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

Desarrollo y evaluación de métodos de selección de características para la predicción de eventos adversos en pacientes polimedicados

Grado en Ingeniería Informática

Trabajo Fin de Grado

Autor: Radostina Spasova Dimitrova

Director: Mikel Galar Idoate

Pamplona, 31-05-2017

upna Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Resumen

El objetivo del proyecto consiste en el desarrollo y evaluación de diferentes métodos de selección de características automáticos para problemas de clasificación. La selección de características es una de las etapas más importantes de la minería de datos y tiene gran influencia en los modelos obtenidos posteriormente por los algoritmos de aprendizaje. Por esta razón resulta de gran interés desarrollar y estudiar el comportamiento de los diferentes modelos existentes en la literatura en cada uno de los problemas reales existentes.

En concreto, el proyecto se centra en el estudio de pacientes polimedicados. Se trata de construir un modelo capaz de predecir cuando un paciente polimedicado sufrirá un evento adverso. Se estudian los eventos cardiovasculares y las hospitalizaciones potencialmente evitables. Este modelo nos permitirá analizar posteriormente, gracias a la selección de características previa, cuando las características del paciente o la toma de un nuevo medicamento pueden llevar a una mayor probabilidad de sufrir un evento adverso. De esta forma los médicos podrán tener un conocimiento extra a la hora de recetar nuevos medicamentos a los pacientes.

Palabras clave: KDD, Minería de datos, clasificación, polimedicado, Selección de Características

1 Introducción

El KDD = 'Knowledge Discovery from Databases' es un proceso que intenta extraer patrones a partir de grandes volúmenes de datos, y su objetivo final es extraer información de estos patrones de forma que sea comprensible y pueda ser utilizado posteriormente.

Informalmente se ha adoptado el término de Minería de datos para el proceso de KDD, sin embargo la Minería de Datos es sólo una etapa de éste.

Las etapas que conforman el KDD son las siguientes:

- **Integración y recopilación**: En esta etapa se produce la familiarización con el dominio del problema y la obtención de conocimiento a priori junto con la unificación de la información, ya que en la mayoría de los casos no se obtiene todos los datos de una sola fuente.

- **Preprocesamiento** Consiste en la preparación de los datos para los posteriores procesos del KDD. La calidad del conocimiento extraído depende directamente de la calidad de los datos, no solo del algoritmo de aprendizaje escogido. Esta etapa se puede dividir en tres pasos:

    - **Limpieza de datos:** Se identificarán datos anómalos (outliers), valores perdidos por una mala recogida o valores inconsistentes, duplicados, etc.

    - **Transformación:** Se construyen nuevos atributos a partir de otros si el problema lo permite y se considera necesario aplicando alguna operación sobre los ya existentes. Y se realiza la discretización de estos.

    - **Reducción de Dimensionalidad:** Se realiza selección de casos y/o selección de atributos.

- **Técnica de Minería de Datos:** En esta etapa seleccionaremos nuestro algoritmo de aprendizaje para construir el modelo a partir de los datos recopilados y preparados anteriormente. El modelo describirá los patrones y relaciones entre estos datos de forma que se podrán entender mejor, realizar predicciones y explicar situaciones pasadas. Es importante seleccionar un algoritmo y un tipo de modelo según las necesidades del problema a resolver. Se pueden dividir los algoritmos en dos grandes grupos: supervisados y no supervisados. En el caso de los supervisados se conoce con anterioridad la clase objetivo de los datos. En cambio en el aprendizaje no supervisado no se conocen.

- **Evaluación:** Para evaluar el conocimiento extraído se utiliza una fase de entrenamiento donde extraemos el conocimiento y una fase de test donde se prueba su validez. Hay diversas técnicas como validaciones simples o cruzadas. Es importante decidir también las medidas de validación, porcentaje de aciertos, media geométrica, error cuadrático medio, etc.

- **Difusión:** Por último se realizara la difusión del conocimiento adquirido con la realización de informes y su utilización.

  Un esquema de las etapas del KDD puede verse en la Figura 1.

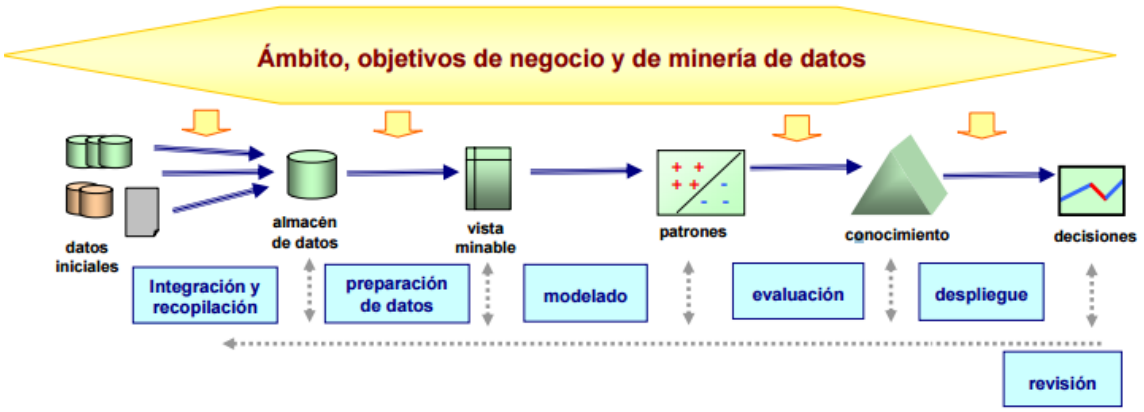

Figura 1- Esquema del proceso KDD

Actualmente estas técnicas tienen diversas aplicaciones. Aplicaciones empresariales como toma de decisiones en banca, seguros, finanzas, marketing… Aplicaciones en investigación científica como medicina, geografía, genética…Aplicaciones en Redes Sociales como minería de textos.

El objetivo de este proyecto se centra en el ámbito de la medicina, más concretamente en el estudio de pacientes polimedicados (aquellos que toman más de 5 medicamentos durante al menos 3 meses). Se trata de predecir episodios adversos principalmente en función de los medicamentos que toman los pacientes. De esta forma queremos estudiar si pueden existir ciertos eventos adversos que vengan dados por la polimedicación o no. En definitiva, nuestro modelo tratará de clasificar si un paciente tendrá un evento adverso en los próximos tres meses o no. Los tipos de eventos estudiados son los eventos cardiovasculares y las hospitalizaciones potencialmente evitables (HPE). El proyecto se centra en el proceso de Selección de Características, que está dentro de la etapa de Preprocesamiento de Datos. Se han desarrollado y evaluado diferentes métodos de selección para este problema de clasificación con el objetivo de realizar el aprendizaje con los datos que son más relevantes para hacer el proceso más eficaz y eficiente. Estos problemas fueron abordados en un proyecto de fin de master [1], por lo que en esta caso se parte de los datos que mejores resultados obtuvieron en ese proyecto.

Cabe destacar que esta etapa es de suma importancia ya que la calidad de los datos que se utilicen para realizar el aprendizaje que nos llevará al “conocimiento”, darán lugar a unos mejores o peores resultados e incluso a un menor tiempo de ejecución. De hecho está demostrado que un mayor número de atributos no da lugar a un mejor porcentaje de acierto. Hay atributos que pueden llegar a perjudicar al clasificador. Por ello el Preprocesamiento es la etapa en la que más tiempo se invierte en todo el proceso de KDD. En la Figura 2 se ven las proporciones de los tiempos empleados en cada etapa.

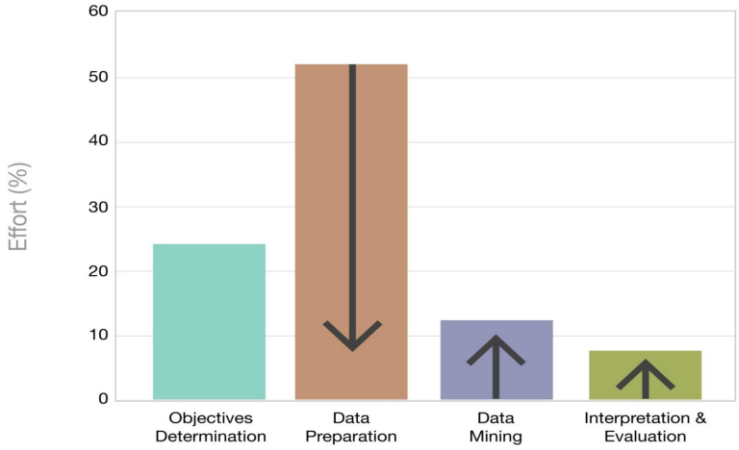

Figura 2- Esfuerzo requerido para el desarrollo de cada una de las fases en un proceso de KDD

Puesto que la Ciencia de Datos nos permite aprender del pasado para tomar mejores decisiones en el futuro las facilidades que se pueden aportar con estas técnicas son diversas en muchos campos. Destacando el ámbito de la salud por su importancia. Con el estudio respecto a si un paciente va a sufrir un episodio no deseado o no en función de los medicamentos administrados se pretende ayudar a prevenir estos episodios, sustituyendo algunos principios activos si el diagnostico lo permite o si esto no es posible llevar un control exhaustivo de los pacientes con mayor riesgo.

Por tanto el objetivo de este proyecto es construir un modelo con los datos proporcionados que permita realizar predicciones sobre posibles eventos adversos, eventos cardiovasculares y hospitalizaciones potencialmente evitables. Para ello se utilizarán varios algoritmos de aprendizaje con todo el conjunto de datos y realizando diferentes métodos de selección de características. Para evaluar estas configuraciones se calcularán distintas medidas de precisión. Además se implementará un algoritmo genético para considerar todos los aspectos configurables del aprendizaje y de la selección de características y así obtener las mejores opciones.

2 Detección de eventos adversos en pacientes polimedicados

Para este proyecto se utilizan datos de pacientes Navarros (2013 - 2015) que se consideren polimedicados y que no sufran o hayan sufrido cáncer ni VIH. Cabe destacar que el concepto de polimendicados viene dado por el proyecto original sobre este tema realizado entre el grupo de investigación GIARA y el Servicio Navarro de Salud donde se acordó que un paciente polimedicado era aquel que toma más de 5 medicamentos durante al menos 3 meses seguidos. El objetivo final en ambos es predecir un evento cardiovascular en los pacientes polimedicados, también se ha evaluado las hospitalizaciones potencialmente evitables (HPE). El número de pacientes con cuyos datos se trabaja es 45.400 después de eliminar una gran parte que no cumplen las condiciones descritas anteriormente.

Para entender mejor el problema se seguirá con una pequeña explicación sobre los datos de los pacientes. Hay una fecha de referencia que separa los datos predictivos de los datos posteriores que se usan para obtener la clase. Los pacientes utilizados evidentemente deben estar vivos en dicha fecha y tienen que cumplir con las condiciones necesarias para ser considerados polimedicados.

Antes de la fecha de referencia consideramos los eventos adversos sufridos como un atributo del paciente, es decir, un antecedente. Después de esta fecha los consideramos como la clase que queremos predecir. Para cada paciente elegido se obtiene si ha sufrido un evento no deseado o no a partir de la fecha de referencia en los próximos 3 meses. Este dato será la clase objetivo. La fecha de referencia se incluye en los datos predictivos. Como datos predictivos se incluye información sobre el paciente, como datos generales, medicamentos tomados, antecedentes, etc. En la Figura 3 se ve un pequeño esquema sobre la fecha de referencia considerando unos datos dentro de un periodo de un año.

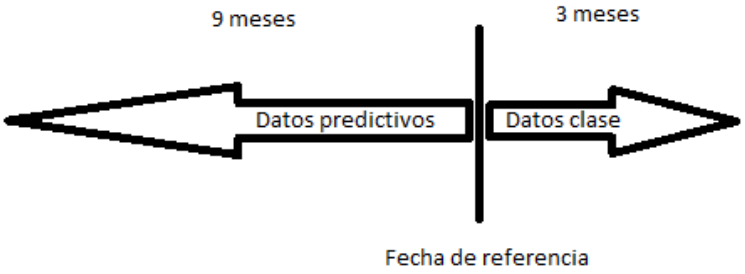

Figura 3- Explicación de la fecha de referencia considerados lo datos de un año

Se ha utilizado la misma metodología para predecir ambos eventos adversos estudiados, los eventos cardiovasculares y eventos de tipo HPE. Por lo que en este proyecto se tratarán en ocasiones como un mismo problema de clasificación llamando a la clase objetivo evento adverso no deseado.

Los datos utilizados originalmente contenían los identificadores de los pacientes, datos generales de los pacientes como la edad, peso, colesterol, si consume tabaco… Los episodios sufridos junto a su descripción, su fecha de inicio y fecha de final, información sobre los ingresos del paciente junto a fechas, motivos , descripción, código CIE9 (Clasificación internacional de enfermedades) y por último los medicamentos que han consumido cada mes.

Los datos que se utilizan en este trabajo están proporcionados por el proyecto anterior por lo que son aquellos que mejores resultados obtuvieron en sus experimentos. En el caso de la predicción de eventos cardiovasculares y de la predicción de HPE se consideró lo siguiente:

- **CMBD:** Se trata del conteo de ingresos agrupado por nivel 1, el nivel que va tiene en cuenta menos detalles con respecto a los ingresos.

