E.T.S. de Ingeniería Industrial,
Informática y de Telecomunicación

Elaboración de un sistema inteligente para la predicción de eventos adversos relacionados con la polimedicación en atención primaria

Máster Universitario en
Ingeniería Informática

Trabajo Fin de Máster

Alumno: Mikel Xabier Uriz Martin

Director: Mikel Galar Idoate

Pamplona, 24 de febrero de 2017

upna Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

RESUMEN

Actualmente todos los datos médicos son almacenados digitalmente para obtener un acceso más rápido al historial clínico de los pacientes. Analizando estos datos se pueden obtener patrones o reglas que nos permitan identificar mejor las causas de una determinada enfermedad para realizar un mejor diagnostico al paciente.

En este proyecto se ha trabajado con los datos de los pacientes Navarros (2013-2015) centrándonos en las personas polimedicadas (aquellas que tomen 5 o más medicamentos durante al menos 3 meses). El objetivo de este trabajo es analizar estos datos médicos, tanto de ingresos hospitalarios y enfermedades como de medicamentos expedidos, para detectar la probabilidad de que un paciente llegue a sufrir un evento adverso. En concreto nos hemos centrado en los eventos adversos cardiovasculares.

Palabras clave: KDD, Data Mining, Minería de datos, clasificación, polimedicado

Contenido

1. Introducción

La ciencia de datos es un campo interdisciplinario que tiene como objetivo extraer el conocimiento de los datos, es decir, emplear diferentes técnicas sobre una serie de datos para extraer información que a priori no es perceptible. Para extraer esta información se emplean técnicas de diferentes áreas como matemáticas, estadística, ciencias de la información, ciencias de la computación, bases de datos, aprendizaje automático, inteligencia artificial,… y minería de datos.

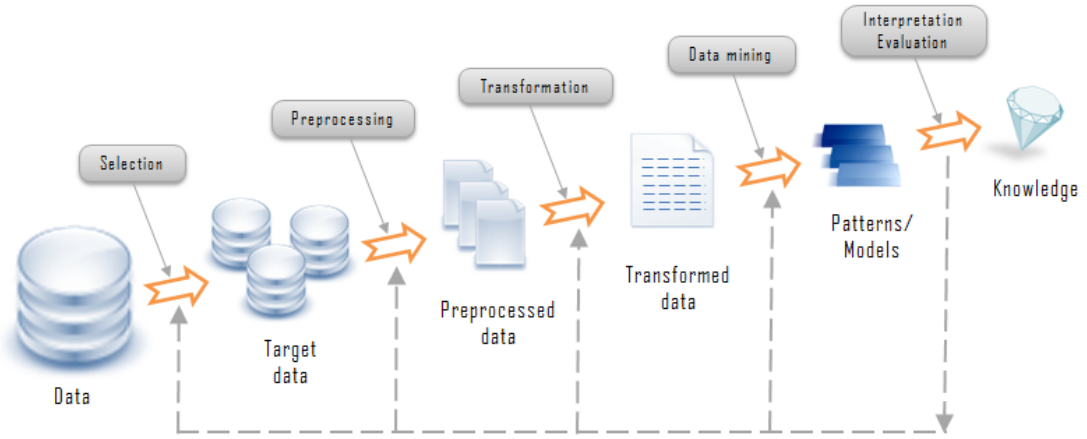


Figura 1. Fases del proceso KDD. Figura extraída de www.zentut.com

Este proyecto se ha tratado como un problema de minería de datos, donde se parte de una serie de datos a los que se les aplican diferentes técnicas con el objetivo de encontrar patrones y extraer el conocimiento. El objetivo es transformar este conocimiento en una estructura comprensible para su uso posterior, como por ejemplo para la predicción de futuros acontecimientos.

La minería de datos en realidad es solo una fase del proceso de descubrimiento de conocimiento en bases de datos (KDD "Knowledge Discovery in Databases") pero a menudo se utiliza el término minería de datos para referirse al proceso completo. Como se puede ver en la Figura 1, el proceso KDD se divide en 5 fases, donde una vez alcanzada la fase final se puede volver a cualquiera de las fases anteriores para realizar alguna modificación con el objetivo de obtener mejores resultados. Lo habitual en los proyectos de minería de datos es realizar esta transición. Las 5 fases del proceso KDD son:

- **Selección:** En esta fase se tienen los datos originales de los que se quiere extraer conocimiento. Sobre todo el conjunto de datos se eligen que datos van a ser utilizados y como se generan, es decir, se escogen las variables objetivo (variables a predecir) y las variables independientes (las que se utilizan para predecir).
- **Pre-procesamiento:** En esta fase se realiza en análisis y limpieza de los datos. Se trata de entender mejor las variables elegidas en la fase anterior mediante resúmenes y gráficos. El objetivo de esta fase es encontrar valores atípicos o ausencia de datos (datos desconocidos sobre alguna variable) para realizar el tratamiento adecuado. Para ello se visualizan los datos mediante histogramas, diagramas de dispersión,… sobre los datos.

- **Transformación:** Con los resultados del análisis de la fase anterior se aplican las técnicas adecuadas para obtener un conjunto de datos preparado para aplicar el o los algoritmos de minería de datos deseados.
- **Minería de datos:** En esta fase se aplican los algoritmos adecuados dependiendo del tipo de problema. Estos algoritmos generan un modelo que representa el conocimiento extraído de los datos.
- **Interpretación/evaluación:** En esta fase se comprueba que el modelo generado en la fase anterior sea adecuado, es decir, que las conclusiones que se obtienen sean suficientemente satisfactorias. En caso de que el modelo no sea suficientemente bueno se vuelve a alguna de las fases anteriores y se realizan modificaciones para obtener nuevos modelos.

Dentro de la minería de datos se encuentra el aprendizaje automático, que trata el diseño y desarrollo de algoritmos que permiten aprender comportamientos, patrones o conceptos sobre una serie de datos. De manera más simple, el aprendizaje automático trata de aprender futuros acontecimientos basándose en experiencias del pasado. El aprendizaje automático se puede dividir en dos categorías principales: aprendizaje supervisado y aprendizaje no supervisado.

En el aprendizaje supervisado los datos están etiquetados por la clase a la que pertenecen. El objetivo de este tipo de aprendizaje es predecir una salida conocida para cada uno de los ejemplos disponibles en el conjunto de entrenamiento. Con estos datos se aprende un modelo, el cual debería de ser capaz de generalizar correctamente lo aprendido. Para esto el modelo debería de ser capaz de predecir las salidas para ejemplos que no se han utilizado en la fase de entrenamiento. Dentro del aprendizaje supervisado se tienen 2 disciplinas principales: clasificación, donde la salida es una etiqueta del conjunto de clases, y regresión, donde la salida a estimar es un valor real.

En los problemas de aprendizaje no supervisado los ejemplos no están etiquetados. El objetivo de este tipo de problemas es modelar la estructura o la distribución en los datos para aprender más de los datos.

Los problemas de clasificación se pueden dividir en dos tipos dependiendo del número de clases que queremos distinguir: problemas binarios, cuando se tienen dos clases, y problemas multi-clase, cuando se tienen más de dos clases.

Este proyecto se ha planteado como un problema de aprendizaje supervisado, donde tratamos de catalogar los pacientes en dos clases posibles, por lo que se plantea un problema de clasificación y más concretamente un problema de clasificación binario.

Los algoritmos de aprendizaje automático pueden producir dos tipos de modelos diferentes: modelos de caja negra y modelos de caja blanca. Los primeros generan un modelo cuya representación interna no permite explicar las decisiones a las salidas dadas. Los de caja blanca, por el contrario, generan una estructura interna que es interpretable, es decir, que es comprensible por el ser humano.

Este proyecto consiste en la utilización de datos clínicos sobre pacientes para predecir futuros eventos sobre ellos. A los médicos no solo les interesa tener un modelo que prediga estos eventos, también están interesados en conocer las decisiones tomadas por el modelo para predecir el evento. Esto les puede ser de gran ayuda, ya que pueden observar las decisiones

tomadas por el algoritmo para llegar a la conclusión. Es por ello que en este proyecto nos hemos centrado más en este tipo de algoritmos aunque se han llegado a realizar pruebas y comparativas con los modelos de caja negra.

Los datos recibidos para la elaboración de este proyecto provienen de diferentes fuentes. Se tienen datos sobre los medicamentos expedidos a los pacientes, ingresos hospitalarios, información general sobre los pacientes,… Una de las características del Big Data es precisamente el hecho de que lo datos provienen de diferentes fuentes. Cuando hablamos del Big Data hablamos de datos con un gran volumen, donde su diversidad y complejidad requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer el valor y conocimiento.

Las técnicas tradicionales no pueden procesar este gran volumen de datos en un tiempo aceptable, es por ello que se requieren nuevas técnicas. Una máquina está limitada en cuanto a potencia de cómputo y aunque se le quiera dotar de mayor capacidad de cómputo muchas veces es imposible o muy caro. Dada esta limitación, la solución que se utiliza es la de emplear varias máquinas conectadas entre sí (normalmente por red), donde cada una de estas máquinas almacena una parte de los datos y estas reciben las técnicas que se quieren aplicar sobre los datos. Cada una realiza el proceso sobre los datos correspondientes y finalmente se agregan los resultados consiguiendo el mismo resultado que utilizando las técnicas tradicionales pero en un menor tiempo.

En este proyecto no se han empleado técnicas de Big Data. Aunque los datos provengan de diferentes fuentes, el volumen final a tratar no es exageradamente grande, pudiendo aplicar las técnicas de minería de datos tradicionales en un tiempo admisible.

Como se ha mencionado, los datos utilizados en este proyecto son datos clínicos, es decir, datos sobre pacientes que provienen de diversos ámbitos de la salud como pueden ser los ingresos hospitalarios o los medicamentos recetados. La recolección y el almacenamiento de este tipo de datos ha crecido mucho en los últimos años, generando una gran cantidad de datos en los historiales clínicos de pacientes. Estos datos contienen mucha información que a simple vista no es apreciable, pudiendo aplicar tecinas de minería de datos para extraer esa información. Por ejemplo, se podría buscar un patrón entre los medicamentos consumidos por los pacientes y una enfermedad concreta, es decir, llegar a la conclusión de que ciertos medicamentos están causando o aumentando la probabilidad de sufrir cierta enfermedad. Actualmente existen diferentes aplicaciones que se emplean sobre este tipo de datos:

- **Efectividad del tratamiento:** Se pueden utilizar la minería de datos para evaluar la efectividad de un tratamiento, ofreciendo un análisis de que acciones resultan eficaces comparando y contrastando las causas, síntomas y tratamientos.
- **Administración de la atención sanitaria:** La minería de datos ayuda a identificar y rastrear los estados de las enfermedades crónicas y de los pacientes de alto riesgo, diseñar intervenciones apropiadas y reducir el número de ingresos hospitalarios.
- **Detectar fraude y abuso:** Utilizando la minería de datos se pueden detectar recetas inapropiadas, seguros fraudulentos y reclamaciones médicas.
- **Identificar diagnósticos:** Se pueden emplear las técnicas de minería de datos para ayudar a identificar el diagnóstico de un paciente basándose en hechos pasados.
- **Identificar riesgos de enfermedades:** La minería de datos puede utilizarse para identificar o calcular el riesgo que tiene un paciente de sufrir alguna enfermedad teniendo en cuenta el historial clínico de este.

En concreto, este proyecto consiste en estudiar el aumento del riesgo de sufrir ciertos ingresos hospitalarios debido al consumo de varios medicamentos conjuntamente. Para ello se utiliza la información de pacientes navarros considerados polimedicados. Debemos destacar que no existe una definición concreta para paciente polimedicado, pero en este proyecto se ha acordado con la parte farmacéutica participante en el proyecto (Servicio Navarro de Salud – Osasunbidea, SNS-O), que un paciente polimedicado es aquel que consuma 5 o más medicamentos durante al menos 3 meses.

El consumo de varios medicamentos conjuntamente puede causar un evento no deseado, es decir, sufrir una reacción no deseada. A este tipo de eventos se les denomina eventos adversos. El objetivo de este proyecto es el de calcular o predecir el riesgo que tiene un paciente polimedicado de sufrir un evento adverso en el futuro, utilizando sobre todo los datos de los medicamentos. En concreto este proyecto se centra en los eventos adversos cardiovasculares, pero se han realizado experimentos extendiendo esta metodología a otro tipo de eventos adversos, en concreto a las hospitalizaciones potencialmente evitables (HPE).

2. Problema: Prevención de eventos adversos en pacientes polimedicados

El objetivo de este proyecto es predecir el riesgo que tiene un paciente de sufrir un evento adverso en base a varios datos clínicos, poniendo especial interés en los datos de los medicamentos. Este proyecto se ha realizado con la colaboración del Servicio Navarro de Salud – Osasunbidea (SNS-O). Estos fueron los que proporcionaron los datos y nos fueron guiando a lo largo del proyecto mediante sus conocimientos clínicos.

El proyecto comienza con la recepción de varios datos clínicos sobre pacientes de Navarra y con un objetivo: predecir el riesgo que tienen de sufrir un evento adverso. Tras la recepción estaríamos al comienzo del proceso KDD anteriormente explicado.

El primer paso fue estudiar los datos para decidir qué datos se van a utilizar y de qué manera. Con los datos dispuestos se planteó un problema de clasificación por lo que también se estableció cual era la variable a predecir. Una de las decisiones fue la de qué parte de la población navarra utilizar.

En este proyecto nos hemos centrado en pacientes navarros que sean polimedicados. De acuerdo a las indicaciones dadas por el equipo del SNS-O, un paciente polimedicado es aquel que tome 5 o más medicamentos durante al menos 3 meses consecutivos. Debemos hacer notar que cuando se detecta que un paciente es polimedicado, se considera que este va a ser polimedicado para todo el estudio, aunque después deje de serlo. Es importante considerar la fecha en la que el paciente empieza a ser polimedicado por la manera de trabajar con las fechas.

En los datos recibidos tenemos información sobre 423.146 pacientes navarros. Pero todos estos pacientes no deben ser considerados en el estudio, tenemos que filtrar quedándonos con aquellos que cumplan la regla para ser polimedicados y eliminar aquellos que sufran o hayan sufrido cáncer o VIH (siguiendo las indicaciones del SNS-O). Finalmente se trabaja con 45.400 pacientes un 10,73% de los pacientes totales.

Los primeros experimentos realizados se centraban en la predicción de eventos adversos cardiovasculares. Una vez terminado con este tipo de eventos adversos se estudió aplicar la misma metodología para otro tipo de eventos. Junto con el SNS-O se decidió aplicar la metodología para hospitalizaciones potencialmente evitables (HPE). Hay que decir que a fecha de hoy aún se están realizando experimentos para este tipo de eventos, por lo que no se han realizado tantos experimentos como para los eventos adversos cardiovasculares.

Este proyecto lo hemos dividido en 4 fases:

- **Pre-procesamiento de datos:** Esta fase trata de la construcción de la base de datos. Antes de insertar todos los datos hay que asegurarse de que los datos no tengan errores, incoherencias,… Para ello se estudia cada uno de los ficheros observando los posibles valores que contienen y eliminando o corrigiendo aquellos que se consideren no válidos.
- **Planteamiento del problema de clasificación:** En esta fase se plantea como afrontar el proyecto. La idea es detectar eventos adversos en un futuro próximo, por lo que se planteó enfocarlo como un problema de clasificación. Para ello se necesita transformar los datos de la BBDD en conjuntos de datos válidos para los algoritmos de minería de datos.

- **Propuestas:** En esta fase se exponen las diferentes transformaciones o técnicas aplicadas sobre el conjunto de entrenamiento original.
- **Marco experimental:** Se presentan todas las pruebas realizadas con sus respectivos parámetros.

3. Herramientas

A lo largo del proyecto se han ido utilizando varios elementos o herramientas para llevarlo a cabo. Se necesitan herramientas de minería de datos con los algoritmos disponibles para la elaboración de modelos, se necesita un modo de almacenamiento de los datos en bruto (sin procesar), una herramienta para extraerlos debidamente,… Aunque se han utilizado más herramientas, estas son las principales:

3.1. Python

Python es un lenguaje de programación interpretado muy utilizado en diferentes ámbitos de la informática. Los lenguajes de programación interpretados "traducen" el código fuente a código máquina (lenguaje hablado por la máquina) a medida que sea necesario (normalmente instrucción a instrucción). Los lenguajes no interpretados en cambio, traducen completamente el código fuente en código maquina generando un binario ejecutable.

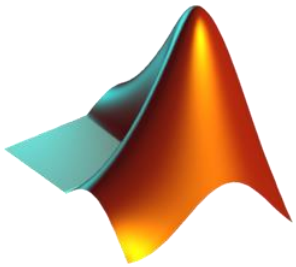
La programación con Python es rápida y sencilla, lo que hace que Python sea adecuado para escribir pequeños programas para tareas simples. Durante el proyecto se ha trabajado con diferentes ficheros, empleando Python para analizarlos, crearlos,… dado que se realiza de manera sencilla. Por ejemplo, se ha utilizado Python para analizar los ficheros de datos originales y generar los ficheros finales realizando el proceso de limpieza.

Además, existen diferentes librerías de Python con algoritmos de minería de datos muy útiles para este proyecto. La librería más popular es Scikit-learn, que es la librería que se ha empleado en este proyecto, aunque no como herramienta de minería de datos principal.

Como herramienta de minería de datos principal se ha utilizado Weka, herramienta fácil de utilizar escrita en Java con múltiples algoritmos de minería de datos. Cada ejecución realizada con Weka genera uno o varios ficheros con los resultados. Para analizar estos ficheros y extraer toda la información deseada se ha utilizado Python.

Python también permite comunicarse con diferentes bases de datos. Por ello se ha desarrollado un pequeño software en Python para generar los conjuntos de datos en los formatos adecuados para los algoritmos de minería de datos. Este software, recibe los parámetros del usuario, que son los atributos a utilizar, realiza las consultas necesarias a la base de datos y escribe los datos en uno o varios ficheros.

3.2. Matlab



Matlab es un software matemático muy utilizado a la hora de trabajar con datos numéricos, sobre todo con matrices. El núcleo de Matlab está preparado para que las operaciones entre matrices sean rápidas, reduciendo el tiempo de ejecución. Además del tratamiento de datos, Matlab permite realizar la representación de datos de manera sencilla, pudiendo realizar diferentes tipos de gráficos.

Por esto, Matlab se ha utilizado para realizar diferentes cálculos entre los conjuntos de datos, como por ejemplo, obtener la correlación entre diferentes atributos. También se ha utilizado para la visualización de datos mediante gráficos o para visualizar la curva ROC (medida de rendimiento explicada posteriormente)

Matlab también dispone de diferentes algoritmos de minería de datos, pero en este proyecto no se han utilizado.

3.3. Weka

Weka es una herramienta de minería de datos con múltiples algoritmos de machine learning escritos en Java. Es una herramienta muy sencilla de emplear con una interfaz gráfica muy intuitiva. Además de los algoritmos, Weka nos ayuda a entender el conjunto de datos pudiendo obtener estadísticas los atributos y permitiendo visualizar los datos.

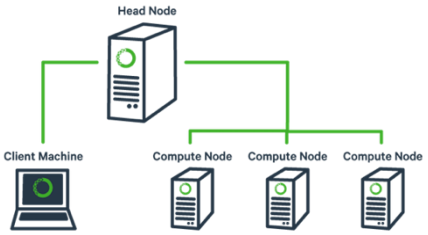
Esta es la herramienta principal utilizada para evaluar los conjuntos de datos. Además de utilizar los algoritmos de clasificación, transformación,… se han modificado algunos algoritmos para adaptarse a nuestra problemática, por ejemplo, añadir la métrica área bajo la curva ROC en ciertos algoritmos.

3.4. KEEL

Keel es una herramienta de minería de datos con múltiples algoritmos de machine learning escritos en Java. Esta herramienta es algo más compleja que Weka por lo que se ha utilizado en menor medida.

Keel se ha utilizado para ejecutar ciertos algoritmos no disponibles en Weka, por ejemplo algoritmos ensemble como UnderBagging.

3.5. Cluster



Un cluster es un conjunto de máquinas conectadas entre sí normalmente por red. Esto nos permite distribuir las ejecuciones por diferentes máquinas (también llamadas nodos) reduciendo el tiempo de ejecución total. Esto lo hace ideal para el proyecto, dado que al tener tanta combinación de variables se necesitan realizar varias ejecuciones, y empleando un cluster se redice el tiempo de espera.

Para las ejecuciones se ha empleado un gestor de colas. Al gestor de colas se le dice que ejecuciones hay realizar y este distribuye las ejecuciones por los diferentes nodos. El gestor conoce en todo momento el estado de los nodos, sabiendo como de "ocupados" están, llevando las ejecuciones a las máquinas más "paradas".

3.6. MySQL

MySQL es un sistema gestor de base de datos relacional, y actualmente es el sistema open source más popular. Ya se ha comentado que todos los datos son almacenados en una base de datos y esta base de datos se ha creado utilizando MySQL.

4. Pre-Procesamiento de datos

Esta fase incluye todo el proceso desde la recepción de los datos hasta la construcción final de la base de datos.

Este proyecto se ha desarrollado en colaboración con el Servicio Navarro de Salud – Osasunbidea (SNS-O), el cual nos proporcionó varios ficheros con toda la información necesaria para el estudio. El trabajo de esta fase consistió en analizar estos ficheros, encontrar elementos incorrectos o no válidos y corregirlos o eliminarlos antes de crear e insertar los datos en la base de datos.

La forma de proceder ha sido siempre la misma: primero se analizaba el fichero tratando de ver posibles fallos. Una vez encontrados los fallos nos poníamos en contacto con el equipo del SNS-O para aclarar de qué se trataba y de cómo proceder. Finalmente, con las indicaciones recibidas, se resolvía el fallo.

Los datos a utilizar en el proyecto provienen de diferentes fuentes, con los datos divididos en varios ficheros:

- **Fichero TIS:** Este fichero contiene la información de todos los pacientes navarros: identificador del paciente, fecha de nacimiento y código TSI.
- **Ficheros DGP:** Estos ficheros contienen los datos generales del paciente como por ejemplo la tensión arterial, el colesterol, el peso,…
- **Fichero Episodios:** Este fichero indica los episodios sufridos por los pacientes. Además del episodio sufrido contiene la fecha de inicio y la fecha de finalización del episodio.
- **Fichero CMBD:** En este fichero se encuentra información sobre los ingresos hospitalarios de los pacientes. Además del motivo del ingreso contiene la fecha de ingreso, la fecha de alta, la forma de ingreso (urgente o programado) y el motivo de alta.
- **Fichero Farmacia:** Este es el fichero que contiene todos los datos sobre medicamentos. Contiene la información sobre que principios activos (medicamentos) han adquirido los pacientes cada mes.

Como ya se ha dicho esta fase consistió en analizar cada uno de estos ficheros y tratarlos de manera adecuada para su posterior inserción en la base de datos:

4.1. Fichero episodios

Este fichero está estructurado en 5 campos: identificador del paciente, fecha de inicio del episodio, fecha de cierre del episodio, código CIAP del episodio y descriptor del episodio. El código CIAP es un código de clasificación internacional de atención primaria. Este código consta de 3 caracteres donde el primer carácter es una letra que representa el aparato o sistema orgánico y los otros dos caracteres son valores numéricos denominados “componentes” que representan, por ejemplo, signos o síntomas (por ejemplo del código K77, la K corresponde al aparato circulatorio y el código completo a insuficiencia cardiaca).

Como ejemplo una de las líneas (un registro) del fichero:

0xb0cc6d15a51acf541ea5d9558,20091105,18001228,R05,TOS

El fichero es un CSV, es decir, un fichero de valores separados por coma. Cada línea es un registro, y los campos están separados por coma. Uno de los problemas encontrados en este fichero fue que el descriptor podía contener comas, por lo que había que tratar todos los últimos campos posibles como uno solo.

Tras analizar el fichero se encontraron ciertos registros donde se desconocía el código CIAP y el descriptor. Estos registros fueron eliminados ya que el resto de campos (fechas e identificador de paciente) no aportaban información relevante para el proyecto. En otros casos se encontraron registros donde el código CIAP era un valor numérico negativo. Los médicos participantes del proyecto indicaron que estos códigos no son válidos, por lo que estos registros también fueron eliminados.

Por otra parte se encontraron registros con fechas de 1800/12/28. Cuando se encuentra esta fecha en la fecha de cierre significa que el episodio sigue abierto (no se ha cerrado) y el registro es correcto. Pero si se encuentra en la fecha de inicio, este registro no es válido por lo que se elimina. Por otro lado se decidió eliminar aquellos registros con fecha de inicio menor a 1900 o con fecha de inicio igual a 9999.

En ciertos casos se tenía que la fecha de cierre era anterior a la fecha de inicio. En estos casos se igualó la fecha de cierre a la fecha de inicio según la recomendación recibida por parte del SNS-O. En otros casos teníamos que el mismo paciente registraba el mismo episodio (mismo código) en la misma fecha de inicio pero en diferentes fechas de cierre. Otra vez siguiendo las indicaciones, se seleccionó aquel registro con la fecha de cierre más alejada, siendo 1800/12/28 (episodios abierto) la fecha más lejana eliminando el resto de registros.

En la Tabla 1 se resume la manera de proceder por cada elemento encontrado.

Condición	Procedimiento
Código CIAP desconocido	Eliminar registros
Código CIAP numérico	Eliminar registros
Fecha de inicio = 1800/12/28	Eliminar registros
Fecha de inicio < 1900	Eliminar registros
Fecha de inicio = 9999	Eliminar registros
Fecha de cierre anterior a fecha de inicio	Igualar la fecha de cierre a la fecha de inicio
Mismos registros con diferentes fechas de cierre	Mantener el registro con la fecha de cierre más alejada y eliminar el resto

Tabla 1. Medidas tomadas para los fallos encontrados en el fichero Episodios

4.2. Ficheros DGP

En este caso la información DGP (Datos Generales del Paciente) vino separado en dos ficheros diferentes. El primer fichero se denomina DGPs y contiene información sobre el colesterol, tensión arterial,… (ver Tabla 3) y el segundo, DGPs2, sobre peso, talla, IMC y tabaco (ver Tabla 5).

Ambos tienen la misma estructura, fichero separado por comas (CSV) con cuatro campos: identificador del paciente, código DGP (talla, colesterol, tabaco,…), fecha de la medida y valor.

Ejemplo DGPs: *0xba5c512684d974e4dc2a11f68,20130522,COL,170*

Ejemplo DGPs2: *0x49bfa0e59b647f94289e8484f,20130306,IMC,23,3*

El primer fichero (DGPs) contiene información sobre códigos cuyos valores solo pueden ser numéricos, pero en ciertos casos se observaron valores anómalos como "*", "+", "M30", ">6",… Los registros con estos valores se eliminaron exceptuando un caso especial: cuando el código era "FITGLOM" y el valor ">60". Este valor tiene sentido clínico por lo que no hay que eliminarlo. En vez de ello, siguiendo las indicaciones recibidas, se recodificó como 80 que es el valor medio entre 60 y 100.

Existen registros que tienen 5 campos en vez de 4. El último campo adicional representa la parte decimal del valor, por lo que hay que se fusionaron los valores de los registros con 5 campos.

Por otra parte se eliminaron los registros cuyos valores estaban fuera de rango. Estos rangos fueron planteados por los farmacéuticos participantes en el proyecto y se presentan en la Tabla 3.

Una vez eliminados los registros no válidos, se tenían registros donde el mismo paciente para el mismo código y misma fecha tenían diferentes valores. En estos casos, primero se analizaron los valores que había que agregar comprobando que la diferencia máxima entre los valores era insignificante. Tras verificar que todos los casos cumplían este requisito, se agregaron utilizando la media aritmética.

En la Tabla 2 se resume la manera de proceder por cada elemento encontrado.

