabla B9: Listado de los códigos CIE-9 y ATC diferentes comparando los listados obtenidos de aplicar la selección de características con el algoritmo *Random Forest* en multi-clase (Tabla B4). Diferenciamos entre la ocurrencia y la presencia de las características.

Ocurrencia de las características				Presencia de las características	
244 (Hipotiroidismo adquirido)	787 (Sintoma aparato digestivo)	463 (Amigdalitis Aguda)	305(Abuso de droga sin dependencia)	244 (Hipotiroidismo adquirido)	558 (Gastritis,enteritis y colitis no infecciosa)
A02AD (Compuestos de aluminio, calcio y magnesio)	N04BA (Dopa)	J01XX (Antibacterianos)	N05AX (Antipsicóticos)	463 (Amigdalitis Aguda)	H03CA (Terapia de iodo)
	R03BA (Glucocorticoides)	N04BC (Agonistas de dopamina)		N04BC (Agonistas de dopamina)	N05AX (Antipsicóticos)

Tabla B10: Listado de los códigos CIE-9 y ATC diferentes comparando los listados obtenidos de aplicar la selección de características con el algoritmo *Random Forest* en multi-etiqueta (Tabla B5). Diferenciamos entre la ocurrencia y la presencia de las características.

Anexo C: Tablas de resultados obtenidos en los experimentos multi-clase y multi-etiqueta

C.1. Resultados de experimentos multi-clase

[H]

Selección de características: Frecuencia - BBDD de pacientes sanos y crónicos - Características clínicas basadas en ocurrencia					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 10,0 Solver: $newton-cg$	$0{,}845 \pm 0{,}015$	$0{,}845 \pm 0{,}015$	$0{,}845 \pm 0{,}015$	$0{,}844 \pm 0{,}015$
Linear SVM	C: 7,5	$0{,}839 \pm 0{,}014$	$0{,}839 \pm 0{,}014$	$0{,}839 \pm 0{,}014$	$0{,}837 \pm 0{,}014$
No Linear SVM	C: 10,0 Gamma: 0,010	$0{,}804 \pm 0{,}016$	$0{,}804 \pm 0{,}016$	$0{,}805 \pm 0{,}016$	$0{,}802 \pm 0{,}016$
*k*NN	Neigh: 1	$0{,}609 \pm 0{,}015$	$0{,}604 \pm 0{,}019$	$0{,}578 \pm 0{,}016$	$0{,}560 \pm 0{,}017$
DT	Max Depth: 18 Min Samples Split: 19	$0{,}814 \pm 0{,}016$	$0{,}814 \pm 0{,}016$	$0{,}814 \pm 0{,}016$	$0{,}812 \pm 0{,}016$
Random Forest	Estimators: 85	$0{,}853 \pm 0{,}014$	$0{,}854 \pm 0{,}014$	$0{,}853 \pm 0{,}014$	$0{,}852 \pm 0{,}014$
MLP	Activation: $logistic$ Solver: sgd Hidden Layers: 150	$0{,}746 \pm 0{,}045$	$0{,}752 \pm 0{,}018$	$0{,}746 \pm 0{,}045$	$0{,}751 \pm 0{,}018$

Tabla C1: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características en ocurrencia, subconjuntos de pacientes sanos y crónicos y características seleccionadas por frecuencia.

Selección de características: Frecuencia - BBDD de pacientes crónicos - Características clínicas basadas en ocurrencia					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 7,5 Solver: $newton-cg$	$0{,}831 \pm 0{,}016$	$0{,}831 \pm 0{,}017$	$0{,}831 \pm 0{,}016$	$0{,}829 \pm 0{,}016$
Linear SVM	C: 0,25	$0{,}825 \pm 0{,}019$	$0{,}825 \pm 0{,}019$	$0{,}825 \pm 0{,}019$	$0{,}822 \pm 0{,}019$
No Linear SVM	C: 10,0 Gamma: 0,010	$0{,}795 \pm 0{,}017$	$0{,}793 \pm 0{,}018$	$0{,}795 \pm 0{,}017$	$0{,}792 \pm 0{,}018$
*k*NN	Neigh: 23	$0{,}612 \pm 0{,}020$	$0{,}652 \pm 0{,}026$	$0{,}564 \pm 0{,}020$	$0{,}519 \pm 0{,}024$
DT	Max Depth: 10 Min Samples Split: 15	$0{,}795 \pm 0{,}020$	$0{,}800 \pm 0{,}020$	$0{,}795 \pm 0{,}020$	$0{,}793 \pm 0{,}020$
Random Forest	Estimators: 70	$0{,}832 \pm 0{,}016$	$0{,}833 \pm 0{,}016$	$0{,}832 \pm 0{,}016$	$0{,}831 \pm 0{,}016$
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 185	$0{,}787 \pm 0{,}018$	$0{,}782 \pm 0{,}019$	$0{,}787 \pm 0{,}018$	$0{,}783 \pm 0{,}019$

Tabla C2: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características en ocurrencia, subconjuntos de pacientes crónicos y características seleccionadas por frecuencia

Selección de características: F-Fisher - BBDD de pacientes sanos y crónicos - Características clínicas basadas en ocurrencia					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 10,0 Solver: $newton-cg$	$0{,}888 \pm 0{,}011$	$0{,}888 \pm 0{,}010$	$0{,}888 \pm 0{,}011$	$0{,}887 \pm 0{,}011$
Linear SVM	C: 0,75	$0{,}884 \pm 0{,}011$	$0{,}887 \pm 0{,}011$	$0{,}884 \pm 0{,}012$	$0{,}884 \pm 0{,}012$
No Linear SVM	C: 10,0 Gamma: 0,010	$0{,}851 \pm 0{,}014$	$0{,}85 \pm 0{,}014$	$0{,}851 \pm 0{,}014$	$0{,}849 \pm 0{,}015$
*k*NN	Neigh: 1	$0{,}713 \pm 0{,}017$	$0{,}700 \pm 0{,}021$	$0{,}664 \pm 0{,}018$	$0{,}655 \pm 0{,}019$
DT	Max Depth: 76 Min Samples Split: 25	$0{,}822 \pm 0{,}015$	$0{,}822 \pm 0{,}015$	$0{,}822 \pm 0{,}015$	$0{,}82 \pm 0{,}016$
Random Forest	Estimators: 80	$0{,}865 \pm 0{,}014$	$0{,}868 \pm 0{,}014$	$0{,}865 \pm 0{,}014$	$0{,}864 \pm 0{,}014$
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 200	$0{,}849 \pm 0{,}014$	$0{,}849 \pm 0{,}014$	$0{,}849 \pm 0{,}014$	$0{,}848 \pm 0{,}014$

Tabla C3: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características en ocurrencia, subconjuntos de pacientes sanos y crónicos y características seleccionadas por prueba F-Fisher.

Selección de características: F-Fisher - BBDD de pacientes crónicos - Características clínicas basadas en ocurrencia					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 10,0 Solver: $newton-cg$	$0{,}884 \pm 0{,}015$	$0{,}885 \pm 0{,}015$	$0{,}884 \pm 0{,}015$	$0{,}884 \pm 0{,}015$
Linear SVM	C: 1,0	$0{,}886 \pm 0{,}014$	$0{,}888 \pm 0{,}014$	$0{,}886 \pm 0{,}014$	$0{,}886 \pm 0{,}015$
No Linear SVM	C: 10,0 Gamma: 0,010	$0{,}863 \pm 0{,}018$	$0{,}863 \pm 0{,}018$	$0{,}863 \pm 0{,}018$	$0{,}862 \pm 0{,}018$
*k*NN	Neigh: 8	$0{,}701 \pm 0{,}020$	$0{,}726 \pm 0{,}021$	$0{,}653 \pm 0{,}020$	$0{,}635 \pm 0{,}022$
DT	Max Depth: 14 Min Samples Split: 33	$0{,}811 \pm 0{,}022$	$0{,}816 \pm 0{,}024$	$0{,}811 \pm 0{,}022$	$0{,}810 \pm 0{,}022$
Random Forest	Estimators: 75	$0{,}848 \pm 0{,}017$	$0{,}851 \pm 0{,}018$	$0{,}848 \pm 0{,}017$	$0{,}847 \pm 0{,}017$
MLP	Activation: $logistic$ Solver: sgd Hidden Layers: 85	$0{,}788 \pm 0{,}018$	$0{,}788 \pm 0{,}019$	$0{,}788 \pm 0{,}018$	$0{,}784 \pm 0{,}019$

Tabla C4: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características en ocurrencia, subconjuntos de pacientes crónicos y características seleccionadas por prueba F-Fisher.