- **Episodios:** Conteo de todos los episodios que están abiertos o han sido cerrados en el último mes agrupados por el primer carácter, es decir la agrupación más general

- **Número de meses para los datos (N_{GEN}):** 6 meses para la obtención de datos

- **DGPs**: Es la media de los resultados de 3 meses a partir del último existente. Es decir la media de los DGPs. Los DGPs definen datos generales del paciente como el colesterol o el tabaco.

- **Antecedentes**: Se tienen antecedente de CMBD y antecedente de episodios

- **TSI**: Grupo TSI del paciente (001, 002A, 002B,…) como atributo numérico. Este código indica su estatus socioeconómico.

- **Farmacia**: Se utilizan los conteos de principios activos según diferentes agrupaciones de estos. Se disponía de 5 agrupaciones, SUB1, SUB3, SUB5, SUB7 Y CODMED. Los cuatro primeros van de menos detalle a más y CODMED es otra forma de agrupar los principios de SUB7 proporcionada por los profesionales de la salud. En este caso se utilizan dos codificaciones: SUB5 y CODMED.

- **Índice de Charlson**: Se trata de un índice de comorbilidad que está asociado a la esperanza de vida. Para su cálculo se utilizan los episodios sufridos sumando algunos puntos dependiendo de estos episodios. En los datos se utiliza 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

Aunque se han considerado diferentes fechas de referencia se ha obtenido el conjunto final de datos uniendo un conjunto con la fecha de referencia del 9/2014 y otro con 12/2013. Se dispone de 5 particiones de estos datos que se han utilizado en este proyecto para los experimentos.

3 Selección de características

3.1 Introducción

Como anteriormente se ha mencionado en la etapa de Preprocesamiento de datos del KDD se encuentran las técnicas de reducción de la dimensionalidad. Dentro de estas técnicas podemos diferenciar dos mecanismos de reducción.

- **Extracción de características:** Consiste en transformar el espacio P de los atributos en otro nuevo que no es un subespacio de P. De esta forma se obtienen nuevos atributos que no están en los datos originales del problema aplicando alguna operación a los atributos originales. Estas técnicas tienden a utilizarse cuando la precisión del modelo es más importante que su interpretabilidad.

- **Selección de características:** Se trata de seleccionar un conjunto de atributos del conjunto original de forma que sean los más adecuados para la tarea a realizar.

El objetivo final en ambas técnicas es conseguir la mayor separabilidad entre clases y un aprendizaje más rápido puesto que en muchas ocasiones se disponen de muchos datos redundantes o que no aportan información a la hora de definir la clase objetivo.

Por la naturalidad del proyecto parece claro que lo adecuado es realizar una selección más que una transformación puesto que interesa saber qué atributos de los pacientes definen un evento adverso, unos principios activos u otros, si consumen tabaco, edad, colesterol, etc. Si se hiciera transformaciones a los atributos el modelo no sería interpretable por los profesionales de la saludo y no se podría sacar conclusiones de él.

Las ventajas de realizar la Selección de características en cualquier proyecto de este tipo son varias, el aumento de la velocidad y los modelos más simples que ayuda a un mejor conocimiento del proceso y entendimiento de los datos. Además mejora los resultados de los algoritmos de aprendizaje y disminuye los requerimientos de almacenaje. Un modelo construido con un mayor número de atributos no supone una mayor probabilidad de éxito.

3.2 Tipos de características

Se distinguen atributos relevantes, irrelevantes y redundantes. Se considera un atributo irrelevante cuando el conocimiento de su valor no aporta nada para predecir la clase objetivo. Los atributos redundantes son aquellos que no aportan la suficiente información y pueden ser eliminados. La redundancia es definida en términos de dependencia que existe entre los atributos, los atributos que están altamente correlacionados se dice que son redundantes. En otras palabras el valor de un atributo redundante puede ser determinado a partir de otros por lo que podemos prescindir de este.

Los atributos relevantes son aquellos que tras su eliminación la precisión del modelo aprendido disminuye y el conocimiento que aportan no puede ser proporcionado por otros atributos.

En términos de tipos de atributos el objetivo de la selección de características es identificar los atributos que son irrelevantes y/o redundantes y eliminarlos obteniendo así las ventajas anteriormente descritas.

3.3 Clasificación de métodos de selección

Los métodos de selección se pueden clasificar de distintas maneras. Las clasificaciones más consideradas son dos. Según el tipo de evaluación que hacemos, individual o por subgrupos y según la relación del método con el algoritmo de aprendizaje. A continuación se explican estas dos clasificaciones.

3.3.1 Clasificación según el modelo de evaluación

Los métodos según el tipo de evaluación que hacemos a los atributos:

- **Evaluación individual:** A cada atributo se le asigna un valor según su grado de relevancia. No se tiene en cuenta la correlación que puede haber entre atributos si no que solo la información que aportan individualmente.

- **Evaluación de subconjuntos:** Se producen subconjuntos de atributos que se evalúan según una medida y se comparan con el mejor anterior subconjunto.

3.3.2 Clasificación según su relación con el algoritmo de aprendizaje

Clasificación basada en la relación entre el algoritmo de selección el método de aprendizaje que se use para inferir el modelo.

- **Técnicas de filtro:** Se calcula la relevancia de los atributos y se eliminan los menos relevantes. Lo bueno o malos que son los atributos se calcula solo en función de las propiedades intrínsecas de los datos sin haber una relación con el algoritmo de aprendizaje. Su complejidad computacional es menor.

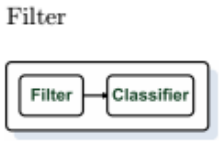

Figura 4 – Descripción gráfica del funcionamiento de los filtros

- **Técnicas de envoltorio (weappers):** La evaluación del subconjunto de variables se obtiene mediante aprendizaje y evaluación de un clasificador. El clasificador es una caja negra, no es consciente de que se está realizando la selección. Hay cierto riesgo de sobreaprendizaje.

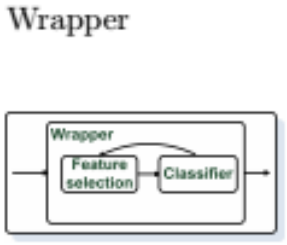

Figura 5 – Descripción gráfica del funcionamiento de los wrappers

- **Técnicas embebidas:** El algoritmo de aprendizaje del clasificador incluye la búsqueda del subconjunto óptimo de variables. El algoritmo de aprendizaje es consciente de que se está realizando la selección. Tiene un menor coste computacional con respecto a las técnicas wrappers.

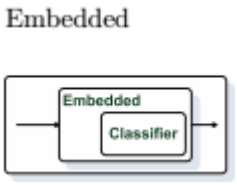

Figura 6 – Descripción gráfica funcionamiento embebidos

3.4 Direcciones de búsqueda

La dirección de búsqueda define como se va a conseguir el subconjunto óptimo. Las más utilizadas y las que están implementadas en este proyecto son las siguientes:

- **Hacia delante (Forward):** Con la dirección de búsqueda hacia delante se comienza con el conjunto vacío de atributos y se van añadiendo según un criterio que distingue el mejor atributo del resto.

- **Hacia atrás (Backward):** La dirección hacia atrás comienza con el conjunto de todos los atributos e iterativamente se van eliminando uno a uno. El criterio de evaluación apunta al peor atributo.

En ambos casos el criterio de parada que se utiliza en este caso es el número de atributos seleccionados, de este modo cuando hayamos alcanzado los N atributos requeridos pararemos la búsqueda.

Existen otras direcciones que combinan las dos anteriores como **Bidirectional Generation,** que empieza la búsqueda en ambas direcciones utilizando la búsqueda hacia delante y hacia atrás de forma concurrente. Otra posible opción es **Random Generation** que también utiliza la búsqueda hacia delante y hacia atrás y la decisión de si añadir o quitar atributos es una decisión aleatoria.

3.5 Discretización

El proceso de discretización de datos es el proceso mediante el cual se transforman los valores continuos de forma que se obtenga un número limitado de estados posibles. En muchas ocasiones es necesario disponer de datos discretizados para utilizar algunos algoritmos de clasificación. En este trabajo se utilizan datos ya discretizados proporcionados por el proyecto fin de máster que comenzó con este estudio.

Los algoritmos de discretización se podrían dividir en dos grandes clases, supervisados y no supervisados. Los supervisados tienen en cuenta la clase y los no supervisados no. En este caso los atributos vienen discretizados con un método supervisado utilizando la entropía.

El procedimiento separa los posibles rangos eligiendo los puntos de corte que menor entropía tengan. El objetivo es que el máximo número de ejemplos que pertenecen a un intervalo sean de la misma clase.

4 Filtros empleados

Los métodos que se han empleado en este proyecto son filtros por lo que no dependen del algoritmo de aprendizaje sino que solo dependen de las propiedades de los datos. En el proyecto se implementa un método de selección general de forma que combinando valores de una serie de parámetros llegamos a distintos filtros. La fórmula se consigue mediante una combinación lineal de la ganancia de información. Nos basamos en un artículo [2] que considera distintos métodos de selección en este marco común.

4.1 Entropía y ganancia de información

La medida de evaluación de los atributos utilizada es la ganancia de información, por ello se deben dar las siguientes explicaciones para la mejor comprensión de los métodos de selección utilizados.

La entropía denota el nivel de incertidumbre que presenta una distribución X, cuanta más entropía más incertidumbre. La entropía que denotamos como $H(X)$ se define como,

$$H(X) = -\sum_{x \in X} p(x) log p(x)$$

Si todos los eventos de una variable aleatoria son igual de probables entonces la entropía es máxima, lo que es lo mismo, una incertidumbre máxima. Puesto que se trabaja con atributos discretos se tiene que, $p(x) = \frac{\#x}{N}$ siendo N el número total de eventos y #x el número de eventos de tipo x.

La información mutua o ganancia de información, I(X;Y) ,de dos variables es una cantidad que mide la dependencia entre dos variables, es decir, mide la reducción de incertidumbre de una variable, X, debido al conocimiento del valor de otra variable o dicho de otra manera la información compartida entre ambas.

La entropía puede verse como un caso particular de la ganancia de información, cuando X e Y son iguales $I(X;X) = H(X)$

Sabemos que,

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log p(x|y)$$

Y se deduce la fórmula que define la ganancia de información de esta manera,

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x \in X} \sum_{y \in Y} p(xy) \log \frac{p(xy)}{p(x)p(y)}$$

Se puede observar que la ganancia de información es 0 si y solo si las variables son independientes, es decir no comparten ninguna información entre sí.

Cabe mencionar también que la ganancia de información puede ser condicional obteniendo la siguiente fórmula,

$$H(X;Y|Z) = H(X|Z) - H(X|YZ)$$

$$= \sum_{z \in Z} p(z) \sum_{x \in X} \sum_{y \in Y} p(xy|z) log \frac{p(xy|z)}{p(x|z)p(y|z)}$$

Poniendo un ejemplo sobre el problema a resolver en este caso, si tenemos una variable COLESTEROL y la clase EVENTO CARDIOVASCULAR, teniendo COLESTEROL alto se disminuye la incertidumbre de sufrir o no un evento (en este caso, a favor de que el evento sea más probable debido al colesterol alto). Cuando un paciente tiene el colesterol alto la probabilidad de que en un futuro sufra problemas cardiovasculares es mayor, es decir, tenemos más certeza sobre el evento cardiovascular o como hemos dicho antes menos incertidumbre sobre dicho evento. De la misma manera si tiene el nivel de colesterol saludable la probabilidad disminuye. Por lo tanto son dos variables que tienen una información mutua importante.

4.2 Fórmula general

Se comenzará explicado la fórmula utilizada para la evaluación de los atributos seleccionados de lo más general a lo más particular. Esta fórmula recoge los métodos de selección de

características basados en la entropía, más concretamente los que son combinaciones lineales de está. El método desarrollado tal y como se ha explicado anteriormente es un método general que nos permite utilizar varios métodos según los valores de los parámetros,

$$J(X_k) = \underset{(1)}{I(X_k;Y)} - \underset{(2)}{\beta \sum_{j \in S} I(X_j;X_k)} + \underset{(3)}{\gamma \sum_{j \in S} I(X_j;X_k|Y)}$$

Siendo Y la clase objetivo, X los atributos, S el conjunto de atributos seleccionados y X_k el atributo que estamos evaluando. Los parámetros β y γ definirán realmente el método que se esté usando.

En primer lugar se debe a explicar el significado de cada uno de los sumandos de la anterior formula,

1) $I(X_k;Y)$: Se trata de la ganancia de información entre el atributo evaluado y la clase objetivo, de forma que si este valor es grande el atributo k es relevante con respecto a la clase. Este valor es independiente de los atributos que se hayan seleccionado, es decir evalúa cada atributo de forma individual.

2) $\sum_{j \in S} I(X_j;X_k)$: En este caso es la ganancia de información del atributo evaluado k con respecto a cada uno de los atributos que ya están seleccionados. Es una propiedad del atributo evaluado que se desea minimizar ya que si la ganancia es muy alta la información que nos aporta dicho atributo nos viene proporcionada por algún atributo o un subconjunto de atributos ya escogidos. Viene a definir la redundancia o correlación entre atributos.

3) $\sum_{j \in S} I(X_j;X_k|Y)$: Este sumando es la ganancia de información entre un par de atributos dada la clase Y. El objetivo es aumentar la información complementaria entre atributos. Si el atributo evaluado comparte mucha información con los atributos ya seleccionados dada la clase objetivo es un dato positivo para seleccionar dicho atributo.