Condición	Procedimiento
Campo "valor" con valores no numéricos excepto "FITGLOM>60"	Eliminar
Valor ">60" en códigos FITGLOM	Recodificar valor como 80
Registros con 5 campos en vez de 4	Unificar los dos últimos campos, siendo el último la parte decimal
Registros con valores fuera de rango	Eliminar
Mismos registros con diferente valor	Agregar valores utilizando la media

Tabla 2. Medidas tomadas para los fallos encontrados en el fichero DGPs

Código	Denominación	Tipo variable	Rango	Normalidad
TAS	Tensión Arterial Sistólica (MMHG)	Cuantitativa	40-270	>140
TAD	Tensión Arterial Diastólica (MMHG)	Cuantitativa	30-150	<90
INR	I.N.R.	Cuantitativa	0-10	
HB A1	Hb A1	Cuantitativa	entre 3 y 19	entre 6 y 9
HB A1C	Hb A1c	Cuantitativa	entre 3 y 19	entre 6 y 9
COL	Colesterol	Cuantitativa	100-600	<220
TGC	Triglicéridos	Cuantitativa	10-1000	<200
LDL	LDL-Colesterol	Cuantitativa	30-300	<160
HDL	HDL-Colesterol	Cuantitativa	1-160	>35
AST	AST - ASPARTATO TRANFERASA	Cuantitativa	1-1000	10--40
ALT	ALT - ALANINA TRANSFERASA	Cuantitativa	1-1000	entre 7 y 41
GGT	GAMMA GT	Cuantitativa	1-1000	entre 6 y 50
FILTGLOM	FILTRADO GLOMERULAR	Cuantitativa	0-150	

Tabla 3. Información y rangos para los diferentes códigos de DGPs

El segundo fichero (DGPs2) contiene información sobre la talla, el peso, el IMC y todo lo relativo al tabaco. Los diferentes códigos se presentan en la Tabla 5.

Analizando los registros con código "TABACO" encontramos diferentes valores cuando, en teoría, tendrían que ser "EXF>1A", "EXF<1A", "FUM" y "NF", que corresponden a exfumador más de un año, exfumador menos de un año, fumador y no fumador respectivamente, pero se encontraron valores numéricos, valores no validos (2xf),… Estos valores anómalos se eliminaron. Como regla general se recodificaron todos los registros que empezaban por "ex" (sin importar las mayúsculas) a "EXF", todos los que empezaban por "n" como "NF" y todos los que empezaban por "s" o "f" a "FUM", dejando tres posibles valores: "FUM", "EXF" y "NF".

Aparte del registro "TABACO" existen otros que hacen referencia al consumo de tabaco. Estos registros tienen los códigos "TABPAQAñ", "TABCIGDI", "TABPURSE" y "TFUMAÑOS" (ver Tabla 5). Estos códigos solo contienen información numérica y fueron recodificados generando un registro con el código TABACO: Si alguno de estos registros contenía un valor mayor que 0, se sustituía por un registro con código TABACO y valor FUM. En caso de que todos fueran 0 se sustituían por un registro de código TABACO y valor NF. Estos nuevos registros solo se generaron cuando no existía un registro de TABACO para el mismo paciente en la misma fecha.

Por último los datos de PESO y TALLA. Estos datos fueron recodificados como registros IMC utilizando la siguiente formula:

$$IMC = \frac{PESO(kg)}{ALTURA(cm)^2}$$

Igual que con el fichero DGPs se eliminaron los registros fuera de rango. También se unificaron los registros utilizando la media de los valores donde teníamos el mismo paciente, código y fecha, exceptuando en el caso de TABACO: si para la misma fecha existían registros indicando que un paciente era fumador, exfumador o no fumador a la vez, se eliminaban estos registros ya que no se puede saber cuál era el bueno.

En la Tabla 4 se resume la manera de proceder por cada elemento encontrado.

Condición	Procedimiento
Registros de TABACO que empiezan por "ex"	Recodificar como "EXF" (Exfumador)
Registros de TABACO que empiezan por "n"	Recodificar como "NF" (No fumador)
Registros de TABACO que empiezan por "s" o por "f"	Recodificar como "FUM" ("Fumador")
Resto de registros de TABACO	Eliminar
Registros de TABPAQAñ, TABCIGDI, TABPURSE o TFUMAÑOS > 0	Generar registro "TABACO FUM"
Registros de TABPAQAñ=TABCIGDI=TABPURSE=TFUMAÑOS=0	Generar registro "TABACO NF"
Registros de PESO y TALLA	Recodificar a IMC empleado la fórmula anterior
Registros con valores fuera de rango	Eliminar
Mismos registros con diferente valor (excepto TABACO)	Agregar valores utilizando la media
Mismo registros de TABACO con diferentes valores	Eliminar

Tabla 4. Medidas tomadas para los fallos encontrados en el fichero DGPs2

Código	Denominación	Tipo variable	Rango
PESO	PESO (KG)	Cuantitativa	(35-200)
TALLA	TALLA (CM)	Cuantitativa	(135-210)
IMC	I.M.C.	Cuantitativa	(14-50)
TABPAQAñ	TABACO, PAQUETES AñO	Cuantitativa	(0-150)
TABCIGDI	CONSUMO DE CIGARRILLOS/DIA	Cuantitativa	(0-90)
TABPURSE	TABACO, PUROS SEMANA	Cuantitativa	(0-75)
TFUMAÑOS	TABACO, FUMADOR AÑOS DE	Cuantitativa	(0-80)
TABACO	TABACO (NF, EXF>1A,EXF<1A,FUM)	Ordinal	(NF, EXF, FUM)

Tabla 5. Información y rangos para los diferentes códigos de DGPs2

4.3. Fichero CMBD

Este fichero contiene toda la información sobre los ingresos de los pacientes y está estructurado en 12 campos separados por coma (CSV): identificador del paciente, código de ingreso (CIE9), descripción del ingreso, mes del ingreso, año del ingreso, mes del alta, año del alta, campo CMA, código del motivo de alta, descripción del motivo de alta, código del motivo de ingreso y descripción del motivo de ingreso.

Ejemplo: *0x399a696c86ead3bccd9641f0e,428.0,INSUFICIENCIA CARDIACA CONGESTIVA, 05,2012,05,2012,CMBD,1,Domicilio,U,Urgente*

Existen dos motivos de ingresos diferentes: Urgente y Programado. En el caso de motivos de alta tenemos 8 motivos diferentes que se pueden consultar en la Tabla 6.

Domicilio
Traslado a otro hospital
Exitus
Otra causa
Alta voluntaria
Fuga o abandono
“Tras asist sociosanitario”
Hospitalización a domicilio

Tabla 6. Motivos de alta

El campo CMA de este fichero contenía siempre el mismo valor por lo que este campo fue ignorado.

Otra cosa a destacar es que se encontraron registros donde un paciente es ingresado el mismo mes del mismo año por el mismo diagnóstico y recibe el alta el mismo mes del mismo año, pero el motivo de ingreso es diferente, un ingreso es programado y el otro urgente. Estos registros son correctos.

4.4. Fichero TIS

Este fichero contiene la información de los pacientes. Es un fichero CSV que se compone del identificador del paciente, la fecha de nacimiento, el grupo TSI al que pertenece y la fecha de TSI.

Ejemplo: *0x75dd28d1232486bb946a48324,1963-12-06 00:00:00,TSI 003,2012-06-21 21:39:29.0882320*

De este fichero solo se eliminaron las horas de las fechas ya que no eran necesarias.

4.5. Fichero FARMACIA

Este fichero contiene toda la información sobre los medicamentos expedidos a los pacientes.

El fichero de farmacia debería de haber tenido 6 campos: identificador del paciente, principio activo del medicamento, número de envases expedidos, dosis, el mes y el año. Pero además de estos campos teníamos información que usaron para validar la extracción y que no son necesarios para el estudio. Estos campos añadieron complejidad para leer y estudiar el fichero dado que alguno de los campos podía contener comas teniendo que tratarlo.

Ejemplo: *0xe5c92c6d8af7b7a6af217810c,ACECLOFENACO |000008|,1,20,02,2014,M01AB16 - Aceclofenaco,PAMPLONA,ELIA JIMENEZ FERNANDO JOSE |60184g9v3|*

Nota: En el ejemplo se observa un nombre y un código. Estos datos no son los que provienen del fichero original, han sido modificados para anonimizados con el objetivo de mantener la protección de datos.

Estudiando el fichero se observaron casos en los que el mismo paciente obtenía el mismo principio activo en el mismo mes y año. En estos casos se unificaron los registros sumando los valores de la dosis y el número de envases.

Se eliminaron los registros donde el principio activo era F0000, Y1000 o Y2000 siguiendo las indicaciones de los farmacéuticos. También se eliminaron los registros con subgrupo “Grupo Desconocido” o “DESCONOCIDO(CARGA)” ya que son efectos y accesorios no medicamentos.

En la Tabla 7 se resume la manera de proceder por cada elemento encontrado.

Condición	Procedimiento
Registros con información extra	Tratar los registros ignorando estos campos
Registros con misma información excepto número de envases dispensados	Unificar registros sumando los valores
Principio activo F0000, Y1000 o Y2000	Eliminar registros
Registros con subgrupo “Grupo Desconocido” o “DESCONOCIDO(CARGA)”	Eliminar registros

Tabla 7. Medidas tomadas para los fallos encontrados en el fichero Farmacia

4.6. Base de Datos (BBDD)

Una vez limpiados los ficheros, se generó una base de datos (BBDD) y se insertaron todos estos datos ya tratados. El objetivo de la BBDD es poder generar el conjunto de datos a utilizar en los algoritmos de una manera más sencilla: utilizando el lenguaje SQL. Con esto se pueden obtener los datos necesarios, realizar estadísticas, obtener información,…

Muchos de los ficheros tenían campos descriptivos, por ejemplo, el fichero de episodios tenía el campo "descriptor" que nos da la descripción del código CIAP asociado. Estos campos no se necesitan para construir el conjunto de datos final, solo para informarnos cuando fuese necesario. Por ello se separaron los campos descriptivos en nuevas tablas:

- EPISODIOS_CIAP: Tabla que contiene los códigos CIAP y sus descriptores
- CMBD_DIAGNOSTICO: Tabla que contiene los códigos de ingresos y sus descripciones
- CMBD_INGRESO: Tabla que contiene los contiene los motivos de ingreso y sus descripciones
- CMBD_ALTA: Tabla que contiene los códigos de alta y sus descripciones

Como se ha comentado existen dos ficheros DGPs, cada uno con datos diferentes. Dado que la estructura de los ficheros es exactamente la misma, se decidió unificar ambos ficheros en una misma tabla (DGPs) para facilitar las consultas a realizar.

Esta BBDD fue formada con todos los pacientes, y como se ha dicho quedan excluidos los pacientes que sufran o hayan sufrido cáncer o VIH. Para ello se eliminaron estos pacientes utilizando la información de ingresos (CMBD) y episodios. El equipo del SNS-O nos proporcionó una lista con los códigos CIAP que representan estas enfermedades en episodios y otra lista con los códigos de ingresos representando estas enfermedades en CMBD (ver Tabla 8 y Tabla 9). Con estos códigos se identificaron los pacientes a eliminar, eliminando todos los registros de estos pacientes en todas las tablas de la base de datos.

CIAP	Descripción
A79	NEOPLASIA M SECUNDARIA DE ORIGEN DESCONOCIDO
B72	OTRO LINFOMA
B73	LEUCEMIA
B90	INFECCION POR VIH, SIDA
D74	NEOPLASIA M DEL ESTOMAGO
D75	NEOPLASIA M DE COLON
D76	NEOPLASIA M PANCREAS
D77	NEOPLASIA M VIAS BILIARES, CONDUCTOS EXTRAHEPATICOS
D78	NEOPLASIA INESPECIFICA/SIN HISTOLOGIA DE APARATO DIGESTIVO
F74	NEOPLASIA INESPECIFICA ANEJOS OCULARES
H75	NEOPLASIA INESPECIFICA APARATO AUDITIVO
K72	NEOPLASIA INESPECIFICA CARDIOVASCULAR
L71	NEOPLASIA M DEL APARATO LOCOMOTOR
L97	NEOPLASIA INESPECIFICA APARATO LOCOMOTOR (CUALQUIER PARTE)
N74	NEOPLASIA M DEL SISTEMA NERVIOSO
N76	NEOPLASIA INESPECIFICA DEL SISTEMA NERVIOSO
R84	NEOPLASIA M DE TRAQUEA/BRONQUIO/PULMON/PLEURA

R85	NEOPLASIA M DE SENOS, RESPIRATORIO
R92	NEOPLASIA INESPECIFICA DEL APARATO RESPIRATORIO
S77	NEOPLASIA M PIEL
S79	NEOPLASIA INESPECIFICA DE LA PIEL
T71	NEOPLASIA M TIROIDES
U75	NEOPLASIA M RIÑON, RENAL
U76	NEOPLASIA M VEJIGA URINARIA
U77	NEOPLASIA M DE URETER
U79	NEOPLASIA NE DEL APARATO URINARIO
W72	NEOPLASIA M EN CONEXION CON EL EMBARAZO
X75	NEOPLASIA M CERVIX UTERINO, CUELLO UTERO
X76	CARCINOMA INTRADUCTAL, MAMA MUJER
X77	NEOPLASIA M ANEJO UTERINO
X81	NEOPLASIA INESPECIFICA APARATO GENITAL FEMENINO
Y77	NEOPLASIA M PROSTATA
Y78	NEOPLASIA M TESTICULO
Y79	NEOPLASIA INESPECIFICA APARATO GENITAL MASCULINO, VARON

Tabla 8. Códigos CIAP correspondientes a cáncer o VIH excluidos

CIE9	Descripción
140	NEOPLASIA MALIGNA DE LABIO
141	NEOPLASIA MALIGNA DE LENGUA
142	NEOPLASIA MALIGNA DE GLANDULAS SALIVARES MAYORES
143	NEOPLASIA MALIGNA ENCIA
144	NEOPLASIA MALIGNA DEL SUELO DE LA BOCA
145	NEOPLASIA MALIGNA OTRAS PARTES BOCA Y PARTES SIN ESPECIFIC.
146	NEOPLASIA MALIGNA OROFARINGE
147	NEOPLASIA MALIGNA NASOFARINGE
148	NEOPLASIA MALIGNA HIPOFARINGE
149	NEOPLASIA MALIGNA OTROS SITIOS Y MAL DEF. LABIO Y OROFARINGE
150	NEOPLASIA MALIGNA ESOFAGO
151	NEOPLASIA MALIGNA ESTOMAGO
152	NEOPLASIA MALIGNA INTESTINO DELGADO, INCLUYENDO DUODENO
153	NEOPLASIA MALIGNA COLON
154	NEOPLASIA MALIGNA RECTO, UNION RECTOSIGMOIDAL Y ANO
155	NEOPLASIA MALIGNA HIGADO Y CANALES BILIARES INTRAHEPAT.
156	NEOPLASIA MALIGNA V.BILIAR Y COND.BIL.EXTRAHEPATICOS
157	NEOPLASIA MALIGNA PANCREAS
158	NEOPLASIA MALIGNA RETROPERITONEO Y PERITONEO
159	N. MALIGNA NEOPLASIA MALIGNA DIGESTIVO/PERITONEO OTROS SITIO
160	NEOPLASIA MALIG FOSAS NASALES, OIDO MEDIO Y SENOS ACCESORIOS
161	NEOPLASIA MALIGNA LARINGE
162	NEOPLASIA MALIGNA TRAQUEA, BRONQUIOS Y PULMON
163	NEOPLASIA MALIGNA PLEURA
164	NEOPLASIA MALIGNA TIMO, CORAZON Y MEDIASTINO
165	NEOPLASIA MALIGNA RESPIRATORIO/IMTRATORACICO OTROS SITIOS
170	NEOPLASIA MALIGNA HUESO Y CARTILAGO ARTICULAR
171	NEOPLASIA MALIGNA TEJIDOS CONECTIVOS Y OTROS TEJIDOS BLANDOS
172	MELANOMA MALIGNO PIEL
173	OTRAS NEOPLASIAS MALIGNAS DE LA PIEL

174	NEOPLASIA MALIGNA MAMA MUJER
175	NEOPLASIA MALIGNA MAMA HOMBRE
179	Neo maligna de útero NEOM
180	NEOPLASIA MALIGNA CERVIX UTERINO
181	Neo maligna de placenta
182	NEOPLASIA MALIGNA CUERPO UTERINO
183	NEOPLASIA MALIGNA OVARIO Y OTROS ANEXOS UTERINOS
184	NEOPLASIA MALIG OTROS ORGANOS GENITALES FEMENINOS Y SIN ESP
185	Neo maligna de próstata
186	NEOPLASIA MALIGNA TESTICULOS
187	NEOPLASIA MALIGNA PENE Y OTROS ORGANOS GENITALES MASCULINOS
188	NEOPLASIA MALIGNA DE LA VEJIGA
189	NEOPLASIA MALIGNA DE RIÑÓN Y OTROS ORG URINARIOS
190	NEOPLASIA MALIGNA OJO
191	NEOPLASIA MALIGNA CEREBRO
192	NEOPLASIA MALIGNA OTRAS PARTES O SIN ESPEC. SISTEMA NERVIOSO
193	NEO MALIGNA DE TIROIDES
194	NEOPLASIA MALIGNA OTRAS GLAN.ENDOCRINAS Y ESTRUC.RELACIONADA
195	NEOPLASIA MALIGNA OTROS SITIOS Y SITIOS MAL DEFINIDOS
196	NEOPLASIA MALIGNA SECUNDARIA Y NEOM DE GANGLIOS LINFATICOS
197	NEOPLASIA MALIGNA SEC. APARATO RESPIRATORIO Y DIGESTIVO
198	NEOPLASIA MALIGNA SECUNDARIA DE OTROS SITIOS ESPECIFICADOS
199	NEOPLASIA MALIGNA SIN ESPECIFICACION DEL SITIO
200	LINFOSARCOMA Y RETICULOSARCOMA
201	ENFERMEDAD DE HODGKIN
202	OTRAS NEOPLASIAS MALIGNAS TEJIDOS LINFOIDES E HISTIOCITICOS
203	NEOPLASIA INMUNOPROLIFERATIVAS Y MIELOMA MULTIPLE
204	LEUCEMIA LINFOIDE
205	LEUCEMIA MIELOIDE
206	LEUCEMIA MONOCITICA
207	OTRAS LUCEMIAS ESPECIFICADAS
208	LEUCEMIA SIN ESPECIFICACION TIPO DE CELULA
209	TUMORES NEUROENDOCRINOS
236	NEO COMPORTAMIENTO INCIERTO ORGANOS GENITOURINARIOS
237	NEOPLASIAS COMPORT. NO DETERM. GLAND. ENDOCR. Y S.NERVIOSO
238	N.COMPOR.NO DETER.DE OTRO SITIOS Y TEJ.DE SITI Y TEJ.NO ESPE
239	NEOPLASIAS DE NATURALEZA NO ESPECIFICADA
042	ENFERMEDAD POR VIRUS DE INMUNODEFICIENCIA HUMANA [VIH]

Tabla 9. Códigos CIE9 correspondientes a cáncer o VIH excluidos

Una vez eliminados, se filtraron aquellos pacientes considerados polimedicados. Como ya se ha mencionado, se acordó con el SNS-O que un paciente polimedicado es aquel que tome 5 o más medicamentos durante al menos tres meses consecutivos. Para identificar estos pacientes se utilizaron los datos de farmacia, fichero que contiene toda la información referente a los medicamentos. En vez de eliminar los pacientes no polimedicados, se creó una nueva base de datos donde solo se mantiene la información de estos pacientes polimedicados, manteniendo la información del resto de pacientes para posibles cambios.

Un paciente empieza a ser polimedicado en una fecha concreta, por lo que se necesita almacenar esta fecha en la base de datos. Dependiendo de esta fecha el paciente formará parte

del conjunto de datos a utilizar o no. Para ello se añadió un nuevo campo en la tabla TIS para almacenar la fecha de polimedicado.

Finalmente se generó una base de datos con el esquema o modelo relacional presentado en la Figura 2.

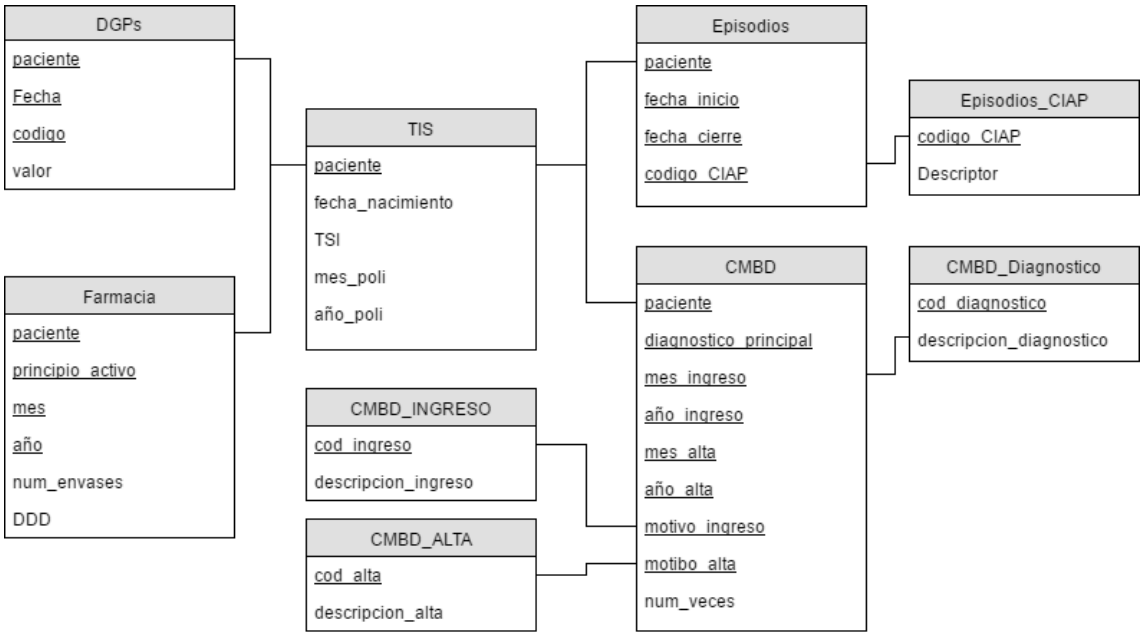


Figura 2. Esquema o modelo relacional de la base de datos inicial

4.7. Ampliación de la BBDD

Como se ha mencionado en la introducción, en el proceso de minería de datos es habitual, y normalmente necesario, repetir alguna de las fases para añadir más información. En este proyecto tras realizar cada una de las fases, se concretaba una reunión para establecer nuevos cambios o mejoras. Algunos de los cambios implicaron añadir información y extender la BBDD.

En la tabla farmacia existe un campo llamado principio activo. Este está formado por 7 caracteres alfanuméricos y es divisible en varios grupos. Si se utiliza únicamente el primer carácter obtenemos un grupo al que llamamos SUB1. Si se utilizan los 3 primeros obtenemos el grupo SUB3, si se utilizan los 5 primeros SUB5 y si se utilizan los 7 caracteres obtenemos el grupo SUB7. A parte de esta agrupación los médicos que trabajan en este proyecto formaron una nueva forma de agrupar a la que llamamos CODMED. Cada uno de los grupos de CODMED está formado por varios principios activos SUB7. Esta información tiene que estar almacenada en la BBDD para poder realizar la conversión. Para ello se creó una tabla “COD_FARMACIA” con la información del nuevo grupo asociado a cada principio activo.

Por otra parte se estudió el campo diagnostico principal (CIE9 o código de ingreso) de la tabla CMBD. Al principio se pensó en agruparlos utilizando la misma técnica que con el principio activo: por caracteres. Pero esta agrupación no era válida desde el punto de vista clínico. Por ello el equipo de clínicos participante en el proyecto formó otra agrupación de 4 niveles. El cuarto nivel corresponde al CIE9 original y con esto se puede conseguir cualquiera de los otros 3 niveles. Esta agrupación se añade a la BBDD como la tabla “CIE9_NIVELES”.

Al principio se utilizaba la información de todos los pacientes sin tener en cuenta que en una fecha dada estos podían haber fallecido ya que no disponíamos de esta información desde el principio. Para ello se nos hizo llegar la fecha de fallecimiento de cada paciente, para lo que tuvo que añadir un campo en la tabla TIS y así descartar los pacientes fallecidos cuando fuese necesario. Además de la fecha de fallecimiento se añadió otro campo que indica el sexo del paciente.

También se añadió información sobre antecedentes. Existe un antecedente de episodios y otro de ingresos (CMBD). Estos campos identifican ciertos episodios o ingresos como antecedente, es decir, se identifican ciertos códigos que representan un antecedente y en aquellos registros con este código se establece el campo antecedente a 1. En caso contrario antecedente 0. Con esto se facilita el cálculo de este atributo a la hora de generar el conjunto de datos a utilizar. Para ello se añadió un campo en la tabla episodios llamado "ant_cardio" y otro en la tabla CMBD llamada también "ant_cardio".

Los códigos de episodios (CIAP) que se identifican como antecedentes se reflejan en la Tabla 10 y los códigos de ingreso (CIE9) en la Tabla 11.

Código CIAP	Descripción
K74	ISQUEMIA CARDIACA CON ANGINA
K75	INFARTO AGUDO DE MIOCARDIO
K76	ISQUEMIA CARDIACA SIN ANGINA
K77	INSUFICIENCIA CARDIACA
K90	ACCIDENTE CEREBROVASCULAR/ICTUS/APOPLEJIA
K91	ENFERMEDAD CEREBROVASCULAR
K92	ATEROSCLEROSIS/ENFERMEDAD ARTERIAL PERIFERICA

Tabla 10. Códigos de episodios (CIAP) que representan un antecedente cardiovascular

Código CIE9	Descripción
398	OTRAS ENFERMEDADES CARDIACAS REUMATICAS
402	ENFERM. CARDIACA HIPERTENSIVA
403	ENFERM. RENAL HIPERTENSIVA
404	ENF.RENAL CRONICA Y CARDIACA HIPERTENSIVA
410	INFARTO AGUDO MIOCARDIO
414	OTRAS FORMAS DE ENFERM. CARDIACAS ISQUEMICAS CRONICAS
415	ENFERMEDAD CARDIOPULMONAR AGUDA
420	PERICARDITIS AGUDA
422	MIOCARDITIS AGUDA
425	MIOCARDIOPATIA
426	ALTERACIONES DE CONDUCCION
427	DISRRITMIAS CARDIACAS
428	INSUFICIENCIA CARDIACA
429	DESCRIP. Y COMPLICACIONES DE ENFERM. CARDIACA
433	OCLUSION Y ESTENOSIS ARTERIAS PRECEREBRALES
434	OCLUSION DE ARTERIAS CEREBRALES
438	EFECTOS TARDIOS DE ENFERMEDAD CEREBROVASCULAR
440	ATEROSCLEROSIS
441	ANEURISMA AORTICO Y DISECCION
442	OTROS ANEURISMAS
443	OTRA ENFERMEDAD VASCULAR PERIFERICA

Tabla 11. Códigos de ingresos (CIE9) que representan un antecedente cardiovascular

De la misma manera que con los antecedentes se añadió un nuevo campo en CMBD: event_adv_cardio. Este campo nos indica si un ingreso es considerado evento adverso cardiovascular o no. El SNS-O fue el que nos hizo llegar que ingresos que son considerados eventos adversos cardiovasculares:

- **Enfermedad cardiaca isquémica:** códigos de diagnóstico desde 410 hasta 414
- **Otras formas de enfermedad cardiaca:** códigos de diagnóstico desde 420 hasta 429
- **Enfermedad cerebrovascular:** códigos de diagnóstico desde 430 hasta 439

Esto facilita la forma de obtener si un paciente ha sufrido evento adverso cardiovascular, generando una consulta SQL con una única condición en vez de una condición por código de ingreso.