Selección de características: Random Forest - BBDD de pacientes sanos y crónicos - Características clínicas basadas en ocurrencia					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 7,5 Solver: $newton-cg$	$0{,}859 \pm 0{,}014$	$0{,}859 \pm 0{,}014$	$0{,}859 \pm 0{,}014$	$0{,}858 \pm 0{,}014$
Linear SVM	C: 0,75	$0{,}853 \pm 0{,}015$	$0{,}854 \pm 0{,}015$	$0{,}853 \pm 0{,}014$	$0{,}851 \pm 0{,}015$
No Linear SVM	C: 10,0 Gamma: 0,010	$0{,}815 \pm 0{,}015$	$0{,}815 \pm 0{,}015$	$0{,}816 \pm 0{,}015$	$0{,}813 \pm 0{,}015$
*k*NN	Neigh: 1	$0{,}620 \pm 0{,}017$	$0{,}613 \pm 0{,}022$	$0{,}580 \pm 0{,}017$	$0{,}565 \pm 0{,}018$
DT	Max Depth: 38 Min Samples Split: 21	$0{,}817 \pm 0{,}014$	$0{,}817 \pm 0{,}014$	$0{,}817 \pm 0{,}014$	$0{,}816 \pm 0{,}014$
Random Forest	Estimators: 81	$0{,}860 \pm 0{,}014$	$0{,}861 \pm 0{,}015$	$0{,}860 \pm 0{,}014$	$0{,}859 \pm 0{,}014$
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 145	$0{,}817 \pm 0{,}014$	$0{,}816 \pm 0{,}015$	$0{,}817 \pm 0{,}014$	$0{,}815 \pm 0{,}015$

Tabla C5: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características en ocurrencia, subconjuntos de pacientes sanos y crónicos y características seleccionadas por *Random Forest*.

Selección de características: Random Forest - BBDD de pacientes crónicos - Características clínicas basadas en ocurrencia					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 10,0 Solver: $newton-cg$	$0{,}849 \pm 0{,}017$	$0{,}849 \pm 0{,}017$	$0{,}849 \pm 0{,}017$	$0{,}848 \pm 0{,}017$
Linear SVM	C: 0,25	$0{,}843 \pm 0{,}018$	$0{,}844 \pm 0{,}018$	$0{,}843 \pm 0{,}018$	$0{,}841 \pm 0{,}019$
No Linear SVM	C: 7,5 Gamma: 0,010	$0{,}810 \pm 0{,}019$	$0{,}808 \pm 0{,}020$	$0{,}810 \pm 0{,}019$	$0{,}807 \pm 0{,}019$
*k*NN	Neigh: 17	$0{,}619 \pm 0{,}020$	$0{,}661 \pm 0{,}023$	$0{,}569 \pm 0{,}020$	$0{,}531 \pm 0{,}025$
DT	Max Depth: 10 Min Samples Split: 31	$0{,}800 \pm 0{,}018$	$0{,}806 \pm 0{,}018$	$0{,}800 \pm 0{,}018$	$0{,}799 \pm 0{,}018$
Random Forest	Estimators: 61	$0{,}838 \pm 0{,}014$	$0{,}838 \pm 0{,}015$	$0{,}838 \pm 0{,}014$	$0{,}836 \pm 0{,}015$
MLP	Activation: $logistic$ Solver: sgd Hidden Layers: 115	$0{,}752 \pm 0{,}039$	$0{,}754 \pm 0{,}022$	$0{,}752 \pm 0{,}039$	$0{,}751 \pm 0{,}022$

Tabla C6: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características en ocurrencia, subconjuntos de pacientes crónicos y características seleccionadas por *Random Forest*.

Selección de características: Frecuencia - BBDD de pacientes sanos y crónicos - Características clínicas binarias					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 0,075 Solver: $newton-cg$	$0{,}850 \pm 0{,}013$	$0{,}85 \pm 0{,}013$	$0{,}850 \pm 0{,}013$	$0{,}849 \pm 0{,}013$
Linear SVM	C: 0,025	$0{,}858 \pm 0{,}013$	$0{,}860 \pm 0{,}014$	$0{,}858 \pm 0{,}013$	$0{,}857 \pm 0{,}013$
No Linear SVM	C: 10,0 Gamma: 0,001	$0{,}858 \pm 0{,}014$	$0{,}860 \pm 0{,}014$	$0{,}858 \pm 0{,}014$	$0{,}857 \pm 0{,}014$
*k*NN	Neigh: 12	$0{,}618 \pm 0{,}022$	$0{,}670 \pm 0{,}019$	$0{,}571 \pm 0{,}022$	$0{,}541 \pm 0{,}025$
DT	Max Depth: 58 Min Samples Split: 33	$0{,}801 \pm 0{,}017$	$0{,}801 \pm 0{,}018$	$0{,}801 \pm 0{,}017$	$0{,}799 \pm 0{,}017$
Random Forest	Estimators: 105	$0{,}849 \pm 0{,}017$	$0{,}850 \pm 0{,}016$	$0{,}849 \pm 0{,}017$	$0{,}847 \pm 0{,}017$
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 135	$0{,}848 \pm 0{,}013$	$0{,}848 \pm 0{,}013$	$0{,}848 \pm 0{,}013$	$0{,}847 \pm 0{,}013$

Tabla C7: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características binarias, subconjuntos de pacientes sanos y crónicos y características seleccionadas por frecuencia.

Selección de características: Frecuencia - BBDD de pacientes crónicos - Características clínicas binarias					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 0,025 Solver: $newton-cg$	$0{,}831 \pm 0{,}017$	$0{,}831 \pm 0{,}018$	$0{,}831 \pm 0{,}017$	$0{,}830 \pm 0{,}017$
Linear SVM	C: 0,025	$0{,}836 \pm 0{,}016$	$0{,}839 \pm 0{,}017$	$0{,}836 \pm 0{,}016$	$0{,}835 \pm 0{,}016$
No Linear SVM	C: 7,5 Gamma: 0,001	$0{,}837 \pm 0{,}016$	$0{,}840 \pm 0{,}017$	$0{,}837 \pm 0{,}016$	$0{,}835 \pm 0{,}016$
*k*NN	Neigh: 12	$0{,}676 \pm 0{,}019$	$0{,}722 \pm 0{,}021$	$0{,}624 \pm 0{,}019$	$0{,}577 \pm 0{,}026$
DT	Max Depth: 12 Min Samples Split: 27	$0{,}785 \pm 0{,}019$	$0{,}789 \pm 0{,}020$	$0{,}785 \pm 0{,}019$	$0{,}785 \pm 0{,}019$
Random Forest	Estimators: 95	$0{,}828 \pm 0{,}016$	$0{,}831 \pm 0{,}016$	$0{,}828 \pm 0{,}016$	$0{,}827 \pm 0{,}016$
MLP	Activation: $logistic$ Solver: sgd Hidden Layers: 90	$0{,}819 \pm 0{,}016$	$0{,}817 \pm 0{,}017$	$0{,}819 \pm 0{,}016$	$0{,}816 \pm 0{,}017$

Tabla C8: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características binarias, subconjuntos de pacientes crónicos y características seleccionadas por frecuencia.

Selección de características: F-Fisher - BBDD de pacientes sanos y crónicos - Características clínicas binarias					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 0,25 Solver: $newton-cg$	$0{,}908 \pm 0{,}012$	$0{,}910 \pm 0{,}013$	$0{,}908 \pm 0{,}012$	$0{,}908 \pm 0{,}013$
Linear SVM	C: 0,075	$0{,}915 \pm 0{,}012$	$0{,}918 \pm 0{,}012$	$0{,}915 \pm 0{,}012$	$0{,}915 \pm 0{,}012$
No Linear SVM	C: 10,0 Gamma: 0,001	$0{,}916 \pm 0{,}012$	$0{,}919 \pm 0{,}012$	$0{,}916 \pm 0{,}012$	$0{,}916 \pm 0{,}012$
*k*NN	Neigh: 1	$0{,}759 \pm 0{,}019$	$0{,}758 \pm 0{,}019$	$0{,}716 \pm 0{,}019$	$0{,}71 \pm 0{,}019$
DT	Max Depth: 12 Min Samples Split: 15	$0{,}832 \pm 0{,}015$	$0{,}839 \pm 0{,}014$	$0{,}832 \pm 0{,}015$	$0{,}831 \pm 0{,}015$
Random Forest	Estimators: 305	$0{,}853 \pm 0{,}017$	$0{,}855 \pm 0{,}016$	$0{,}853 \pm 0{,}016$	$0{,}852 \pm 0{,}017$
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 125	$0{,}903 \pm 0{,}010$	$0{,}904 \pm 0{,}010$	$0{,}903 \pm 0{,}010$	$0{,}903 \pm 0{,}010$

Tabla C9: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características binarias, subconjuntos de pacientes sanos y crónicos y características seleccionadas por la prueba F-Fisher.