Según los parámetros β y γ que se utilicen el método dará más o menos relevancia en disminuir la redundancia entre atributos y en aumentar la información complementaría entre ambos.

4.2.1 MIM: Mutual Information Maximisation

Se trata del método más sencillo, el método 'MIM', consiste simplemente en el primer sumando de la formula general, o dicho de otra forma obtenemos dicho método cuando los valores de los parámetros β y γ en la fórmula general valen 0.

Se trata de realizar un ranking de atributos según la ganancia de información con la clase objetivo Y. De esta forma se evalúa cada atributo de forma independiente con respecto a los

demás, se realiza una evaluación individual. En este caso se le indica al algoritmo cuantos atributos deseamos seleccionar y nos quedamos con los N mejores.

$$J_{mim}(X_k) = I(X_k; Y)$$

En este método no se tiene en cuenta la redundancia entre atributos, por lo que podemos estar seleccionando dos atributos aparentemente muy buenos, pero que nos aportan la misma información por lo que uno de ellos sería innecesario. Aparentemente no se cree que sea el método que mejores resultados vaya a aportar.

4.2.2 MIFS: Mutual Information Maximisation

El método llamado 'Mutual Information Maximisation' continúa utilizando la ganancia de información para calcular la relevancia, pero incluye otro término para penalizar la redundancia entre atributos. De esta forma se intenta conseguir una baja correlación entre los atributos que se van seleccionando. La fórmula del 'MIFS' se define de la siguiente manera,

$$J_{mifs}(X_k) = I(X_k; Y) - \beta \sum_{j \in S} I(X_j; X_k)$$

Con S el conjunto de atributos si el atributo evaluado tiene unas ganancias de información altas con los atributos en S obtendremos una puntación baja. Con el parámetro beta se regula la importancia que se le da a la correlación entre atributos. De esta forma si β = 1 supone que a las dependencias entre atributos se les da una máxima transcendencia.

Otra forma de entender el parámetro es la creencia en la independencia entre los datos con los que se trabaja. Si elegimos un valor cercano a 0 supone una fuerte creencia en la independencia de los datos, por lo que se puede decir que el método 'MIM' asume la independencia total de datos.

En este método β puede tomar valores entre 0 y 1, sin incluir el 0.

4.2.3 MRMR: Minimum Redundacy Maximum Relevance

Este método es una forma particular del 'MIFS', la fórmula es la misma la diferencia es el valor de β = $\frac{1}{|S|}$. De esta forma el valor del parámetro es inversamente proporcional al tamaño del conjunto de atributos seleccionados. Cuanto más grande es el conjunto mayor creencia se tiene de la independencia entre atributos.

$$J_{jmi}(X_k) = I(X_k; Y) - \frac{1}{|S|} \sum_{j \in S} I(X_j; X_k)$$

4.2.4 JMI: Join Mutual Information

Este caso utiliza β = $\frac{1}{|S|}$ y $\gamma = \frac{1}{|S|}$, por lo que cuanto mayor es nuestro subconjunto menos relevancia se le da a estos dos sumandos y asumimos que habrá más correlación entre los atributos seleccionados (al aumentar el conjunto es cada vez más probable que sea necesaria cierta redundancia ya que de otra forma podemos estar eliminando características útiles por penalizarlas demasiado). A su vez, también reducimos la importancia de la ganancia condicionada más independencia entre atributos.

$$J_{jmi}(X_k) = I(X_k;Y) - \frac{1}{|S|}\sum_{j \in S} I(X_j;X_k) + \frac{1}{|S|}\sum_{j \in S} I(X_j;X_k|Y)$$

4.2.5 CMI: Conditional Mutual Information

El CMI se define con la fórmula general utilizando β = (0,1) y γ = (0,1). Se da una importancia según el valor que queramos asignarles a los parámetros a la redundancia entre atributos y la ganancia de información entre atributos dada la clase objetivo.

$$J_{cmi}(X_k) = I(X_k;Y) - \beta \sum_{j \in S} I(X_j;X_k) + \gamma \sum_{j \in S} I(X_j;X_k|Y)$$

4.2.6 CIFE: Condicional Infomax Feature Extraction

El CIFE es una forma particular del CMI se define con la fórmula general utilizando β = 1 y γ = 1. Se da una importancia máxima a la redundancia entre atributos y la ganancia de información entre atributos dada la clase objetivo.

$$J_{cife}(X_k) = I(X_k;Y) - \sum_{j \in S} I(X_j;X_k) + \sum_{j \in S} I(X_j;X_k|Y)$$

4.2.7 CONDRED: Conditional Redundancy

CONDRED utiliza β = 0 y γ = 1, por tanto la fórmula queda de la siguiente manera,

$$J_{cmi}(X_k) = I(X_k;Y) + \sum_{j \in S} I(X_j;X_k|Y)$$

No parece una forma intuitiva de evaluación ya que tiene en cuenta positivamente la ganancia entre atributos condicionada a la clase pero no penaliza de ninguna manera la redundancia. En el Paper Conditional Likelihood Maximisation utilizado para este proyecto presenta a esa opción como un método poco explorado ya que tiene poca justificación.

4.3 Resumen

Por último cabe destacar que todos los métodos explicados se fueron proponiendo de forma independiente y el artículo en el que se basa este proyecto crea un marco común que permite unificar todos los métodos y por eso se ha visto conveniente explicarlos desde este marco común para que quede más simple.

Como resumen se aporta la siguiente tabla en que vemos los posibles valores de los parámetros de la fórmula que define cada método particular.

Método	β	γ
CMI	(0 , 1)	(0 , 1)
CIFE	1	1
JMI	$\frac{1}{\|S\|}$	$\frac{1}{\|S\|}$
MIM	0	0
MIFS	(0 , 1]	0
MRMR	$\frac{1}{\|S\|}$	0
CONDRED	0	1

Tabla 1 – Resumen de los métodos de selección y sus

Cabe destacar que no se ha seguido estrictamente la idea de la generalización con una formula dados dos parámetros puesto que en ocasiones estos son 0.Hay métodos de selección que no requieren muchos cálculos por ejemplos en el caso del MIM no se necesita realmente rellenar la estructura de los conteos de cada atributo con los demás puesto que no utiliza los sumandos que necesitan estos conteos, o dicho de otra forma beta y gamma son 0. Por lo que es totalmente innecesario calcular estos sumandos para posteriormente multiplicaros por un parámetro con valor 0.

Como curiosidades en la siguiente tabla sacada del artículo Conditional Likelihood Maximisation [2] se presentan algunos datos sobre varios métodos basados en la ganancia de información, en azul aparecen algunos de los que hemos utilizado en el proyecto.

Método	Nombre completo	Autor
MIM	Mutual Information Maximisation	Lewis(1992)
MIFS	Mutual Information Feature Selection	Battiti(1994)
KS	Koller-Sahami metric	Koller and Sahami(1996)
JMI	Joint Mutual Information	Yang and Moody(1999)
MIFS-U	MIGS-'Uniform'	Kwak and Choi(2002)
IF	Informative Fragments	Vidal-Naquet and Ullman(2003)
FCBF	Fast Correlation Based Filter	Yu and Liu (2004)
AMIFS	Adaptive MIFS	Tesmer and Estevez(2004)
CMIM	Conditional Mutual Info Maximisation	Fleuret(2004)
MRMR	Max-Relevance Min-Redundancy	Peng et al.(2005)
ICAP	Interaction Capping	Jakulin(2005)
CIFE	Conditional Infomax Feature Extraction	Lin and Tang(2006)
DISR	Double Input Symmetrical Relevance	Meyer and Bontempi (2006)
MINRED	Minimum Redundancy	Duch(2006)
IGFS	Interaction Gain Feature Selection	El Akadi et al.(2008)
SOA	Second Order Aproximation	Guo and Nixon(2009)
CMIFS	Conditional MIFS	Cheng et al.(2011)

Tabla 2 –Ejemplos de métodos de selección basados en la ganancia de información junto a sus autores.

4.4 Obtención de la fórmula general

El método general utilizado se obtiene a partir de uno de los que hemos explicado anteriormente como método particular. Concretamente deriva de "Conditional Mutual Information".

Se define como la evaluación de un atributo k con respecto a la clase habiendo ya seleccionado el conjunto de atributos S.

$$J_{cmi}(X_k) = I(X_k; Y \mid S)$$

A partir de las definiciones de ganacia de información y de ganancia de información condicionada se deduce que $I(A;B|C) - I(A;B) = I(A;C|B) - I(A;C)$ de la siguiente manera,

$$I(A;B|C) - I(A;B) = (\, H(A|C) + H(A|BC)\,) - (\, H(A) - H\,(A|B)\,)$$

$$I(A;C|B) - I(A;C) = (\, H(A|B) + H(A|CB)\,) \; - \; (\, H(A) - H(A|C)\,)$$

Usando esta propiedad se reinscribe la fórmula del método dando lugar a,

$$J_{cmi}(X_k) = I(X_k; Y \mid S) = I\,(X_k\,; Y) - I(X_k;\, S) + I\,(X_k; S|Y)$$

Se aprecia que esta forma de representar del "CMI" se parece mucho a la formula general que utilizamos para definir todos los métodos. Puesto que se trabaja con un espacio de muestra finito y discreto, tenemos N atributos y el conjunto S tiene M atributos seleccionados asumimos lo siguiente,

Para todo atributo X_k que no está seleccionado, es decir $X_k \notin S$

$$p(S|x_k) = \prod_{j \in S} p(x_j|x_k)$$

$$p(S|x_k y) = \prod_{j \in S} p(x_j|x_k y)$$

Es decir, los atributos seleccionados son independientes y condicionalmente independientes con la clase a un atributo X_k aún no seleccionado.

Esta suposición nos permite calcular la ganancia de un atributo no escogido, X_k con todos los que sí están en el conjunto S sumando cada ganancia de información de cada atributo de S con el atributo X_k. Ocurre lo mismo con la ganancia condicionada.

Para explicar más en detalle la transformación desglosamos un poco más la formula, concretamente los dos últimos sumandos obtenemos lo siguiente,

$$\begin{aligned} J_{cmi}(X_k) &= I\,(X_k\,; Y) \\ &\quad -H(\,S) + H\,(S|X_k) \\ &\quad +H(S|Y) - H(S|X_k Y) \end{aligned}$$

Gracias a lo asumido sobre la independencia de los atributos nos queda,

$$\begin{aligned} J_{cmi}(X_k) &= I\,(X_k\,; Y) \\ &\quad -H(\,S) + \sum_{j \in S} H\left(X_j \middle| X_k\right) \\ &\quad +H(S|Y) - H \sum_{j \in S} H\left(X_j \middle| X_k Y\right) \end{aligned}$$

Puesto que no podemos convertir en sumatorio los dos términos que quedan de entropía se introducen dos términos neutros para recuperar la ganancia de información del candidato X_k con todos los miembros de S, $\sum_{j \in S} H(X_j) - \sum_{j \in S} H(X_j)$ y $\sum_{j \in S} H\left(X_j \middle| Y\right) - \sum_{j \in S} H(X_j|Y)$. La fórmula queda de la siguiente manera,

$$\begin{aligned} J_{cmi}(X_k) &= I\,(X_k\,; Y) \\ &\quad -\sum_{j \in S} I\left(X_j \middle| X_k\right) + \sum_{j \in S} H(X_j) - H(\,S) \\ &\quad \sum_{j \in S} I\left(X_j \middle| X_k Y\right) - \sum_{j \in S} H\left(X_j \middle| Y\right) + H(S|Y) \end{aligned}$$

De esta fórmula podemos eliminar los términos que son constantes con respecto a X_k porque no tienen efecto a la hora de elegir atributo.

De esta manera llegamos a la fórmula con la que generalizamos todos los métodos utilizados en este proyecto junto a los dos parámetros β y γ.

$$J'_{cmi} = I(X_k;Y) - \sum_{j \in S} I(X_j;X_k) + \sum_{j \epsilon S} I(X_j;X_k|Y)$$

4.5 Ejemplo de utilización de los métodos

Para entender un poco mejor los métodos de selección explicados anteriormente se va a proceder a realizar un ejemplo utilizando un conjunto de datos con 3 atributos y 4 ejemplos. El ejemplo trata de ilustrar el procedimiento del método general en la dirección de búsqueda hacia delante. En la siguiente tabla se observan los datos que se van a utilizar para ilustrar los métodos.

DATASET			
Atributo1	Atributo2	Atributo3	Clase
2	1	2	1
1	2	3	1
1	2	1	0
1	1	1	0

Tabla 3 – Conjunto de datos de ejemplo.

Se van a calcular los 3 términos de la fórmula general para a continuación combinarlas para ver las puntuaciones que obtendríamos en cada caso.

Los resultados obtenidos calculando la ganancia de información de cada uno de los atributos son los siguientes:

$$I(Atributo1;Clase) = 0{,}2156$$

$$I(Atributo2;Clase) = 0$$

$$I(Atributo3;Clase) = 0{,}6930$$

Recordamos la fórmula de la ganancia de información,

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(xy) \log \frac{p(xy)}{p(x)p(y)}$$

El concepto de ganancia de formación es la información que comparten dos variables, en nuestros métodos son cada atributo con la clase objetivo. En este caso tenemos un atributo que tiene ganancia 0 con respecto a la clase. Fijándonos solamente en el valor de ese atributo y en el de la clase, no podemos sacar ninguna conclusión, tenemos una incertidumbre máxima. El

Atributo2 puede tomar dos posibles valores y toma el valor 1 con la clase 1 y con la clase 0 de la misma forma que lo hace cuando toma el valor 2. En otras palabras el Atributo2 no nos aporta absolutamente nada sobre el valor de la clase objetivo.