Para finalizar se añadió otra tabla: HPE. Como se ha comentado además de los eventos adversos cardiovasculares se han tenido en cuenta otro tipo de eventos, en concreto, las hospitalizaciones potencialmente evitables (HPE). Para identificar que pacientes han sufrido este tipo de evento, el SNS-O no hizo llegar la información indicando que pacientes (identificador) habían sufrido qué tipo de HPE en qué fecha (mes y año). Toda esta información fue añadida a la base de datos en la tabla HPE.

Los datos con los que se empezaron a trabajar eran de los años 2013 y 2014. El proyecto empezó a principios del 2015 por lo que los datos de 2015 no estaban disponibles todavía. Avanzando con el proyecto llegamos a 2016, lo que supuso poder utilizar los datos de 2015.

El equipo del SNS-O nos proporcionó los datos de 2015. Los ficheros recibidos fueron analizados en busca de posibles errores no encontrados en los ficheros de 2013 y 2014. Tras el análisis se observó que la manera de tratarlos podía ser la misma, por lo que se reutilizaron los procedimientos aplicados a esos ficheros, sin requerir ninguno nuevo. Estos datos tampoco supusieron añadir nuevos campos a la base de datos ya que la estructura de los ficheros era la misma.

Al tener datos de 2015 tuvimos que volver a detectar pacientes polimedicados. Dada la posibilidad de que pacientes del año 2014 no habrían sido detectados, se tuvo que utilizar la base de datos original (sin filtro de polimedicados). Por ejemplo, puede que un paciente empezase a consumir 5 medicamentos el 11/2014, por lo que al no tener los datos de 2015 no había estado consumiendo estos 5 medicamentos durante 3 meses.

Con todos los cambios indicados se forma una base de datos cuyo esquema o modelo relacional se presenta en la Figura 3.

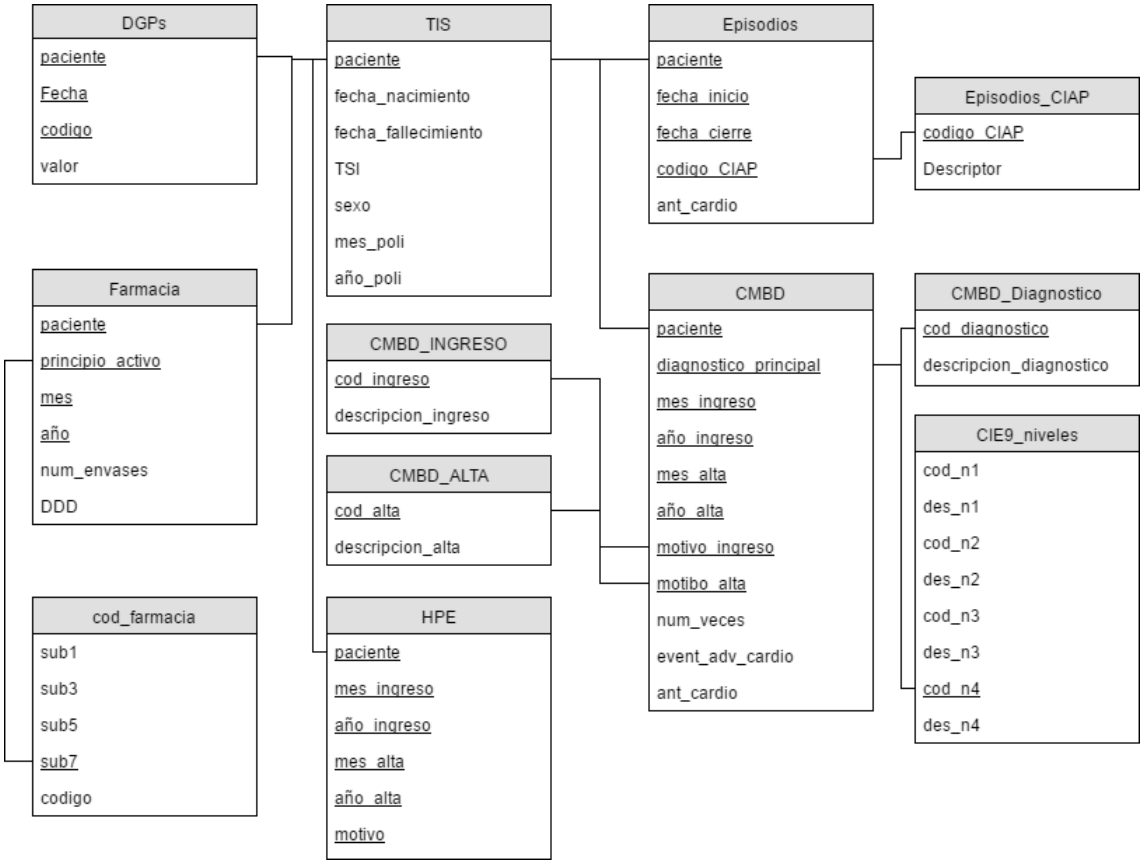


Figura 3. Esquema o modelo relacional de la base de datos final (ampliada)

4.8. Análisis de la base de datos

Con la base de datos montada, se procedió a sacar estadísticas y resúmenes sobre los datos con el objetivo de entender los datos y procesar nuevos posibles errores. Este análisis se podría haber realizado sobre los ficheros sin necesidad de montar la base de dato pero utilizando la base de datos se facilita la forma de sacar estadísticas complejas, utilizando el lenguaje SQL.

Se realizaron dos informes sobre la base de datos: uno con la base de datos completa (sin excluir a los polimedicados) y la otra sobre la base de datos de los polimedicados. Estos informes fueron estudiados junto al equipo del SNS-O para detectar posibles errores en los datos.

El informe completo puede consultarse en el Anexo I y el informe con los datos de pacientes polimedicados en el Anexo II.

5. Planteamiento del problema de clasificación

Una vez dispuesta la BBDD con los datos ya procesados, nos planteamos explotar la información contenida. El problema es que no se pueden aplicar técnicas de minería de datos directamente sobre la BBDD, es necesario construir un conjunto de datos adecuado para poder extraer la información relevante.

El problema fue propuesto como un problema de clasificación: se utilizan los datos de unos pacientes de los que se sabe si han sufrido evento o no para entrenar un modelo y se trata de predecir para nuevos pacientes si van a sufrir evento o no en un futuro próximo.

Como el objetivo es predecir eventos basándose en hechos del pasado se plantea coger ciertos datos como datos predictivos (atributos del conjunto de datos) y coger datos posteriores, que no se solapen con los anteriores, para obtener la clase. Para hacer esto se plantea escoger una fecha, a la que llamaremos fecha de referencia, la que se utiliza para separar los datos a utilizar como atributos y los datos de donde obtener la clase. Los datos anteriores a esta fecha son los que utilizan para crear los atributos predictivos y la clase se obtiene de los datos posteriores a la fecha de referencia. Como se puede ver los datos que se utilizar para predecir y los datos para calcular las clases a predecir no se solapan. En la Figura 4 podemos observar de manera visual el funcionamiento de las fechas utilizando 9/2014 como fecha de referencia.

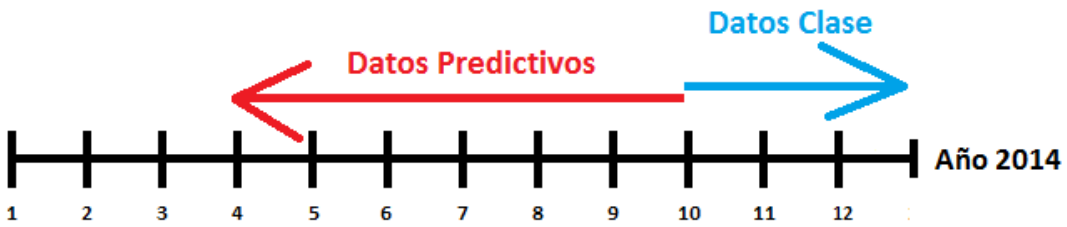


Figura 4. Ejemplo visual del funcionamiento de las fechas de referencia utilizando 9/2014

Por ejemplo supongamos que elegimos 9/2014 como fecha de referencia y las tablas CMBD y Farmacia de donde obtener los datos (para simplificar solo se cogen datos de estas dos). Se obtienen el número de medicamentos asociados a cada paciente buscando en los datos anteriores a la fecha 9/2014 (el mes de la fecha de referencia entra en los datos predictivos) y los ingresos sufridos antes del 9/2014 como variables predictivas (atributos) y considerar si ha sufrido o no evento adverso mirando el campo “event_adv_cardio” de la tabla CMBD en las fechas posteriores a la fecha de referencia. Con esto se podría generar un conjunto de datos válido para utilizarlo con los algoritmos de clasificación.

El proceso general para generar un conjunto de datos valido para las herramientas de minería de datos a partir de los datos almacenados en BBDD es el siguiente:

1- Se obtienen los identificadores de los pacientes válidos dada la fecha de referencia. Hay que tener en cuenta que el paciente no haya fallecido y que para la fecha dada el paciente sea considerado polimedicado. Por esto era importante almacenar la fecha en la que el paciente se empieza a considerar polimedicado.
2- Por cada paciente se obtienen los datos deseados realizando varias consultas a la base de datos montada para obtenerlos. En esta parte hay que tener en cuenta que estos datos se obtienen considerando que sean anteriores a la fecha de referencia.

3- Por cada paciente se calcula si ha sufrido o no evento adverso en los próximos N_{clase} meses (N_{clase} parámetro configurable por defecto a 3) a partir de la fecha de referencia. En este paso se está calculando la clase a predecir.
4- Finalmente se escriben estos datos en uno o varios ficheros, en el formato específico para la herramienta. Existen tres formatos posibles: CSV (valores separados por coma), ARFF (para la herramienta weka) y otro para la herramienta KEEL, aunque CSV no se ha empleado tanto como los otros.

5.1. Variables consideradas

En el paso 2 se ha dicho que se obtienen los datos deseados. En la BBDD existe mucha información que puede ser obtenida de múltiples formas. En el proyecto se han considerado las siguientes opciones por cada una de las tablas.

- **DGPs:** En esta tabla se almacenan todos los datos generales del paciente, como el colesterol, datos sobre tabaco,… De esta tabla se pueden generar los siguientes atributos:
    - Conteo de valores dentro de rango: En una primera versión utilizando los rangos de normalidad previamente mostrados (Tablas 3 y 5). Se contaba por cada uno de los códigos DGPs el número de registros que excedían los rangos saludables para cada paciente en los N_{DGP} (por defecto 3 según lo acordado con el SNS-O) meses anteriores a la fecha de referencia.
    - Media de los valores: Esta opción consiste en obtener por cada código DGPs la media de los valores en los N_{DGP} meses anteriores. Los datos de los DGPs que se recogen por los médicos no son muy frecuentes por lo que con esta técnica se encuentran muchos casos en los que no se tiene valor. Para solucionar esto primero se busca la fecha más reciente (desde la fecha de referencia hacia atrás) en la que se tenga algún dato. Con esta fecha se miran los N_{DGP} meses anteriores y se calcula la media, eliminando así muchos casos sin valor. Dado que esta opción presentaba mejores resultados se descartó la anterior.
- **EPISODIOS:** Esta tabla contiene la información de los episodios sufridos por los pacientes, así como el antecedente cardiovascular de episodios del que se ha hablado anteriormente. De esta tabla se pueden generar los siguientes atributos:
    - Conteo de episodios. En este caso solo se cuenta el número de episodios asociados al paciente pero pudiendo agruparlos de diferentes maneras. Se coge por cada paciente los datos de los episodios y se cuenta el número de episodios asociados. Ya se ha mencionado que el código CIAP consta de 3 caracteres donde el primer carácter representa el aparato o sistema orgánico, por ejemplo del código K77 la K corresponde al aparato circulatorio y el código completo a insuficiencia cardiaca. Este conteo se puede agrupar por los caracteres del código de episodios (1 o 2 caracteres). Es decir, por cada código agrupado, se calcula el conteo de esos episodios generando así tantos atributos como códigos distintos.
        - Existen tres versiones para obtener el conteo dependiendo de las fechas de inicio y cierre: La primera opción es coger todos los episodios que están abiertos o han sido cerrados en el último mes. La segunda es coger todos los episodios abiertos en los últimos 6 meses y la tercera

opción es una combinación de las dos anteriores sin duplicidad de datos, es decir, los episodios que están abiertos, los que han sido cerrados en el último mes o han sido abiertos en los últimos 6 meses pero no en el último mes. Tras realizar varias pruebas se observó que la primera opción obtenía mejores resultados.

- Aparte del conteo de episodios existe el antecedente antes mencionado que asocia ciertos episodios como antecedentes cardiovasculares. Para ello se mira si el paciente tiene algún registro con antecedente a 1 en los N_{GEN} (por defecto a 6 según lo acordado con el SNS-O) meses anteriores a la fecha de referencia.
- Índice de Charlson. Este índice es un índice de comorbilidad que está asociado a la esperanza de vida. Para el cálculo de este se utilizan los episodios sufridos por el paciente sumando ciertos puntos dependiendo de los episodios (estos puntos pueden consultarse en la Tabla 12). Además de estos puntos se añaden puntos por edad: se añade un punto por cada década por encima de los 40 años.
  - Existen dos maneras de utilizar esta información. La primera es utilizar la puntuación total en un único atributo. La segunda es utilizar un atributo por cada componente que da puntos como atributo binario junto con la edad como atributo numérico. De esta forma el algoritmo puede dar más peso a componentes que se consideren más importantes y descartar las no influyentes.

Episodio	Puntos
Enfermedad coronaria	1
Insuficiencia cardiaca	1
Otra enfermedad vascular periférica	1
Enfermedad cerebrovascular	1
Demencia	1
Enfermedad pulmonar crónica	1
Enfermedad tejido conectivo	1
Ulcera	1
Enfermedad hepática	1
Diabetes	1
Diabetes con daño órgano diana	2
Hemiplejia	2
Enfermedad renal	2
Linfoma	2

Tabla 12. Puntos asignados a cada componente del índice de Charlson

- **CMBD:** Esta tabla contiene la información de los ingresos de los pacientes. De esta tabla se pueden generar los siguientes atributos:
  - Conteo de ingresos. Al igual que con los episodios se cuenta el número de ingresos pudiéndolo agrupar en varias maneras. Como se ha explicado anteriormente se pueden agrupar los ingresos en 4 niveles diferentes, siendo el nivel 4 el propio código de ingreso (CIE9). Para obtener la cuenta, se calcula el número de ingresos diferentes sufridos en los N_{GEN} meses anteriores a la fecha de referencia.
  - De la misma forma que con los episodios existe información sobre antecedentes en CMBD. Para utilizarlo como atributo se mira si el paciente tiene algún registro

con el campo antecedente a 1 en los N_{GEN} meses anteriores a la fecha de referencia.

- **FARMACIA:** Esta tabla contiene toda la información sobre los medicamentos asociados a los pacientes. Esta es la tabla más importante para los miembros del SNS-O participantes en el proyecto ya que uno de los objetivos del proyecto es estudiar la interacción de varios medicamentos y ver si existe relación con algún evento adverso como el evento adverso cardiovascular. De esta tabla se pueden generar los siguientes atributos:
    - Conteo de principios activos. En este caso se cuenta por cada paciente el número de principios activos diferentes que ha tomado durante los N_{GEN} meses anteriores a la fecha de referencia. Este conteo se puede agrupar por cada uno de los subgrupos: SUB1 (primer carácter), SUB3 (tres primeros caracteres), SUB5 (cinco primeros caracteres), SUB7 (el principio activo completo) o CODMED (agrupación recibida).
    - Conteo de principios por cada elemento de los grupos. En la opción anterior se calcula el conteo agrupado, resultando en una única variable. En este caso se calcula el conteo por cada elemento del grupo elegido, resultando en varias variables. Por ejemplo suponiendo que el subgrupo 1 (SUB1) se compone de los elementos A, B, C y D, esta opción devolvería el conteo para cada uno de estos elementos. Este conteo también se utiliza con las diferentes agrupaciones: SUB1, SUB3, SUB5, SUB7 y CODMED.
- **TIS:** Esta tabla contiene información personal sobre los pacientes como la fecha de nacimiento o sexo. De esta tabla se pueden obtener los siguientes atributos:
    - TSI que nos indica a que grupo pertenece el paciente dependiendo de su estatus socioeconómico.
    - Edad del paciente. Es en esta tabla donde se almacena la fecha de nacimiento del paciente para calcular su edad. Hay que tener en cuenta que si la opción de índice Charlson separado es elegida, la edad ya se utiliza y no es necesario volver a utilizarla.
    - Sexo (género). El sexo del paciente pudiendo ser hombre o mujer.

Todas las opciones anteriores se utilizan como variables predictivas. Queda por determinar la variable a predecir. En este caso, como se ha mencionado anteriormente, se trata de predecir eventos adversos cardiovasculares. Para facilitar las cosas se añadió a la base de datos un campo en la tabla CMBD el cual indica si un ingreso es debido a la ocurrencia de un evento cardiovascular o no.

Para determinar la clase hay que hacer unas comprobaciones. Un ingreso no se considera como evento si el mes anterior ha sufrido un ingreso similar (evento del mismo tipo). Como la clase se determina mirando los N_{CLASE} siguientes meses a la fecha de referencia, primero se mira si el paciente ha sufrido un evento en la fecha de referencia, para saber si el mes siguiente es válido, es decir, que se puede utilizar para calcular la clase. Si ha sufrido evento el siguiente mes se considera no válido y no se utiliza para calcular la clase. Esta comprobación se hace para todos los meses (N_{CLASE}) para detectar los meses válidos.

Una vez detectados los meses válidos se busca el campo “event_adv_cardio” en esos meses, y si alguno de los registros tiene el valor 1, entonces se establece como case de clase positiva (1) y negativa (0) en caso contrario. Por último si no encontramos ningún mes válido, el paciente queda excluido ya que no tiene sentido considerarlo en el sistema.

Con esto ya estaría todo lo necesario para construir un conjunto de datos válido para utilizarlo con algunas de las herramientas consideradas en este trabajo.

5.2. Medidas de rendimiento

Con estos datos se plantea un problema de clasificación, utilizando unos datos etiquetados (con un valor de clase conocido) se entrena un sistema y este sistema predice la clase de datos de la que se desconoce la clase. Para este tipo de problemas se utilizan diferentes medidas como evaluación. La medida más habitual es la precisión, es decir, el porcentaje de instancias correctamente clasificadas.

En este proyecto esta medida no es adecuada debido a que el número de ejemplos que tenemos por cada una de las clases: problema del no balanceo. El problema del no balanceo surge cuando se tienen muchos más ejemplos de una clase que de otra. A la clase con mayor número de ejemplos se le denomina clase mayoritaria o clase negativa y a la case con menos ejemplos clase minoritaria o clase positiva.

Supongamos un problema con 50.000 instancias (ejemplos) de las cuales 49.000 pertenecen a la clase negativa (clase mayoritaria) y 1.000 a la clase positiva (clase minoritaria). Un clasificador que siempre predijese clase negativa tendría una precisión del 98%, dado que acertaría las 49.000 instancias negativas y fallaría las 1.000 positivas. Este porcentaje es un valor alto pero el clasificador no es bueno, dado que no acierta ningún ejemplo positivo.

En este proyecto se trabaja con un ratio de 0.99 para la clase negativa y 0.01 para la positiva aproximadamente. Es por esto que la precisión no es adecuada y por ello se consideran otras medidas que introducimos a continuación.

5.2.1. Matriz de confusión

Los algoritmos de clasificación son entrenados con un conjunto de datos donde cada ejemplo esta etiquetado, es decir, se sabe a qué clase pertenece cada ejemplo. Una vez entrenado el clasificador, este recibe nuevos datos no etiquetados y les asigna una clase a cada uno. Para saber cómo se comporta el clasificador se puede construir una matriz, llamada matriz de confusión, donde se visualiza las predicciones realizadas frente a las clases reales, observando los aciertos y errores cometidos por el clasificador.

Esta matriz es una matriz cuadrada (que tiene el mismo número de filas y de columnas) donde el número de filas (y de columnas) es el mismo que el número de clases. Las columnas corresponden a las clases predichas de los ejemplos y las a filas las clases reales. Con esto se puede ver de manera sencilla las diferentes clases asignadas por el clasificador. Además utilizando los valores presentes en la matriz se pueden calcular diferentes medidas de evaluación.

	Gato	Perro	Conejo
Gato	5	3	0
Perro	2	3	1
Conejo	0	2	11

Figura 5. Ejemplo de matriz de confusión con 3 clases. Figura extraída de Wikipedia.org

En el ejemplo de la Figura 5 se tienen que clasificar 27 ejemplos de los cuales 8 son gatos, 6 son perros y 13 son conejos. De los 8 gatos, 5 se han identificado correctamente y 3 incorrectamente, confundiéndolos con perros. De los 6 perros, 3 han sido identificado correctamente y 3 incorrectamente indicando que 2 son gatos y 1 es conejo. Y de los 13 conejos se han identificado 11 correctamente y 2 incorrectamente indicando que son perros.

Como se ha mencionado anteriormente, este proyecto es un problema de clasificación binaria, es decir, los ejemplos solo pertenecen a dos clases (clase positiva y clase negativa), por ello la matriz de confusión a generar es una matriz de 2 filas por 2 columnas. Con los problemas de dos clases se trabaja con una nomenclatura especial, que se puede observar en la Figura 6.

- True Positive (TP): cuantos ejemplos positivos han sido bien clasificados
- True Negative (TN): cuantos ejemplos negativos han sido bien clasificados
- False Negative (FN): cuantos ejemplos positivos se han mal-clasificado como negativos
- False Positive (FP): cuantos ejemplos negativos se han mal-clasificado como positivos

Actual \ Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Figura 6. Representación de la matriz de confusión para problemas de clasificación binaria.

5.2.2. Medidas basadas en la matriz de confusión

- **True Positive Rate (TPR):** También conocido como **sensibilidad (o recall)**, es el número de ejemplos de la clase positiva correctamente predichos.

$$TPR = \frac{TP}{TP + FN}$$

- **True Negative Rate (TNR):** También conocido como **especificidad**, es el número de ejemplos de la clase negativa correctamente predichos.

$$TNR = \frac{TN}{TN + FP}$$

- **GM:** Es la media geométrica entre el TPR y el TNR.

$$GM = \sqrt{TPR * TNR}$$

- **Precisión:** También conocida como **Valor Predictivo Positivo (PPV)** es la probabilidad de que una prueba positiva verdaderamente sea positiva.

$$Precision(PPV) = \frac{TP}{TP + FP}$$

- **F-Measure:** es la media armónica entre la precisión (PPV) y la sensibilidad (recall o TPR).

$$F - Measure(F1) = 2\frac{Precision\ x\ Recall}{Precision + Recall}$$

En la Figura 7 se puede observar una representación gráfica entre el PPV y la sensibilidad (TPR).

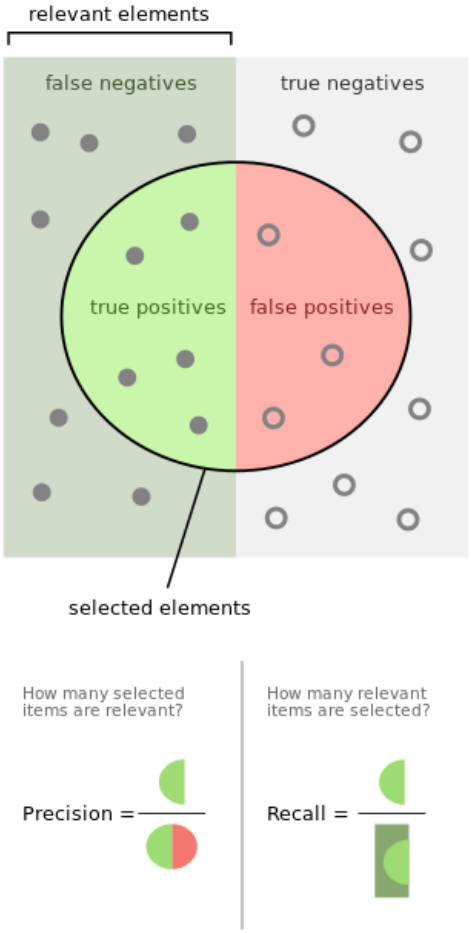


Figura 7. Representación gráfica de las medidas PPV y Recall. Figura extraída de Wikipedia.org

5.2.3. Medidas de clasificadores probabilísticos

Los clasificadores devuelven la clase a la que pertenece un ejemplo dado, es decir, si es positivo o negativo en este caso. Los clasificadores probabilísticos en cambio, devuelven la probabilidad de que un ejemplo pertenezca a cada una de las clases, es decir, cuál es la probabilidad de que sea negativo y cuál es la probabilidad de que sea positivo. Para decidir a qué clase pertenece el ejemplo se necesita un umbral θ. Cuando la probabilidad de ser positivo es mayor o igual al umbral, la instancia se clasifica como positiva $p_+ \geq \theta$. Al principio del proyecto se utilizaba 0.5 como umbral pero dado que los clasificadores dan mayor probabilidad a la clase negativa, por

el problema del no balanceo, se obtenía un TPR bajo. Para ofrecer unos resultados más equilibrados se busca el punto de corte estudiando la GM. Primero, con el clasificador entrenado, se clasifican todas las instancias de entrenamiento para obtener su probabilidad. Con estas probabilidades se busca el punto (umbral θ) de corte que maximice la GM. Ese punto de corte es el que se utilizará en los datos de test.

5.2.4. Curva ROC

La curva ROC (Receiver Operator Characteristics) modela la compensación entre el TPR y el FPR. La curva se dibuja en un espacio de dos dimensiones, donde los FPR van en el eje X y los TPR en el eje Y. Cada punto de la curva corresponde a un par (FPR, TPR) donde la situación perfecta es tener el punto (0,1), tener una separación de las clases perfecta. La curva ROC es insensible a los cambios de distribución de clases, lo que la hace perfecta para situaciones de datos no balanceados. En la Figura 8 se puede observar un ejemplo de la curva ROC.

Class	p_+	Class	p_+
+	0.9	-	0.7
+	0.9	+	0.7
+	0.9	-	0.7
+	0.9	-	0.3
-	0.9	+	0.3
+	0.9	-	0.3
-	0.9	+	0.1
+	0.7	-	0.1
+	0.7	-	0.1
-	0.7	-	0.1

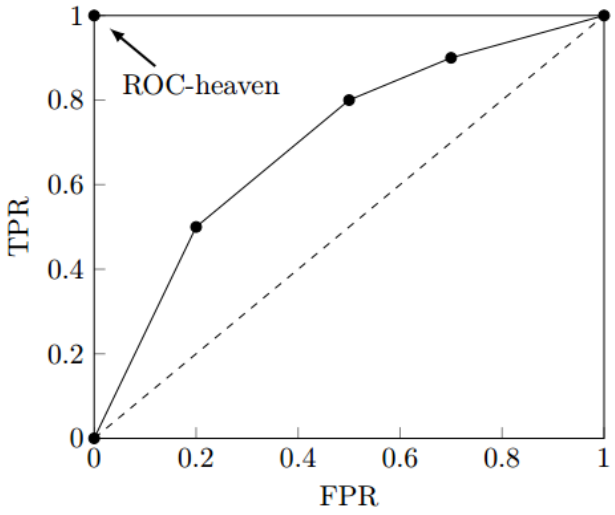


Figura 8. Ejemplo de la curva ROC

La información de la curva ROC se puede resumir en un valor: el área bajo la curva ROC (AUC-Area Under ROC-Curve). Este valor corresponde al área contenida entre la curva y el eje horizontal en el intervalo [0,1]. Para calcular este valor se utilizan todos los puntos empleados (pares de FPR y TPR) para dibujar la curva. Estos puntos se calculan variando el umbral de probabilidad θ y calculando el FPR y TPR con cada umbral diferente. Con esto el AUC se puede calcular como la suma de todas las áreas de los trapezoides formados. A este método se le denomina regla del trapezoide. Primero se ordenan los puntos de manera descendente por la probabilidad p_+. Con las probabilidades se va variando el umbral θ calculando los FPR y TPR en cada valor de probabilidad. Con los FPR y TPR actuales y previos se forma un trapezoide del cual se calcula su área y se suma al área total. En la Figura 9 tenemos el código que se utiliza para calcular el valor ROC AUC y en la Figura 10 se presenta el ejemplo de la regla del trapezoide.