Selección de características: F-Fisher - BBDD de pacientes crónicos - Características clínicas binarias					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 0,075 Solver: $newton-cg$	$0{,}903 \pm 0{,}014$	$0{,}904 \pm 0{,}014$	$0{,}903 \pm 0{,}014$	$0{,}903 \pm 0{,}014$
Linear SVM	C: 0,025	$0{,}907 \pm 0{,}013$	$0{,}910 \pm 0{,}013$	$0{,}907 \pm 0{,}013$	$0{,}907 \pm 0{,}013$
No Linear SVM	C: 10,0 Gamma: 0,001	$0{,}909 \pm 0{,}012$	$0{,}911 \pm 0{,}012$	$0{,}909 \pm 0{,}012$	$0{,}909 \pm 0{,}012$
*k*NN	Neigh: 7	$0{,}783 \pm 0{,}019$	$0{,}815 \pm 0{,}018$	$0{,}750 \pm 0{,}019$	$0{,}739 \pm 0{,}021$
DT	Max Depth: 14 Min Samples Split: 11	$0{,}818 \pm 0{,}018$	$0{,}826 \pm 0{,}018$	$0{,}818 \pm 0{,}018$	$0{,}818 \pm 0{,}018$
Random Forest	Estimators: 100	$0{,}852 \pm 0{,}016$	$0{,}858 \pm 0{,}016$	$0{,}852 \pm 0{,}016$	$0{,}852 \pm 0{,}016$
MLP	Activation: $logistic$ Solver: sgd Hidden Layers: 100	$0{,}880 \pm 0{,}015$	$0{,}879 \pm 0{,}016$	$0{,}880 \pm 0{,}015$	$0{,}879 \pm 0{,}016$

Tabla C10: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características binarias, subconjuntos de pacientes crónicos y características seleccionadas por la prueba F-Fisher.

Selección de características: Random Forest - BBDD de pacientes sanos y crónicos - Características clínicas binarias					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 0,05 Solver: $newton-cg$	$0{,}877 \pm 0{,}014$	$0{,}877 \pm 0{,}015$	$0{,}877 \pm 0{,}014$	$0{,}876 \pm 0{,}014$
Linear SVM	C: 0,025	$0{,}888 \pm 0{,}013$	$0{,}890 \pm 0{,}013$	$0{,}888 \pm 0{,}013$	$0{,}887 \pm 0{,}013$
No Linear SVM	C: 10,0 Gamma: 0,001	$0{,}889 \pm 0{,}013$	$0{,}892 \pm 0{,}013$	$0{,}889 \pm 0{,}013$	$0{,}889 \pm 0{,}013$
*k*NN	Neigh: 1	$0{,}673 \pm 0{,}018$	$0{,}672 \pm 0{,}019$	$0{,}627 \pm 0{,}018$	$0{,}616 \pm 0{,}020$
DT	Max Depth: 14 Min Samples Split: 11	$0{,}816 \pm 0{,}015$	$0{,}817 \pm 0{,}016$	$0{,}816 \pm 0{,}015$	$0{,}814 \pm 0{,}015$
Random Forest	Estimators: 92	$0{,}856 \pm 0{,}017$	$0{,}858 \pm 0{,}016$	$0{,}856 \pm 0{,}017$	$0{,}855 \pm 0{,}017$
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 135	$0{,}874 \pm 0{,}013$	$0{,}875 \pm 0{,}013$	$0{,}874 \pm 0{,}013$	$0{,}874 \pm 0{,}013$

Tabla C11: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características binarias, subconjuntos de pacientes sanos y crónicos y características seleccionadas por *Random Forest*

Selección de características: RandomForest - BBDD de pacientes crónicos - Características clínicas binarias					
Algoritmo	Parámetros libres	Tasa de acierto	Tasa de precisión	Tasa de exactitud	Tasa de F1 Score
RLM	C: 0,25 Solver: $newtong-cg$	$0{,}869 \pm 0{,}016$	$0{,}87 \pm 0{,}016$	$0{,}869 \pm 0{,}016$	$0{,}868 \pm 0{,}016$
Linear SVM	C: 0,05	$0{,}874 \pm 0{,}014$	$0{,}877 \pm 0{,}014$	$0{,}874 \pm 0{,}014$	$0{,}874 \pm 0{,}014$
No Linear SVM	C: 10,0 Gamma: 0,001	$0{,}877 \pm 0{,}014$	$0{,}881 \pm 0{,}014$	$0{,}877 \pm 0{,}014$	$0{,}877 \pm 0{,}014$
*k*NN	Neigh: 7	$0{,}699 \pm 0{,}018$	$0{,}743 \pm 0{,}019$	$0{,}658 \pm 0{,}018$	$0{,}627 \pm 0{,}023$
DT	Max Depth: 14 Min Samples Split: 47	$0{,}801 \pm 0{,}021$	$0{,}807 \pm 0{,}024$	$0{,}801 \pm 0{,}021$	$0{,}801 \pm 0{,}021$
Random Forest	Estimators: 71	$0{,}842 \pm 0{,}016$	$0{,}845 \pm 0{,}016$	$0{,}842 \pm 0{,}016$	$0{,}841 \pm 0{,}016$
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 140	$0{,}861 \pm 0{,}016$	$0{,}862 \pm 0{,}016$	$0{,}861 \pm 0{,}016$	$0{,}861 \pm 0{,}016$

Tabla C12: Media y desviación típica para cuatro medidas de prestaciones cuando se aplican diferentes algoritmos multi-clase con características binarias, subconjuntos de pacientes crónicos y características seleccionadas por *Random Forest*

C.2. Resultados de experimentos multi-etiqueta

Selección de características: Frecuencia - BBDD de pacientes sanos y crónicos - Características clínicas basadas en ocurrencia									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 1,0	0,843 ± 0,014	0,062 ± 0,007	0,916 ± 0,013	0,951 ± 0,009	0,865 ± 0,018	0,914 ± 0,012	0,889 ± 0,013	0,932 ± 0,008
Linear SVM OneVsRest	C: 5,0	0,852 ± 0,015	0,053 ± 0,006	0,937 ± 0,014	0,963 ± 0,008	0,862 ± 0,018	0,921 ± 0,010	0,895 ± 0,013	0,941 ± 0,007
No Linear SVM Label Powerset	C: 7,5 Gamma: 0,01	0,803 ± 0,018	0,08 ± 0,008	0,87 ± 0,015	0,918 ± 0,010	0,874 ± 0,017	0,910 ± 0,011	0,871 ± 0,013	0,914 ± 0,008
No Linear SVM OneVsRest	C: 10,0 Gamma: 0,01	0,813 ± 0,016	0,070 ± 0,007	0,894 ± 0,013	0,936 ± 0,009	0,859 ± 0,017	0,913 ± 0,011	0,875 ± 0,012	0,924 ± 0,008
kNN	Neigh: 7	0,670 ± 0,018	0,174 ± 0,008	0,908 ± 0,027	0,928 ± 0,011	0,573 ± 0,018	0,680 ± 0,015	0,665 ± 0,022	0,785 ± 0,011
DT	Max Depth: 74 Min Samples Split: 25	0,816 ± 0,016	0,071 ± 0,006	0,892 ± 0,013	0,938 ± 0,009	0,848 ± 0,018	0,908 ± 0,011	0,868 ± 0,012	0,923 ± 0,007
Random Forest	Estimators: 50	0,847 ± 0,014	0,054 ± 0,005	0,945 ± 0,013	0,960 ± 0,007	0,852 ± 0,016	0,924 ± 0,008	0,888 ± 0,012	0,942 ± 0,005
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 185	0,799 ± 0,016	0,074 ± 0,007	0,904 ± 0,015	0,942 ± 0,008	0,838 ± 0,016	0,896 ± 0,011	0,868 ± 0,013	0,919 ± 0,007

Tabla C13: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características en ocurrencia, subconjuntos de pacientes sanos y crónicos y características seleccionadas por frecuencia.