Ahora poniendo nuestra atención sobre el mejor de los atributos, en este caso el Atributo3. En los datos se ve justificado este hecho ya que cuando el Atributo3 toma el valor 1 la clase objetivo toma el valor 0 y si el atributo toma otro valor que no sea el 1 la clase objetivo es 1. Se ve de forma simple que nos aporta mucha información sobre la clase.

De este primer paso se deduce que el mejor atributo es el atributo 3 por ello se comenzará con el conjunto formado por este atributo. Para calcular el segundo término de la fórmula que corresponde a la redundancia entre atributos se va a calcular la ganancia de información de los demás con respecto al mejor, en este caso el 3. Los resultados obtenidos son los siguientes,

$$I(Atributo1; Atributo3) = 0{,}5683$$
$$I(Atributo2; Atributo3) = 0{,}3465$$

Fijándonos en los valores de los atributos se puede llegar a deducir que el Atributo1 y el Atributo3 son más redundantes que el 2 y el 3. Solo se da en un ejemplo en el que el Atributo 1 y el 3 tienen distintos valores.

Con respecto a la tercera parte de la fórmula, es decir, la información compartida entre atributos dada la clase, no hay una relación clara entre dos atributos fijándonos en si la clase es 1 o 0. Los resultados de la información compartida son los siguientes,

$$I(Atributo1; Atributo3|Y) = 0{,}3465$$
$$I(Atributo2; Atributo3|Y) = 0{,}3465$$

Recordamos la fórmula utilizada para calcular esta ganancia condicionada,

$$I(X;Y|Z) = \sum_{z \in Z} p(z) \sum_{x \in X} \sum_{y \in Y} p(xy|z) log \frac{p(xy|z)}{p(x|z)p(y|z)}$$

Ahora después de calcular todos los valores posibles veamos cómo se procederá a evaluar cada atributo. Al comienzo del algoritmo tenemos en conjunto de los atributos escogido S vacío y se elige para empezar cualquier método el atributo que mayor ganancia de información tiene respecto a la clase.

Vamos a considerar el método MIM, que se trata solo de la ganancia de información con respecto a la clase, el método MIFS con el parámetro β=1 que considera la ganancia de información y la redundancia, el método CIFE que considera la ganancia de información, la redundancia y la información compartida con β=1 y γ=1 y por último CONDRED que considera la ganancia de información y la información compartida con γ=1.

1. **MIM:** En la tabla 4 vienen los valores de puntuación del método MIM, en este caso si se tuvieran que elegir 2 atributos descartaríamos el atributo 2 puesto que tiene la menor puntuación.

$$J_{MIM} = I(Atributo; Clase)$$

	MIM
Atributo1	**0,2156**
Atributo2	0
Atributo3	**0,693**

Tabla 4 –Puntación de los atributos de ejemplos según el método MIM.

2. **MIFS:** Como primer paso se elige el atributo que mejor ganancia de información con respecto a la clase haya dado y este es el 3 y a continuación se resta la redundancia de los atributos no elegidos con respecto al elegido. En este caso elegimos el atributo 3 y el 2.

$$J_{MIFS} = I(Atributo; Clase) - \beta\, I(Atributo; Atributo)$$

Atributos ∈ S Atributos ∉ S

Atributo3 →

	MIFS
Atributo1	-0,3527
Atributo2	**-0,3465**

Tabla 5 –Puntación de los atributos de ejemplos según el método MIFS.

3. **CIFE:** En este método utilizamos los tres sumandos con el mismo procedimiento de elegir el primer atributo de S calculando la ganancia de información con respecto a la clase. En este caso como la ganancia de información entre el atributo 3 y el 1 y entre el atributo 3 y el 2 da el mismo valor los atributos escogidos son los mismos que en el método MIFS.

$$J_{CIFE} = I(Atributo; Clase) - \beta\, I(Atributo; Atributo) + \gamma I(Atributo; Atributo|Clase)$$

Atributos ∈ S Atributos ∉ S

Atributo3 →

	CIFE
Atributo1	-0,0062
Atributo2	**0**

Tabla 6 –Puntación de los atributos de ejemplos según el método CIFE.

4. **CONDRED:** Este método utiliza la ganancia de información con respecto a la clase y la información compartida entre atributos, como ya hemos dicho este último valor es el mismo para ambos casos por lo que los atributos escogidos serán los mismos que en el método MIM.

$$\boldsymbol{J_{CONDRED} = I(Atributo; Clase) + \gamma I(Atributo; Atributo|Clase)}$$

Atributos ∈ S Atributos ∉ S

	CONDRED
Atributo1	**0,5621**
Atributo2	0,3465

Tabla 7 –Puntación de los atributos de ejemplos según el método CONDRED.

En el caso de tener más atributos se deberán hacer los cálculos con respecto a todos los atributos que están elegidos, es decir que están dentro del conjunto S.

5 Implementación de los métodos

Para la implementación de los filtros de selección de características se han utilizado dos estructuras para calcular las distintas ganancias de información necesarias en el método general. Para realizar estos cálculos el primer paso es calcular las distintas probabilidades que componen la definición de entropía y por tanto la de ganancia de información.

Cómo ya sabemos se trata de un espacio muestral discreto por lo que la probabilidad de suceso de cada variable aleatoria que en nuestro caso es un atributo se estima como $p(x) = \frac{\#x}{N}$, siendo N el número de sucesos totales y #x el número de sucesos de x.

La implementación completa se ha realizado en Matlab, un entorno optimizado para el cómputo con matrices.

5.1.1 Estructura probabilidades Atributo – Clase

Recordemos la definición final de ganancia de información entre dos variables que se corresponde con el primer sumando del método general,

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(xy) \log \frac{p(xy)}{p(x)p(y)}$$

En nuestro caso X son los atributos de los que disponemos e Y es la clase objetivo. Para calcular este valor en primer lugar se deben calcular las probabilidades que lo definen ,$p(xy)$, $p(y)$ y $p(x)$. Para ello como se trata de datos discretos se va a proceder a realizar conteos.

Puesto que este término de la combinación lineal se utiliza siempre en todos los métodos se ha implementado de forma que se calcula la ganancia de información de todos los atributos con respecto a la clase en una vez utilizando cálculos matriciales y al principio de forma que estas operaciones se hacen solo una vez.

La estructura es una matriz de tres dimensiones *(x,y,z)* en la que las filas *x* son todos los atributos posibles, las columnas *y* los valores de la clase objetivo, en este caso tenemos solo dos valores, y la profundidad *z* son los posibles valores que pueden tomar los atributos. Por lo que el tamaño de la matriz será (N,nClases,K) donde N es el número de atributos nClases es el número de posibles clases y K es el máximo valor que puede tomar cualquier atributo. Si algún atributo no toma menos de K valores los restantes quedarán a 0. Aun que se pierde algo de memoria que no va a ser útil se agilizan los cálculos puesto que de esta forma se realizan operaciones matriciales que es el punto fuerte de la herramienta utilizada.

Los atributos puedan tomar varios valores y como se va a proceder a calcular la ganancia de información de todos los atributos a la vez tenemos que pensar que disponemos de los atributos $\{At_1\ At_2\ At_3 \dots At_N\}$ y cada uno de ellos puede tomar k valores distintas por lo que tenemos que calcular la probabilidad de cada valor de cada atributo que será nuestro suceso.

En nuestra estructura la fila 1, columna 1 y profundidad 1 representa el número de ejemplos cuyo atributo 1 (fila 1), tiene como valor 1 (profundidad 1) y además el valor de la clase es 1 (columna 1).

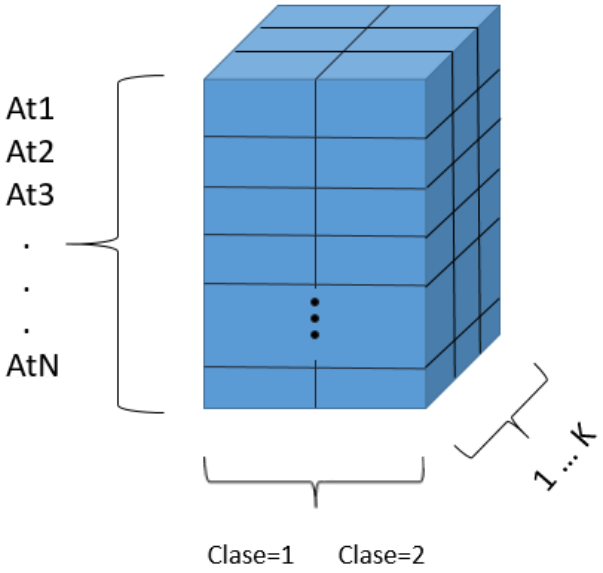

Figura 7: Matriz de conteos 1

Trabajamos con los valores 1 y 2 para la clase objetivo ya que los índices en las matrices en Matlab empiezan en 1 por lo que en la implementación la clase 1 se corresponde a la 0 en los datos y la clase 2 a la 1.

Para rellenar la estructura se recorren todos los atributos y se calculan los conteos de cada uno de ellos con respecto a cada valor que puedan tomar y al valor de la clase, es decir en cada iteración rellenaremos (At_i ,nClases ,K) siendo i = 1,2,…Nº atributos. Para obtener los valores correspondientes a esta matriz se calcula un histograma tridimensional con la función de Matlab *accumarray()* con cada atributo e insertaremos los valores obtenidos en la fila correspondiente.

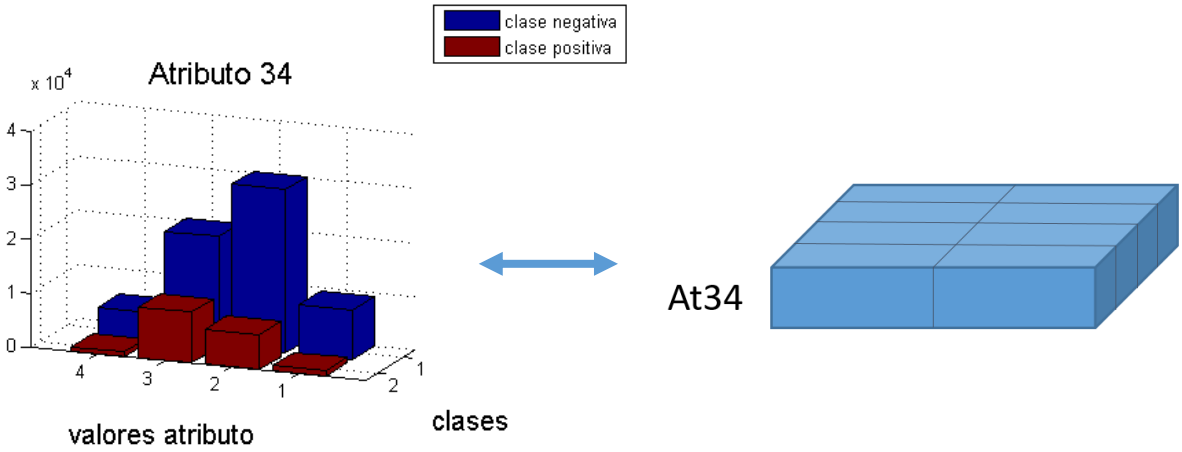

Figura 8 - A la derecha un ejemplo en gráfica 3D de los datos que obtenemos con la función accumarray y a la izquierda la parte de la estructura final donde colocamos estos datos.

Tras obtener estos conteos y tener la estructura lista el siguiente paso es realizar los cálculos oportunos para obtener el ranking MIM. Entendidos estos cálculos los demás términos de la fórmula se calculan siguiendo el mismo patrón.

Dividiendo cada valor obtenido en la estructura por el número total de ejemplos obtendremos una matriz de las mismas características con las probabilidades de que ocurra un valor determinado de x y un valor determinado de y.

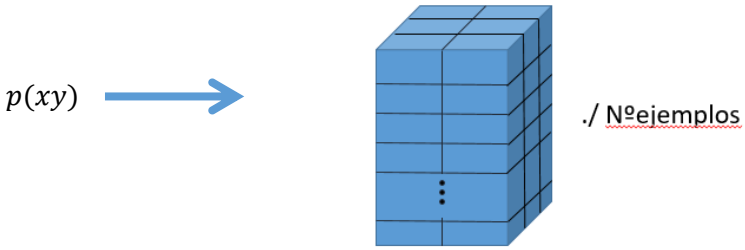

Figura 9- Matriz que contiene las probabilidades p(xy)

Si sumamos la matriz original por las filas obtendremos los conteos sin tener en cuenta las clases, es decir solamente fijándonos en los atributos y sus valores. A continuación dividiendo entre el número de ejemplos obtendremos p(x).