Algorithm 1 Calculation of the AUC

Input: For each instance $\mathbf{x}_i \in T$, $i = 1, \ldots, n$, the estimated probability p^i_+ and its true class $l(\mathbf{x}_i)$.
Output: The AUC
$T_{sort} \leftarrow T$ sorted by decreasing values of p^i_+
$AUC \leftarrow 0$
$TP \leftarrow 0$, $FP \leftarrow 0$
$p_{prev} \leftarrow 0$
$tpr_{prev} \leftarrow 0$, $fpr_{prev} \leftarrow 0$
for $i = 1, \ldots, n$ **do**
  if $p^i_+ \neq p_{prev}$ **then**
    $tpr_{new} \leftarrow \frac{TP}{n_+}$
    $fpr_{new} \leftarrow \frac{FP}{n_-}$
    $area \leftarrow \frac{(tpr_{prev}+tpr_{new})(fpr_{new}-fpr_{prev})}{2}$
    $AUC \leftarrow AUC + area$
    $tpr_{prev} \leftarrow tpr_{new}$, $fpr_{prev} \leftarrow fpr_{new}$
    $p_{prev} \leftarrow p^i_+$
  if $l(\mathbf{x}_i) = Pos$ **then**
    $TP \leftarrow TP + 1$
  else
    $FP \leftarrow FP + 1$
$AUC \leftarrow AUC + \frac{(tpr_{prev}+1)(1-fpr_{prev})}{2}$
return AUC

Figura 9. Pseudo-código para calcular el AUC de la curva ROC

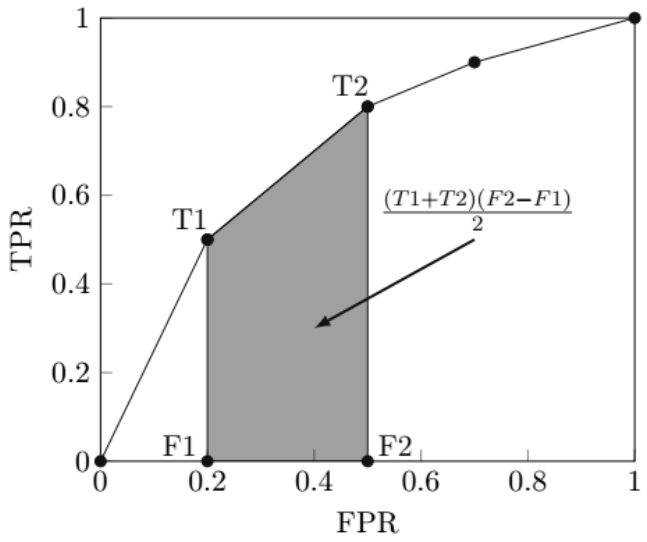


Figura 10. Ejemplo de la regla del trapezoide para calcular la AUC de la curva ROC

6. Propuestas

En el proyecto se han empleado diferentes técnicas para abordar el proyecto. Como se ha mencionado se ha planteado un problema de clasificación por lo que se utilizaron diferentes técnicas y algoritmos de clasificación.

Es los problemas de minería de datos el algoritmo de clasificación a utilizar es de importancia pero no tanto como los datos a utilizar. A menudo estos datos tienen que ser tratados y transformados para que el algoritmo de clasificación obtenga mejores resultados. En esta sección describimos las técnicas utilizadas para transformar el conjunto de datos inicial en un conjunto de datos con el que obtener mejores resultados.

6.1. Técnicas de clasificación

El proyecto se ha planteado como un problema de clasificación, tratar de etiquetar o predecir la clase de los pacientes basándonos en eventos del pasado. Para esta tarea existen multitud de algoritmos diferentes. En este proyecto hemos utilizado los siguientes algoritmos de clasificación:

- **Regresión Logística (Logistic)** – Modelo de regresión logística con regularización (para no ajustarse excesivamente a los datos de entrenamiento). Este algoritmo se utiliza mucho dentro del ámbito médico ya que su funcionamiento es fácil de entender y se puede conocer la influencia que tiene cada atributo.
- **Árbol de decisión (C4.5)** – Es un árbol de decisión que se construye un particionado recursivo de los ejemplos y donde se utiliza la ganancia de información normalizada (diferencia en entropía) para seleccionar el atributo que mejor divide los datos en cada nodo. Este tipo de algoritmos (árboles) se utilizan mucho ya que son muy fáciles de interpretar, pudiendo saber qué decisiones han llevado al algoritmo a predecir la clase.
- **Adaboost.M1 (con decision stump)** – Es un ensemble (conjunto de clasificadores) formados en este caso por clasificadores muy simples (decision stumps) que simplemente seleccionan el mejor atributo para la clasificación y establecen un umbral en dicho atributo para clasificar el ejemplo como de una clase u otra. Adaboost sigue un modelo iterativo por el cual se va centrando en los ejemplos más difíciles. Este algoritmo también se puede interpretar ya que conocemos en cada iteración que atributo se escoge y con qué peso. Pero además de ser interpretable, este se ha utilizado ya que los resultados que se obtienen con este clasificador suelen ser buenos.
- **Naïve Bayes** – Modelo probabilístico basado en el teorema de Bayes. Asume la independencia de las variables. Este algoritmo es muy simple y fácil de entender.
- **XGBoost** – Es un algoritmo de tipo boosting. Estos emplean un conjunto de algoritmos "débiles" que al unir las capacidades predictivas de todos se consigue un algoritmo de clasificación robusto. XGBoost emplea arboles CART como algoritmo de clasificación "débil". Este algoritmo es más complejo de entender pero obtiene buenos resultados. No se ha utilizado de manera primordial por no ser del todo interpretable, pero al ser muy preciso en una gran variedad de problemas se ha utilizado para ver la diferencia con los algoritmos más interpretables.

- **UnderBagging** – Es un ensemble formado por diferentes clasificadores. El algoritmo se repite un número fijado de iteraciones realizando undersampling en cada iteración, es decir, en cada iteración se obtiene un subconjunto de datos aleatorio. Este subconjunto se utiliza para entrenar el clasificador de la iteración correspondiente para finalmente combinar las salidas de cada uno de los clasificadores entrenados. Este clasificador obtiene buenos resultados al utilizarlo en conjuntos de datos no balanceados (tener muchos más ejemplos de una clase que de la otra), es por ello que se ha decido emplearlo.
- **FARC-HD** – Método de clasificación basado en reglas de asociación difusas. Este genera reglas de clasificación difusas, se preseleccionan una serie de reglas basándose en una medida de evaluación y finalmente se utiliza un algoritmo genético para hacer selección de reglas y ajuste de estas. Al ser un buen algoritmo que genera reglas difusas se ha decidido emplearlo en alguno de los experimentos.
- **SVM** – Algoritmo de clasificación que trata de separar los ejemplos de las diferentes clases mediante un hiperplano. Se busca el hiperplano que tanga la máxima distancia (margen) con los puntos que estén más cerca de él mismo. Este se ha utilizado ya que es uno de los mejores algoritmos de la literatura.

6.2. Feature selection

La selección de atributos se utiliza para reducir la dimensionalidad del conjunto de datos, es decir, para seleccionar un subconjunto del conjunto inicial de atributos. El objetivo de estas técnicas es mejorar el resultado reduciendo la complejidad del problema. Además se consigue identificar los atributos más relevantes para el problema. En este proyecto se han empleado tres técnicas diferentes:

6.2.1. RANKER

En la selección de atributos tipo “Ranker”, se evalúa la capacidad predictiva de cada atributo respecto a la clase de manera independiente y se crea un ranking de los atributos sobre el cual se debe aplicar un umbral (cantidad de atributos o valor de mínimo evaluación obtenido) para seleccionar los atributos a utilizar en el modelo. Para evaluar los atributos existen diferentes opciones. En este proyecto al utilizar la herramienta Weka, hemos utilizado los evaluadores descritos en su libro [2]:

- ChiSquared – Calcula el estadístico Chi cuadrado de cada atributo respecto a la clase.
- GainRatio – Evalúa la calidad del atributo en base al ratio de ganancia de información.
- InfoGain – Evalúa la calidad del atributo en base a la ganancia de información.
- ReliefF – Evalúa el atributo basado en los vecinos más cercanos (ejemplos con valores similares para el mismo atributo).

6.2.2. FILTRO

En este tipo de selección de atributos en vez de obtener una lista ordenada de los atributos en base a su utilidad para la clasificación, directamente obtiene un subconjunto de atributos a utilizar por el modelo (no es necesario aplicar un umbral). Para ello, se utiliza por un lado un método de búsqueda y por otro lado un evaluador capaz de evaluar grupos de atributos. El método de búsqueda se encarga de ir seleccionando atributos iterativamente en base al evaluador seleccionado. Existen diferentes métodos de búsqueda y evaluadores, que al igual que con los de tipo Ranker se ha sacado la información del libro de Weka [2].

- Métodos de búsqueda
    - LinearForwarding – Realiza una búsqueda hacia adelante añadiendo un atributo en cada iteración. Tiene un sistema de back-tracking (para eliminar atributos que no aportan mejora) y otro mecanismo para limitar la búsqueda.
    - RankSearch – Utiliza primero un "ranker" (GainRatio) para evaluar cada atributo de manera independiente y añade al subconjunto de atributos cada atributo en el orden establecido por el ranker, seleccionando el mejor de los subconjuntos.
    - Scatter – Es un método de búsqueda basado en algoritmos evolutivos.
- Evaluadores de subconjuntos de atributos
    - Cfs – Considera lo bueno que es cada atributo de manera independiente para predecir la clase, además del grado de redundancia entre los atributos en el subconjunto. Es decir, se fomentan atributos correlacionados con la clase pero que tengan poca interrelación entre ellos.
    - Consistency – Se evalúa la consistencia del subconjunto de atributos proyectando los ejemplos sobre los valores de clase.

6.2.3. WRAPPER

Este tipo de selección de atributos actúa igual que el anterior difiriendo en el evaluador de atributos. Como evaluador se utiliza un algoritmo de clasificación y una medida de rendimiento (precisión, área bajo la curva ROC,…). El algoritmo de clasificación recibe el conjunto de datos reducidos con el conjunto de atributos a evaluar y devuelve la medida deseada. Finalmente el wrapper escoge aquel conjunto de atributos con mejor resultado. Hay que decir que este tipo de selección es la más costosa (computacionalmente hablando) ya que tiene que evaluar varias veces diferentes conjuntos de datos. En el proyecto se ha utilizado la ROC-AUC como medida dado el problema de balanceo de clases antes mencionado.

6.3. Discretización

Algunos algoritmos de clasificación obtienen mejores resultados cuando los atributos son nominales. Incluso existen algoritmos que no admiten atributos numéricos y todos tienen que ser atributos nominales. Para ello los atributos numéricos tienen que ser "discretizados" en diferentes rangos.

La discretización es una técnica que convierte los atributos numéricos en discretos (rangos). Existen multitud de algoritmos que realizan este proceso los cuales se pueden dividir en dos categorías: supervisados y no supervisados. Los primeros tienen en cuenta la clase para realizar el proceso mientras que los no supervisados no. En el proyecto se ha utilizado un algoritmo supervisado.

La idea es ir separando el atributo numérico en diferentes rangos basándose en la medida de la entropía (ecuación siguiente) hasta llegar a una condición de parada. Primero se ordenan los valores del atributo. Por cada posible corte se calcula la entropía, y se elige el punto de corte con menor entropía. El algoritmo se vuelve aplicar sobre las partes restantes hasta encontrarse en una condición de parada, que puede ser que todos (o casi todos) los intervalos sean "puros", es decir, que todas las instancias del intervalo pertenezcan a la misma clase.

$$E(S) = \sum_{i=1}^{c} -p_i * log_2 p_i$$

Por ejemplo en la Figura 11 se discretiza el espacio continuo en 3 partes. Para ello se han encontrado dos puntos de corte, c1 y c2, generando tres bloques o intervalos:

- Bloque A. Este bloque contiene todos los datos del intervalo $(-\infty, c1]$
- Bloque B. Este bloque contiene todos los datos del intervalo $(c1, c2]$
- Bloque C. Este bloque contiene todos los datos del intervalo $(c2, \infty)$

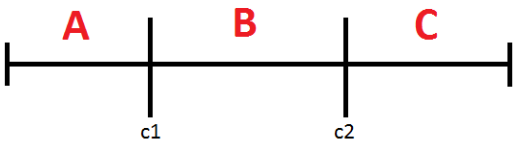


Figura 11. Representación gráfica de la discretización

6.4. Binarización

Esta técnica es utilizada para convertir los atributos nominales en atributos binarios, es decir, atributos con dos posibles valores. Esta transformación suele resultar en una mejora de la precisión ya que consigue modelar mejor los atributos nominales para algunos clasificadores. Dado un atributo nominal con N opciones, se generan N nuevos atributos binarios donde cada uno de estos nuevos atributos representan las opciones del atributo original. Las instancias tomaran valor 1 en los atributos binarios correspondientes al valor del atributo original. Por ejemplo, supongamos el siguiente conjunto de datos (Figura 12) centrándonos en el atributo *GRUPO*.

ATT1	ATT2	GRUPO
5	8	G1
1	4	G3
2	9	G1
3	6	G2

Figura 12. Ejemplo de conjunto de datos original a binarizar

Podemos observar que el atributo *GRUPO* toma tres valores diferentes: G1, G2 y G3. Al binarizar este conjunto de datos se transforma el atributo *GRUPO* generando tres nuevos atributos, uno por cada posible valor de *GRUPO*. Estos atributos son atributos binarios, es decir, solo pueden tener los valores 0 o 1 (ver Figura 13).

ATT1	ATT2	GRUPO_G1	GRUPO_G2	GRUPO_G3
5	8	1	0	0
1	4	0	0	1
2	9	1	0	0
3	6	0	1	0

Figura 13. Conjunto de datos binarizado

En ciertos casos de este proyecto la binarización no es adecuada. Esto sucede, por ejemplo, con el atributo TSI. Este atributo representa el estatus socio-económico del paciente y puede tomar diferentes valores: TSI 001, TSI 002-01, TSI 002-02, TSI 003, TSI 004 y TSI 005. Estos valores tienen un orden lógico para los farmacéuticos del SNS-O. Al binarizar este atributo, cada opción puede tener un peso diferente, pudiendo no ser lineal (TSI 002-01 < TSI 003 < TSI 001 por ejemplo) lo que no tiene lógica para ellos. Este caso ha sido codificado de manera diferente y se explica en secciones posteriores.

6.5. Instance selection

En el conjunto de datos generados se pudieron observar instancias duplicadas, es decir, se encontraron varios ejemplos (instancias o pacientes) con todos los valores iguales. Es más, en ciertos casos se encontraron instancias donde para los mismos datos tenían diferentes clases. Esto quiere decir que un algoritmo de clasificación nunca va a discriminar estos pacientes, si tienen los mismos valores exceptuando la clase, el algoritmo se tendrá que decantar por una clase u otra, clasificando incorrectamente alguna instancia.

Para tratar este problema se plantearon dos técnicas. La primera y la más sencilla era eliminar todas las instancias duplicadas. La segunda consistía en eliminar solo las de la clase negativa, dejando los ejemplos de la clase positiva.

Este planteamiento no era lógico desde el punto de vista clínico de acuerdo al equipo del SNS-O, por lo que se descartó continuar con dicho trabajo y se permitieron duplicados.

6.6. Feature engineering

Como se ha dicho, tener un buen conjunto de datos es muy importante para obtener buenos resultados y en muchas ocasiones los atributos empleados no son lo suficientemente buenos. Para esto se utiliza el feature engineering, que lo que trata es de generar nuevos atributos utilizando los datos y el conocimiento con el objetivo de mejorar los resultados obtenidos con el algoritmo de aprendizaje automático.

6.6.1. MuSer

Esta técnica es la adaptación de [5], que trata de buscar aquellas secuencias de medicamentos que más diferencian a los pacientes de diferentes clases para utilizarlas como atributos. Se busca por cada clase las secuencias de medicamentos que más se repiten, para posteriormente generar un atributo por cada una de estas secuencias indicando si dentro del conjunto de medicamentos del paciente existe la secuencia correspondiente o no.

En el artículo original utilizaban esta técnica con datos de análisis. Como el proyecto se centra en pacientes polimedicados se decidió emplear los datos de medicamentos, aunque también se han realizado pruebas con los datos DGPs.

Para construir el conjunto de datos para los datos de farmacia se realiza el siguiente proceso:

1- Se utilizan los medicamentos en codificación SUB7 o SUB5. A cada uno de los elementos de esta codificación se le asigna un número entero único, el que será su identificador.
2- Por cada paciente se obtienen todos los principios activos asociados (en codificación subgrupo 7 o 5) en los N (parámetro fijado a 6) meses anteriores a la fecha base.
3- En este paso se tiene por cada paciente una seguida de números enteros. En estos datos se separan los pacientes según su clase (sufren evento adverso o no).
4- En este paso se requiere de un algoritmo para obtener patrones frecuentes. A este algoritmo se le da la lista de secuencias (seguidas numéricas de los pacientes), y devuelve aquellas secuencias más frecuentes. Como algoritmo se ha utilizado PrefixSpan. Este algoritmo se ejecuta sobre cada separación, obteniendo las secuencias frecuentes para cada una de las clases.
5- En el paso anterior se pueden generar muchas secuencias frecuentes. Por ello se filtran aquellas que sean más discriminantes. Por cada secuencia frecuente se tiene el número de repeticiones existente. Habría que quedarse con aquellas con mayor número de repeticiones, pero dado que existe un no balanceo entre clases, las secuencias de la clase 0 (no sufre evento) tienes más repeticiones que las de la clase 1 (sufre evento). Para equilibrar se utiliza un peso en vez del número de repeticiones. Este peso se calcula como la división entre el número de repeticiones y el número de pacientes de la clase correspondiente.
6- Una vez se calculan los pesos se filtran las secuencias. Por cada secuencia de la clase 1 se mira su peso, se mira el peso de la misma secuencia pero de la clase 0, y si el peso es mayor en la clase 1 esta secuencia es seleccionada. En el caso que la secuencia no aparezca en las secuencias de la clase 0, siempre es seleccionada.
7- Cada una de las secuencias seleccionadas van a ser los nuevos atributos, donde tomarán valor 1 si la secuencia es una sub-secuencia del paciente y 0 en caso contrario.

Estos nuevos atributos generados se combinan con el resto de atributos.

6.6.2. Combinación de medicamentos

La idea es combinar diferentes atributos de medicamentos para ayudar al algoritmo de clasificación a obtener mejores resultados. Al principio se combinaban todos los medicamentos 2 a 2 de diferentes formas:

- $f(x,y) = x * y$
- $f(x,y) = x^2 * y^2$
- $f(x,y) = \begin{cases} 1, & x,y > 0 \\ 0, & en\ caso\ contrario \end{cases}$

Pero combinando los medicamentos 2 a 2 se generan muchas combinaciones: $\frac{n*(n-1)}{2}$. Para evitar tanta combinación se hace una preselección de atributos de farmacia: se utiliza un algoritmo de selección de atributos sobre el conjunto de datos original. Este algoritmo devuelve una lista con los atributos más representativos, de donde se obtendrán los atributos sobre medicamentos para combinarlos.

Por otra parte se estudia la prevalencia de la clase positiva (% de ejemplos) combinando atributos. El objetivo es encontrar combinación de medicamentos que tengan una alta prevalencia, es decir, encontrar combinaciones de medicamentos que se observen en más casos positivos que negativos. Se llega a calcular la prevalencia combinando hasta 4 medicamentos. Para reducir el número de combinaciones, antes de realizar las combinaciones se eliminan aquellos atributos (o combinaciones) que tengan una baja prevalencia. Finalmente con las combinaciones interesantes se genera un atributo binario donde tomará el valor 1 si el paciente consume todos los medicamentos combinados o 0 en caso contrario.

6.7. No balanceo

Ya se ha comentado que este trabajo presenta el problema del no balanceo, es decir, que existen muchas más instancias de la clase negativa que de la positiva. Tenemos que aproximadamente el 99% de los ejemplos son de la clase negativa (no sufre evento) y solo un 1% de la clase positiva (sufre evento). Para hacer frente a este problema se han utilizado las siguientes técnicas:

6.7.1. Oversampling

El objetivo del oversampling es aumentar el número de ejemplos de la clase minoritaria (positiva). El modelo más simple y quizás el más adecuado para este proyecto es el de replicar aleatoriamente los ejemplos de la clase positiva hasta obtener un conjunto de datos balanceado (mismo número de ejemplos de la clase positiva que de la negativa). Este proceso solo se realiza para el entrenamiento, el conjunto de test permanece intacto.

6.7.2. Undersampling

A diferencia del Oversampling lo que se trata es de eliminar instancias de la clase mayoritaria (negativa) hasta tener el mismo número que de la clase minoritaria (positiva). El proceso más simple que existe es el de eliminar instancias aleatoriamente, que es el que se ha empleado en este proyecto. Esta técnica puede ofrecer peores resultados que el oversampling ya que se está eliminando mucha información que podría ser útil.

6.7.3. Híbrido Oversampling Undersampling

En este proyecto se ha utilizado conjuntamente estas dos técnicas para balancear el conjunto de datos sin eliminar mucha información ni añadir ejemplos en exceso. Para ello se fija un número mayor que el número de ejemplos de la clase minoritaria y menor que el número de ejemplos de la clase mayoritaria, generando y eliminando ejemplos de cada clase hasta balancear el conjunto de datos al número dado.

6.7.4. OneClass Classification

Esta técnica utiliza únicamente los datos de una clase para la clasificación. Los algoritmos de OneClass reciben los datos de la clase minoritaria (positiva) y este busca la frontera de estos datos.

6.7.5. CostSensitive

Esta técnica asigna diferentes costes a cada clase. De esta manera se le asigna a la clase minoritaria un coste mayor para que el algoritmo sufra mayor penalización por clasificar mal un ejemplo de esta clase. El problema de esta técnica es que es difícil encontrar los pesos adecuados de cada clase.

7. Marco experimental

Los algoritmos de clasificación suelen tener parámetros configurables. En esta sección se presentan las configuraciones elegidas para los algoritmos empleados en el proyecto.

7.1. Validación cruzada

Para evaluar un conjunto de datos hay que separarlo de tal manera que una parte se utilice para entrenar el modelo (clasificador) y la otra para observar el comportamiento de este, de tal manera que los datos que se utilizan para evaluar no se utilicen para entrenar. Una de las técnicas más utilizadas es separar el conjunto de datos utilizando un porcentaje. Por ejemplo 70-30 donde el 70% de los ejemplos se utilizan para entrenar y el 30% para evaluar.

En este proyecto se ha utilizado otra técnica más avanzada: validación cruzada en k particiones (k-fold cross validation). Esta técnica divide el conjunto de datos en *k* partes tratando que cada parte tenga el mismo número de instancias de cada clase. Con esta división se generan k conjuntos de entrenamiento y k conjuntos de test. Para generar estos conjuntos se realizan iteraciones desde 1 hasta k, de tal manera que el conjunto que coincida con la iteración se utiliza como conjunto evaluador y la unión del resto como conjunto de entrenamiento. Hay que hacer notar que utilizando la validación cruzada los datos de los pacientes utilizados para entrenar nunca se utilizan para evaluar.

En este proyecto se ha fijado el valor de k a 5. El funcionamiento de la validación cruzada se puede observar de manera visual en la Figura 14.

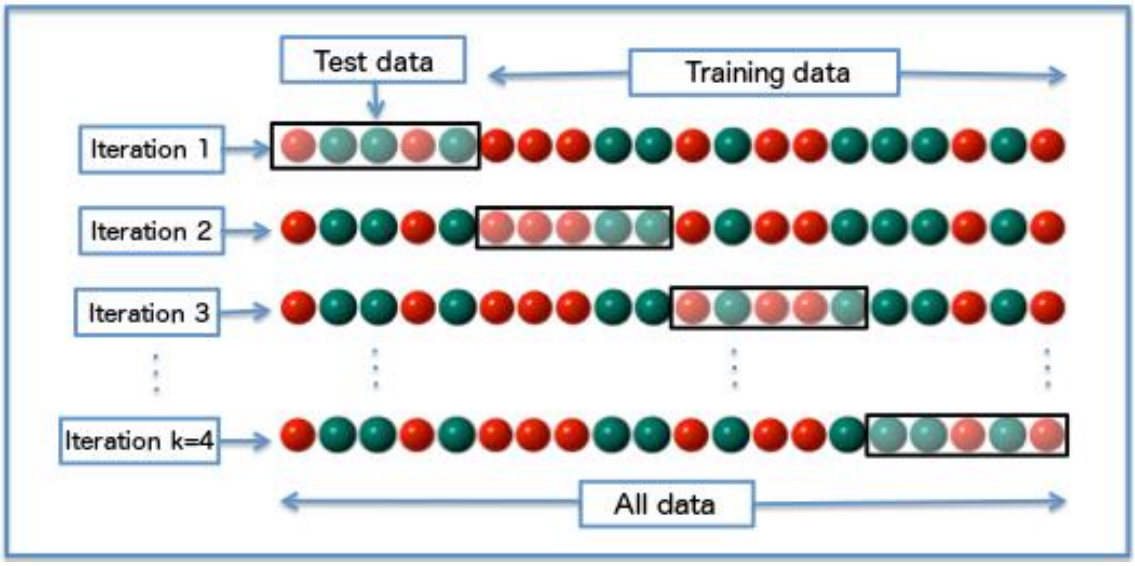


Figura 14. Ejemplo de validación cruzada con k=4. Figura extraída de Wikipedia.org

7.2. Fechas conjuntos de datos

Al principio se realizaban las pruebas utilizando 9/2014 como fecha de referencia, ya que los datos que disponíamos llegaban hasta 12/2014. Con esto se utilizaban los meses 10, 11 y/o 12 para calcular la clase.

Los datos a utilizar para predecir la clase eran los datos de los N_{GEN} meses anteriores a la fecha de referencia, pero como N_{GEN} no superaba los 6 meses existían datos de 2013 y 2014 que no se utilizaban. Además el número de ejemplos positivos era muy bajo. Por estas dos razones se planteó una nueva fecha de referencia: 12/2013. Con esta fecha se utilizan los datos de 1/2014, 2/2014 y 3/2014 para calcular la clase, que no interfiere con los datos obtenidos por la otra fecha de referencia (4/2014-9/2014). Esto se puede ver en la Figura 15.

Con las dos fechas de referencias se obtienen dos conjuntos de datos diferentes. Al principio se ejecutaban los algoritmos en cada uno de los conjuntos, para comprobar que los resultados no eran muy diferentes. Una vez comprobado esto, las siguientes ejecuciones se centraron en la unión de estos dos conjuntos de datos.

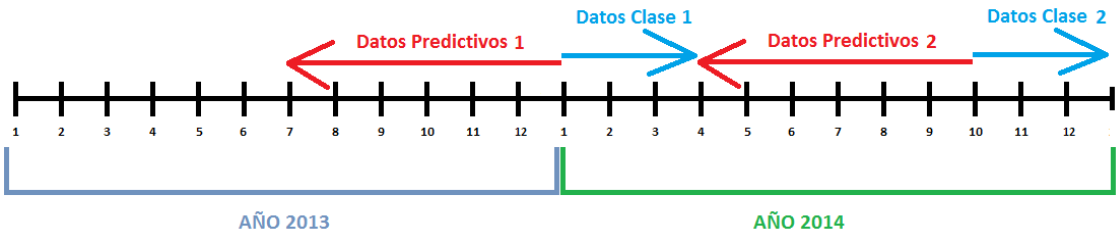


Figura 15. Ejemplo para las fechas de referencia 9/2014 y 12/2013

7.3. Unión de conjuntos de datos

Con el fin de aumentar el número de ejemplos positivos se decidió también utilizar cada mes entre 12/2013 y 9/2014 como fechas de referencia (ambos incluidos) y unir todos los conjuntos en uno solo. Esto genera un problema al evaluar el conjunto de datos, ya que los datos de un paciente en el mes 1/2014 se utiliza para entrenar el modelo y los datos del mismo paciente pero el mes 2/2014 se utiliza como testeo. Dado que en dos meses consecutivos los datos de un paciente no difieren mucho esto ayuda al sistema a obtener mejores resultados pero no son resultados del todo reales.