Selección de características: Frecuencia - BBDD de pacientes crónicos - Características clínicas basadas en ocurrencia									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 0,25	0,827 ± 0,019	0,07 ± 0,008	0,925 ± 0,016	0,956 ± 0,009	0,874 ± 0,017	0,923 ± 0,011	0,897 ± 0,012	0,939 ± 0,007
Linear SVM OneVsRest	C: 2,5	0,822 ± 0,015	0,065 ± 0,006	0,945 ± 0,015	0,965 ± 0,008	0,867 ± 0,014	0,922 ± 0,009	0,901 ± 0,010	0,943 ± 0,005
No Linear SVM Label Powerset	C: 10,0 Gamma: 0,01	0,792 ± 0,018	0,087 ± 0,007	0,876 ± 0,013	0,923 ± 0,010	0,889 ± 0,014	0,929 ± 0,008	0,882 ± 0,010	0,926 ± 0,006
No Linear SVM OneVsRest	C: 10,0 Gamma: 0,01	0,784 ± 0,019	0,080 ± 0,008	0,901 ± 0,014	0,939 ± 0,009	0,869 ± 0,015	0,923 ± 0,010	0,884 ± 0,011	0,931 ± 0,007
*k*NN	Neigh: 36	0,638 ± 0,026	0,167 ± 0,010	0,934 ± 0,019	0,912 ± 0,012	0,638 ± 0,013	0,79 ± 0,013	0,669 ± 0,015	0,846 ± 0,010
DT	Max Depth: 28 Min Samples Split: 23	0,797 ± 0,019	0,079 ± 0,008	0,9 ± 0,016	0,943 ± 0,011	0,856 ± 0,017	0,919 ± 0,011	0,875 ± 0,012	0,931 ± 0,007
Random Forest	Estimators: 80	0,825 ± 0,018	0,062 ± 0,007	0,948 ± 0,013	0,960 ± 0,008	0,862 ± 0,017	0,933 ± 0,008	0,895 ± 0,013	0,946 ± 0,006
MLP	Activation: $relu$ Solver: $adam$ Hidden Layers: 180	0,753 ± 0,023	0,093 ± 0,009	0,891 ± 0,021	0,927 ± 0,015	0,850 ± 0,023	0,913 ± 0,012	0,867 ± 0,015	0,920 ± 0,007

Tabla C14: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características en ocurrencia, subconjuntos de pacientes crónicos y características seleccionadas por frecuencia.

Selección de características: Fisher - BBDD de pacientes sanos y crónicos - Características clínicas basadas en ocurrencia									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 5,0	0,892 ± 0,012	0,042 ± 0,005	0,958 ± 0,012	0,969 ± 0,008	0,913 ± 0,014	0,939 ± 0,009	0,934 ± 0,009	0,954 ± 0,006
Linear SVM OneVsRest	C: 0,5	0,895 ± 0,013	0,038 ± 0,005	0,974 ± 0,007	0,979 ± 0,005	0,917 ± 0,014	0,94 ± 0,009	0,944 ± 0,008	0,959 ± 0,005
No Linear SVM Label Powerset	C: 10,0 Gamma: 0,01	0,85 ± 0,015	0,059 ± 0,007	0,916 ± 0,012	0,945 ± 0,009	0,908 ± 0,014	0,927 ± 0,011	0,911 ± 0,010	0,936 ± 0,007
No Linear SVM OneVsRest	C: 10,0 Gamma: 0,01	0,864 ± 0,018	0,051 ± 0,007	0,948 ± 0,014	0,963 ± 0,010	0,930 ± 0,012	0,949 ± 0,008	0,939 ± 0,009	0,956 ± 0,006
*k*NN	Neigh: 3	0,743 ± 0,019	0,138 ± 0,009	0,924 ± 0,015	0,934 ± 0,009	0,684 ± 0,020	0,759 ± 0,016	0,773 ± 0,018	0,837 ± 0,012
DT	Max Depth: 32 Min Samples Split: 25	0,831 ± 0,015	0,066 ± 0,006	0,908 ± 0,012	0,945 ± 0,008	0,856 ± 0,021	0,912 ± 0,011	0,879 ± 0,014	0,928 ± 0,007
Random Forest	Estimators: 100	0,857 ± 0,012	0,050 ± 0,004	0,955 ± 0,012	0,965 ± 0,006	0,860 ± 0,014	0,928 ± 0,007	0,897 ± 0,010	0,946 ± 0,005
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 180	0,842 ± 0,015	0,058 ± 0,006	0,941 ± 0,011	0,957 ± 0,008	0,886 ± 0,015	0,917 ± 0,011	0,912 ± 0,011	0,937 ± 0,007

Tabla C15: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características en ocurrencia, subconjuntos de pacientes sanos y crónicos y características seleccionadas por prueba F-Fisher.

Selección de características: Fisher - BBDD de pacientes crónicos - Características clínicas basadas en ocurrencia									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 1,0	0,886 ± 0,015	0,045 ± 0,007	0,964 ± 0,013	0,973 ± 0,008	0,925 ± 0,015	0,949 ± 0,009	0,943 ± 0,010	0,961 ± 0,006
Linear SVM OneVsRest	C: 2,5	0,878 ± 0,015	0,044 ± 0,006	0,968 ± 0,011	0,975 ± 0,007	0,931 ± 0,011	0,950 ± 0,007	0,948 ± 0,008	0,962 ± 0,005
No Linear SVM Label Powerset	C: 10,0 Gamma: 0,01	0,864 ± 0,016	0,056 ± 0,007	0,935 ± 0,015	0,957 ± 0,011	0,930 ± 0,011	0,948 ± 0,008	0,933 ± 0,008	0,952 ± 0,006
No Linear SVM OneVsRest	C: 10,0 Gamma: 0,01	0,832 ± 0,014	0,063 ± 0,006	0,911 ± 0,012	0,942 ± 0,008	0,880 ± 0,013	0,922 ± 0,010	0,895 ± 0,010	0,932 ± 0,006
*k*NN	Neigh: 3	0,750 ± 0,022	0,139 ± 0,010	0,927 ± 0,015	0,936 ± 0,009	0,739 ± 0,018	0,817 ± 0,013	0,808 ± 0,015	0,872 ± 0,009
DT	Max Depth: 30 Min Samples Split: 7	0,807 ± 0,019	0,076 ± 0,009	0,910 ± 0,017	0,948 ± 0,010	0,865 ± 0,017	0,920 ± 0,011	0,886 ± 0,013	0,934 ± 0,008
Random Forest	Estimators: 95	0,842 ± 0,017	0,056 ± 0,006	0,958 ± 0,013	0,968 ± 0,008	0,869 ± 0,014	0,936 ± 0,007	0,904 ± 0,011	0,951 ± 0,005
MLP	Activation: $relu$ Solver: $adam$ Hidden Layers: 120	0,823 ± 0,025	0,067 ± 0,010	0,937 ± 0,017	0,949 ± 0,015	0,904 ± 0,015	0,936 ± 0,010	0,919 ± 0,010	0,943 ± 0,008

Tabla C16: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características en ocurrencia, subconjuntos de pacientes crónicos y características seleccionadas por prueba F-Fisher.