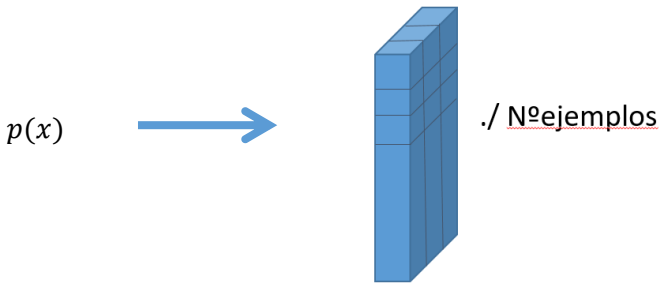

Figura 10- Matriz que contiene las probabilidades p(x)

Sumando la matriz original en profundidad y dividiendo el resultado entre el número de ejemplos totales obtendremos p(y). Todas las filas del resultado tendrán los mismos valores p(

y= 1) y p(y = 2). Gracias a esto podremos calcular p(x) x p(y) de forma eficiente duplicando p(x) y multiplicando esta con p(y) gracias a la función bsxfun de Matlab.

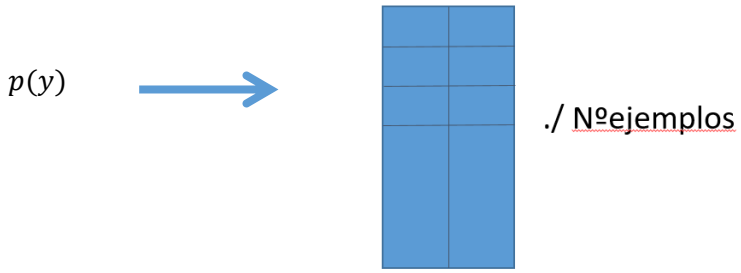

Figura 11- Matriz que contiene las probabilidades p(y)

Llegados a este punto disponemos de dos matrices del mismo tamaño que presentan p(xy) y p(x) x p(y) , se procede a realizar los cálculos matriciales de la ganancia de información $p(xy)\log\frac{p(xy)}{p(x)p(y)}$, obteniendo también una matriz de las mismas dimensiones que las anteriores. Por último se realizan los sumatorios, por columnas y a continuación en profundidad, obteniendo un vector que serán las ganancias de información de cada atributo con respecto a la clase. Si se ordena este vector junto con los atributos se sabrá de más a menos cuales tiene mejor puntuación.

5.1.2 Estructura probabilidades Atributo - Atributo

La estructura que se construye para los conteos de los valores de los atributos con respecto a los demás atributos es algo más compleja pero sigue el mismo patrón. En este caso dispondremos de una matriz por cada atributo.

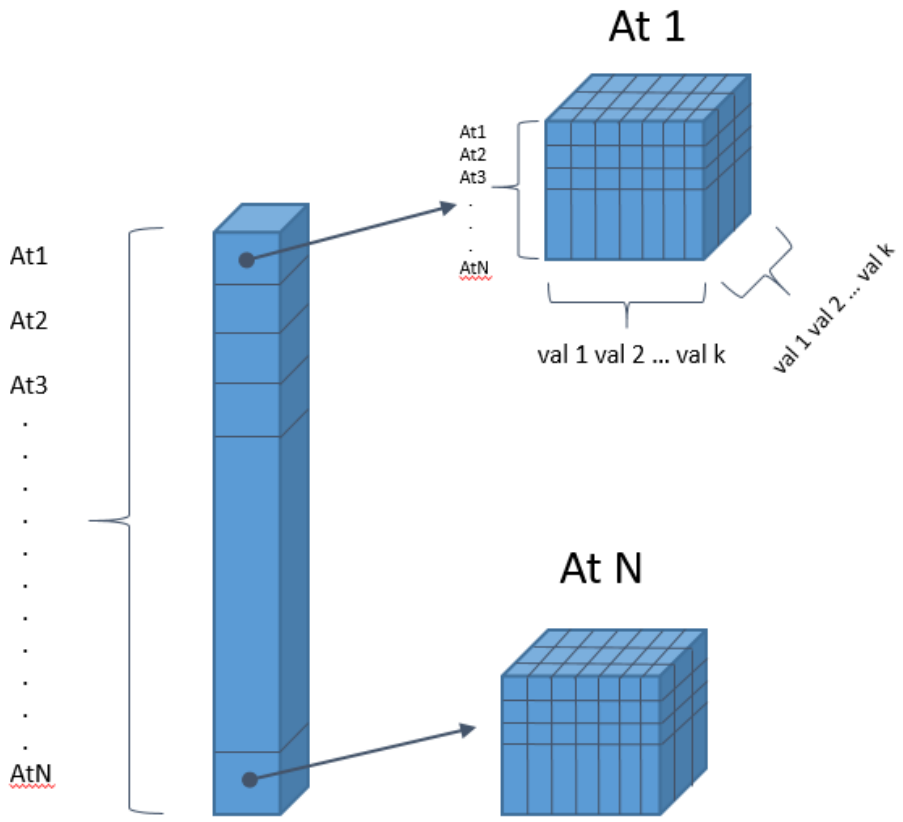

Figura 12- Estructura que contienen los conteos de los valores de los atributos con respecto a los demás atributos.

Tal y como se ve en la figura disponemos de un vector en el que cada elemento se corresponde a un atributo. Se ha construido esta estructura con las celdas de Matlab, de forma que tenemos un vector de N celdas, donde N es el número de atributos. Cada una de las celdas apunta a una matriz de cuatro dimensiones. En la figura 12 no aparece ilustrada la cuarta dimensión por simplificar el dibujo. Imaginemos que en cada celda de las matrices existen dos valores en vez de uno y esta será nuestra cuarta dimensión.

Cada matriz representa a un atributo X_k y sus dimensiones son (Nº atributos, Atributo K, maxK, Nº clases), las filas representan a los posibles atributos, las columnas "Atributo K" son los posibles valores que puede tomar el atributo X_k, la profundidad "maxK" es el máximo de los posibles valores que puede tomar cualquier atributo y la cuarta dimensión, "Nº clases", es el valor de la clase.

Gracias a esta estructura se puede sacar la redundancia o la información compartida de un Atributo X_k con respecto a todos los demás con los mismos cálculos matriciales que se utilizan para calcular su ganancia de información con respecto a la clase.

Cabe destacar que esta estructura se utiliza y se rellena de formas distintas según la dirección de búsqueda y se sirve para calcular los dos últimos sumandos del método general.

5.1.2.1 Optimizado para la dirección de búsqueda hacia delante

En esta opción se comienza con el conjunto vacío y se va añadiendo el atributo que mejor puntuación tenga. Para optimizar el método no se realizan todos los conteos a priori si no que se van calculando los conteos del atributo que se elige cada vez y se suma la puntuación del atributo elegido a la puntación del resto de atributos que ya estaban seleccionados. Comenzando con el atributo que mayor puntuación ha dado en el MIM.

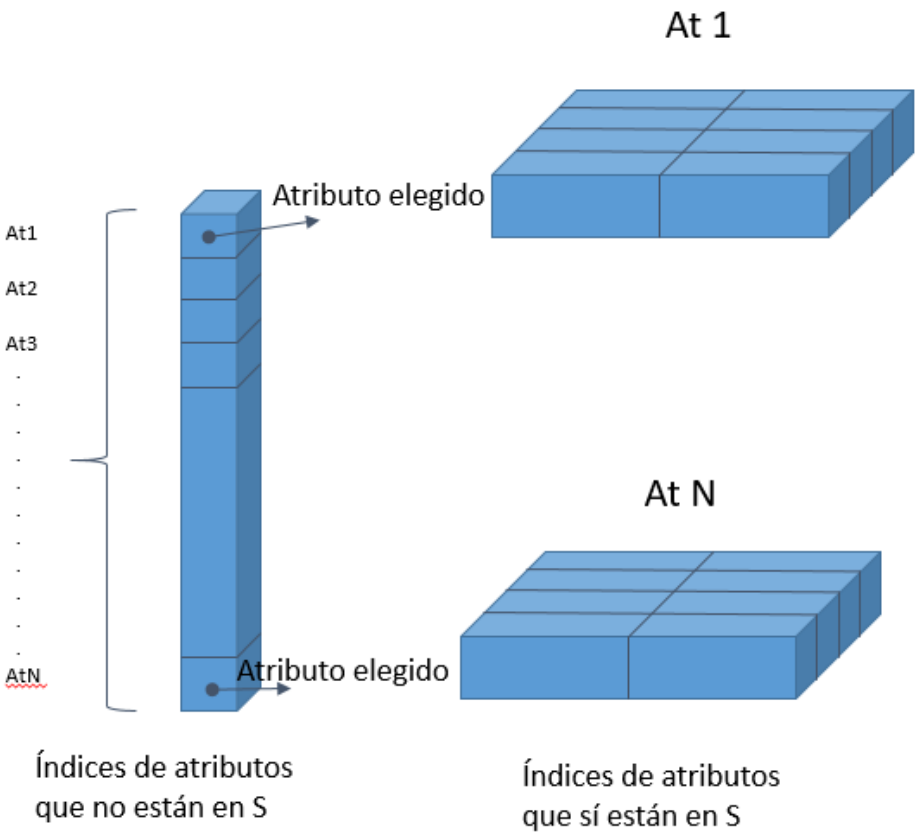

Figura 13- Estructura que contienen los conteos de los valores de los atributos con respecto a los demás atributos en la dirección de búsqueda hacia delante.

En cada iteración para los atributos que aún no están en S calculamos los conteos pertinentes para obtener las probabilidades solo del atributo elegido en dicha iteración con respecto a los atributos que no están en S. Tras realizar el conteo calculamos el término que corresponda, por

ejemplo la redundancia que se trata del segundo término y se la sumamos a la anterior redundancia calculada. De la misma forma con el tercer sumando del método.

Para calcular el segundo sumando del método general, es decir, la redundancia entre atributos, utilizaremos los mismos pasos que para el primer término añadiendo una dimensión más, o lo que es lo mismo, un sumatorio más. Que corresponde al número de atributos que ya están seleccionados. De modo que par a cada atributo considerado que aún no está en S calculamos la redundancia con respecto a todos los que sí lo están.

$$\sum_{j \in S} I(X_j; X_k) = \sum_{j \in S} \sum_{x \in X} \sum_{x \in X} p(X_j X_k) \log \frac{p(X_j X_k)}{p(X_j) p(X_k)}$$

En el caso del tercer sumando, es decir, la información compartida entre atributos dada la clase, se añade un sumatorio más. Ya que calcularemos las probabilidades según el valor de la clase. Se realizarán los mismos cálculos pero Nº de clases veces, en este caso dos veces puesto que tenemos dos clases y se multiplicará por la probabilidad de cada clase.

$$\sum_{j \in S} I(X_j; X_k|Y) = \sum_{j \in S} \sum_{y \in Y} p(y) \sum_{x \in X} \sum_{x \in X} p(x_j x_k|y) log \frac{p(x_j x_k|y)}{p(x_j|y) p(x_k|y)}$$

5.1.2.2 Optimizado para la dirección de búsqueda hacia atrás

En el caso de la búsqueda hacia atrás disponemos de todos los atributos inicialmente y elegimos el que peor puntuación tenga para eliminarlo. El procedimiento en sí mismo requiere de realizar los conteos de todos los atributos respecto a todos inicialmente para poder elegir el que menor puntuación obtenga. Por lo que se rellena la estructura anteriormente descrita al inicio del método y en cada iteración se utilizan los índices de los atributos que corresponda según el atributo elegido para salir del conjunto S.

Para optimizar el conteo en este caso se utilizan los conteos de los atributos anteriores de la siguiente manera. Recorremos los atributos y calculamos los conteos con la función de Matlab accumarray de la misma manera que si utilizamos la dirección de búsqueda hacia delante pero cuando calculamos estos conteos en el Atributok con respecto a los demás, solamente se calculan los atributos desde k+1 hasta Nºatributos cogiendo los conteos de los atributos anteriores a k de sus correspondientes matrices. A las partes de las matrices de 1 a k-1 que reutilizamos se les debe realizar sobre el eje y para que concuerden correctamente los valores para los cálculos posteriores.

Esta forma de reutilizar los datos, aunque parece costosa ha reducido el tiempo de ejecución de los conteos a la mitad con respecto a realizar todos los cálculos sin reutilizar los datos.