Para solventar esto se realiza un particionado de los datos con conocimiento. Se trata de que todos los datos de un paciente estén en una única partición, de tal manera que sus datos se utilicen para entrenar o para ser testeados pero no para ambas cosas a la vez.

Para realizar este particionado se necesita obtener cada conjunto de datos con el identificador del paciente. De estos conjuntos de datos se separan los datos por paciente, de tal manera que se tengan por cada paciente todos sus datos asociados en un bloque. Como se ha comentado el proyecto emplea la validación cruzada de 5 particiones, por lo que hay que repartir estos bloques en las 5 particiones de tal manera que cada partición tenga aproximadamente el mismo número de ejemplos. La solución que empleamos fue ordenar los bloques según su número de ejemplos, e ir cogiendo uno a uno los bloques con mayor número de ejemplos asociándolos a la

partición con menor ocupación. De esta manera se consigue una distribución casi uniforme entre las particiones (todas las pariciones tienen aproximadamente el mismo número de ejemplos) y que todos los datos del mismo paciente estén en una única partición.

7.4. Datos de 2015

Posteriormente se obtuvieron los datos de 2015, los cuales se utilizaron únicamente para validar el modelo, es decir, para comprobar cómo se comporta el modelo con datos totalmente nuevos.

En este caso se obtienen los datos cada mes de 2015 hasta llegar a septiembre (incluido). No podemos coger más meses ya que se necesitan 3 meses para el cálculo de la clase. Estos 9 conjuntos de datos se evalúan de manera independiente, y también se genera otro conjunto de datos: la unión de los 9 conjuntos en uno solo.

Como entrenamiento se utiliza el conjunto de datos resultante de la unión de los conjuntos generados utilizando las fechas 9/2014 y 12/2013.

7.5. Configuraciones de los métodos utilizados

7.5.1. Regresión Logística (Logistic)

Este algoritmo tiene un parámetro configurable: *ridge*. Este valor es el parámetro de regularización que nos permite no ajustarnos demasiado a los datos de entrenamiento para generalizar mejor.

En muchos experimentos este parámetro se ha dejado por defecto: 10^{-8}. En otros experimentos se ha realizado un proceso de búsqueda para ajustar este parámetro. Para ello se realiza un proceso de búsqueda en el intervalo $[10^{-4}, 10^{4}]$ tratando de encontrar el valor que mayor ROC AUC obtenga. Para calcular la ROC AUC se ejecuta una validación cruzada de 5 particiones sobre la partición de entrenamiento, realizando la media aritmética entre los 5 resultados.

7.5.2. Árbol de decisión (C4.5)

Este algoritmo tiene 4 parámetros configurables:

- Poda. Especifica si podar el árbol o no una vez construido. Se han utilizado las dos versiones.
- Valor de confianza para la poda. Se ha dejado el valor por defecto: 0.25
- Número mínimo de instancias por hoja. Se ha dejado el valor por defecto: 2
- Suavizado de Laplace. Utilizar la técnica del suavizado de Laplace para suavizar las probabilidades. Se han empleado las dos opciones: con y sin suavizar.

7.5.3. Adabost.M1

Este algoritmo tiene dos parámetros configurables: el clasificador a utilizar y el número de iteraciones a realizar. Como se ha comentado antes se utiliza el algoritmo Decision Stump como clasificador. El número de iteraciones se ha establecido a 50.

7.5.4. Naïve Bayes

Este algoritmo no tiene parámetros configurables.

7.5.5. XGBoost

Este algoritmo tiene una gran variedad de parámetros de configuración. Utilizamos un algoritmo de búsqueda exhaustiva para encontrar la combinación de parámetros que utilizaremos. Los parámetros a los que encontrar valor son:

- **max_depth.** Máxima profundidad de los árboles.
- **min_child_weight.** Suma mínima de los pesos de las instancias necesarias en los hijos
- **gamma.** Reducción de pérdida mínima necesaria para hacer una partición adicional en un nodo hoja del árbol.
- **subsample.** Ratio de ejemplos a utilizar para el entrenamiento (1 = 100%).
- **colsample_bytree.** Ratio de características a utilizar a la hora de construir cada árbol.
- **reg_alpha.** Término de regularización.

Para la búsqueda, primero se buscan los valores para los parámetros *max_depth* y *min_child_weight* dentro de los conjuntos {3,5,7,9} y {1,3,5} respectivamente. Una vez tenemos estos parámetros miramos si los valores siguientes y anteriores dan mejores resultados (ya que en los conjuntos anteriores solo buscamos sobre números impares). Es decir, si *max_depth* nos da un valor *a* y *min_child_weight* un valor *b* se realiza otra búsqueda sobre los conjuntos {*a-1, a, a+1}* para *max_depth* y *{b-1,b,b+1}* para *min_child_weight*.

El siguiente parámetro a ajustar es *gamma*. Se buscan valores en el conjunto {0, 0.1, 0.2, 0.3, 0.4}.

Una vez tenemos *gamma* pasamos a *subsample* y *colsample_bytree*. Ambos parámetros se buscan en el conjunto {0.6, 0.7, 0.8, 0.9}. Una vez tenemos el valor realizamos lo mismo que con *max_depth* y *min_child_weight* restando/sumando 0.05 a los valores devueltos. Finalmente vemos si es mejor utilizar el conjunto entero (*subsample*=1) o el valor devuelto por la búsqueda.

Por último tratamos de ajustar el valor para *reg_alpha*. Los valores para este parámetro se buscan en el conjunto {0.000001, 0.001, 0.1, 1, 100}.

Aparte de los parámetros a ajustar se han establecido los siguientes valores para los siguientes parámetros. El resto de parámetros:

- **objective: 'binary:logistic'.** Función objetivo a utilizar junto con el tipo de problema (binario o multiclase).

- **scale_pos_weight: 1.** Balance de los pesos entre las instancias de clase positiva y negativa.
- **learning_rate: 0.1.** El tamaño de paso utilizado en el proceso de actualización para impedir el sobreajuste.
- **n_estimators: 1000.** Número de árboles a utilizar.

7.5.6. UnderBagging

Este algoritmo requiere de un clasificador. En este proyecto se ha utilizado el clasificador C4.5. Los parámetros a configurar son:

- **prune:** parámetro que indica si podar el árbol o no después de construirlo. En el proyecto si se ha utilizado la poda.
- **confidence:** Nivel de confianza. Se ha establecido a 0.25
- **minItemsets:** número mínimo de instancias por hoja. Se ha establecido a 2
- **nClassifiers:** número de iteraciones a realizar. Se ha establecido a 100

7.5.7. FARC-HD

- **Número de valores lingüísticos:** 5
- **Mínimo soporte:** 0.05
- **Maximum confidence:** 0.8
- **Profundidad de los árboles:** 3
- **Parámetro K de la prescreening (kt):** 2
- **Máximo número de evaluaciones: 15000**
- **Tamaño de la población: 50**
- **Parámetro alpha:** 0.15
- **Bits por gen:** 30

8. Resultados experimentales

En este capítulo se presentan los experimentos realizados a lo largo del proyecto. Los experimentos se exponen en un orden cronológico siendo el primero el experimento más antiguo.

Los experimentos realizados se centran en el evento adverso cardiovascular, explicado anteriormente. Excepto el último experimento. El último experimento se extiende la metodología para la predicción de hospitalizaciones potencialmente evitables (HPE).

Para ese experimento el SNS-O nos hizo llegar la información de que ingresos se consideran HPE. Con la información de los HPE se establece la clase del conjunto de datos, es decir, la variable a predecir.

Tenemos diferentes motivos de HPE (ver Tabla 13), pero en este caso se han utilizado todos los motivos como si de uno solo se tratase, viendo el problema como un problema de clasificación binaria, es decir, si va a sufrir un HPE o no.

Motivo	Nº registros
Angina	80
Asma	116
Deshidrata	124
Diabetes	21
EPOC	1288
FalloCardCong	219

Tabla 13. Motivos HPE y el número de registros de cada motivo

Hay que tener en cuenta que el número de registros de la Tabla 13 no es exactamente el número de pacientes de la clase positiva que tendremos debido a que utilizamos la fecha de referencia para establecer los pacientes positivos y solo serán positivos aquellos que tengan un evento adverso en los tres meses posteriores a la fecha de referencia.

8.1. Reducción de combinaciones

Como se ha dicho existen muchas combinaciones posibles para generar el conjunto de datos: varias opciones de CMBD, de farmacia,… Para determinar que combinación ofrece mejores resultados, se ejecutaron diferentes algoritmos de clasificación sobre todas las combinaciones posibles. Los resultados se ordenaron utilizando la ROC-AUC de manera descendente para observar que opciones aparecen más arriba, es decir, que opciones ofrecen mejores resultados.

Para observar las opciones de manera rápida se crea una tabla (ver Tabla 14) con tantas columnas como diferentes opciones y una fila por cada resultado. La tabla tiene el valor 1 si en el resultado (fila) se ha utilizado la opción columna. Por ejemplo, si el resultado de la fila 4 utiliza SUB7 (columna 5), el valor de la fila 4 y columna 5 será 1. Para una mejor visualización se colorea en rojo estos valores.

Se estudian las siguientes combinaciones:

- Farmacia: Conteo de principios activos (PA), conteo agrupado por los diferentes subgrupos (S1, S3, S5 y S7)
- DGPs: Utilizar los datos de DGPs
- Episodios: Episodios agrupados por el primer carácter (EPI1) o por los dos primeros caracteres (EPI2)
- CMBD: Agrupado por el primer carácter (C1) o por los dos primeros caracteres (C2).
- Charlson: Utilizar el índice de Charlson (NUM) o cada componente de Charlson de manera individual (SEP).
- Antecedentes (ANT): Utilizar únicamente el antecedente de episodios (E), utilizar únicamente el antecedente de CMBD (C) o utilizar ambos antecedentes simultáneamente (EC).
- Mes: Utilizar 3 o 6 meses para la recolección de datos.

Estos resultados solo se ejecutaron con el conjunto de datos generado con la fecha 9/2014.

Tras observar la Tabla 14 se llegó a las siguientes conclusiones:

- Seguir utilizando las diferentes opciones de farmacia podía ser de interés para los médicos
- Utilizar los DGPs mejor que no utilizarlos
- Utilizar los episodios agrupado por el primer carácter
- Utilizar CMBD agrupado por los 2 primeros caracteres
- Utilizar Charlson separado
- Utilizar ambos antecedentes simultáneamente
- Utilizar 6 meses para la obtención de datos

Además el SNS-O nos indicó que utilizar 3 meses para la obtención de datos no era del todo correcta. Al utilizar 3 meses se estaba utilizando la información de los meses 7, 8 y 9 de 2014. Estos meses corresponden a la época de verano por lo que sugirieron utilizar 6 meses para la obtención de los datos.

Índice	FARMACIA					DGPS	EPISODIOS		CMBD		CHARLSON		ANT			MES	
	PA	S1	S3	S5	S7		EPI1	EPI2	C1	C2	NUM	SEP	E	C	EC	3	6
1	0	0	0	0	1	1	0	0	0	1	0	1	0	1	0	0	1
2	0	0	0	0	1	1	0	0	0	1	0	1	0	0	1	0	1
3	0	0	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1
4	0	0	0	0	1	0	1	0	0	1	0	1	0	0	1	0	1
5	0	0	0	0	1	1	1	0	0	1	0	1	0	1	0	0	1
6	0	0	0	0	1	1	1	0	0	1	0	1	0	0	1	0	1
7	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0
8	0	0	0	0	1	1	0	1	0	1	0	1	0	1	0	0	1
9	0	0	0	0	1	1	0	1	0	1	0	1	0	0	1	0	1
10	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1	0	1
11	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	0	1
12	0	0	0	0	1	0	0	1	0	1	0	1	0	0	1	0	1
13	0	0	0	0	1	0	0	0	0	1	0	1	0	1	0	0	1
14	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0
15	0	0	0	0	1	1	1	0	0	1	0	1	1	0	0	0	1
16	0	0	0	0	1	1	1	0	0	1	0	1	0	0	0	0	1
17	0	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1
18	0	0	0	0	1	1	1	0	0	0	0	1	0	1	0	0	1
19	0	0	0	0	1	1	0	1	0	1	0	1	1	0	0	0	1
20	0	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	1
21	0	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	1
22	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0	1
23	0	0	0	0	1	1	0	1	0	1	0	1	0	0	1	1	0
24	0	0	0	0	1	1	0	1	0	1	0	1	0	1	0	1	0
25	0	0	0	1	0	0	0	1	0	1	0	1	0	0	1	1	0
26	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1	0
27	0	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	1
28	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0
29	0	0	0	1	0	0	0	1	0	1	0	1	0	1	0	1	0
30	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	1
31	0	0	0	0	1	0	0	1	0	1	0	1	1	0	0	0	1
32	0	0	1	0	0	0	0	1	0	1	0	1	1	0	0	1	0
33	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0
34	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0	1	0
35	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	1	0
36	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0
37	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	1
38	0	0	0	0	1	0	1	0	0	1	0	1	1	0	0	0	1
39	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	1	0

Tabla 14. Mejores opciones para los conjuntos de datos

8.2. Agrupación CMBD

La agrupación realizada para CMBD inicialmente, por caracteres, no era adecuada hablando desde el punto de vista clínico. Los médicos realizaron una agrupación por niveles. Esta agrupación consta de 4 niveles, donde el nivel 4 es el código de ingreso completo. Tras realizar los experimentos se observó que los resultados obtenidos eran mejores al utilizar la agrupación por el primer nivel (de los cuatro niveles propuestos).

8.3. Selección de características

8.3.1. Selección de mejores métodos

Los conjuntos de datos que se generan tienen muchos atributos, superando los 1000 atributos al utilizar SUB7 como farmacia. Esto provoca que los algoritmos sean más lentos y que los resultados no sean buenos del todo. Por ello se decide realizar una selección de atributos. Con los métodos antes mencionados se reduce la dimensionalidad del conjunto de datos y se ejecutan los algoritmos de clasificación sobre el nuevo conjunto de datos.

El objetivo de esta prueba es reducir el número de algoritmos de selección a utilizar, ya que la lista es bastante extensa. Una vez ejecutados los algoritmos, solo se recogen los datos de aquellos que tengan un resultado "bueno" (mayores que un umbral) mirando la ROC AUC como medida, llegando a considerar 179 resultados. Con estos datos se estudia que algoritmos de selección son los que mejores resultados obtienen. Para ello se mira cuantas veces aparece cada uno de ellos, quien es el que obtiene mejor resultado, quien el peor y la media por cada método. Con cada una de estas opciones se ordenan los resultados de mejor a peor (mayor a menor) generando un ranking.

Por ejemplo, para el ranking del conteo. Miramos cuantas veces ha sido seleccionado cada algoritmo, teniendo un valor de aparición por cada uno. Estos se ordenan de manera descendente (el que más veces ha sido elegido el primero) generando una especie de ranking. De esta manera el algoritmo de la primera posición tendrá ranking 1, el de la segunda posición ranking 2 y así sucesivamente.

Los métodos utilizados con su respectivo ranking por cada prueba se pueden ver en la Tabla 15.

En este experimento solo se utilizaron los algoritmos de selección tipo "Ranker" y tipo "Filtro", excluyendo a los de tipo "Wrapper". Estos últimos se llegaron a ejecutar en alguna ocasión pero dado que requieren mucho tiempo de ejecución sin ofrecer mejora en los resultados, se dejaron de utilizar.

Tras este proceso se determinan los algoritmos de selección a utilizar en las siguientes pruebas. Estos aparecen subrayados en la Tabla 15.

Algoritmo de selección	Ranking Conteo	Ranking Media	Ranking Max	Ranking Min
Filtro Scatter con Cfs	5	4	3	2
Ranker con InfoGain	**2**	**1**	**2**	**1**
Ranker con ChiSquared	**1**	**2**	**4**	**3**
Filtro RankSearch con Cfs	6	5	1	2
Ranker con RelieF	7	7	6	3
Filtro RankSearch con Consistency	**3**	**3**	**5**	**2**
Filtro Linear Fordward con 1	4	6	3	2

Tabla 15. Ranking de los métodos de selección

8.3.2. Variantes de los algoritmos de selección tipo RANKER

Este tipo de algoritmos no realiza una selección de atributos, este asigna a cada atributo un valor que se calcula utilizando un evaluador. Lo que se consigue al ejecutar este algoritmo es un ranking con los atributos ordenados de manera descendente por el valor de cada atributo. Con esta lista se elige un corte y todos aquellos atributos que tengan mayor puntuación que el corte son los atributos que se seleccionan. Para la elección de este corte se han probado varias opciones:

- Calcular la media de los valores y seleccionar aquellos atributos con valor mayor o igual a la media
- En ocasiones los atributos obtenían puntuación 0 (dependiendo del evaluador). Seleccionar todos aquellos atributos con valor mayor a 0
- Calcular la media de los valores sin tener en cuenta los atributos con valor 0, y seleccionar aquellos con valor mayor o igual a esa media.

El problema con estos métodos es que el número de atributos seleccionados es mayor cuantos más atributos tenga el conjunto de datos original. Para evitar este problema se planteó elegir los *K* mejores atributos. Tras realizar pruebas con todas las variantes se observó que los mejores resultados se obtenían utilizando la técnica de la media, por lo que en las pruebas se utiliza este método.

8.4. Varias opciones de farmacia en el mismo conjunto de datos

Tras realizar varios experimentos con diferentes opciones de farmacia, se observa que hay ciertos medicamentos que son seleccionados por los algoritmos de selección. Dependiendo de la agrupación de farmacia elegida los medicamentos varían, por lo que se decide unir diferentes grupos (como SUB5 y CODMED) y ejecutar los experimentos. Lo que se consigue es tener un conjunto de datos con atributos de varios grupos de farmacia simultáneamente. Estos resultados no fueron buenos, ya que existía mucha redundancia entre estos atributos, por lo que se descartó esta opción.

8.5. Número de meses para determinar la clase

En los primeros experimentos se determinaba la clase mirando el siguiente mes o los tres meses siguientes a la fecha de referencia. En los resultados se apreciaba una pequeña mejora utilizando 1 mes para determinar la clase que utilizando 3 meses, ya que es más fácil de predecir. Tras esto se acordó con el SNS-O centrase en 3 meses para reducir la combinatoria sabiendo que los resultados a 1 mes son parecidos (a 1 mes mejora levemente). Con esto se consiguió reducir las posibles combinaciones a la mitad.

8.6. Farmacia

Existen 5 opciones diferentes para farmacia: SUB1, SUB3, SUB5, SUB7 y CODMED. Con el objetivo de reducir combinaciones y tiempo de espera entre resultados y resultados, se decidió, junto con el equipo del SNS-O, utilizar dos opciones en base a los resultados: SUB5 y CODMED.

8.7. Combinación de clasificadores

Otra prueba realizada fue combinar los resultados de diferentes clasificadores en uno. Cada clasificador nos dice para cada ejemplo a clasificar la probabilidad de pertenecer a una clase o a otra. Lo que hicimos fue calcular la media de estas probabilidades generando así una nueva salida.

8.8. Discretización manual

Ya se ha comentado que en algunos experimentos se utiliza la discretización, dividir los atributos continuos en varios rangos, con el objetivo de "ayudar" al clasificador a obtener mejores resultados. El SNS-O indicó que en el caso del DGPs sería más adecuado discretizarlos en rangos con conocimiento, es decir, que tengan sentido clínico. Para ello nos hicieron llegar los rangos por cada uno de los códigos DGPs, que se pueden consultar en la Tabla 16. En la tabla podemos ver los puntos de corte a utilizar para generar los intervalos. Por ejemplo utilizando IMC se generarían los siguientes intervalos:

$$\{(-\infty, C1], (C1, C2], (C2, C3], (C3, \infty)\} =>_{IMC} \{(-\infty, 18], (18{,}24], (24{,}30], (30, \infty)\}$$

En los experimentos se han empleado tres combinaciones posibles de discretización:

- Solo discretizar los DGPs con los rangos establecidos por el SNS-O, sin modificar el resto de atributos
- Discretizar todos los atributos numéricos pero utilizando el proceso de discretización automática. Hay que tener en cuenta que en este caso los DGPs se discretizan de manera automática, no con los rangos recibidos
- Primero discretizar los atributos DGPs con los rangos recibidos y posteriormente discretizar el resto de atributos con el proceso automático.

	C1	C2	C3
IMC	18	24	30
TAD	80	90	100
TAS	120	140	160
COL	200	240	350
HDL	40		60
LDL	100	130	190
TGC	150	200	500
ALT	40	80	120
AST	35	70	110
HB A1C	5,7	6,5	8
FILTGLOM	30	60	100
GGT	50	100	150
INR	2	4	6

Tabla 16. Rangos establecidos por el SNS-O para los datos de DGPs

8.9. TSI

Cada paciente tiene asociado un código TSI que especifica el estatus socio-económico del paciente, que pueden ser TSI 001, TSI 002-00, TSI 002-01, TSI 002-02, TSI 003, TSI 004, TSI 005 y TSI 006. Esto se llevaba al conjunto de datos como un atributo discreto con esas opciones. Los médicos comentaron que a mayor TSI menor riesgo debería haber, por lo que sería más adecuado codificar el TSI de manera numérica. Tras esto se eliminaron los TSI 002-00 y TSI 006, y se empezó a codificar los TSI de la siguiente manera:

- TSI 001 pasa a ser 1
- TSI 002-01 pasa a ser 2
- TSI 002-02 pasa a ser 3
- TSI 003 pasa a ser 4
- TSI 004 se fusiona con TSI 005 y pasan a ser 5

Con el conjunto de datos de la unión de las fechas de referencia 12/2013 y 9/2014 se consigue la distribución entre clases mostrada en la Tabla 17.

	Clase		
TSI	**Positiva**	**Negativa**	**Total**
1	19 (0,62%)	3.044 (99,38%)	3.063 (4,01%)
2	549 (1,05%)	51.940 (98,95%)	52.489 (68,80%)
3	91 (0,70%)	12.968 (99,30%)	13.059 (17,12%)
4	27 (0,53%)	5.027 (99,47%)	5.054 (6,62%)
5	6 (0,23%)	2.622 (99,77%)	2.628 (3,44%)
Total 2015	692 (0,91%)	75.601 (99,09%)	**76.293 (100%)**

Tabla 17. Distribución de los datos entre las clases con la nueva representación de TSI

En la Tabla 18 podemos ver los pesos asignados por la regresión Logística para cada uno de los valores de TSI. Como se puede ver no se cumple del todo que a mayor TSI menor riesgo (cuanto mayor es el peso menor es el riesgo). Pero al codificarlo como atributo numérico se obtiene un único peso para todo el TSI: *0,0562*. De esta manera si se cumple que a mayor TSI menor riesgo.

TSI	Peso
TSI 001	-0,0899
TSI 002-00	17,9719
TSI 002-01	-0,1566
TSI 002-02	-0,0023
TSI 003	-0,4716
TSI 004	0,3361
TSI 005	18,412
TSI 006	34,1047

Tabla 18. Pesos que asigna la regresión logística al atributo TSI discreto

8.10. Episodios

En el apartado 8.1 se escogió la agrupación por el primer carácter para los episodios ya que es la opción que mejores resultados obtenía. Para calcular las variables de episodios se cuenta el número de episodios (agrupado por el primer carácter) asociados al paciente donde el episodio este abierto o se haya cerrado en el último mes, pero que la fecha de inicio no sea menor a seis meses anteriores a la fecha de referencia.

Tras recibir los comentarios de los médicos se decidió añadir una nueva opción para los episodios. Esta opción se calcula igual que la opción anterior pero eliminando la restricción de la fecha de inicio. Tras realizar pruebas se decidió utilizar estas dos opciones conjuntamente, es decir, tener el doble de columnas de episodios.

Posteriormente, dado que las dos opciones que se utilizaban eran muy similares se propuso cambiar la forma de codificar los episodios. Con la colaboración de los médicos se propusieron tres maneras diferentes:

1. Episodios que están abiertos o han sido cerrados en el último mes. Esta es la misma opción que la primera (sin la restricción de la fecha de inicio)
2. Todos los episodios abiertos en los últimos 6 meses.
3. La tercera opción es la combinación de las dos anteriores, pero a diferencia que antes sin tener el doble de columnas: episodios que están abiertos, los que han sido cerrados en el último mes o han sido abiertos en los últimos 6 meses pero no en el último mes.

Tras realizar las pruebas con cada una de estas opciones se llegó a la conclusión que la primera opción es la más adecuada de acuerdo a los resultados (eran similares y el modelo final era más simple).

8.11. Resultados final

Tras realizar los experimentos anteriores y reunirnos con el SNS-O se llegó a realizar un último informe con experimentos para la predicción de eventos adversos cardiovasculares. Para estos experimentos se generaron conjuntos de datos con las siguientes opciones basándonos en los resultados de experimentos anteriores:

- **CMBD:** Conteo de ingresos agrupado por nivel 1.
- **Episodios:** Conteo de todos los episodios que están abiertos o han sido cerrados en el último mes agrupados por el primer carácter.
    - Se han considerado los episodios que estén **entre el 7 y el 9 después de la letra** correspondiente al nivel, **salvo para la Z** donde se cuentan **todos**.
- **Número de meses para los datos (N_{GEN}):** 6 meses
- **DGPs**: Media de los resultados de 3 meses a partir del último existente
- **Antecedentes**: Tanto antecedente de CMBD como antecedente de episodios
- **TSI**: Grupo TSI del paciente (001, 002A, 002B,…) como atributo numérico.
- **Farmacia**: Se utilizan dos codificaciones: SUB5 y CODMED
- **Índice de Charlson**: Charlson separado – Utilizar cada componente de Charlson como atributo
- **Conteo de los principios activos diferentes que toma el paciente**. El conteo se realiza agrupando de la misma manera que se agrupan los datos de farmacia. Si se utiliza farmacia CODMED, el conteo se realizará agrupando en CODMED. Además siempre se añade el conteo agrupando por SUB1.
- **Clase** (variable a predecir): Se ha utilizado 3 meses para la predicción del evento
- **Clasificador:** En este experimento solo se utiliza la regresión logística (Logistic). Se utilizan los parámetros por defecto en caso de que no se especifique.

Para este caso se generaron dos conjuntos de datos, uno con la fecha de referencia 9/2014 y otro con la fecha de referencia 12/2013. Estos dos conjuntos se unieron formando uno único, sobre el que se ejecutó el clasificador.

También se generaron todos los conjuntos de datos variando la fecha de referencia mes a mes entre 12/2013 y 9/2014 (ambos incluidos, en total 10 conjuntos) uniendo todos en uno único. Sobre este conjunto de datos se realizó la validación cruzada (teniendo en cuenta que los datos un paciente estén en una única partición de la forma explicada anteriormente).

Además de estos datos se generaron 9 conjuntos de datos de 2015, desde 1/2015 hasta 9/2015. Estos se utilizaron individualmente y uniendo los 9 en un único conjunto. El clasificador se entrenó con el conjunto de datos de la unión de 9/2014 y 12/2013 y se calcularon los resultados utilizando los conjuntos de 2015.

En la Tabla 19 se pueden consultar en número de ejemplos por cada uno de los conjuntos de datos generados.