Selección de características: Random Forest - BBDD de pacientes sanos y crónicos - Características clínicas basadas en ocurrencia									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 0,5	0,862 ± 0,012	0,054 ± 0,005	0,941 ± 0,013	0,962 ± 0,008	0,884 ± 0,014	0,921 ± 0,009	0,910 ± 0,009	0,941 ± 0,006
Linear SVM OneVsRest	C: 2,5	0,877 ± 0,014	0,044 ± 0,005	0,957 ± 0,011	0,971 ± 0,006	0,891 ± 0,014	0,933 ± 0,009	0,921 ± 0,009	0,952 ± 0,006
No Linear SVM Label Powerset	C: 10,0 Gamma: 0,01	0,821 ± 0,015	0,073 ± 0,007	0,881 ± 0,014	0,923 ± 0,011	0,892 ± 0,014	0,920 ± 0,010	0,886 ± 0,011	0,921 ± 0,008
No Linear SVM OneVsRest	C: 10,0 Gamma: 0,01	0,832 ± 0,014	0,063 ± 0,006	0,911 ± 0,012	0,942 ± 0,008	0,880 ± 0,013	0,922 ± 0,010	0,895 ± 0,010	0,932 ± 0,006
*k*NN	Neigh: 5	0,670 ± 0,019	0,171 ± 0,007	0,922 ± 0,019	0,928 ± 0,010	0,600 ± 0,016	0,688 ± 0,014	0,702 ± 0,017	0,79 ± 0,010
DT	Max Depth: 74 Min Samples Split: 5	0,812 ± 0,017	0,075 ± 0,007	0,888 ± 0,016	0,935 ± 0,010	0,847 ± 0,020	0,903 ± 0,012	0,866 ± 0,015	0,919 ± 0,008
Random Forest	Estimators: 37	0,855 ± 0,013	0,051 ± 0,005	0,948 ± 0,012	0,963 ± 0,006	0,860 ± 0,015	0,925 ± 0,008	0,895 ± 0,011	0,944 ± 0,005
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 135	0,815 ± 0,018	0,069 ± 0,007	0,919 ± 0,013	0,946 ± 0,008	0,861 ± 0,014	0,905 ± 0,010	0,888 ± 0,011	0,925 ± 0,008

Tabla C17: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características en ocurrencia, subconjuntos de pacientes sanos y crónicos y características seleccionadas por *Random Forest*.

Selección de características: Random Forest - BBDD de pacientes crónicos - Características clínicas basadas en ocurrencia									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 0,25	0,864 ± 0,019	0,056 ± 0,008	0,952 ± 0,014	0,966 ± 0,009	0,906 ± 0,015	0,937 ± 0,010	0,927 ± 0,012	0,952 ± 0,007
Linear SVM OneVsRest	C: 1,0	0,862 ± 0,019	0,050 ± 0,007	0,966 ± 0,013	0,973 ± 0,008	0,909 ± 0,015	0,94 ± 0,010	0,935 ± 0,011	0,956 ± 0,007
No Linear SVM Label Powerset	C: 10,0 Gamma: 0,01	0,826 ± 0,018	0,072 ± 0,008	0,949 ± 0,013	0,966 ± 0,009	0,869 ± 0,016	0,908 ± 0,010	0,906 ± 0,012	0,936 ± 0,007
No Linear SVM OneVsRest	C: 10,0 Gamma: 0,01	0,822 ± 0,021	0,067 ± 0,008	0,924 ± 0,014	0,948 ± 0,009	0,902 ± 0,017	0,937 ± 0,011	0,913 ± 0,013	0,943 ± 0,007
*k*NN	Neigh: 7	0,678 ± 0,020	0,168 ± 0,008	0,919 ± 0,014	0,914 ± 0,010	0,684 ± 0,018	0,787 ± 0,013	0,751 ± 0,017	0,846 ± 0,008
DT	Max Depth: 46 Min Samples Split: 43	0,805 ± 0,017	0,076 ± 0,007	0,914 ± 0,017	0,948 ± 0,009	0,855 ± 0,020	0,920 ± 0,012	0,880 ± 0,014	0,934 ± 0,007
Random Forest	Estimators: 66	0,827 ± 0,017	0,061 ± 0,006	0,955 ± 0,012	0,961 ± 0,007	0,863 ± 0,015	0,934 ± 0,007	0,897 ± 0,012	0,947 ± 0,005
MLP	Activation: $relu$ Solver: $adam$ Hidden Layers: 135	0,786 ± 0,023	0,081 ± 0,009	0,918 ± 0,018	0,939 ± 0,013	0,874 ± 0,022	0,922 ± 0,014	0,893 ± 0,014	0,930 ± 0,008

Tabla C18: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características en ocurrencia, subconjuntos de pacientes crónicos y características seleccionadas por *Random Forest*.

Selección de características: Frecuencia - BBDD de pacientes sanos y crónicos - Características clínicas binarias									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 0,025	0,854 ± 0,012	0,056 ± 0,005	0,916 ± 0,012	0,953 ± 0,007	0,876 ± 0,017	0,926 ± 0,010	0,895 ± 0,012	0,939 ± 0,006
Linear SVM OneVsRest	C: 0,025	0,854 ± 0,012	0,053 ± 0,005	0,930 ± 0,012	0,959 ± 0,007	0,865 ± 0,015	0,927 ± 0,008	0,893 ± 0,012	0,943 ± 0,005
No Linear SVM Label Powerset	C: 10,0 Gamma: 0,001	0,856 ± 0,013	0,055 ± 0,005	0,920 ± 0,011	0,955 ± 0,006	0,876 ± 0,018	0,926 ± 0,010	0,896 ± 0,012	0,940 ± 0,006
No Linear SVM OneVsRest	C: 10,0 Gamma: 0,001	0,855 ± 0,013	0,052 ± 0,005	0,937 ± 0,012	0,962 ± 0,007	0,860 ± 0,016	0,925 ± 0,008	0,892 ± 0,013	0,943 ± 0,005
*k*NN	Neigh: 7	0,676 ± 0,024	0,167 ± 0,010	0,962 ± 0,015	0,967 ± 0,008	0,569 ± 0,022	0,666 ± 0,021	0,681 ± 0,023	0,788 ± 0,015
DT	Max Depth: 14 Min Samples Split: 21	0,807 ± 0,018	0,075 ± 0,007	0,885 ± 0,018	0,933 ± 0,012	0,843 ± 0,019	0,905 ± 0,011	0,862 ± 0,013	0,919 ± 0,007
Random Forest	Estimators: 40	0,833 ± 0,013	0,059 ± 0,005	0,939 ± 0,013	0,955 ± 0,006	0,843 ± 0,017	0,918 ± 0,009	0,879 ± 0,013	0,936 ± 0,005
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 155	0,844 ± 0,014	0,057 ± 0,005	0,915 ± 0,012	0,950 ± 0,007	0,873 ± 0,017	0,927 ± 0,009	0,892 ± 0,012	0,938 ± 0,006

Tabla C19: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características binarias, subconjuntos de pacientes sanos y crónicos y características seleccionadas por frecuencia.

Selección de características: Frecuencia - BBDD de pacientes crónicos - Características clínicas binarias									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 0,025	0,838 ± 0,014	0,063 ± 0,006	0,924 ± 0,013	0,957 ± 0,007	0,884 ± 0,014	0,934 ± 0,008	0,902 ± 0,010	0,946 ± 0,005
Linear SVM OneVsRest	C: 0,0075	0,828 ± 0,017	0,063 ± 0,006	0,949 ± 0,013	0,967 ± 0,007	0,858 ± 0,017	0,924 ± 0,008	0,895 ± 0,013	0,945 ± 0,006
No Linear SVM Label Powerset	C: 7,5 Gamma: 0,001	0,840 ± 0,014	0,062 ± 0,006	0,928 ± 0,013	0,960 ± 0,007	0,881 ± 0,015	0,932 ± 0,009	0,903 ± 0,011	0,946 ± 0,005
No Linear SVM OneVsRest	C: 5,0 Gamma: 0,001	0,830 ± 0,015	0,062 ± 0,006	0,95 ± 0,012	0,968 ± 0,006	0,859 ± 0,017	0,925 ± 0,008	0,896 ± 0,013	0,946 ± 0,005
*k*NN	Neigh: 17	0,731 ± 0,018	0,131 ± 0,006	0,975 ± 0,009	0,968 ± 0,006	0,66 ± 0,012	0,803 ± 0,009	0,716 ± 0,015	0,878 ± 0,006
DT	Max Depth: 12 Min Samples Split: 13	0,790 ± 0,020	0,082 ± 0,008	0,893 ± 0,020	0,940 ± 0,012	0,854 ± 0,018	0,919 ± 0,009	0,871 ± 0,013	0,929 ± 0,007
Random Forest	Estimators: 95	0,817 ± 0,017	0,064 ± 0,006	0,947 ± 0,013	0,958 ± 0,006	0,855 ± 0,015	0,930 ± 0,008	0,889 ± 0,013	0,944 ± 0,005
MLP	Activation: $logistic$ Solver: sgd Hidden Layers: 180	0,803 ± 0,016	0,071 ± 0,006	0,915 ± 0,013	0,945 ± 0,007	0,877 ± 0,016	0,932 ± 0,008	0,893 ± 0,012	0,938 ± 0,005

Tabla C20: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características binarias, subconjuntos de pacientes crónicos y características seleccionadas por freceuncia.