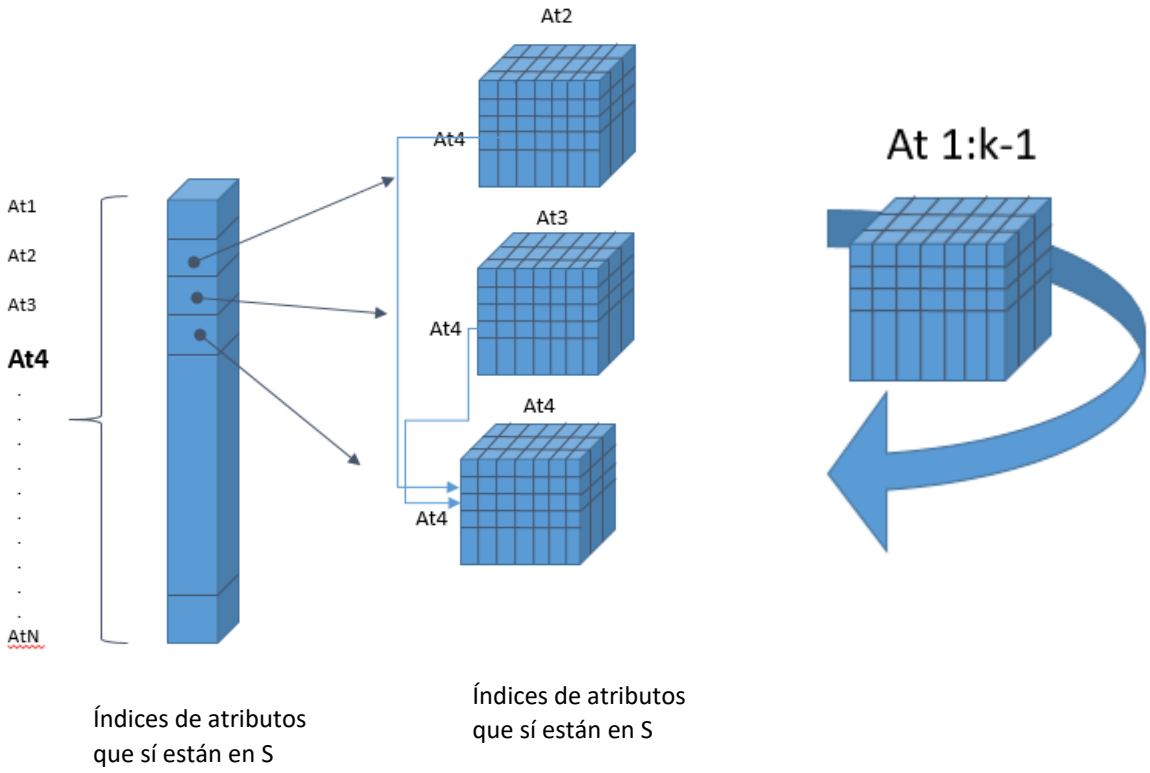

Figura 14- Figura explicativa sobre el método para rellenar la estructura que contienen los conteos de los valores de los atributos con respecto a los demás atributos en la búsqueda hacia delante.

Para realizar los cálculos de cada uno de los métodos por lo tanto de cada uno de los sumandos de la fórmula general se consideran los índices del vector de celdas y las filas de las matrices como los atributos que sí están en el conjunto S de mejores atributos. Se realiza de esta forma puesto que se comienza con todos los atributos y lo que se quiere evaluar es el peor de los que sí está seleccionado con respecto a los demás que también lo están.

Cada vez que se elige el peor atributo se recalculan los términos correspondientes al método de selección utilizado, es decir, la redundancia y/o la información compartida dada la clase, sin tener en cuenta el atributo elegido.

6 Evaluación de los métodos de selección

Para evaluar cada uno de los métodos de selección implementados en este proyecto se han utilizado varios algoritmos de clasificación supervisados diferentes ya implementados en Matlab.

Para saber cómo se comporta cada clasificador se utilizan métricas basadas en la matriz de confusión junto al porcentaje de acierto. Puesto que se utilizan algoritmos supervisados la matriz de confusión es una forma visual y sencilla de analizar las clases predichas con respecto a las reales.

De este modo se ha medido la precisión del clasificador con el conjunto de atributos seleccionado en cada caso.

6.1 Algoritmos de clasificación

- **Regresión Logística:** Se trata de un clasificador fácilmente interpretable por lo que se utiliza a menudo en el ámbito de la salud. Además es el algoritmo son el que mejores resultados se obtuvo en el proyecto original por lo que la mayor parte de los experimentos se han realizado con este clasificador.

- **SVM:** Support Vector Machines es un clasificador que trata de separar los ejemplos del conjunto de datos por un hiperplano de forma que lo que están a un lado serán de una clase y los que están a otro de la otra con un margen de separación máximo. Se ha utilizado el kernel lineal para los experimentos.

- **Naive Bayes:** Es un clasificador basado en el teorema de Bayes y se dice que es "ingenuo" puesto que asume la independencia entre las variables.

- **KNN:** Llamado k-vecinos más cercanos, se basa en el conjunto de ejemplos para realizar la clasificación. De forma que se elegirá la clase para un ejemplo según la clase de los k ejemplos que más se aproximen a este.

Hay clasificadores que de forma original proporcionan la clase predicha directamente y otros que devuelven la probabilidad de que cada ejemplo sea de cada clase. Sin embargo en este trabajo se han calculado siempre las probabilidades antes de hacer las predicciones con distintas funciones que nos aporta Matlab para calcular estas probabilidades a posteriori.

6.2 Matriz de confusión

La matriz de confusión es una herramienta muy utilizada para visualizar el desempeño de los algoritmos de clasificación supervisados. Se han calculado las siguientes medidas de precisión:

- **True Positive Rate (TPR):** Es el ratio de ejemplos positivos que han sido clasificados correctamente, es decir con clase positiva.

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

- **True Negative Rate (TNR):** Se trata del rato de ejemplos negativos que han sido clasificados correctamente en la clase negativa.

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

- **GM:** Es la media geométrica entre los dos valores anteriores, TNR y TPR.

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

- **ACC:** Es el porcentaje de acierto que simplemente suma todos los ejemplos bien predichos de todas las clases y los divide entre en número total de ejemplos.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

Otra medida de precisión a la que se va a dar mucha importancia en este proyecto es el **AUC**, que define el área bajo la curva ROC. La curva ROC es una representación gráfica de los verdaderos positivos frente a los falsos positivos. En la figura 15 se observa el espacio ROC, la línea roja representa el azar, el objetivo es siempre obtener puntos por encima de esta línea, o lo que es lo mismo obtener un valor de AUC mayor a 0.5.

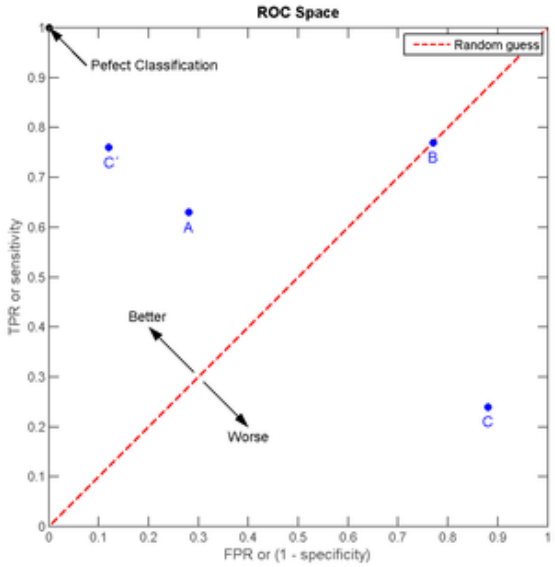

Figura 15- Espacio de la curva ROC

Puesto que siempre se trabaja con datos no balanceados y se tiene muchos más datos de la clase negativa que de la positiva se calcula un umbral en el que la media geométrica es máxima. Este umbral se aplica a las probabilidades que se han obtenido y se decide la clase predicha. El umbral se calcula realizando una predicción en el conjunto de entrenamiento y posteriormente se aplica a las probabilidades predichas en el conjunto de test para decidir la clase obtenida.

6.3 Algoritmo genético

Uno de los experimentos realizados es la ejecución de un algoritmo genético para determinar la mejor configuración total para la predicción de eventos adversos. Se ha implementado un genético muy sencillo para maximizar la media geométrica en la clasificación. Como ya se ha explicado anteriormente la media geométrica se optimiza al mover el umbral de la clasificación. Un ejemplo no se clasifica de una clase u otra si la probabilidad de esa clase es mayor que 0.5 si no que se utiliza dicho umbral para decidir la clase de forma que la media geométrica sea máxima. Esta búsqueda del mejor umbral se realiza con la clasificación en el conjunto de entrenamiento.

Se comienza generando una población de N cromosomas. Para realizar la evaluación de los N cromosomas se separa el conjunto de entrenamiento en dos subconjuntos, extrayendo el 30% de los ejemplos como conjunto de validación. Se separan estos dos conjuntos de forma que los dos cumplan el mismo ratio de ejemplos de la clase positiva y de la clase negativa que el conjunto original de entrenamiento. Se construye el modelo con el conjunto de entrenamiento y se clasifica con el de validación. Esta separación es necesaria para que una vez encontrada la población óptima o acabadas las iteraciones máximas del genético evaluar el genético con el conjunto de test evitando así el sobreaprendizaje.

Una vez evaluados todos los individuos con el conjunto de validación se eligen los M mejores cromosomas con el método del torneo. Estos individuos con mejor fitness se cruzarán dando lugar a cromosomas nuevos que sustituirán los que peor puntación tengan. Tras el cruce los hijos podrán mutar o no según una probabilidad.

Puesto que para calcular el fitness de cada cromosoma se debe llevar a cabo la selección de características y la clasificación de ejemplos es un proceso muy costoso. En este caso la función de fitness que se ha utilizado es la media geométrica ajustada con el umbral adecuado. Como condiciones de parada están las iteraciones máximas y un valor máximo de la media geométrica.

Cromosomas

Los cromosomas en este caso constan de 4 genes individuales, aunque uno de ellos se podría dividir en tres partes dependientes entre sí.

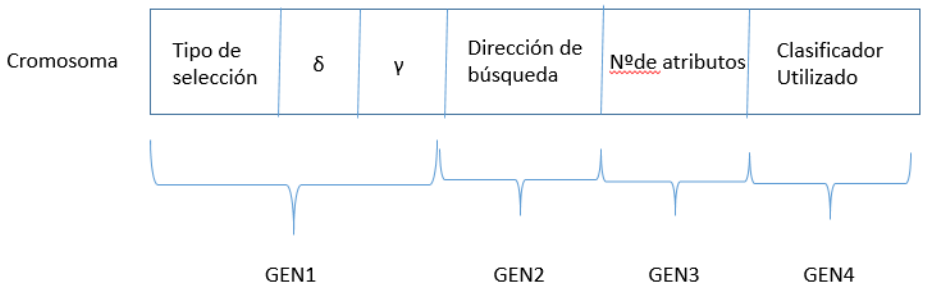

Figura 16- Esquema visual de cómo están formados los cromosomas en el algoritmo genético..

El primer gen compondría el tipo de selección y sus parámetros beta y gamma ya que estos dependen del tipo de selección. El segundo gen es la dirección de búsqueda que en este caso puede ser hacia delante o hacia atrás, el tercero es el número de atributos que queremos seleccionar y por último el cuarto gen que es el algoritmo de clasificación.

Torneo

Para elegir los mejores cromosomas que van a dar lugar a los nuevos individuos se utiliza el algoritmo del torneo. Consiste en la selección de X cromosomas de forma aleatoria y nos quedamos con los dos que mayor fitness tengan. Se define un número de torneo que determina el número de cromosomas que se seleccionan aleatoriamente. De esta forma cuanto mayor sea este número más probabilidad de escoger los mejores individuos para el cruce. En la experimentación se han usado distintos números de torneo.

Cruce

El cruce que se utiliza en este caso es el cruce uniforme. Consiste en generar un vector de tamaño Nºde genes con probabilidades generadas aleatoriamente. De dos progenitores se obtienen dos hijos. El primer hijo coge los genes con probabilidades mayores que 0.5 del padre y los que son menores de 0.5 de la madre. El segundo hijo se obtiene haciendo lo contrario. De esta forma obtenemos dos individuos nuevos.

Mutación

Para tener cierta variabilidad se ha implementado una mutación sencilla aplicada a los hijos obtenidos del cruce. Se calcula una probabilidad y si esta es mayor a una dada (0.99 o 0.95) se muta un gen aleatorio.

7 Experimentación

Se han realiza pruebas con cada uno de los clasificadores y con cada uno de los métodos de selección con las dos direcciones posibles, hacia delante y hacia atrás. Cada una de estas configuraciones se ha repetido para 10, 20, 30, 40 y 50 atributos. En el caso del clasificador KNN se han utilizado 15, 20 y 50 vecinos.

En caso de los métodos de selección que pueden utilizar valores de los parámetros β y γ variables se ha optado a utilizar en ambos casos el valor de 0.5 para los siguientes experimentos. En caso del algoritmo genético estos valores tendrán más variabilidad.

Respecto a los parámetros configurables para ejecutar el genético, se ha utilizado una población de 20 cromosomas, cruzando los 10 mejores individuos elegidos con el método del Torneo utilizando un torneo de 8 cromosomas. La probabilidad de mutación es 0.99. Como condiciones de parada están las 50 iteraciones máximas y un valor máximo de la media geométrica.

En la tabla 8 vienen dados como un resumen recordatorio los métodos que se van a utilizar con los valores de β y γ. Debajo de esta en la tabla 9 vienen los algoritmos de aprendizaje utilizados con los valores de sus parámetros configurables junto con los valores que los definen en Matlab 2017 que es la versión utilizada.

Método	β	γ
CMI	0,5	0,5
CIFE	1	1
JMI	$\frac{1}{\|S\|}$	$\frac{1}{\|S\|}$
MIM	0	0
MIFS	0,5	0
MRMR	$\frac{1}{\|S\|}$	0
CONDRED	0	1

Tabla 8- Tabla resumen para los métodos de selección y los parámetros que los definen.

Clasificador	Párametros	Matlab 2017	Valor
Regresión Logística			
Naïve Bayes	Distribución multinomial	"DistributionNames"	"mn"
KNN	Nº de vecinos	"NumNeighbors"	10,15,20,50
SVM	parámetro C	"BoxContraint"	1

Tabla 9- Tabla resumen para los algoritmos de aprendizaje.