Fecha de referencia	Clase			
	Positiva	%	Negativa	%
09/2014	347	0,83%	41.556	99,17%
12/2013	345	1,00%	34.045	99,00%
12/2013 + 09/2014	692	0,91%	75.601	99,09%
Unión 12/2013 a 09/2014	3.389	0,88%	383.548	99,12%
01/2015	350	0,82%	42.271	99,18%
02/2015	347	0,81%	42.282	99,19%
03/2015	359	0,84%	42.280	99,16%
04/2015	333	0,78%	42.410	99,22%
05/2015	318	0,74%	42.699	99,26%
06/2015	309	0,72%	42.889	99,28%
07/2015	332	0,77%	42.936	99,23%
08/2015	332	0,77%	42.996	99,23%
09/2015	315	0,73%	43.071	99,27%
Total 2015	2.995	0,77%	383.834	99,23%

Tabla 19. Número de ejemplos de cada clase por cada conjunto de datos

En los experimentos se realizaron diferentes transformaciones antes de obtener los resultados. En las tablas las transformaciones realizadas aparecen en la columna MODELO. Se utilizaron algoritmos de selección de características y la técnica del oversampling. En la Tabla 20 encontramos las diferentes combinaciones de transformaciones y como se han denominado en las tablas de resultados.

MODELO	DESCRIPCIÓN
Inicial	No se realiza ninguna transformación sobre el dataset
Over.	Se aplica Oversampling sobre el dataset de entrenamiento
Sel. Att.	Se realiza el proceso de selección de atributos. En otra columna aparecerá el método de selección utilizado pudiendo ser “RANKER_1” o “RANKER_3”.
Over. + Sel. Att.	Primero realizamos Oversampling sobre el dataset y se aplica la selección de atributos. Igual que antes tenemos dos métodos de selección posibles: “RANKER_1” y “RANKER_3”
Sel. Att. + Over.	En este caso primero realizamos la selección de atributos utilizando el “RANKER_1” o el “RANKER_3” y posteriormente se aplica Oversampling sobre el dataset reducido.

Tabla 20. Modelos utilizados con su denominación y descripción

Para la selección de características se han utilizaron 2 algoritmos diferentes. Ambos son métodos de selección tipo RANKER, donde nos devuelven la lista de los atributos ordenadas con una puntuación y se escogen aquellos atributos con puntuación mayor a la media de todas las puntuaciones. Estos se pueden consultar en la Tabla 21 junto con su denominación en las tablas de resultados.

METODO	DESCRIPCIÓN
RANKER 1	Ranker con ChiSquared como evaluador.
RANKER 3	Ranker con InfoGain como evaluador

Tabla 21. Algoritmos de selección utilizados junto con su denominación en las tablas de resultados

8.11.1. Validación cruzada sobre el conjunto de datos de la unión de 9/2014 y 12/2013

En este experimento se utilizan los conjuntos generados uniendo el conjunto con fecha de referencia 9/2014 y el conjunto con fecha de referencia 12/2013 tanto para farmacia SUB5 como para farmacia CODMED. Sobre este conjunto de datos se realiza la validación cruzada con 5 particiones, realizando la media aritmética para presentar los resultados en las tablas.

FARMACIA CODMED

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,7662**	0,7075	0,6864	0,6961	0,0211	0,0409	0,0430	39,8	Over. + Sel. Att.
RANKER_1	**0,7662**	0,6912	0,7024	0,6960	0,0204	0,0397	0,0432	41,6	Over. + Sel. Att.
RANKER_3	**0,7660**	0,7139	0,6850	0,6978	0,0215	0,0416	0,0438	42,6	Sel. Att. + Over.
RANKER_1	**0,7654**	0,7326	0,6706	0,6992	0,0224	0,0433	0,0426	41,2	Sel. Att. + Over.
RANKER_3	**0,7654**	0,7172	0,6749	0,6951	0,0214	0,0414	0,0424	42,6	Sel. Atr.
RANKER_1	**0,7646**	0,7268	0,6792	0,7019	0,0223	0,0432	0,0416	41,2	Sel. Atr.
-	**0,7280**	0,7411	0,6041	0,6681	0,0209	0,0405	0,0346	214	Inicial
-	**0,7260**	0,7524	0,5997	0,6703	0,0216	0,0418	0,0372	214	Over.

Tabla 22. Resultados para farmacia CODMED

FARMCIA SUB5

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,7692**	0,7300	0,6821	0,7045	0,0226	0,0438	0,0432	56	Sel. Att. + Over.
RANKER_1	**0,7688**	0,7154	0,6778	0,6944	0,0214	0,0415	0,0436	50,4	Over. + Sel. Att.
RANKER_3	**0,7682**	0,7406	0,6677	0,7016	0,0230	0,0444	0,0408	56	Sel. Atr.
RANKER_3	**0,7682**	0,7257	0,6647	0,6913	0,0218	0,0423	0,0448	48,4	Over. + Sel. Att.
RANKER_1	**0,7678**	0,7209	0,6763	0,6956	0,0218	0,0421	0,0378	51,6	Sel. Att. + Over.
RANKER_1	**0,7660**	0,7374	0,6589	0,6948	0,0227	0,0438	0,0396	51,6	Sel. Atr.
-	**0,6970**	0,7515	0,5780	0,6585	0,0210	0,0405	0,0314	325	Inicial
-	**0,6910**	0,7475	0,5751	0,6552	0,0208	0,0401	0,0333	325	Over.

Tabla 23. Resultados para farmacia SUB5

Como podemos ver en los resultados, realizar la selección de atributos en necesaria. Viendo el área bajo la curva ROC vemos que SUB5 obtiene resultados algo mejores. También se observa que utilizando SUB5 se utilizan más características que con CODMED.

Se puede observar que el F-measure es bajo ya que el PPV también lo es y esto es debido a la baja prevalencia de la clase positiva. Como el número de ejemplos de la clase positiva es mucho más bajo que el de la clase negativa, el número de verdaderos positivos (TP) que conseguimos, aun siendo un porcentaje alto (respecto al número total de positivos), es mucho menor que los falsos positivos (FP).

8.11.2. Validación cruzada sobre el conjunto de datos de la unión desde 12/2013 hasta 9/2014

Este experimento se utiliza la unión de todos los conjuntos de datos generados entre las fechas 12/2013 y 9/2014 (10 conjuntos diferentes). Los datos presentados en las tablas también son la media aritmética de las 5 particiones (pero las particiones están realizadas con conocimiento, teniendo en cuenta que los datos de un paciente van a estar en una única partición).

FARMACIA CODMED

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,7646**	0,7242	0,6742	0,6986	0,0212	0,0411	0,0446	46,2	Sel. Atr.
RANKER_1	**0,7644**	0,7238	0,6716	0,6969	0,0211	0,0409	0,0438	44,2	Sel. Atr.
RANKER_3	**0,7644**	0,7293	0,6601	0,6936	0,0211	0,0410	0,0438	46,2	Sel. Att. + Over.
RANKER_1	**0,7638**	0,7124	0,6784	0,6951	0,0204	0,0397	0,0430	44,2	Sel. Att. + Over.
RANKER_1	**0,7638**	0,7248	0,6642	0,6937	0,0209	0,0406	0,0442	42,4	Over. + Sel. Att.
RANKER_3	**0,7634**	0,7248	0,6610	0,6920	0,0208	0,0404	0,0434	40,4	Over. + Sel. Att.
-	**0,7458**	0,7327	0,6388	0,6840	0,0207	0,0401	0,0406	214	Inicial
-	**0,7450**	0,7415	0,6350	0,6856	0,0213	0,0412	0,0396	214	Over.

Tabla 24. Resultados para farmacia CODMED

FARMACIA SUB5

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,7648**	0,7049	0,6828	0,6937	0,0201	0,0390	0,0432	48,4	Over. + Sel. Att.
RANKER_1	**0,7644**	0,7041	0,6819	0,6928	0,0200	0,0389	0,0432	49,8	Over. + Sel. Att.
RANKER_3	**0,7640**	0,7211	0,6683	0,6942	0,0207	0,0402	0,0424	57,6	Sel. Att. + Over.
RANKER_1	**0,7640**	0,7265	0,6610	0,6927	0,0209	0,0406	0,0422	55,8	Sel. Att. + Over.
RANKER_1	**0,7632**	0,7406	0,6465	0,6917	0,0216	0,0418	0,0436	55,8	Sel. Atr.
RANKER_3	**0,7624**	0,7403	0,6474	0,6920	0,0216	0,0417	0,0430	57,6	Sel. Atr.
-	**0,7306**	0,7237	0,6279	0,6737	0,0197	0,0383	0,0376	325	Inicial
-	**0,7288**	0,7263	0,6276	0,6746	0,0199	0,0385	0,0362	325	Over.

Tabla 25. Resultados para farmacia SUB5

Comparando estos resultados con el experimento anterior, no se observa una fuerte mejora a pesar de utilizar más ejemplos. Esto puede deberse a la repetición de ejemplos, es decir, que los ejemplos de los diferentes meses son demasiado parecidos lo que no nos lleva a un aumento de información.

8.11.3. Validación cruzada con discretización

En este caso se utiliza el mismo conjunto de datos que en el experimento 8.11.1 (unión de 12/2013 y 9/2014). Este se transforma utilizando la discretización, que como se ha comentado tenemos la opción de realizar un proceso automático o utilizar los rangos establecidos por el SNS-O (solo los atributos de DGPs). Se han ejecutado 3 combinaciones de discretización:

- Solo discretizar los DGPs con los rangos establecidos por el SNS-O
- Discretización del conjunto de datos completo con el proceso automático
- Primero discretizar los DGPs con los rangos establecidos y posteriormente discretizar el resto de atributos con el proceso automático

Para reducir la combinatoria solo se ha empleado CODMED como farmacia.

SOLO DGPs CON RANGOS ESTABLECIDOS

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_1	**0,7670**	0,7241	0,6791	0,6998	0,0221	0,0428	0,0442	43	Over + Sel. Att.
RANKER_3	**0,7656**	0,7290	0,6777	0,7024	0,0225	0,0435	0,0440	43,8	Over + Sel. Att.
RANKER_3	**0,7654**	0,7361	0,6634	0,6975	0,0225	0,0435	0,0442	40,4	Sel. Att. + Over
RANKER_1	**0,7654**	0,7328	0,6662	0,6972	0,0223	0,0431	0,0422	40,2	Sel. Att. + Over
RANKER_3	**0,7644**	0,7275	0,6590	0,6918	0,0217	0,0421	0,0432	40,4	Sel. Att.
RANKER_1	**0,7642**	0,7265	0,6690	0,6951	0,0220	0,0426	0,0422	40,2	Sel. Att.
-	**0,7232**	0,7704	0,5681	0,6596	0,0224	0,0430	0,0348	214	Inicial
-	**0,7208**	0,7515	0,5968	0,6687	0,0215	0,0415	0,0366	214	Over

Tabla 26. Resultados con discretización utilizando los rangos establecidos por el SNS-O (solo DGPs)

TOTALMENTE AUTOMÁTICO

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_1	**0,7658**	0,7114	0,6834	0,6956	0,0212	0,0411	0,0412	41,2	Sel. Att.
RANKER_1	**0,7658**	0,7337	0,6604	0,6953	0,0222	0,0430	0,0440	44,2	Over + Sel. Att.
RANKER_1	**0,7656**	0,7262	0,6678	0,6956	0,0219	0,0424	0,0404	41,2	Sel. Att. + Over
RANKER_3	**0,7654**	0,7340	0,6575	0,6939	0,0221	0,0428	0,0434	44,4	Over + Sel. Att.
RANKER_3	**0,7620**	0,7319	0,6604	0,6944	0,0221	0,0427	0,0420	42,6	Sel. Att.
RANKER_3	**0,7616**	0,7308	0,6503	0,6888	0,0217	0,0420	0,0410	42,6	Sel. Att. + Over
-	**0,7572**	0,7338	0,6518	0,6902	0,0220	0,0425	0,0404	214	Inicial
-	**0,7548**	0,7417	0,6359	0,6840	0,0222	0,0428	0,0408	214	Over

Tabla 27. Resultados con discretización total utilizando el proceso automático

DGPs CON RANGOS ESTABLECIDOS + AUTOMÁTICO

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_1	**0,7664**	0,7185	0,6836	0,6996	0,0218	0,0423	0,0396	40,2	Sel. Att.
RANKER_1	**0,7652**	0,7233	0,6878	0,7040	0,0223	0,0432	0,0394	40,2	Sel. Att. + Over
RANKER_1	**0,7648**	0,7470	0,6460	0,6938	0,0229	0,0443	0,0430	42,8	Over + Sel. Att.
RANKER_3	**0,7642**	0,7372	0,6547	0,6935	0,0224	0,0433	0,0426	42,4	Over + Sel. Att.
RANKER_3	**0,7626**	0,7257	0,6604	0,6918	0,0215	0,0417	0,0406	40,4	Sel. Att.
RANKER_3	**0,7612**	0,7287	0,6532	0,6892	0,0216	0,0417	0,0402	40,4	Sel. Att. + Over
-	**0,7516**	0,7288	0,6575	0,6907	0,0218	0,0421	0,0382	214	Inicial
-	**0,7480**	0,7376	0,6302	0,6806	0,0216	0,0418	0,0378	214	Over

Tabla 28. Resultados con discretización utilizando los rangos establecidos para DGPs y el proceso automático para el resto

En los resultados podemos ver como discretizando los DGPs utilizando los rangos recibidos se obtienen mejores resultados que utilizando la discretización automática aunque la mejora es pequeña. La discretización mejora ligeramente los resultados obtenidos que sin utilizarla. Sin embargo, en el caso en el que solo usamos selección de características el modelo no hay diferencias importantes con el modelo sin discretización.

8.11.4. Ajuste del parámetro de regularización para el algoritmo de regresión logística

En los experimentos anteriores el parámetro de la regresión logística no era establecido y se utilizaba el valor por defecto (10^{-8}). En este caso utiliza la técnica de búsqueda para este valor en el rango $[10^{-4}, 10^{4}]$. Como conjunto de datos se utiliza la unión de los conjuntos con fechas de referencia 9/2014 y 12/2013 (validación cruzada), y como farmacia tanto SUB5 como CODMED.

FARMACIA CODMED

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
-	**0,7686**	0,7466	0,6547	0,6979	0,0231	0,0446	0,0438	214	Inicial
RANKER_3	**0,7678**	0,7092	0,6850	0,6960	0,0212	0,0411	0,0442	42,6	Sel. Atr.
RANKER_1	**0,7662**	0,6919	0,7024	0,6964	0,0205	0,0397	0,0432	41,6	Over. + Sel. Att.
RANKER_3	**0,7662**	0,7076	0,6864	0,6961	0,0211	0,0409	0,0430	39,8	Over. + Sel. Att.
RANKER_3	**0,7660**	0,7139	0,6850	0,6978	0,0215	0,0416	0,0438	42,6	Sel. Att. + Over.
RANKER_1	**0,7658**	0,7409	0,6530	0,6940	0,0227	0,0439	0,0442	41,2	Sel. Atr.
RANKER_1	**0,7654**	0,7326	0,6706	0,6992	0,0224	0,0433	0,0426	41,2	Sel. Att. + Over.
-	**0,7260**	0,7524	0,5997	0,6703	0,0216	0,0418	0,0372	214	Over.

Tabla 29. Resultados con ajuste del parámetro de regularización para CODMED

FARMACIA SUB5

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,7694**	0,7278	0,6705	0,6973	0,0223	0,0431	0,0434	56	Sel. Atr.
RANKER_3	**0,7692**	0,7279	0,6835	0,7044	0,0225	0,0435	0,0432	56	Sel. Att. + Over.
RANKER_1	**0,7688**	0,7154	0,6778	0,6944	0,0214	0,0415	0,0436	50,4	Over. + Sel. Att.
RANKER_3	**0,7682**	0,7257	0,6647	0,6913	0,0218	0,0423	0,0448	48,4	Over. + Sel. Att.
RANKER_1	**0,7680**	0,7035	0,6979	0,6992	0,0212	0,0411	0,0426	51,6	Sel. Att. + Over.
RANKER_1	**0,7676**	0,7678	0,6242	0,6918	0,0241	0,0464	0,0428	51,6	Sel. Atr.
-	**0,7660**	0,7735	0,6243	0,6945	0,0246	0,0474	0,0430	325	Inicial
-	**0,6910**	0,7475	0,5751	0,6552	0,0208	0,0401	0,0320	325	Over.

Tabla 30. Resultados con ajuste del parámetro de regularización para SUB5

Se puede destacar que realizar selección de características es lo que mejor funciona de manera global, y en concreto el algoritmo de selección RANKER_3 es el que obtiene mejores resultados en ambos casos.

Aunque SUB5 obtiene mejores resultados, se puede destacar que el mejor resultado de CODMED se obtiene sin realizar ninguna transformación lo que podría indicar una buena agrupación de los medicamentos.

La regularización en cualquier caso es beneficiosa y su ajuste permite obtener resultados más robustos.

8.11.5. Validación sobre 2015

En este caso no se utiliza la técnica de validación cruzada. En este caso se entrena el modelo con el conjunto de la unión de 9/2014 y 12/2013 entero, y se utiliza 2015 para evaluar y obtener los resultados. Los conjuntos de datos desde 1/2015 hasta 9/2015 se unen formando un único conjunto de datos, y es este el que se utiliza para obtener los resultados.

FARMACIA CODMED

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,7760**	0,7221	0,6864	0,7040	0,0187	0,0365	0,0340	45	Sel. Att. + Over
RANKER_3	**0,7760**	0,7204	0,6816	0,7007	0,0185	0,0360	0,0350	45	Sel. Att.
RANKER_1	**0,7750**	0,7225	0,6786	0,7002	0,0186	0,0361	0,0350	38	Sel. Att.
RANKER_1	**0,7730**	0,7076	0,6905	0,6990	0,0179	0,0350	0,0340	38	Sel. Att. + Over
RANKER_1	**0,7680**	0,7250	0,6701	0,6970	0,0185	0,0360	0,0340	43	Over + Sel. Att.
RANKER_3	**0,7670**	0,7395	0,6510	0,6939	0,0190	0,0368	0,0340	40	Over + Sel. Att.
-	**0,7350**	0,7747	0,5711	0,6651	0,0192	0,0372	0,0300	214	Inicial
-	**0,7310**	0,7620	0,5748	0,6618	0,0183	0,0355	0,0290	214	Over

Tabla 31. Resultados validación 2015 para CODMED

FARMACIA SUB5

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_1	**0,7740**	0,7166	0,6905	0,7034	0,0185	0,0360	0,0330	52	Sel. Att.
RANKER_3	**0,7730**	0,6699	0,7333	0,7009	0,0169	0,0330	0,0340	55	Sel. Att.
RANKER_3	**0,7720**	0,7018	0,6973	0,6995	0,0178	0,0346	0,0330	55	Sel. Att. + Over
RANKER_1	**0,7720**	0,7389	0,6605	0,6986	0,0192	0,0373	0,0320	52	Sel. Att. + Over
RANKER_3	**0,7660**	0,7517	0,6330	0,6898	0,0193	0,0375	0,0330	49	Over + Sel. Att.
RANKER_1	**0,7630**	0,7647	0,6133	0,6848	0,0198	0,0383	0,0320	52	Over + Sel. Att.
-	**0,7320**	0,7731	0,5718	0,6649	0,0191	0,0370	0,0280	325	Inicial
-	**0,7260**	0,7628	0,5711	0,6600	0,0183	0,0354	0,0270	325	Over

Tabla 32. Resultados validación 2015 para SUB5

Como podemos ver CODMED obtiene mejores resultados, lo que significa que este generaliza mejor que SUB5 y además utiliza menos características. También se observa que RANKER_3 funciona mejor con CODMED. Utilizando SUB5 el método de selección es indiferente.

8.11.6. Validación de la unión desde 12/2013 hasta 9/2014 sobre 2015

Este experimento difiere del anterior en el conjunto de entrenamiento. En este caso se utiliza la unión de todos los conjuntos de datos desde 12/2013 hasta 9/2014 (10 conjuntos). Como farmacia solo se ha utilizado CODMED.

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_1	**0,7770**	0,7377	0,6646	0,7002	0,0192	0,0374	0,0350	45	Sel. Att.
RANKER_3	**0,7760**	0,7376	0,6650	0,7004	0,0192	0,0374	0,0350	48	Sel. Att.
RANKER_1	**0,7750**	0,7333	0,6714	0,7017	0,0191	0,0371	0,0350	45	Sel. Att. + Over
RANKER_3	**0,7740**	0,7376	0,6639	0,6998	0,0192	0,0373	0,0350	48	Sel. Att. + Over
RANKER_1	**0,7720**	0,7462	0,6473	0,6950	0,0193	0,0376	0,0340	43	Over + Sel. Att.
RANKER_3	**0,7710**	0,6975	0,7068	0,7021	0,0177	0,0346	0,0340	41	Over + Sel. Att.
-	**0,7480**	0,7568	0,6112	0,6801	0,0191	0,0370	0,0320	214	Inicial
-	**0,7460**	0,7573	0,6146	0,6823	0,0192	0,0373	0,0310	214	Over

Tabla 33. Resultados validación 2015 con la unión desde 12/2013 hasta 9/2014 con farmacia CODMED

Al igual que con la validación cruzada (8.11.2), no se aprecia una mejora al unir todos los conjuntos de datos desde 12/2013 hasta 9/2014.

8.11.7. Validación sobre 2015 con discretización

Para este experimento se utilizan las mismas opciones de discretización que en el experimento 8.10.3 pero aplicándolas sobre el conjunto de datos de 2015 (unión de los 9 meses). Igual que en el experimento 8.11.3 se utiliza farmacia CODMED y la unión de 9/2014 y 12/2013 como conjunto de datos de entrenamiento.

SOLO DGPs CON RANGOS ESTABLECIDOS

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,7740**	0,7263	0,6762	0,7008	0,0187	0,0365	0,0350	41	Sel. Att.
RANKER_3	**0,7730**	0,7286	0,6731	0,7003	0,0188	0,0366	0,0340	41	Sel. Att. + Over.
RANKER_3	**0,7730**	0,7243	0,6752	0,6993	0,0186	0,0362	0,0340	40	Over. + Sel. Att.
RANKER_1	**0,7730**	0,7256	0,6673	0,6959	0,0185	0,0359	0,0350	38	Sel. Att.
RANKER_1	**0,7720**	0,7242	0,6769	0,7002	0,0186	0,0363	0,0340	41	Over. + Sel. Att.
RANKER_1	**0,7710**	0,7172	0,6782	0,6974	0,0182	0,0355	0,0340	38	Sel. Att. + Over.
-	**0,7330**	0,7936	0,5510	0,6613	0,0202	0,0390	0,0300	214	Inicial
-	**0,7190**	0,7769	0,5534	0,6557	0,0188	0,0364	0,0290	214	Over

Tabla 34. Resultados 2015 con discretización de DGPs utilizando los rangos del SNS-O

TOTALMENTE AUTOMÁTICO

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_1	**0,7770**	0,6886	0,7224	0,7053	0,0176	0,0344	0,0330	38	Sel. Att.
RANKER_3	**0,7770**	0,7627	0,6466	0,7023	0,0206	0,0400	0,0340	45	Sel. Att.
RANKER_3	**0,7740**	0,7263	0,6776	0,7015	0,0188	0,0366	0,0330	45	Sel. Att. + Over.
RANKER_1	**0,7720**	0,7159	0,6908	0,7032	0,0185	0,0359	0,0330	38	Sel. Att. + Over.
RANKER_1	**0,7690**	0,7498	0,6527	0,6996	0,0198	0,0384	0,0330	45	Over. + Sel. Att.
RANKER_3	**0,7690**	0,7281	0,6697	0,6983	0,0187	0,0364	0,0330	43	Over. + Sel. Att.
-	**0,7630**	0,7635	0,6282	0,6926	0,0201	0,0390	0,0320	214	Inicial
-	**0,7600**	0,7681	0,6201	0,6901	0,0203	0,0392	0,0320	214	Over

Tabla 35. Resultados 2015 utilizando el proceso de discretización automática sobre todos los atributos

DGPs CON RANGOS ESTABLECIDOS + AUTOMÁTICO

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,7760**	0,7308	0,6724	0,7010	0,0190	0,0369	0,0340	41	Sel. Att.
RANKER_1	**0,7740**	0,6916	0,7150	0,7032	0,0176	0,0344	0,0340	38	Sel. Att.
RANKER_3	**0,7720**	0,7400	0,6605	0,6992	0,0193	0,0374	0,0330	41	Sel. Att. + Over.
RANKER_1	**0,7690**	0,7283	0,6711	0,6991	0,0187	0,0365	0,0340	38	Sel. Att. + Over.
RANKER_1	**0,7670**	0,7464	0,6469	0,6949	0,0193	0,0376	0,0330	43	Over. + Sel. Att.
RANKER_3	**0,7670**	0,7464	0,6469	0,6949	0,0193	0,0376	0,0330	43	Over. + Sel. Att.
-	**0,7550**	0,7666	0,6112	0,6845	0,0198	0,0384	0,0300	214	Inicial
-	**0,7400**	0,7921	0,5527	0,6617	0,0201	0,0389	0,0300	214	Over

Tabla 36. Resultados 2015 discretizando los DGPs con los rangos establecidos y discretizando el resto de atributos con el proceso automático

Se puede ver que con la discretización no obtenemos malos resultados, pero tampoco nos provoca una gran mejora, por lo que no es completamente necesaria.

8.11.8. Validación sobre 2015 con estimación del parámetro de la regresión logística

En este caso utilizando la unión de 9/2014 y 12/2013 y se ha buscado el valor para el parámetro de regularización (igual que en el experimento 8.11.4). Con este valor se ha evaluado el conjunto de datos de 2015 (unión de los 9 meses).

Vistos los resultados en el experimento 8.11.4, se decidió utilizar únicamente el modelo inicial y el modelo con selección de atributos.

FARMACIA CODMED

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,7760**	0,7125	0,6939	0,7031	0,0183	0,0357	0,0350	45	Sel. Att.
RANKER_1	**0,7750**	0,7600	0,6279	0,6908	0,0198	0,0384	0,0350	38	Sel. Att.
-	**0,7730**	0,7461	0,6548	0,6990	0,0196	0,0380	0,0340	214	Inicial

Tabla 37. Resultados 2015 ajustando el parámetro de regularización para farmacia CODMED

FARMACIA SUB5

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,7740**	0,7069	0,6939	0,7004	0,0180	0,0350	0,0350	55	Sel. Att.
RANKER_1	**0,7740**	0,7009	0,6980	0,6994	0,0177	0,0346	0,0340	52	Sel. Att.
-	**0,7690**	0,7472	0,6480	0,6958	0,0194	0,0377	0,0340	325	Inicial

Tabla 38. Resultados 2015 ajustando el parámetro de regularización para farmacia SUB5

Como se puede ver en los resultados ajustando el parámetro conseguimos un modelo más robusto. Con selección de atributos obtenemos mejores resultados pero el modelo inicial también consigue un valor alto de ROC-AUC.

8.11.9. Coeficientes de la regresión logística

El algoritmo regresión logística asigna un peso a cada atributo. Este peso nos ayuda a ver que atributos están siendo más influyentes a sufrir evento adverso y cuales a no sufrirlo. Esto para el SNS-O es muy importante ya que pueden observar que medicamentos, ingresos, episodios,… son los que están siendo más influyentes y con qué peso.

En este apartado se exponen los pesos que la regresión logística asigna a cada atributo. Como conjunto de datos se utiliza la unión de 9/2014 y 12/2013, y se obtienen los pesos al realizar selección de características, y selección de características más oversampling. Viendo que el algoritmo de selección RANKER_3 es el que mejores resultados obtiene, este es el que se utiliza.

Hay que tener en cuenta que un valor positivo implica aumentar la probabilidad a la clase 0**.** Por el contrario un valor negativo implica aumentar la probabilidad la clase 1 (positiva, cardiovascular).