Selección de características: Fisher - BBDD de pacientes sanos y crónicos - Características clínicas binarias									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 0,025	0,91 ± 0,009	0,033 ± 0,003	0,968 ± 0,010	0,975 ± 0,006	0,926 ± 0,011	0,953 ± 0,007	0,946 ± 0,006	0,964 ± 0,004
Linear SVM OneVsRest	C: 0,1	0,912 ± 0,010	0,031 ± 0,003	0,964 ± 0,011	0,971 ± 0,007	0,935 ± 0,009	0,963 ± 0,005	0,948 ± 0,006	0,967 ± 0,004
No Linear SVM Label Powerset	C: 7,5 Gamma: 0,001	0,911 ± 0,010	0,032 ± 0,004	0,971 ± 0,008	0,977 ± 0,005	0,928 ± 0,011	0,953 ± 0,007	0,948 ± 0,007	0,965 ± 0,004
No Linear SVM OneVsRest	C: 5,0 Gamma: 0,001	0,909 ± 0,010	0,032 ± 0,003	0,969 ± 0,008	0,973 ± 0,006	0,931 ± 0,010	0,959 ± 0,006	0,948 ± 0,006	0,966 ± 0,004
*k*NN	Neigh: 5	0,81 ± 0,016	0,108 ± 0,005	0,97 ± 0,007	0,964 ± 0,006	0,736 ± 0,014	0,798 ± 0,011	0,826 ± 0,011	0,873 ± 0,007
DT	Max Depth: 18 Min Samples Split: 13	0,835 ± 0,013	0,063 ± 0,005	0,913 ± 0,014	0,946 ± 0,008	0,864 ± 0,016	0,917 ± 0,010	0,886 ± 0,010	0,931 ± 0,006
Random Forest	Estimators: 100	0,855 ± 0,015	0,051 ± 0,005	0,953 ± 0,011	0,963 ± 0,006	0,860 ± 0,016	0,927 ± 0,009	0,897 ± 0,013	0,945 ± 0,006
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 155	0,903 ± 0,011	0,035 ± 0,004	0,955 ± 0,012	0,968 ± 0,007	0,934 ± 0,011	0,958 ± 0,006	0,944 ± 0,008	0,963 ± 0,004

Tabla C21: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características binarias, subconjuntos de pacientes sanos y crónicos y características seleccionadas por la preuba F-Fisher.

Selección de características: Fisher - BBDD de pacientes crónicos - Características clínicas binarias									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 0,1	0,900 ± 0,012	0,038 ± 0,005	0,970 ± 0,009	0,977 ± 0,005	0,931 ± 0,012	0,957 ± 0,007	0,949 ± 0,007	0,967 ± 0,004
Linear SVM OneVsRest	C: 0,05	0,889 ± 0,013	0,039 ± 0,005	0,968 ± 0,007	0,974 ± 0,005	0,936 ± 0,013	0,959 ± 0,007	0,951 ± 0,007	0,967 ± 0,004
No Linear SVM Label Powerset	C: 10,0 Gamma: 0,001	0,903 ± 0,012	0,036 ± 0,005	0,974 ± 0,007	0,978 ± 0,005	0,935 ± 0,011	0,959 ± 0,006	0,953 ± 0,007	0,968 ± 0,004
No Linear SVM OneVsRest	C: 5,0 Gamma: 0,01	0,889 ± 0,015	0,040 ± 0,006	0,954 ± 0,010	0,965 ± 0,007	0,944 ± 0,011	0,966 ± 0,006	0,949 ± 0,008	0,966 ± 0,005
*k*NN	Neigh: 5	0,837 ± 0,019	0,093 ± 0,007	0,970 ± 0,006	0,965 ± 0,006	0,806 ± 0,017	0,873 ± 0,011	0,869 ± 0,013	0,917 ± 0,007
DT	Max Depth: 12 Min Samples Split: 5	0,820 ± 0,016	0,069 ± 0,006	0,929 ± 0,014	0,957 ± 0,008	0,864 ± 0,017	0,923 ± 0,009	0,892 ± 0,013	0,940 ± 0,006
Random Forest	Estimators: 60	0,835 ± 0,016	0,058 ± 0,006	0,962 ± 0,010	0,965 ± 0,005	0,866 ± 0,017	0,935 ± 0,008	0,902 ± 0,014	0,95 ± 0,005
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 145	0,883 ± 0,015	0,042 ± 0,005	0,958 ± 0,008	0,969 ± 0,005	0,938 ± 0,012	0,960 ± 0,007	0,947 ± 0,008	0,964 ± 0,005

Tabla C22: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características binarias, subconjuntos de pacientes crónicos y características seleccionadas por la prueba F-Fisher.

Selección de características: *Random Forest* - BBDD de pacientes sanos y crónicos - Características clínicas binarias									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 0,025	0,886 ± 0,013	0,043 ± 0,005	0,942 ± 0,012	0,965 ± 0,007	0,901 ± 0,015	0,942 ± 0,009	0,920 ± 0,010	0,953 ± 0,005
Linear SVM OneVsRest	C: 0,075	0,891 ± 0,011	0,038 ± 0,004	0,947 ± 0,012	0,966 ± 0,006	0,905 ± 0,014	0,951 ± 0,007	0,923 ± 0,010	0,959 ± 0,005
No Linear SVM Label Powerset	C: 10,0 Gamma: 0,001	0,889 ± 0,013	0,042 ± 0,005	0,948 ± 0,012	0,969 ± 0,007	0,900 ± 0,016	0,941 ± 0,009	0,922 ± 0,011	0,955 ± 0,005
No Linear SVM OneVsRest	C: 2,5 Gamma: 0,01	0,883 ± 0,012	0,042 ± 0,005	0,930 ± 0,013	0,955 ± 0,008	0,914 ± 0,014	0,956 ± 0,007	0,921 ± 0,010	0,955 ± 0,005
*k*NN	Neigh: 3	0,71 ± 0,020	0,147 ± 0,008	0,959 ± 0,010	0,958 ± 0,009	0,647 ± 0,022	0,716 ± 0,017	0,757 ± 0,019	0,819 ± 0,011
DT	Max Depth: 10 Min Samples Split: 17	0,823 ± 0,013	0,068 ± 0,005	0,923 ± 0,014	0,952 ± 0,009	0,837 ± 0,019	0,901 ± 0,012	0,874 ± 0,011	0,926 ± 0,006
Random Forest	Estimators: 72	0,833 ± 0,013	0,058 ± 0,005	0,953 ± 0,010	0,959 ± 0,006	0,838 ± 0,017	0,914 ± 0,009	0,881 ± 0,013	0,936 ± 0,005
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 170	0,877 ± 0,015	0,044 ± 0,005	0,935 ± 0,014	0,960 ± 0,008	0,906 ± 0,014	0,945 ± 0,008	0,919 ± 0,011	0,952 ± 0,006

Tabla C23: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características binarias, subconjuntos de pacientes sanos y crónicos y características seleccionadas por *Random Forest*.

Selección de características: *Random Forest* - BBDD de pacientes crónicos - Características clínicas binarias									
Algoritmo	Parámetros libres	Tasa de acierto	Hamming Loss	Tasa de precisión (macro)	Tasa de precisión (micro)	Tasa de exactitud (micro)	Tasa de exactitud (micro)	Tasa de F1-Score (macro)	Tasa de F1-Score (micro)
Linear SVM Label Powerset	C: 0,025	0,883 ± 0,010	0,044 ± 0,004	0,960 ± 0,012	0,974 ± 0,006	0,913 ± 0,013	0,949 ± 0,007	0,935 ± 0,008	0,961 ± 0,004
Linear SVM OneVsRest	C: 0,075	0,882 ± 0,012	0,042 ± 0,005	0,962 ± 0,009	0,974 ± 0,006	0,914 ± 0,012	0,953 ± 0,007	0,936 ± 0,008	0,964 ± 0,004
No Linear SVM Label Powerset	C: 10,0 Gamma: 0,001	0,886 ± 0,010	0,043 ± 0,004	0,964 ± 0,010	0,976 ± 0,006	0,913 ± 0,013	0,950 ± 0,007	0,937 ± 0,007	0,963 ± 0,004
No Linear SVM OneVsRest	C: 7,5 Gamma: 0,001	0,881 ± 0,013	0,042 ± 0,005	0,972 ± 0,007	0,977 ± 0,005	0,908 ± 0,014	0,950 ± 0,007	0,936 ± 0,009	0,963 ± 0,004
*k*NN	Neigh: 3	0,777 ± 0,019	0,123 ± 0,008	0,961 ± 0,009	0,958 ± 0,007	0,749 ± 0,016	0,825 ± 0,011	0,827 ± 0,014	0,887 ± 0,008
DT	Max Depth: 10 Min Samples Split: 17	0,812 ± 0,017	0,072 ± 0,007	0,927 ± 0,016	0,956 ± 0,008	0,854 ± 0,019	0,920 ± 0,011	0,885 ± 0,013	0,937 ± 0,006
Random Forest	Estimators: 50	0,818 ± 0,016	0,064 ± 0,006	0,956 ± 0,009	0,959 ± 0,006	0,856 ± 0,017	0,931 ± 0,008	0,891 ± 0,014	0,945 ± 0,005
MLP	Activation: $relu$ Solver: sgd Hidden Layers: 190	0,868 ± 0,012	0,047 ± 0,005	0,947 ± 0,011	0,966 ± 0,007	0,921 ± 0,012	0,953 ± 0,007	0,933 ± 0,008	0,959 ± 0,004