7.1 Resultados eventos cardiovasculares

Se han realizado las pruebas anteriormente descritas sin embargo nos hemos centrado en los clasificadores que mejores resultados nos ha aportado. En la tabla 10 se muestran las medidas de precisión obtenidas en cada clasificador sin realizar algoritmos de selección de características. Para este problema los mejores clasificadores de los considerados son la Regresión Logística seguido por Naïve Bayes.

	Precisiones sin Selección de Atributos				
	MG	TNR	TPR	ACC	AUC
SVM	0,4304	0,4028	0,6636	0,4051	0,5481
KNN10	0,3692	0,9311	0,1474	0,9240	0,5396
KNN15	0,4228	0,9015	0,1994	0,8951	0,5514
KNN20	0,4515	0,8796	0,2326	0,8737	0,5585
KNN50	0,5877	0,7370	0,4697	0,7346	0,6180
BAYES	0,6762	0,6861	0,6691	0,6859	0,7412
REGLOG	0,6957	0,7714	0,6302	0,7701	0,7618

Tabla 10- Medidas de precisión para los distintos algoritmos de aprendizaje sin realizar selección de atributos en el problema de eventos cardiovasculares.

7.1.1 Regresión logística

En el caso de la Regresión logística hemos obtenido un AUC de 0,7618 consideradas todas las características. Tal y como se ha dicho anteriormente se han realizado todas las selecciones en las dos direcciones implementadas. En la tabla 11 se muestran el valor de AUC en cada caso para la dirección hacia delante.

	Forward						
NºAtributos	MIM	MIFS	MRMR	CIFE	CONDRED	CMI	JMI
10	0,7620	0,6654	0,6512	0,7412	0,7520	0,7579	0,7412
20	0,7687	0,6760	0,6624	0,7475	0,7694	0,7604	0,7475
30	0,7711	0,6877	0,6762	0,7488	0,7719	0,7604	0,7488
40	0,7684	0,7063	0,7064	0,7492	0,7692	0,7574	0,7492
50	0,7714	0,7247	0,7244	0,7498	0,7703	0,7571	0,7498

Tabla 11- Tabla que recoge los valores de AUC de los diferentes métodos de selección con la dirección de búsqueda hacia delante utilizando como algoritmo de aprendizaje la Regresión logística.

Se observa que el método CIFE y el JMI dan los mismos valores. Una muy posible causa es el hecho de que en la búsqueda hacia delante se comienza por elegir un atributo y se va aumentando el tamaño del subconjunto. Por ello JMI y CIFE es muy probable que comiencen eligiendo los mismos atributos. Recordamos que δ = 1 y γ = 1 en el método CIFE y δ = $\frac{1}{|S|}$ y γ = $\frac{1}{|S|}$, puesto que comenzamos con $|S| = 1$ parece lógico obtener unos resultados muy parecidos.

En la siguiente tabla se presentan los valores de AUC para la dirección hacia atrás. En este caso cabe destacar que JMI y CIFE no dan los mismos resultados por la misma razón explicada anteriormente. En este caso comenzamos con $|S| = Nº Atributos$ por lo que estos dos métodos comienzan diferenciándose más entre sí.

	Backward						
NºAtributos	MIM	MIFS	MRMR	CIFE	CONDRED	CMI	JMI
10	0,7620	0,5807	0,7220	0,7419	0,7419	0,7419	0,7420
20	0,7687	0,6130	0,7258	0,7670	0,7663	0,7663	0,7668
30	0,7711	0,6629	0,7263	0,7688	0,7697	0,7697	0,7685
40	0,7684	0,6927	0,7266	0,7698	0,7692	0,7692	0,7695
50	0,7714	0,7166	0,7367	0,7702	0,7702	0,7702	0,7702

Tabla 12- Tabla que recoge los valores de AUC de los diferentes métodos de selección con la dirección de búsqueda hacia atrás utilizando como algoritmo de aprendizaje la Regresión logística.

A continuación en las figura 17 y 18 se pueden observar las tendencias de todas las selecciones según la dirección de búsqueda. Observando ambas gráficas de AUC se ve claramente que los métodos que tienen en cuenta solo la redundancia y no la ganancia de información obtienen resultados peores que el resto de métodos.

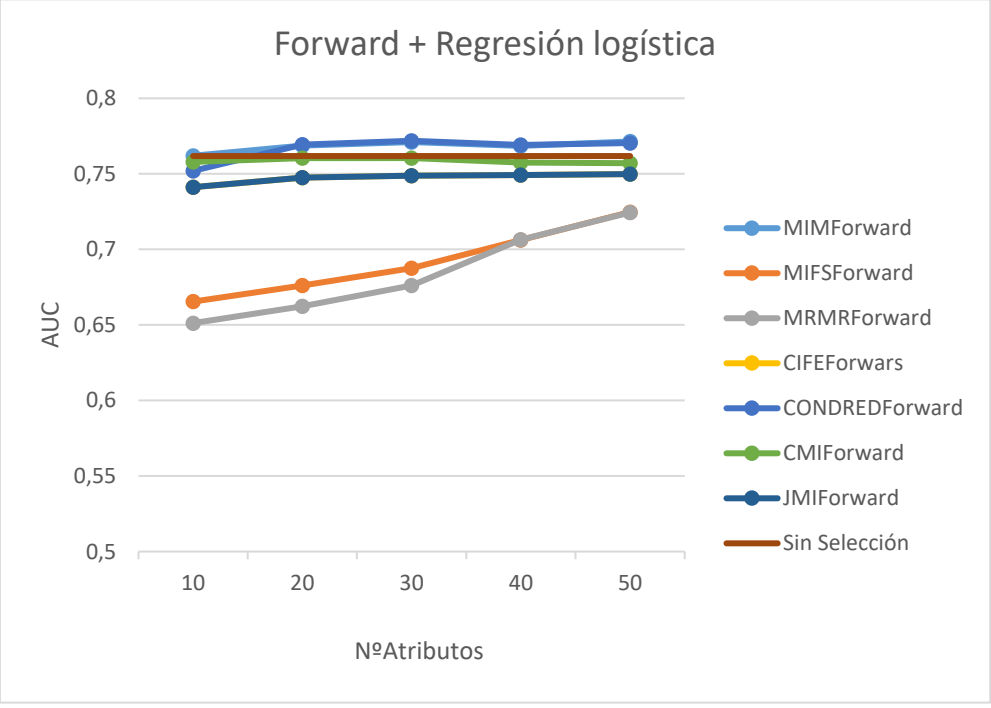

Figura 17- Gráfica comparativa de todos los métodos de selección de características con la dirección de búsqueda hacia delante. Se muestra la evolución del AUC con respecto al número de atributos seleccionados.

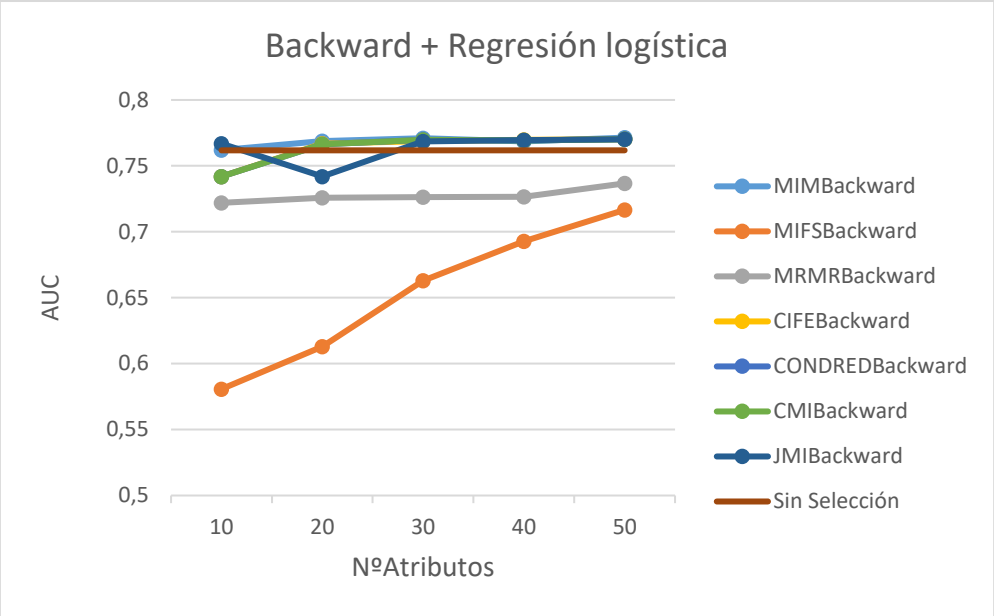

Figura 18- Gráfica comparativa de todos los métodos de selección de características con la dirección de búsqueda hacia atrás. Se muestra la evolución del AUC con respecto al número de atributos seleccionado.

Nos encontramos con unas tendencias parecidas en ambas direcciones. Los métodos MRMR y MIFS que son los que únicamente tienen en cuenta las redundancias entre atributos tienen los peores resultados. Este hecho puede deberse a que los datos con los que se trabaja son ya muy poco redundantes entre sí y al tener en cuenta la redundancia y no la información compartida dada la clase parte de la información compartida se calcule como si fuera redundancia, siendo una redundancia "buena" y necesaria.

Extrayendo los mejores método en cada tamaño del conjunto de atributos obtenemos los datos de la tabla 13 para la búsqueda hacia delante y la tabla 14 para la búsqueda hacia atrás El mejor de los resultados viene dado por CONDRED y la dirección hacia delante con 0,7719 mejorando un 1% con respecto a realizar el aprendizaje y la clasificación con todos los atributos.

Forward		
Método	NºAtr	AUC
MIM	10	0,7620
CONDRED	20	0,7694
CONDRED	30	0,7719
CONDRED	40	0,7692
MIM	50	0,7714
Sin Selección	todos	0,7618

Tabla 13- Mejores resultados de AUC en la dirección de búsqueda hacia delante

Backward		
Método	NºAtr	AUC
MIM	10	0,7620
MIM	20	0,7687
MIM	30	0,7711
CIFE	40	0,7698
MIM	50	0,7714
Sin Selección	todos	0,7618

Tabla 14- Mejores resultados de AUC en la dirección de búsqueda hacia atrás.

Cabe destacar que aunque en ambas tablas aparece el método de selección MIM, este no pertenece a ninguna dirección ya que es un método de tipo ranking, la selección no depende de los atributos ya seleccionados por lo que la dirección no altera el resultado.

Para resumir obtenemos la gráfica siguiente con las tendencias de los 3 métodos que consideramos mejores, CONDRED y CIFE que son los mejores en cada dirección de búsqueda y el ranking MIM que también ha dado unos muy buenos resultados. Se observa que solo en el subconjunto de 10 atributos no se ha podido superar el valor de AUC sin selección lo que indica que 10 atributos no son suficientes para este problema. Además a continuación se presenta la tabla 15 con todas las medidas de precisión obtenidas al realizar estas selecciones.

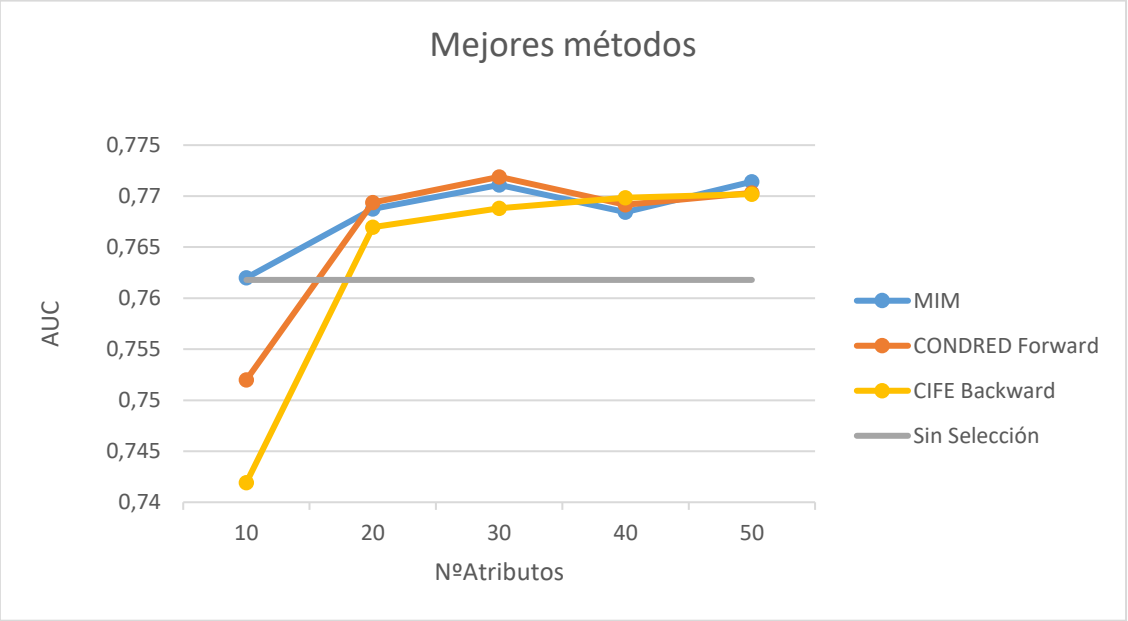

Figura 19- Gráfica comparativa del valor de AUC obtenido para cada tamaño de subconjunto de atributos en los mejores métodos de selección para la Regresión Logística.