FARMACIA CODMED

Atributo	Descripción	Coef. Selección	Coef. Selección y over
ant_cmbd=1	Antecedente cardiovascular CMBD	-0,6254	-0,6401
ant_epi=1	Antecedente cardiovascular episodios	-0,2633	-0,4562
charlson_age	Componente edad de Charlson	-0,0194	-0,0233
charlson_diabetes=1	Componente diabetes de Charlson	-0,009	-0,0247
charlson_enf_cerebrovascular=1	Componente Charlson de enf_cerebrovascular	-0,0965	-0,0675
charlson_enf_coronaria=1	Componente Charlson de enf_coronaria	-0,2076	-0,1573
charlson_enf_renal=1	Componente Charlson de enf_renal	0,0923	0,133
charlson_insuf_card=1	Componente Charlson insuf_card	-0,1324	-0,1633
dgps_COL	Valores de los registros de DGPs con código COL	-0,0008	-0,0015
dgps_FILTGLOM	Valores de los registros de DGPs con código FILTGLOM	0,002	0,002
dgps_HDL	Valores de los registros de DGPs con código HDL	0,0019	0,0033
dgps_TAD	Valores de los registros de DGPs con código TAD	0,0048	0,0055
dgps_TAS	Valores de los registros de DGPs con código TAS	-0,0043	-0,004
num_cod_A02BC__Inhibidores_de_la_bomba_de_protones	Numero de principios activos de A02BC__Inhibidores_de_la_bomba_de_protones	-0,0216	-0,0199
num_cod_ACA__Dihidropiridinas	Numero de principios activos de ACA__Dihidropiridinas	-0,0413	-0,0533
num_cod_Agonistas_muscarinicos_de_larga_duracion	Numero de principios activos de Agonistas_muscarinicos_de_larga_duracion	-0,0041	-0,015
num_cod_AINES	Numero de principios activos de AINES	0,0371	0,01
num_cod_Antiagreg_Inhib_recept	Numero de principios activos de Antiagreg_Inhib_recept	-0,0492	-0,0442
num_cod_B01AA__Antagonistas_de_la_vitamina_K	Numero de principios activos de B01AA__Antagonistas_de_la_vitamina_K	-0,0394	-0,0421
num_cod_Betabloqueantes	Numero de principios activos de Betabloqueantes	-0,04	-0,0458
num_cod_C01DA__Nitratos_organicos	Numero de principios activos de C01DA__Nitratos_organicos	-0,0349	-0,0701
num_cod_C03DA__Antagonistas_de_la_aldosterona	Número de principios activos de C03DA__Antagonistas_de_la_aldosterona	-0,0854	-0,104
num_cod_Corticoides	Numero de principios activos de Corticoides	-0,0397	-0,0739
num_cod_Digoxina	Numero de principios activos de Digoxina	-0,0332	-0,0368
num_cod_Diureticos_del_asa	Numero de principios activos de Antiagreg_Inhib_recept	-0,0849	-0,0831
num_cod_Heparinas	Numero de principios activos de Heparinas	-0,1198	-0,098
num_cod_Hierro	Numero de principios activos de Hierro	-0,0339	-0,0531
num_cod_IECA	Numero de principios activos de IECA	-0,0096	-0,0144

num_cod_Inhibidores_del_acido_urico	Numero de principios activos de Inhibidores_del_acido_urico	0,0534	0,073
num_cod_Insulinas	Numero de principios activos de Insulinas	-0,0507	-0,0456
num_cod_Quinolonas	Numero de principios activos de Quinolonas	-0,0719	-0,1566
num_cod_Salicilatos_y_dipirid	Numero de principios activos de Salicilatos_y_dipirid	-0,0219	-0,035
num_epi_all_B	Conteo de episodios de sangre, órganos hematopoyéticos y sistema inmunitario	-0,101	-0,0812
num_epi_all_K	Conteo de episodios de aparato circulatorio	-0,13	-0,1588
num_epi_all_L	Conteo de episodios de aparato locomotor	0,1125	0,1093
num_ingr_03	Conteo de ingresos de Enf endocrinas/nutricional/metabólica/inmunitaria	-0,0896	-0,3551
num_ingr_04	Conteo de ingresos de enf sangre y órganos hematopoyético	-0,4571	-0,4413
num_ingr_07	Conteo de ingresos de enf sistema circulatorio	-0,1318	-0,2218
num_ingr_08	Conteo de ingresos de enf sistema respiratorio	-0,0435	-0,0364
num_ingr_09	Conteo de ingresos de Enf sistema digestivo	-0,103	-0,1249
num_ingr_16	Conteo de ingresos de signos, síntomas y estados mal definidos	-0,1711	-0,4496
num_pac	Conteo de principios activos agrupados por CODMED	-0,0185	-0,0231
num_pas1	Conteo de principios activos en SUB1	-0,0448	-0,0477
TSI	TSI al que pertenece el paciente	0,0476	-0,0029
Intercept	Coeficiente del algoritmo Logistic	76.704	36.643

Tabla 39. Coeficientes de la regresión logística para farmacia CODMED

FARMACIA SUB5

Atributo	Descripción	Coef. Selección	Coef. Selección y over
ant_cmbd=1	Antecedente cardiovascular CMBD	-0,635	-0,6493
ant_epi=1	Antecedente cardiovascular Episodios	-0,2656	-0,4439
charlson_age	Componente edad de Charlson	-0,0182	-0,0215
charlson_diabetes_con_dano=1	Componente diabetes con daño de Charlson	-0,1821	-0,0548
charlson_diabetes=1	Componente diabetes de Charlson	-0,0051	-0,0511
charlson_enf_cerebrovascular=1	Componente enf_cerebrovascular Charlson	-0,0877	-0,0325
charlson_enf_coronaria=1	Componenete enf_coronaria Charlson	-0,2169	-0,1931
charlson_enf_renal=1	Componenete enf_renal de Charlson	0,0175	0,0401
charlson_insuf_card=1	Componente insuf_card de Charlson	-0,1258	-0,1757
dgps_COL	Valores de los registros de DGPs con código COL	-0,0009	-0,0017
dgps_FILTGLOM	Valores de los registros de DGPs con código FILTGLOM	0,0022	0,0021
dgps_HDL	Valores de los registros de DGPs con código HDL	0,0018	0,0032
dgps_INR	Valores de los registros de DGPs con código INR	0,0018	-0,0373
dgps_TAD	Valores de los registros de DGPs con código TAD	0,0051	0,0058
dgps_TAS	Valores de los registros de DGPs con código TAS	-0,0045	-0,004
num_epi_all_B	Conteo de episodios de sangre, órganos hematopoyéticos y sistema inmunitario	-0,1076	-0,0982
num_epi_all_K	Conteo de episodios de aparato circulatorio	-0,1349	-0,1573
num_epi_all_L	Conteo de episodios de aparato locomotor	0,1042	0,0977
num_epi_all_U	Conteo de episodios de aparato urinario	0,1251	0,1493
num_ingr_03	Conteo de ingresos de Enf endocrinas/nutricional/metabólica/inmunitaria	-0,0833	-0,3466
num_ingr_04	Conteo de ingresos de enf sangre y órganos hematopoyético	-0,4417	-0,4871
num_ingr_07	Conteo de ingresos de enf sistema circulatorio	-0,1329	-0,2189
num_ingr_08	Conteo de ingresos de enf sistema respiratorio	-0,0363	-0,0246
num_ingr_09	Conteo de ingresos de Enf sistema digestivo	-0,1001	-0,1016
num_ingr_16	Conteo de ingresos de signos, síntomas y estados mal definidos	-0,1747	-0,4619
num_paA02BC	Numero de principios activos de Inhibidores de la bomba de protones	-0,0228	-0,0203
num_paA10AB	Numero de principios activos de Insulinas y anlogos de accin rpida	-0,0253	0,0686
num_paA10AE	Numero de principios activos de Insulinas y anlogos de accin prolongada	-0,085	-0,1259
num_paB01AA	Numero de principios activos de Antagonistas de la vitamina K	-0,0364	-0,0318
num_paB01AB	Numero de principios activos de Grupo de la heparina	-0,1203	-0,093

num_paB01AC	Numero de principios activos de Inhibidores de la agregacin plaquetaria,	-0,0297	-0,0398
num_paB03AA	Numero de principios activos de Hierro bivalente, preparados orales	-0,014	-0,0308
num_paB03AB	Numero de principios activos de Hierro trivalente, preparados orales	-0,0862	-0,1103
num_paC01AA	Numero de principios activos de Glucsidos digitlicos	-0,037	-0,051
num_paC01BC	Numero de principios activos de Antiarrtmicos de clase Ic	-0,1385	-0,1303
num_paC01DA	Numero de principios activos de Nitratos orgnicos	-0,0396	-0,076
num_paC03CA	Numero de principios activos de Sulfonamidas, monofrmacos	-0,0885	-0,0891
num_paC03DA	Numero de principios activos de Antagonistas de la aldosterona	-0,0904	-0,1136
num_paC07AB	Numero de principios activos de Agentes beta-bloqueantes selectivos	-0,028	-0,0273
num_paC08CA	Numero de principios activos de Derivados de la dihidropiridina	-0,035	-0,0479
num_paC09AA	Numero de principios activos de Inhibidores de la ECA, monofrmacos	-0,0101	-0,0075
num_paC09CA	Numero de principios activos de Antagonistas de angiotensina II, monofrma	-0,0113	-0,0153
num_paH02AB	Numero de principios activos de Glucocorticoides	-0,0435	-0,0635
num_paJ01DD	Numero de principios activos de Cefalosporinas de tercera generacin	-0,2401	-0,3764
num_paJ01MA	Numero de principios activos de Fluoroquinolonas	-0,0877	-0,1685
num_paM01AB	Número de principios activos de Derivados del cido actico y sustancias r	-0,0031	-0,0246
num_paM01AE	Numero de principios activos de Derivados del cido propinico	0,096	0,0587
num_paM01AX	Numero de principios activos de Otros agentes antiinflamatorios y antirreu	0,0218	0,0417
num_paM04AA	Numero de principios activos de Preparados que inhiben la produccin de c	0,0553	0,0682
num_paN05BA	Numero de principios activos de Derivados de la benzodiazepina	0,0191	0,0163
num_paR03BB	Numero de principios activos de Anticolinrgicos	0,0109	0,0022
num_pas1	Conteo de principios activos en SUB1	-0,0618	-0,0592
num_pas5	Conteo de principios activos en SUB5	-0,0062	-0,0137
TSI	TSI al que pertenece el paciente	0,0578	0,0109
Intercept	Coeficiente del algoritmo Logistic	75.028	33.985

Tabla 40. Coeficientes de la regresión logística para farmacia SUB5

En las Tablas 39 y 40 podemos observar que coeficientes asigna la regresión logística a cada atributo. En ambas se puede observar que los atributos con mayor peso a sufrir evento adverso (mayor peso negativo) son los antecedentes, las componentes de Charlson enfermedad coronaria e insuficiencia cardiaca, y el ingreso por sangre y órganos hematopoyético entre otros.

Además de estos se pueden observar diferentes medicamentos también influyentes a sufrir evento adverso, lo que es de interés para el SNS-O.

8.11.10. Resumen

Finalmente se presentan unas tablas (Tabla 41 y Tabla 42) con los mejores resultados de todos los experimentos anteriores a modo de resumen. Los resultados en negrita corresponden al mejor resultado de cada método y subrayado en mejor resultado global.

VALIDACIÓN CRUZADA

MÉTODO	12/2013 y 9/2014	De 12/2013 a 9/2014	Ajuste Logistic	Discretización	
Inicial	0,7280	0,7306	**0,7686**	0,7572	Total automática
Oversampling	0,7260	0,7450	0,7260	**0,7548**	**Total automática**
Selección de atributos	0,7682	0,7646	**0,7694**	0,7664	DGPS recib. + aut.
Selección + Oversampling	**0,7692**	0,7644	**0,7692**	0,7656	Total automática
Oversampling + Selección	**0,7688**	0,7648	**0,7688**	0,7670	Solo DGPs recib.

Tabla 41. Resumen resultados para validación cruzada

DATOS 2015

MÉTODO	12/2013 y 9/2014	De 12/2013 a 9/2014	Ajuste Logistic	Discretización	
Inicial	0,7350	0,7480	**0,7730**	0,7630	Total automática
Oversampling	0,7310	0,7460	-	**0,7600**	**Total automática**
Selección de atributos	0,7760	**0,7770**	0,7760	**0,7770**	**Total automática**
Selección + Oversampling	**0,7760**	0,7750	-	0,7740	Total automática
Oversampling + Selección	0,7680	0,7720	-	**0,7730**	**Solo DGPs recib.**

Tabla 42. Resumen resultados con 2015 como validación

En las tablas de resumen se puede observar que los modelos con selección de atributos son los más robustos. Obtienen un AUC medio en todos los meses similar, por lo que podemos decir que los modelos generalizan bien con datos desconocidos.

8.12. Resultados Hospitalizaciones Potencialmente Evitables (HPE)

Como se ha comentado se ha extendido esta metodología para otro tipo de eventos: Hospitalizaciones Potencialmente Evitables (HPE). En esta sección se realizan experimentos similares a los realizados con los eventos adversos cardiovasculares.

Se ha mostrado en la Tabla 13 los diferentes motivos de HPE que se tienen, aunque se consideran todos los motivos como uno para tratar este problema como un problema de clasificación binaria.

Para la generación de conjunto de datos se utilizan las siguientes opciones:

- **CMBD:** Conteo de ingresos agrupado por nivel 1.
- **Episodios:** Conteo de todos los episodios que están abiertos o han sido cerrados en el último mes agrupados por el primer carácter.
    - Se han considerado los episodios que estén **entre el 7 y el 9 después de la letra** correspondiente al nivel, **salvo para la Z** donde se cuentan **todos**.
- **Número de meses para los datos (N_{GEN}):** 6 meses
- **DGPs**: Media de los resultados de 3 meses a partir del último existente
- **Antecedentes**: En este caso se utiliza el antecedente de HPE de CMBD. Este antecedente indica si el paciente ha sufrido HPE en los 6 meses anteriores a la fecha de referencia.
- **TSI**: Grupo TSI del paciente (001, 002A, 002B,…) como atributo numérico.
- **Farmacia**: Se utilizan dos codificaciones: SUB5 y CODMED
- **Índice de Charlson**: Charlson separado – Utilizar cada componente de Charlson como atributo
- **Conteo de los principios activos diferentes que toma el paciente**. El conteo se realiza agrupando de la misma manera que se agrupan los datos de farmacia. Si se utiliza farmacia CODMED, el conteo se realizará agrupando en CODMED. Además siempre se añade el conteo agrupando por SUB1.
- **Clase** (variable a predecir): Se trata de predecir si se sufre una HPE en los 3 meses siguientes a la fecha de referencia.
- **Clasificador:** En este experimento solo se utiliza la regresión logística (Logistic). Se utilizan los parámetros por defecto en caso de que no se especifique.

Para este experimento se generaron dos conjuntos de datos, uno con la fecha de referencia 9/2014 y otro con la fecha de referencia 12/2013. Estos conjuntos de datos se unieron formando uno único, sobre el que se ejecutó el clasificador.

Además de estos datos se generaron 9 conjuntos de datos de 2015, desde 1/2015 hasta 9/2015. Estos se utilizaron individualmente y uniendo los 9 en un único conjunto. El clasificador se entrena con el conjunto de datos de la unión de 9/2014 y 12/2013 y se calculan los resultados utilizando los conjuntos de 2015.

En la Tabla 43 se pueden consultar en número de ejemplos por cada uno de los conjuntos de datos generados.

Fecha de referencia	Clase			
	Positiva	%	Negativa	%
09/2014	132	0,32%	41.771	99,68%
12/2013	153	0,44%	34.238	99,56%
12/2013 + 09/2014	285	0,37%	76.009	99,63%
01/2015	154	0,36%	42.421	99,64%
02/2015	124	0,29%	42.451	99,71%
03/2015	98	0,23%	42.477	99,77%
04/2015	97	0,23%	42.478	99,77%
05/2015	104	0,24%	42.471	99,76%
06/2015	117	0,27%	42.458	99,73%
07/2015	123	0,29%	42.452	99,71%
08/2015	127	0,30%	42.448	99,70%
09/2015	144	0,34%	42.431	99,66%
Unión 2015	1.088	0,28%	382.087	99,72%

Tabla 43. Numero de ejemplos de cada clase por cada uno de los conjuntos de datos generados

En los experimentos se realizaron diferentes transformaciones antes de obtener los resultados. En las tablas las transformaciones realizadas aparecen en la columna MODELO. Se utilizaron algoritmos de selección de características y la técnica del oversampling. En la Tabla 44 encontramos las diferentes combinaciones de transformaciones y como se han denominado en las tablas de resultados.

MODELO	DESCRIPCIÓN
Inicial	No se realiza ninguna transformación sobre el dataset
Over.	Se aplica Oversampling sobre el dataset de entrenamiento
Sel. Att.	Se realiza el proceso de selección de atributos. En otra columna aparecerá el método de selección utilizado pudiendo ser “RANKER_1” o “RANKER_3”.
Over. + Sel. Att.	Primero realizamos Oversampling sobre el dataset y se aplica la selección de atributos. Igual que antes tenemos dos métodos de selección posibles: “RANKER_1” y “RANKER_3”
Sel. Att. + Over.	En este caso primero realizamos la selección de atributos utilizando el “RANKER_1” o el “RANKER_3” y posteriormente se aplica Oversampling sobre el dataset reducido.

Tabla 44. Modelos utilizados con su denominación y descripción

Para la selección de atributos se utilizaron 2 algoritmos diferentes. Ambos son métodos de selección tipo RANKER, donde nos devuelven la lista de los atributos ordenadas con una puntuación y se escogen aquellos con puntuación mayor a la media de todas las puntuaciones. Estos se pueden consultar en la Tabla 45 junto con su denominación en las tablas de resultados.

METODO	DESCRIPCIÓN
RANKER 1	Ranker con ChiSquared como evaluador.
RANKER 3	Ranker con InfoGain como evaluador

Tabla 45. Algoritmos de selección utilizados junto con su denominación en las tablas de resultados

8.12.1. Validación cruzada sobre la unión de 9/2014 y 12/2013

En este experimento se utiliza el conjunto de datos resultante de la unión de 9/2014 y 12/2013, realizando la validación cruzada de 5 particiones. Se presenta una tabla por cada transformación realizada, donde se añade una columna (FARMACIA) indicando la opción de farmacia utilizada.

CODMED

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,8862**	0,8504	0,8073	0,8283	0,0198	0,0387	0,1028	31,6	Over + Sel. Att.
RANKER_1	**0,8854**	0,8653	0,7759	0,8188	0,0213	0,0414	0,0982	32,2	Over + Sel. Att.
RANKER_1	**0,8818**	0,8509	0,7794	0,8139	0,0193	0,0377	0,1164	21,6	Sel. Att. + Over
RANKER_1	**0,8794**	0,8658	0,7656	0,8134	0,0219	0,0425	0,1102	21,6	Sel. Att.
RANKER_3	**0,8770**	0,8636	0,7825	0,8220	0,0214	0,0416	0,1018	27,4	Sel. Att. + Over
RANKER_3	**0,8724**	0,8565	0,7655	0,8080	0,0202	0,0393	0,1104	27,4	Sel. Att.
-	**0,7932**	0,8641	0,6356	0,7406	0,0176	0,0341	0,0724	213	Over
-	**0,7906**	0,8745	0,6360	0,7445	0,0187	0,0363	0,0782	213	Inicial

Tabla 46. Resultados validación cruzada para CODMED

SUB5

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER_3	**0,8854**	0,8783	0,7550	0,8135	0,0230	0,0446	0,1000	41,4	Sel. Att. + Over
RANKER_1	**0,8854**	0,8284	0,7935	0,8099	0,0175	0,0342	0,0964	42,4	Over + Sel. Att.
RANKER_3	**0,8818**	0,8420	0,7971	0,8177	0,0192	0,0374	0,0960	40,4	Over + Sel. Att.
RANKER_1	**0,8810**	0,8773	0,7549	0,8134	0,0228	0,0442	0,1056	28,6	Sel. Att. + Over
RANKER_3	**0,8748**	0,8762	0,7619	0,8165	0,0227	0,0440	0,1096	41,4	Sel. Att.
RANKER_1	**0,8734**	0,8773	0,7621	0,8170	0,0228	0,0443	0,1088	28,6	Sel. Att.
-	**0,7656**	0,8736	0,6117	0,7288	0,0182	0,0352	0,0670	324	Inicial
-	**0,7608**	0,8823	0,6112	0,7337	0,0193	0,0373	0,0662	324	Over

Tabla 47. Resultados validación cruzada para SUB5

Como se puede apreciar en las tablas, la selección de características hace que los resultados mejoren mucho. En general CODMED funciona mejor que SUB5 sobre todo sin realizar ningún tipo de procesamiento sobre los datos (modelo inicial).

Si comparamos los resultados de los HPE con los eventos cardiovasculares, podemos apreciar una notable mejora. Como la metodología utilizada es la misma para ambos eventos, podemos decir que predecir un HPE es más fácil que predecir un evento adverso cardiovascular.

Se puede observar que el PPV es bajo y esto es debido a la baja prevalencia de la clase positiva. Como el número de ejemplos de la clase positiva es mucho más bajo que el de la clase negativa, el número de verdaderos positivos (TP) que conseguimos, aun siendo un porcentaje alto (respecto al número total de positivos), es mucho menor que los falsos positivos (FP).

8.12.2. Ajuste del parámetro de regularización de la regresión logística

En este caso se busca el parámetro de regularización de la regresión logística empleando la técnica de búsqueda explicada anteriormente. El conjunto de datos a utilizar es el mismo que en el experimento anterior (8.12.1)

CODMED

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
-	**0,8956**	0,9053	0,7167	0,8026	0,0294	0,0561	0,1158	213	Inicial
RANKER_3	**0,8880**	0,8767	0,7205	0,7897	0,0248	0,0474	0,1176	27,4	Sel. Att.
RANKER_3	**0,8862**	0,8504	0,8073	0,8283	0,0198	0,0387	0,1028	31,6	Over + Sel. Att.
RANKER_1	**0,8856**	0,8999	0,7168	0,7996	0,0288	0,0548	0,1184	21,6	Sel. Att.
RANKER_1	**0,8854**	0,8653	0,7759	0,8188	0,0213	0,0414	0,0982	32,2	Over + Sel. Att.
RANKER_1	**0,8818**	0,8509	0,7794	0,8139	0,0193	0,0377	0,1164	21,6	Sel. Att. + Over
RANKER_3	**0,8770**	0,8636	0,7825	0,8220	0,0214	0,0416	0,1018	27,4	Sel. Att. + Over
-	**0,7968**	0,8641	0,6356	0,7406	0,0176	0,0341	0,0724	213	Over

Tabla 48. Resultados validación cruzada con ajuste de regularización para CODMED

SUB5

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
-	**0,8954**	0,9122	0,6956	0,7928	0,0365	0,0683	0,1144	213	Inicial
RANKER_3	**0,8920**	0,9119	0,7099	0,8000	0,0334	0,0629	0,1184	41,4	Sel. Att.
RANKER_1	**0,8866**	0,8969	0,6963	0,7840	0,0302	0,0567	0,1178	28,6	Sel. Att.
RANKER_3	**0,8854**	0,8783	0,7550	0,8135	0,0230	0,0446	0,1000	41,4	Sel. Att. + Over
RANKER_1	**0,8854**	0,8284	0,7935	0,8099	0,0175	0,0342	0,0964	42,4	Over + Sel. Att.
RANKER_3	**0,8818**	0,8420	0,7971	0,8177	0,0192	0,0374	0,0960	40,4	Over + Sel. Att.
RANKER_1	**0,8810**	0,8773	0,7549	0,8134	0,0228	0,0442	0,1056	28,6	Sel. Att. + Over
-	**0,7660**	0,8823	0,6112	0,7337	0,0193	0,0373	0,0663	213	Over

Tabla 49. Resultados validación cruzada con ajuste de regularización para SUB5

En las tablas de resultados se puede apreciar una mejora utilizando el parámetro de regularización que sin utilizarlo. Sobre todo cuando no realizamos ningún tipo de procesamiento sobre los datos (modelo inicial) llegando a aumentar el área bajo la curva ROC de forma notable. Es más, el mejor resultado que se obtiene en este experimento es utilizando CODMED con el modelo inicial.

8.12.3. Validación sobre 2015

En este experimento se utiliza el conjunto de 2015 (unión de los 9 meses: desde 1/2015 hasta 9/2015) como conjunto de test y el conjunto resultante de la unión de 9/2014 y 12/2013 como conjunto de entrenamiento.

CODMED

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER3	**0,8894**	0,8569	0,7696	0,8119	0,0151	0,0296	0,0861	32	Sel. Att. + Over
RANKER3	**0,8807**	0,8944	0,6986	0,7903	0,0185	0,0361	0,0874	27	Sel. Att.
RANKER1	**0,8784**	0,8359	0,7700	0,8022	0,0132	0,0260	0,0950	32	Sel. Att. + Over
RANKER1	**0,8778**	0,8671	0,7357	0,7985	0,0155	0,0304	0,0911	20	Over + Sel. Att.
RANKER3	**0,8778**	0,8671	0,7357	0,7985	0,0155	0,0304	0,0911	27	Over + Sel. Att.
RANKER1	**0,8678**	0,8226	0,7712	0,7964	0,0122	0,0241	0,0873	20	Sel. Att.
-	**0,8244**	0,8664	0,6941	0,7753	0,0146	0,0286	0,0724	213	Inicial
-	**0,8133**	0,8735	0,6678	0,7635	0,0148	0,0289	0,0694	213	Over

Tabla 50. Resultados de 2015 para CODMED

SUB5

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER3	**0,8809**	0,8645	0,7300	0,7943	0,0151	0,0296	0,0854	44	Sel. Att. + Over
RANKER1	**0,8791**	0,8739	0,7304	0,7989	0,0162	0,0318	0,0947	44	Sel. Att. + Over
RANKER3	**0,8716**	0,8950	0,6837	0,7822	0,0182	0,0355	0,0820	42	Sel. Att.
RANKER1	**0,8696**	0,8690	0,7229	0,7923	0,0155	0,0303	0,0888	28	Over + Sel. Att.
RANKER3	**0,8683**	0,8706	0,7163	0,7895	0,0155	0,0304	0,0866	42	Over + Sel. Att.
RANKER1	**0,8628**	0,8683	0,7167	0,7888	0,0153	0,0299	0,0796	28	Sel. Att.
-	**0,7289**	0,8991	0,5647	0,7124	0,0158	0,0307	0,0573	213	Inicial
-	**0,7136**	0,8865	0,5351	0,6885	0,0133	0,0260	0,0588	213	Over

Tabla 51. Resultados de 2015 para SUB5

Utilizando los datos de 2015 se aprecia que CODMED obtiene mejores resultados que SUB5 en cuanto a ROC AUC se refiere. Hay que destacar el modelo inicial y el modelo con oversampling donde la diferencia entre las dos opciones de farmacia es muy amplia. Esto se puede deber a que el modelo con CODMED generaliza mejor que el modelo con SUB5.

También es apreciable que el uso de selección de atributos es necesaria, no solo se aumenta el valor del área bajo la curva ROC, además nos genera un modelo más interpretable ya que utiliza menos atributos.

8.12.4. Ajuste del parámetro de regularización sobre 2015

Este experimento utiliza los mismos conjuntos que el anterior (8.12.3). En este caso se busca el parámetro de regularización de la regresión logística empleando la técnica de búsqueda explicada anteriormente.