Tabla C24: Media y desviación típica para ocho medidas de prestaciones cuando se aplican diferentes algoritmos multi-etiqueta con características binarias, subconjuntos de pacientes crónicos y características seleccionadas por *Random Forest*.

Anexo D: Diagrama de Gantt y presupuesto del trabajo

En este anexo se muestra la cronología del trabajo y se calculan, de forma aproximada, los costes que han supuesto la realización del trabajo.

D.1. Diagrama de Gantt

En la Figura D1 se muestra la evolución cronológica de este trabajo con un diagrama de Gantt.

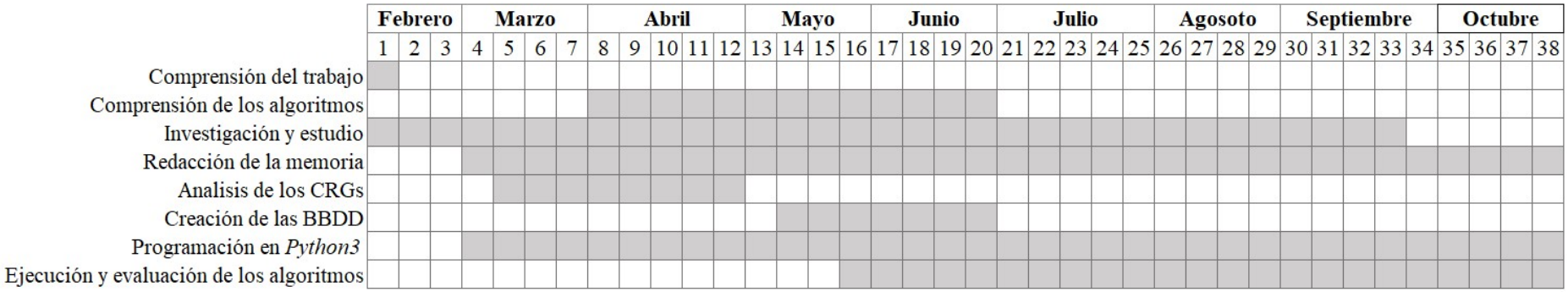


Figura D1: Diagrama de Gantt.

D.2. Presupuesto del trabajo

En esta sección intentaremos estimar el coste del trabajo realizado si hubiera sido desarrollado profesionalmente. El coste total lo vamos a dividir en costes directos e indirectos.

D.2.1. Costes directos

Los costes directos son aquéllos que atañen directamente a la realización del proyecto, entre ellos están los costes materiales, de personal y los de dirección.

D.2.1.1. Costes Materiales

Abarcamos los costes asociados al *hardware* y al *sotfware* empleados en este trabajo. En la Tabla D1 se muestran los diferentes conceptos y su valor monetario estimado. Una aclaración sobre la Tabla D1 es que se han usado 4 ordenadores a pleno rendimiento para la ejecución de los algoritmos.

Concepto	Coste €
4 ordenadores sobremesa	800×4
Licencia Ubuntu 18.04.3	0
Licencia *Python* 3	0
Licencia LaTeX	0
Coste total material	3220

Tabla D1: Coste del material utilizado.

D.2.1.2. Coste Personal

Las horas invertidas en este trabajo se estiman como una jornada a tiempo parcial ($30h/semana$), es decir, el $75\,\%$ de una jornada laboral completa ($40h/semana$). Consultando en el mercado sobre el salario actual de un Ingeniero Junior de Telecomunicaciones recién titulado es de unos $21{,}000$ €brutos anuales. Tal y como se indica en la Sección D.1, el periodo de tiempo invertido han sido de 10 meses ($83{,}3\,\%$ de un años). La Fórmula ID.1 nos muestra el coste personal:

$$\begin{aligned} Coste_{personal} = €\ brutos\ anuales \times Tiempo\ Total\ Invertido(\,\%) \times \\ Tiempo\ Semanal\ Invertido(\,\%) = 21000 \times 0{,}833 \times 0{,}75 = 13120€ \end{aligned} \tag{ID.1}$$

D.2.1.3. Coste Dirección

En un proyecto de ingeniería, el salario del director de proyecto se pondera con el $7\,\%$ de la suma resultante de los costes material y personal. El resultado se muestra en la Fórmula ID.2. Hay que tener en cuenta que en este trabajo han invertido su tiempo y conocimiento dos tutoras.

$$\begin{aligned} Coste_{direccion} = 2 \times 0{,}07 \times (Coste_{material} + Coste_{personal}) = \\ 2 \times 0{,}07 \times (3220 + 13120) = 2288€ \end{aligned} \tag{ID.2}$$

D.2.2. Costes Indirectos

Los costes indirectos agrupan gastos que no son directamente imputables como producción, es decir, costes de fibra óptica, desplazamiento, material de oficina, calefacción, aire acondicionado y un largo etcétera. Se calcularán considerando un porcentaje estimado en un 12 % de los costes directos. La Fórmula ID.3 nos muestra su valor.

$$\begin{aligned} Costes_{indirectos} = 0{,}12 \times Costes_{directos} = \\ 0{,}12 \times 18628 = 2236€ \end{aligned} \tag{ID.3}$$

D.2.3. Coste total

En la Tabla D2 se desglosan los costes asociados a la realización de este trabajo. Hemos dejado un margen de un 2 % como previsión de algún imprevisto.

Concepto		Coste €
Costes Directos	Coste Material	3220
	Coste Personal	13120
	Coste Dirección	2288
Costes Indirectos		2236
Costes Imprevistos		418
Coste total trabajo		21282

Tabla D2: Costes totales trabajo.

Bibliografía

[1] Organización Mundial de la Salud (OMS), "ENT Perfiles de países en 2018," *Redacción Organización Mundial de la Salud*, Jan. 2019.

[2] ——, "Lucha contra las ENT - Mejores Inversiones," *Redacción Organización Mundial de la Salud*, 2018.

[3] Organización para la Cooperación y el Desarrollo Económicos (OCDE) y el European Observatory on Health Systems and Policies, "Perfil Sanitario de España en 2017," *Comisión Europea*, 2017.

[4] H. Silva-Cárcamo, "Comorbilidades en los pacientes con Diabetes Mellitus Tipo 2 del Instituto Nacional del Diabético, Tegucigalpa, Honduras," *Archivos de medicina*, 2016.

[5] 3M™, "3M™ Clinical Risk Groups: Measuring risk, managing care," *3M™ - Health Information Systems*, 2016.

[6] J. Landa. ¿Qué es KDD y Minería de Datos? Acceso: 2019-04-30. [Online]. Available: http://fcojlanda.me/es/ciencia-de-los-datos/kdd-y-mineria-de-datos-espanol/

[7] Machine Learning in Python. Choosing the right estimator. Acceso: 2019-03-02. [Online]. Available: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

[8] Phil Chang. Stackoverflow: how to implement walk forward testing in sklearn? Acceso: 2019-03-10. [Online]. Available: https://stackoverflow.com/questions/31947183/how-to-implement-walk-forward-testing-in-sklearn

[9] I. Katakis, G. Tsoumakas, and I. P. Vlahavas, *Data Mining and Knowledge Discovery Handbook*. Springer, 2009.