CODMED

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
RANKER3	**0,8894**	0,8569	0,7696	0,8119	0,0151	0,0296	0,0861	27	Over + Sel. Att.
-	**0,8870**	0,9224	0,6540	0,7766	0,0236	0,0455	0,1011	213	Inicial
RANKER3	**0,8823**	0,9335	0,6292	0,7663	0,0264	0,0506	0,1034	27	Sel. Att.
RANKER1	**0,8784**	0,8359	0,7700	0,8022	0,0132	0,0260	0,0950	20	Over + Sel. Att.
RANKER1	**0,8734**	0,9348	0,6279	0,7660	0,0269	0,0515	0,1034	20	Sel. Att.
RANKER1	**0,8696**	0,8690	0,7229	0,7923	0,0155	0,0303	0,0888	44	Sel. Att. + Over
RANKER3	**0,8683**	0,8706	0,7163	0,7895	0,0155	0,0304	0,0866	44	Sel. Att. + Over
-	**0,8162**	0,8735	0,6678	0,7635	0,0148	0,0289	0,0694	213	Over

Tabla 52. Resultados de 2015 con ajuste de regularización para CODMED

SUB5

AlgSel	ROC-AUC	TNR	TPR	GM	PPV	F-Measure	PRC-AUC	Nº Att	MODELO
-	**0,8857**	0,9415	0,6090	0,7571	0,0290	0,0553	0,1033	213	Inicial
RANKER3	**0,8850**	0,9517	0,5852	0,7461	0,0335	0,0633	0,1009	42	Sel. Att.
RANKER3	**0,8809**	0,8645	0,7300	0,7943	0,0151	0,0296	0,0854	42	Over + Sel. Att.
RANKER1	**0,8791**	0,8739	0,7304	0,7989	0,0162	0,0318	0,0947	28	Over + Sel. Att.
RANKER1	**0,8778**	0,8671	0,7357	0,7985	0,0155	0,0304	0,0911	32	Sel. Att. + Over
RANKER3	**0,8778**	0,8671	0,7357	0,7985	0,0155	0,0304	0,0911	32	Sel. Att. + Over
RANKER1	**0,8770**	0,9535	0,5793	0,7430	0,0345	0,0650	0,1019	28	Sel. Att.
-	**0,7203**	0,8865	0,5415	0,6927	0,0135	0,0263	0,0610	213	Over

Tabla 53. Resultados de 2015 con ajuste de regularización para SUB5

Como se puede ver en los resultados ajustando el parámetro conseguimos un modelo más robusto. Se puede ver que se consigue un valor de ROC AUC alto, comparable a los modelos de selección de atributos, utilizando el modelo inicial, es decir, se consiguen resultados similares realizando la selección de atributos que sin realizarla.

8.12.5. Coeficientes de la regresión logística

Como se ha comentado en el apartado 8.11.9, el algoritmo regresión logística asigna un peso a cada atributo. En este apartado se exponen, de la misma manera que en 8.10.9, los pesos que la regresión logística asigna a cada atributo.

Como conjunto de datos se utiliza la unión de 9/2014 y 12/2013, y se obtienen los pesos al realizar selección de características, y selección de características más oversampling. Viendo que el algoritmo de selección RANKER_3 es el que mejores resultados obtiene, este es el que se utiliza.

Hay que tener en cuenta que un valor positivo implica aumentar la probabilidad a la clase 0**.** Por el contrario un valor negativo implica aumentar la probabilidad la clase 1 (positiva, cardiovascular).

FARMACIA CODMED

Atributo	Descripción	Coef. Selección	Coef. Selección y over
ant_cmbd=1	Antecedente cardiovascular CMBD	-1,5618	-1,9339
charlson_age	Componente edad de Charlson	-0,0143	-0,0251
charlson_enf_pulm_cronica=1	Componente enf_pulm_cronica de Charlson	-0,7426	-0,5335
charlson_insuf_card=1	Componente Charlson insuf_card	0,1929	0,0033
dgps_ALT	Valores de los registros de DGPs con código ALT	0,0089	-0,0051
dgps_TABACO=0	Tabaco no fumadores	0,2868	0,227
dgps_TABACO=1	Tabaco exfumadores	-0,1141	-0,0935
dgps_TABACO=-1	Tabaco registro desconocido	0,0148	0,1162
dgps_TABACO=2	Tabaco fumador	-0,6288	-0,4365
num_cod_A01AB__Antiinfecciosos_y_antisepticos	Numero de principios activos de A01AB__Antiinfecciosos_y_antisepticos	-0,0376	0,0413
num_cod_Agonistas_muscarinicos_de_corta_duracion	Numero de principios activos de Agonistas_muscarinicos_de_corta_duracion	-0,02	-0,1473
num_cod_Agonistas_muscarinicos_de_larga_duracion	Numero de principios activos de Agonistas_muscarinicos_de_larga_duracion	-0,1765	-0,2261
num_cod_Betaagonistas_de_corta_duracion	Numero de principios activos de Betaagonistas_de_corta_duracion	-0,0642	-0,0854
num_cod_Betaagonistas_de_larga_duracion	Numero de principios activos de Betaagonistas_de_larga_duracion	-0,1013	-0,2356
num_cod_Betalactamicos	Numero de principios activos de Betalactamicos	-0,1529	-0,3561
num_cod_Corticoides	Numero de principios activos de Corticoides	-0,1238	-0,1493
num_cod_Corticoides_inhalados	Numero de principios activos de Corticoides_inhalados	-0,1063	-0,0504
num_cod_Diureticos_del_asa	Numero de principios activos de Diureticos del asa	-0,0571	-0,0948
num_cod_Ivabradina	Numero de principios activos de Ivabradina	-0,0871	-0,2944
num_cod_Quinolonas	Numero de principios activos de Quinolonas	-0,1634	-0,0493
num_cod_R03DA__Xantinas	Numero de principios activos de R03DA__Xantinas	-0,0957	-0,1604
num_cod_Roflumilast	Numero de principios activos de Roflumilast	0,0646	0,1283
num_epi_all_K	Conteo de episodios de aparato circulatorio	-0,125	-0,108
num_epi_all_R	Conteo de episodios de aparato respitatorio	-0,0305	-0,0435
num_ingr_07	Conteo de ingresos de enf sistema circulatorio	-0,102	-0,3977
num_ingr_08	Conteo de ingresos de enf sistema respiratorio	-0,1776	-0,4049

num_ingr_16	Conteo de ingresos de signos, síntomas y estados mal definidos	-0,6407	-0,4387
num_pac	Conteo de principios activos agrupados por CODMED	-0,0082	-0,0592
num_pas1	Conteo de principios activos en SUB1	0,0559	0,1018
Intercept	Coeficiente del algoritmo Logistic	79,703	41,032

Tabla 54. Coeficientes de la regresión logística para HPE y farmacia CODMED

FARMACIA SUB5

Atributo	Descripción	Coef. Selección	Coef. Selección y over
ant_cmbd=1	Antecedente cardiovascular CMBD	-1,6028	-1,8151
charlson_age	Componente edad de Charlson	-0,0101	-0,0208
charlson_enf_pulm_cronica=1	Componente enf_pulm_cronica de Charlson	-0,7831	-0,5691
charlson_insuf_card=1	Componente insuf_card de Charlson	0,228	0,116
charlson_otra_enf_vasc_perif=1	Componente otra_enf_vasc_perif de Charlson	-0,1392	0,1306
dgps_ALT	Valores de los registros de DGPs con código ALT	0,0149	0,0007
dgps_AST	Valores de los registros de DGPs con código AST	-0,0092	-0,0072
dgps_TABACO=0	Tabaco no fumador	0,2532	0,268
dgps_TABACO=1	Tabaco exfumador	-0,1226	-0,1515
dgps_TABACO=-1	Tabaco registro desconocido	0,0727	0,2123
dgps_TABACO=2	Tabaco fumador	-0,6271	-0,5482
dgps_TAD	Valores de los registros de DGPs con código TAD	0,0016	0,0023
num_epi_all_A	Conteo de episodios de Problemas generales e inespecíficos	-0,0423	0,0763
num_epi_all_B	Conteo de episodios de sangre, órganos hematopoyéticos y sistema inmunitario	-0,0867	-0,1857
num_epi_all_F	Conteo de episodios de Ojo y anejos	-0,0394	0,0389
num_epi_all_K	Conteo de episodios de aparato circulatorio	-0,1178	-0,1251
num_epi_all_R	Conteo de episodios de aparato respiratorio	-0,0261	-0,0406
num_epi_all_X	Conteo de episodios de Aparato genital femenino y mamas	0,7759	0,763
num_ingr_07	Conteo de ingresos de enf sistema circulatorio	-0,0953	-0,365
num_ingr_08	Conteo de ingresos de enf sistema respiratorio	-0,1677	-0,4481
num_ingr_16	Conteo de ingresos de signos, síntomas y estados mal definidos	-0,6283	-0,3139
num_paA01AB	Numero de principios activos de Insulinas y anlogos de accin rpida	-0,0003	0,2179
num_paB01AA	Numero de principios activos de Antagonistas de la vitamina K	-0,0378	-0,0904
num_paC01AA	Numero de principios activos de Glucsidos digitlicos	0,0768	0,1345
num_paC03CA	Numero de principios activos de Sulfonamidas, monofrmacos	-0,0499	-0,0676

num_paC03DA	Numero de principios activos de Antagonistas de la aldosterona	-0,0458	-0,2027
num_paC10AA	Numero de principios activos de Inhibidores de la HMG CoA reductasa	0,0471	0,1099
num_paG04CA	Numero de principios activos de Antagonistas de los receptores alfa adren	-0,0148	0,0066
num_paH02AB	Numero de principios activos de Glucocorticoides	-0,0971	-0,0651
num_paJ01CR	Numero de principios activos de Combinaciones de penicilinas, incluyendo i	-0,0195	-0,0495
num_paJ01DD	Numero de principios activos de Cefalosporinas de tercera generacin	-0,3599	-0,7778
num_paJ01EE	Numero de principios activos de Combinaciones de sulfonamidas y trimetopri	-0,4726	-0,9291
num_paJ01MA	Numero de principios activos de Fluoroquinolonas	-0,136	-0,0292
num_paR03AC	Numero de principios activos de Agonistas selectivos de receptores beta-2	-0,0895	-0,1585
num_paR03AK	Numero de principios activos de Adrenrgicos y otros agentes contra padeci	-0,1998	-0,2751
num_paR03BA	Numero de principios activos de Glucocortocoides	-0,1417	-0,3062
num_paR03BB	Numero de principios activos de Anticolinrgicos	-0,1685	-0,2245
num_paR03DA	Numero de principios activos de Xantinas	-0,0767	0,0154
num_paR03DC	Numero de principios activos de Antagonistas del receptor de leucotrienos	0,0677	0,066
num_paR03DX	Numero de principios activos de Otros agentes contra padecimientos obstruc	0,061	0,0973
num_paR06AC	Numero de principios activos de Etilendiaminas sustituidas	-0,2752	0,0426
num_pas1	Conteo de principios activos en SUB1	0,0661	0,0592
num_pas5	Conteo de principios activos en SUB5	-0,0133	-0,063
TSI	TSI al que pertenece el paciente	0,0268	-0,0329
Intercept	Coeficiente del algoritmo Logistic	7,5048	3,5759

Tabla 55. Coeficientes de la regresión logística para HPE y farmacia SUB5

En las Tablas 54 y 55 podemos consultar los pesos que la regresión logística asigna a cada atributo. Se puede destacar que el antecedente es el atributo que mayor peso (negativo) obtiene influyendo a sufrir un evento tipo HPE.

Si comparamos estos coeficientes con los coeficientes del evento adverso cardiovascular, podemos destacar, entre otros, el atributo de la componente de Charlson enfermedad pulmonar crónica. Este atributo no era seleccionado (por los métodos de selección de atributos) para los cardiovasculares mientras que para los HPE obtiene un valor alto a sufrir evento HPE. Es más, en este caso aparecen también los datos de tabaco, dando mayor riesgo a los fumadores. Esto se puede deber a que entre los motivos HPE el más frecuente es el EPOC, una enfermedad relacionada con los pulmones (es una enfermedad pulmonar crónica).

Algo que se puede detectar en los coeficientes es que al aplicar el oversampling en ciertos atributos el paso pasa de ser positivo a negativo, es decir, pasa de ser protector a aumentar el riesgo de sufrir evento. Esto se debe que al replicar ejemplos de la clase positiva aleatoriamente (oversampling) hacemos que el algoritmo de más peso a esta clase.

8.12.6. Resumen

Finalmente se presenta en la Tabla 56 los mejores resultados de todos los experimentos anteriores a modo de resumen. Los resultados en negrita corresponden al mejor resultado de cada método y subrayado en mejor resultado global.

MÉTODO	VALIDACIÓN CRUZADA		DATOS 2015	
	12/2013 y 9/2014	Ajuste Logistic	12/2013 y 9/2014	Ajuste Logistic
Inicial	0,7906	**0,8956**	0,8244	**0,8870**
Oversampling	0,7932	**0,7968**	0,8133	**0,8162**
Selección de atributos	0,8794	**0,8920**	0,8807	**0,8850**
Selección + Oversampling	**0,8854**	**0,8854**	**0,8894**	0,8778
Oversampling + Selección	**0,8862**	**0,8862**	0,8778	**0,8894**

Tabla 56. Resumen resultados

Los modelos con selección de atributos son los más robustos. Obtienen un AUC medio en todos los meses similar, por lo que podemos decir que los modelos generalizan bien con datos desconocidos.

Comparando los resultados de los eventos HPE y los eventos cardiovasculares podemos observar una notable mejora en los resultados. Dado que la metodología empleada para ambos tipos de eventos es la misma, podemos concluir que loe eventos HPE son más fáciles de predecir que los eventos cardiovasculares.

9. Conclusiones y líneas futuras

Este proyecto se ha desarrollado en colaboración con el Servicio Navarro de Salud – Osasunbidea (SNS-O) cuyo objetivo es elaborar un sistema inteligente para la predicción de eventos adversos sobre la población Navarra. En concreto nos hemos centrado en aquellos pacientes de Navarra que sean considerados polimedicados, que ,junto al SNS-O, se definió un paciente polimedicado como aquel que tome 5 o más medicamentos durante al menos 3 meses consecutivos.

El proyecto se centró originalmente en desarrollar el sistema para la predicción de eventos adversos cardiovasculares, con capacidad para extender el sistema a otro tipo de eventos. Posteriormente, el sistema se adaptó para otro tipo de eventos llamados Hospitalizaciones Potencialmente Evitables (HPE).

Con estos dos tipos de eventos se realizaron diversos experimentos variando la forma de codificar los datos, realizando pre-procesamiento sobre los datos (selección de atributos, oversampling,…), utilizar diversos clasificadores,…

En los resultados se puede observar que llegamos a detectar una gran cantidad de estos eventos adversos, siendo los resultados de los HPE mejores que los eventos cardiovasculares. Observando el área bajo la curva ROC y la media geométrica (GM) se puede decir que conseguimos una detección aceptable. Sí que el sistema detecta riesgo de sufrir evento a muchos pacientes que en realidad no lo sufren (PPV bajo), pero esto es debido a la baja prevalencia de la clase positiva, es decir, que el número de pacientes que sufren evento adverso es mucho menor que el número de pacientes que no lo sufren.

Otra cosa a destacar son los coeficientes asignados por la regresión logística a cada uno de los atributos. En ellos se puede observar un conjunto de medicamentos que aumentan la probabilidad de sufrir evento adverso además de otro tipo de atributos, como por ejemplo el hecho de ser fumador aumenta el riesgo de sufrir un evento tipo HPE. Estos coeficientes son de gran ayuda para SNS-O ya que a parir de estos, ellos pueden realizar sus propias conclusiones.

Dado que los experimentos con los eventos tipo HPE han sido escasos (debido al tiempo), quedan pendientes la realización de varios experimentos como líneas futuras:

- En ocasiones los coeficientes que la regresión logística asigna a los atributos, cambian de signo al hacer oversampling. Para obtener un modelo más robusto se propone utilizar el algoritmo UnderBagging con la regresión logística como clasificador. El UnderBagging realiza una serie de iteraciones, donde en cada iteración obtiene un subconjunto de los datos aleatoriamente (undersampling). Sobre este subconjunto se ejecuta la regresión logística asignando un peso a cada atributo. Al realizar varias iteraciones tendremos que cada atributo obtiene un coeficiente por iteración. Lo que proponemos es realizar la media de esos coeficientes generando un nuevo modelo más robusto. Además se podría obtener el intervalo de confianza del coeficiente de cada atributo, lo que el SNS-O agradecería.
- Variar los valores de los medicamentos para observar el comportamiento del modelo. Se propone observar la probabilidad de sufrir un evento adverso aumentando o reduciendo los valores de cada medicamento consumido con el objetivo de identificar el aumento de qué medicamentos aumentan la probabilidad de sufrir un evento adverso.

- En ocasiones puede que el coeficiente asignado por la regresión logística a un atributo no es adecuado para el SNS-O. Lo que se propone es realizar un modelo en el cual los pesos puedan ser actualizados de manera manual utilizando el conocimiento del SNS-O para observar los cambios de probabilidad de sufrir evento.
- Realizar experimentos con cada motivo de HPE (Hospitalización Potencialmente Evitable) por separado. Hasta ahora se ha realizado una clasificación binaria, donde se considera si el paciente sufre o no un HPE. Dado que el EPOC es el HPE más frecuente (casi un 70% sobre el resto), el clasificador da más peso a este tipo de HPE que al resto. La idea que generar un modelo con cada HPE para observar que medicamentos, ingresos,… son los influyentes por cada motivo.
- Realizar experimentos con todos los motivos HPE en un mismo conjunto pero viendo el problema como un problema multi-clase. En esta ocasión se trata de generar un modelo que nos indique el riesgo que tiene cada paciente de sufrir cada diferente motivo de HPE.

10. Bibliografía y referencias

[1] X. Wu and V. Kumar, "The Top Ten Algorithms in Data Mining", CRC Press, ISBN: 978-1420089646, 2009

[2] I. H. Witten, E. Frank and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques" Morgan Kaufmann; Edición 3. ISBN: 978-0123748560

[3] S. Garcia, J. Luengo and F. Herrera, "Data Preprocessing in Data Mining", Springer International Publishing; Edición 1. ISBN: 978-3-319-10246-7

[4] M. Galar, A. Fernández, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews, vol. 42, no. 4, 2012.

[5] C. A. Ferreira, J. Gamma and V. S. Costa, "Exploring Multi-Relational Temporal Databases with a Propositional Sequence Miner"

[6] F. Herrera, C. J. Carmona, P. González and M. J. del Jesus, "An overview on subgroup discovery: foundations and applications" Knowl Inf Syst (2011) 29: 495. doi:10.1007/s10115-010-0356-2

[7] A. Bafna and J. Wiens, "Automated Feature Learning: Mining Unstructured Data for Useful Abstractions", 2015 IEEE International Conference on Data Mining

[8] A. J. Knobbe and E. K. Y. Ho, "Maximally Informative k-Itemsets and their Efficient Discovery"

[9] S. Salah, R. Akbarinia and F. Masseglia, "Fast Parallel Mining of Maximally Informative k-Itemsets in Big Data", 2015 IEEE International Conference on Data Mining

[10] J. M. Reps, J. M. Garibaldi, U.Aickelin, D. Soria, J. Gibson and R. Hubbard, "Comparision of algorithms that detect drug side effects using electronic healthcare databases", Sof Comput, vol. 17, nº 12, pp. 2381-2397, 2013.

[11] Z. J. Eapen, L. Liang, G. C. Fonarow, P. A. Heidenreich, L. H. Curtis, E. D. Peterson and A. F. Hernandez, "Validated, Electronic Health Record Deployable Prediction Models for Assessing Patient Risk of 30-Day Rehospitalization and Mortality in Older Heart Failure Patients", JACC: Heart Failure, vol. 1, nº 3, pp. 245-251, 2013.

[12] J. M. Reps, J. M. Garibaldo, U. Aickelin, D. Soria, J. E. Gibson and R. B. Hubbard, "A Novel Semisupervised Algorithm for Rare Prescription Side Effect Discovery", IEEE Journal of Biomedical and Health Informatics, vol. 18, nº 2, pp. 537-547, 2014.

[13] F. Rossi, H. Carneiro, C. da Silva and P. de Carvalho, "Possible adverse drug events leading to hospital admission in a Brazilian teaching hospital", Clinics, vol. 69, nº 3, pp. 163-167, 2014.

[14] V. Koutkias, V. Kilintzis, G. Stalidis, K. Lazou, J. Niès, L. Durand-Texte, P. McNair, R. Beuscart and N. Maglaveras, "Knowledge engineering for adverse drug event prevention: On the design and development of a uniform, contextualized and sustainable knowledge-based framework", Journal of Biomedical Informatics, vol. 45, nº 3, pp. 495-506, 2012.

[15] Y. Ji, H. Ying, P. Dews, A. Mansour, J. Tran, R. E. Miller and R. M. Massanari, "A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance", IEEE Transactions Information Technology Biomedicine, vol. 15, nº 3, pp. 428-437, 2011.

[16] J. Pathak, R. C. Kiefer and C. G. Chute, "Using linked data for mining drug-drug interactions in electronic health records", Stud Health Technol Inform, vol. 192, nº 1, pp. 682-686, 2013.

[17]J. Huang, J. Huan, A. Tropsha, J. Dang, H. Zhang and M. Xiong, “Semantics-Driven Frequent Data Pattern Minning on Electronic Health Records for Effective Adverse Drug Event Monitoring”, 2013 IEEE International Conference on Bioinformatics and Biomedicine

[18] S. L. Brilleman and C. Salisbury, “Comparing measures of multimorbidity to predict outcomes in primary care: a cross sectional study”, Family Parctice, vol. 30, nº 2, pp. 172-178, 2013.

[19] J. M. Quail, L. M. Lix, B. A. Osman and G. F. Teare, “Comparing comorbidity measures for predicting mortality and hospitalization in three population-based cohorts”, BMC Health Services Research, vol. 11, nº 146, pp. 1-12, 2011.

[20] L. G. Amrock, M. D. Neuman, H. Lin and S. Deiner, “Can routine preoperative data predict adverse outcomes in the elderly? Development and validation of a simple risk model incorporating a chart-derived frailty score”, American College of Surgeons, vol. 219, nº 4, pp. 684-694, 2014.

[21] C. Walsh and G. Hripcsak, “The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions”, Journal of Biomedical Informatics, vol. 52, nº 1, pp. 418-426, 2014.

[22] J Hippisley-Cox and C. Couoland, “Development and validation of risk prediction equations to estimate future risk of blindness and lower limb amputation in patients with diabetes: cohort study”, BMJ, vol. 351, nº 1, pp. 1-11, 2015.

[23] P. Bohórquez, M. D. Nieto, B. Pascual, M. J. García, M. A. Ortiz and M. Bernabéu, “Validación de un modelo pronóstico para pacientes pluripatológicos en atención primaria: Estudio PROFUND en atención primaria”, vol. 46, nº 3, pp. 41-48, 2014.

[24] R. J. B. Loymans, P. J. Honkoop, E. H. Termeer, J. B. Snoeck-Stroband, W. J. J. Assendelft, T. R. J. Schermer, K. F. Chung, A. R. Sousa, P. J. Sterk, H. K. Reddel, J. K. Sont and G. ter Riet, “Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model”, Thorax, vol. 71, nº 9, pp. 838-846, 2016.

[25] S. R. Sukumar, N. Ramachandran and R. K. Ferrell, “Quality of Big Data in health care”, International Journal of Health Care Quality Assurance, vol. 28, nº 6, pp. 621-634, 2015.

[26] M. Lamain-de Ruiter, A. Kwee, C. A. Naaktgeboren, I. de Groot, I. M. Evers, F. Groenendaal, Y. R. Hering, A. J. M. Huisjes, C. Kirpestein, W. M. Monincx, J. E. Siljee, A. Van’t Zelfde, C. M. van Oirschot, S. A. Vankan-Buitelaar, M. A. A. W. Vonk, T. A. Weigers, J. J. Zwart, A. Franx, K. G. M. Moons and M. P. H. Koster, “External validation of prognostic models to predict risk of gestational diabetes mellitus in one Dutch cohort: prospective multicentre cohort study”, BMJ, vol. 354, nº 1, pp. 1-11, 2016.

[27] L. C. Kobayashi, S. E. Jackson, S. J. Lee, J. Wardle and A. Steptoe, “The development and validation of an index to predict 10-year mortality risk in a longitudinal cohort of older English adults”, Age and Ageing, vol. 0, nº 1, pp. 1-6, 2016.

[28] Y. Barak-Corren, V. M. Castro, S. Javitt, A. G. Hoffnagle, Y. Dai, R. H. Perlis, M. K. Nock, J. W. Smoller and B. Y. Reis, “Predicting Suicidal Behavior From Longitudinal Electronic Health Records”, The American Journal of Psychiatry, vol. 174, nº 2, pp. 154-162, 2017.

[29] H. Yu, H. Lo, H. Hsieh, J. Lou, T. G. McKenzie, J. Chou, P. Chung, C. Ho, C. Chang, T. Kuo, Y. Lo, P. T. Chang, C. Po, C. Wang, Y. Huang, C. Hung, Y. Ruan, Y. Lin, S. Lin, H. Lin and C. Lin, “Feature Engineering and Classifier Ensembre for KDD Cup 2010”, JMLR: Workshop and Conference Proceedings 1, pp. 1-16, 2010.

[30] P. Brierley, D. Vogel and R. Axelrod, “Heritage Provider Network Healg Prize, Round 1 Milestone Prize, How We Did It – Team ‘Market Makers’”, HPN Health Prize, 2011

11. Anexos

11.1. Anexo I: Informe base de datos completa

En este informe se presenta la información almacenada en la base de datos completa (sin filtro de polimedicados) mediante gráficos y estadísticas.

11.2. Anexo II: Informe base de datos polimedicados

En este informen se presenta la información almacenada en la base de datos de polimedicados mediante gráficos y estadísticas.

11.3. Anexo III: Informe experimentos 28/04/2016

Este informe incluye experimentos realizados uniendo todos los conjuntos variando la fecha de referencia mes a mes entre 12/2013 y 9/2014 (10 conjuntos de datos en uno). Además incluye experimentos utilizando el método Oversampling Undersampling explicado anteriormente.

11.4. Anexo IV: Informe experimentos 05/05/2016

Este informe incluye los experimentos realizados modificando la forma de codificar los episodios. Se añade a los conjuntos de datos todos los episodios de los últimos 6 meses (a partir de la fecha de referencia) generando el doble de columnas de episodios que en los experimentos anteriores.

11.5. Anexo V: Informe experimentos 04/07/2016

En los experimentos anteriores no se conocía la fecha de fallecimiento. Este informe incluye los experimentos realizados excluyendo los pacientes fallecidos antes de la fecha de referencia. Además incluye los primeros resultados utilizando los datos de 2015

11.6. Anexo VI: Informe experimentos 09/09/2016

Informe complementario al anterior con experimentos que quedaron pendientes.

11.7. Anexo VII: Informe experimentos 26/10/2016

Informe con los experimentos cambiando la forma de codificar la información TSI de los pacientes. En los experimentos anteriores se utilizaba como atributo nominal y en este informe se utiliza como atributo numérico.

11.8. Anexo VIII: Informe experimentos 08/11/2016

Resumen del informe anterior. En el informe anterior contiene muchas tablas que pueden confundir al SNS-O. Por ejemplo los experimentos de 2015 se realizaron obteniendo resultados mes a mes. En el resumen se incluye la media de los meses reduciendo la complejidad del informe.

11.9. Anexo IX: Informe experimentos 09/11/2016

Informe complementario al experimento anterior (Anexo VIII). En este se exponen resultados utilizando la combinación de medicamentos explicada anteriormente.

11.10. Anexo X: Informe experimentos 02/12/2016

Este informe incluye los experimentos realizados con las tres opciones de episodios finales a utilizar:

1. Episodios que están abiertos o han sido cerrados en el último mes.
2. Todos los episodios abiertos en los últimos 6 meses.
3. Combinación de las 2 anteriores sin duplicidad de datos, es decir, los episodios que están abiertos, los que han sido cerrados en el último mes o han sido abiertos en los últimos 6 meses pero no en el último mes.

11.11. Anexo XI: Informe experimentos 19/12/2016

Este informe incluye todos los experimentos realizados para finalizar con el evento adverso cardiovascular (sección 8.11)

11.12. Anexo XII: Informe experimentos 05/01/2017

Este documento es un resumen del informe anterior (Anexo XI). Se trata de reducir la complejidad de lectura para mejor entendimiento con el SNS-O.

11.13. Anexo XIII: Informe experimentos 19/01/2017

Este informe incluye los primeros experimentos realizados con las hospitalizaciones potencialmente evitables (HPE, sección 8.12).