[10] F. S. Caparrini. Redes Neuronales: una visión superficial. Acceso: 2019-02-30. [Online]. Available: http://www.cs.us.es/~fsancho/?e=72

[11] Ministerio de Sanidad, Consumo y Bienestar Social, *Edición electrónica de la CIE-9-MC*.

[12] C. C. Pérez, "Usuarios generan 2,5 quintillones de bytes en datos diarios, dice comScore," *El Financiero*, 2014.

[13] I. Katakis, G. Tsoumakas, and I. P. Vlahavas, "Multilabel text classification for automated tag suggestion," in *Proceedings of the ECML/PKDD 2008 Discovery Challenge, Antwerp, 2008*, 2008.

[14] S. Diplaris, G. Tsoumakas, P. Mitkas, and I. Vlahavas, "Protein Classification with Multiple Algorithms," in *Adv. Informatics*, vol. 3746, 2005, pp. 448–456.

[15] M. Boutell, J. Luo, X. Shen, and C. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, pp. 1757–1771, 09 2004.

[16] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, Z. Lei, W. Cukierski, S. Frey, A. Hadley, M. Betts, X. Fern, J. Irvine, L. Neal, A. Thomas, G. Fodor, G. Tsoumakas, H. Wei Ng, T. Nguyen, H. Huttunen, P. Ruusuvuori, and M. Milakov, "The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment," in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 09 2013, pp. 1–8.

[17] M. Ufuk Dalmış, S. Vreemann, T. Kooi, R. Mann, N. Karssemeijer, and A. Gubern-Mérida, "Fully automated detection of breast cancer in screening MRI using convolutional neural networks," *Journal of Medical Imaging*, vol. 5, p. 1, 01 2018.

[18] A. Conner-Simons, "Using artificial intelligence to improve early breast cancer detection," *MIT News*, 2017.

[19] L. E. Juarez-Orozco, T. Maaniitty, J. Benjamins, M. Niemi, P. Van Der Harst, A. Saraste, and J. Knuuti, "Refining the long-term prognostic value of hybrid PET/CT through machine learning," *ICNC, Nuclear Cardiology and Cardiac CT 2019, Session Young Investigator Award Abstracts*, 2019.

[20] N. Barrett and J. Weber, "Applying natural language processing toolkits to electronic health records - an experience report," *Studies in health technology and informatics*, vol. 143, pp. 441–6, 02 2009.

[21] J.-S. Lin, P. Ligomenides, S.-C. Lo, M. Freedman, and S. Mun, "A hybrid neural-digital computer-aided diagnosis system for lung nodule detection on digitized chest radiographs," in *Computer-Based Medical Systems*, 07 1994, pp. 207 – 212.

[22] H. Wang, L. Chengde, Y. Peng, and X. Hu, "Application of Improved Random Forest Variables Importance Measure to Traditional Chinese Chronic Gastritis Diagnosis," in *IT in Medicine and Education, 2008. ITME 2008. IEEE International Symposium*, 01 2009, pp. 84 – 89.

[23] M. Mammadov, A. M Rubinov, and J. Yearwood, "The study of drug–reaction relationships using global optimization techniques," *Optimization Methods and Software*, vol. 22, pp. 99–126, 03 2007.

[24] Organización Mundial de la Salud (OMS). Enfermedades crónicas. Acceso: 2019-04-15. [Online]. Available: https://www.who.int/topics/chronic_diseases/es/

[25] ——. Enfermedades no transmisibles (o crónicas). Acceso: 2019-04-15. [Online]. Available: https://www.who.int/features/chronic_disease/es/

[26] Redacción EsCrónicos, "IV Barómetro: Encuesta sobre la calidad de la asistencia sanitaria a los pacientes crónicos en España," *EsCrónicos*, 2017.

[27] National Institutes of Health (NIH). High Blood Pressure. Acceso: 2019-04-16. [Online]. Available: https://medlineplus.gov/highbloodpressure.html

[28] ——. Diabetes. Acceso: 2019-04-16. [Online]. Available: https://medlineplus.gov/diabetes.html

[29] National Institute of Diabetes and Digestive and Kidney (NIDDK). High Blood Pressure and Kidney Disease. Acceso: 2019-04-16. [Online]. Available: https://www.niddk.nih.gov/health-information/kidney-disease/high-blood-pressure

[30] Organización Mundial de la Salud (OMS), "La Organización Mundial de la Salud (OMS) publica hoy su nueva Clasificación Internacional de Enfermedades (CIE-11)," *Redacción Organización Mundial de la Salud*, 2018.

[31] V. N. Slee, "The International Classification of Diseases: Ninth Revision (ICD-9)," *Ann Intern Med*, vol. 88, p. 424–426, 03 1978.

[32] VIDAL Group. Clasificación ATC. Acceso: 2019-04-15. [Online]. Available: https://www.vademecum.es/atc

[33] Sacyl. Estratificación de pacientes. Acceso: 2019-04-15. [Online]. Available: https://www.saludcastillayleon.es/transparencia/es/transparencia/informacion-datos-publicos/datos-interes/estratificacion-pacientes-clasificacion-grado-complejidad

[34] F. J. Gutiérrez Expósito, *Estratificación de pacientes crónicos diabéticos. Análisis estadístico de datos de dispensación farmacológica*, 2016, Trabajo Fin de Máster, Universidad Rey Juan Carlos.

[35] J. Fernández Sánchez, *Statistical Analysis of Demographic and Clinical Data from Healthy and Chronic Hypertensive Patients*, 2016, Trabajo Fin de Grado, Universidad Rey Juan Carlos.

[36] A. Alberca Díaz-Plaza, *Clasificación de pacientes crónicos diabéticos e hipertensos usando árboles de decisión*, 2017, Trabajo Fin de Grado, Universidad Rey Juan Carlos.

[37] E. Morales, "Clasificación Multi-Etiqueta," *Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)*, 2017.

[38] Francisco Dionicio Pichardo Morales, *Apuntes de Desbalance de clases en clasificación inteligente*. Clasificación Inteligente de Patrones, 2012.

[39] N. Spolaôr, E. Cherman, M.-C. Monard, and H. Lee, "A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach," *Electronic Notes in Theoretical Computer Science*, vol. 292, p. 135–151, 03 2013.

[40] G. Doquire and M. Verleysen, "Feature Selection for Multi-label Classification Problems," in *Advances in Computational Intelligence*, vol. 6691, 06 2011, pp. 9–16.

[41] O. Gharroudi, H. Elghazel, and A. Aussem, "A Comparison of Multi-Label Feature Selection Methods Using the Random Forest Paradigm," in *Electronic Notes in Theoretical Computer Science*, 05 2014, pp. 95–106.

[42] R. Shaikh, "Cross validation explained: Evaluating estimator performance." *Towards Data Science*, 11 2018.

[43] OneVsRestClassifier Method in Python. OneVsRestClassifier Method. Acceso: 2019-03-02. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

[44] Multi-Label Classification in Python. Label Powerset Method. Acceso: 2019-03-02. [Online]. Available: http://scikit.ml/api/skmultilearn.problem_transform.lp.html

[45] S. Polamuri, "How the multinomial logistic regression model works in machine learning," *Data Aspirant*, 03 2017.

[46] S. Wan, M.-W. Mak, and S.-Y. Kung, "Adaptive thresholding for multi-label SVM classification with application to protein subcellular localization prediction," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 10 2013, pp. 3547–3551.

[47] D. Srivastava and L. Bhambhu, "Data classification using support vector machine," *Journal of Theoretical and Applied Information Technology*, vol. 12, pp. 1–7, 02 2010.

[48] S. Raschka, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition*. Packt Publishing, 2017.

[49] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label leaming," *Pattern Recognition*, vol. 40, pp. 2038–2048, 07 2007.

[50] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," in *Lecture Notes in Computer Science*, vol. 2168, 11 2002.

[51] M.-L. Zhang and Z.-H. Zhou, "Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization," in *IEEE Transactions on Knowledge and Data Engineering (Volume: 18 , Issue: 10)*, 08 2006, pp. 1338 – 1351.

[52] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*, 12 2014.

[53] Jupyter. Proyect Jupyter. Acceso: 2019-03-02. [Online]. Available: https://jupyter.org/

[54] Python. Python 3.0 Releaser. Acceso: 2019-03-02. [Online]. Available: https://www.python.org/download/releases/3.0/

[55] Linuxize. How To Use Linux Screenr. Acceso: 2019-03-02. [Online]. Available: https://linuxize.com/post/how-to-use-linux-screen/