En la tabla 15 vienen dadas todas las medidas de precisión de los mejores métodos con el número de atributos que mejor valor de AUC ha obtenido. Para realizar mejor la comparación también vienen los datos sin realizar la selección de atributos.

	MG	TNR	TPR	ACC	AUC
CONDRED 30 Forward	0,7073	0,7313	0,6879	0,7309	0,7719
CIFE 50 Backward	0,6954	0,7190	0,6776	0,7187	0,7702
MIM 50	0,7047	0,7154	0,6951	0,7153	0,7711
Sin Selección	0,6957	0,7714	0,6302	0,7701	0,7618

Tabla 15- Las medidas de precisión de los mejores métodos para la Regresión Logística.

7.1.2 Naïve Bayes

Dados los resultados de las clasificaciones sin realizar selección de características el siguiente clasificador con el que se obtienen unos resultados buenos es Naïve Bayes, por ello se ha realizado también con este clasificador los estudios a fondo, sin embargo, por las características del clasificador los mismos métodos de selección no han podido mejorar apenas el valor de AUC con respecto a la clasificación sin selección de atributos.

El mejor resultado obtenido en las pruebas ha sido con el método MIM y 50 atributos obteniendo un 0,7422 de valor de AUC mejorando solo un 0,1% respecto al clasificador sin selección. En las siguientes dos tablas se ven las mejores selecciones hacia delante y hacia atrás.

Entre los mejores resultados de todos los tamaños de conjuntos de atributos solamente se ha mejorado con 50.

Forward		
Método	NºAtr	AUC
CMI	10	0,6735
CONDRED	20	0,7098
MIM	30	0,7292
MIM	40	0,7384
MIM	50	0,7422
Sin Selección	todos	0,7412

Tabla 16- Tabla con los mejores valores de AUC en la dirección hacia delante obtenidos en cada tamaño del conjunto de atributos

Backward		
Método	NºAtr	AUC
JMI	10	0,6671
MRMR	20	0,7026
MIM	30	0,7292
MIM	40	0,7384
MIM	50	0,7422
Sin Selección	todos	0,7412

Tabla 17- Tabla con los mejores valores de AUC en la dirección hacia atrás obtenidos en cada tamaño del conjunto de atributos

En este caso también se da que las selecciones que tienen en cuenta solo la ganancia de información y la redundancia son los que tienen a obtener peores resultados. En las siguientes graficas viene dado el valor de AUC de cada selección con respecto al número de atributos seleccionados y se observa claramente que con esta clase de métodos de selección y Naïve Bayes no es posible mejorar mucho los resultados.

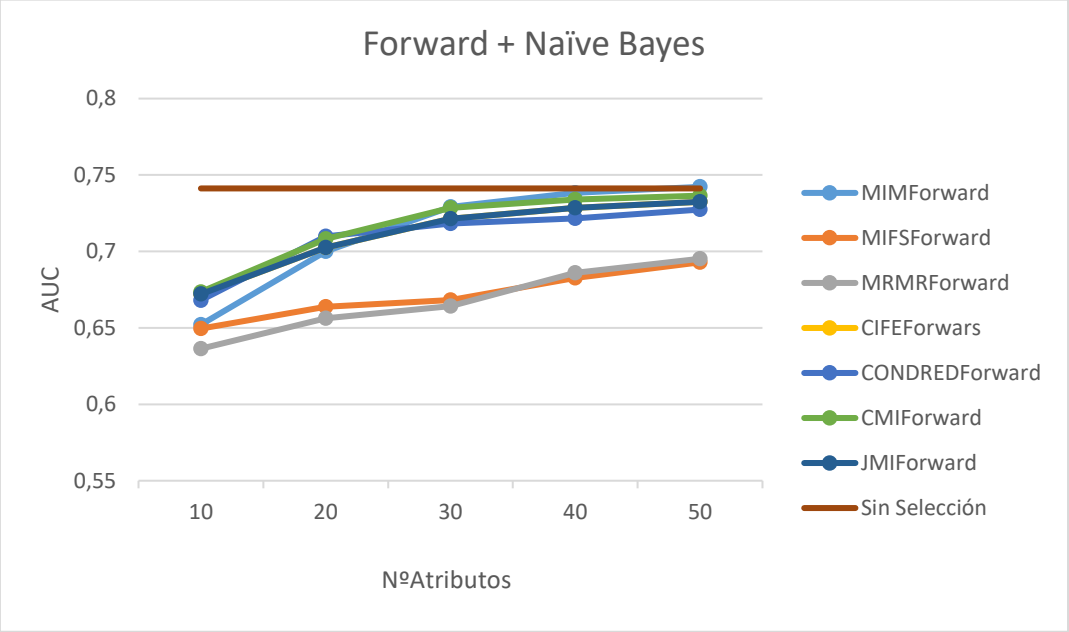

Figura 20- Gráfica comparativa de todos los métodos de selección de características con la dirección de búsqueda hacia delante. Se muestra la evolución del AUC con respecto al número de atributos seleccionado.

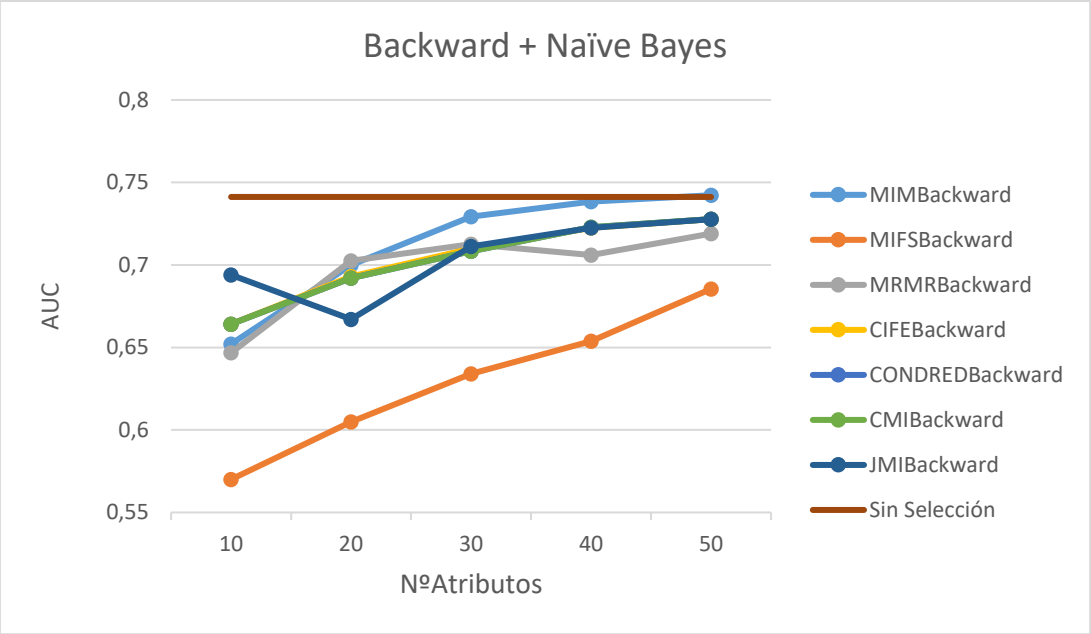

Figura 21- Gráfica comparativa de todos los métodos de selección de características con la dirección de búsqueda hacia atrás. Se muestra la evolución del AUC con respecto al número de atributos seleccionado.

7.1.3 Otros clasificadores

Cómo ya hemos comentado antes se han utilizado dos algoritmos de clasificación más KNN y SVM aun qué no se han obtenido unos resultados muy buenos con las selecciones de características se ha mejorado bastante con respecto a la clasificación sin selección, pero en ningún caso se ha superado a la Regresión Logística ni a Naïve Bayes.

En KNN 50 vecinos hemos obtenido cómo mejor selección CIFE, en la tabla 18 se ven las medidas de precisión de esta selección y de la clasificación con el conjunto original de atributos.

	MG	TNR	TPR	ACC	AUC
CIFE 20 Forward	0,58491514	0,62844456	0,57213012	0,6279358	0,63251393
CIFE 20 Backward	0,58075454	0,65881513	0,5387238	0,65773082	0,62842762
Sin Selección	0,58769621	0,73704057	0,46971119	0,73461514	0,61799557

Tabla 18- Tabla de con las mejores configuraciones de selección con el algoritmo KNN con 50 vecinos.

En SVM prácticamente todos los métodos de selección de características nos han hecho obtener un mejor resultado que el conjunto original. En la siguiente tabla se ven los mejores resultados de las selecciones con la SVM.

	MG	TNR	TPR	ACC	AUC
MRMR 20 Backward	0,47428473	0,56682812	0,60136586	0,56713643	0,62165219
MIM 40	0,56882803	0,65200109	0,50005213	0,65062058	0,60596557
SVM	0,43041029	0,4027848	0,66360129	0,40514495	0,54814992

Tabla 19- Tabla de con las mejores configuraciones de selección con el algoritmo SVM.

7.1.4 Algoritmo Genético

El algoritmo genético se ha probado generando 20 individuos de población y 50 iteraciones máximas. Se han considerado 50 iteraciones puesto que la evaluación de cada individuo supone la selección de características, entrenar un modelo y realizar la clasificación por lo que se hace muy costoso. Sin embargo las poblaciones han llegado a converger a los mejores individuos. Como algoritmos de aprendizaje se han utilizado solo la Regresión Logística y Naïve Bayes ya que se ha considerado innecesario realizar las pruebas con los demás clasificadores por sus bajos resultados.

Como ya se ha comentado anteriormente se dispone de 5 conjuntos de datos con los cuales se trabaja. El algoritmo genético se ha ejecutado para cada dataset calculando los mejores cromosomas para cada uno de ellos.

En la tabla 20 vemos los mejores resultados para cada subconjunto. En la columna configuración se ven el tipo de selección los valores de beta y gamma cuando es necesario, la dirección de búsqueda y el número de atributos considerado. No se ha indicado el clasificador ya que en todos los casos es la Regresión Logística.

	Configuración	MG	TNR	TPR	ACC	AUC
Dataset1	MIM 39-Atributos	0,7267	0,7159	0,7269	0,7050	0,7705
Dataset2	CMI 0,59 0,97 Backward 63-Atributos	0,7552	0,6881	0,7564	0,6259	0,7444
	CIFE Backward 63-Atributos	0,7552	0,6881	0,7564	0,6259	0,7444
Dataset3	MIM 31-Atributos	0,6779	0,7110	0,6773	0,7464	0,7980
Dataset4	CMI 0,81 0,96 Forward 37-Atributos	0,7160	0,7344	0,7157	0,7536	0,8048
Dataset5	CIFE Backward 35-Atributos	0,6994	0,6903	0,6996	0,6812	0,7323

Tabla 20- Tabla de con las mejores configuraciones para realizar la clasificación para cada dataset según el algoritmo genético.

Cada una de estas configuraciones se ha lanzado individualmente para todos los conjuntos para comprobar su funcionamiento de forma general. En la siguiente tabla se ven los resultados de cada configuración obtenida para todos los datasets.

Configuración	MG	TNR	TPR	ACC	AUC
MIM 39-Atributos	0,6961	0,7295	0,6661	0,7289	0,7704
CMI 0,5973 0,9710 Backward 63-Atributos	0,7004	0,7421	0,6619	0,7414	0,7689
CIFE Backward 63-Atributos					
MIM 31-Atributos	0,7109	0,7046	0,7181	0,7047	0,7721
CMI 0,8104 0,9675 Forward 37-Atributos	0,6769	0,7044	0,6518	0,7039	0,7463
CIFE Backward 35-Atributos	0,7063	0,7141	0,6994	0,7140	0,7710

Tabla 21- Valores de las medidas de precisión de las mejores configuraciones obtenidas para cada dataset evaluadas para todos los dataset calculando su media aritmética.

Las filas sombreadas en verde son los mejores resultados obtenidos. Con respecto a las mejores pruebas anteriores realizadas sin utilizar el algoritmo genético la selección MIM con 31 atributos ha superado mínimamente los resultados anteriores.

7.2 Resultados HPE

En el caso de las hospitalizaciones potencialmente evitables los resultados de los que se parte sin selección de características son mucho mejores ya que se trata de un problema más fácil. Esta vez hemos repetido los experimentos con Naïve Bayes, Regresión Logística, SVM y KNN con 50 vecinos pero nos hemos centrado en los mejores clasificadores que en este caso también son la Regresión Logística y Naïve Bayes. En la figura siguiente se observan los valores de AUC sin selección de características.

	MG	TNR	TPR	ACC	AUC
BAYES	0,7720	0,9265	0,6455	0,9254	0,8589
REGLOG	0,7920	0,8877	0,7097	0,8871	0,8551
KNN50	0,7264	0,9126	0,5790	0,9113	0,7587
svm	0,5263	0,6255	0,4454	0,6248	0,5571

Tabla 22 - Tabla de con las medidas de precisión obtenidas sin realizar selección de características.

7.2.1 Regresión Logística

En el problema de HPE utilizando la Regresión Logística se ha visto más mejoras que en el problema de los eventos cardiovasculares. En las siguientes dos tablas vienen dados los resultados de AUC de cada dirección.

	Forward						
NºAtributos	MIM	MIFS	MRMR	CIFE	CONDRED	CMI	JMI
10	0,8793	0,8074	0,8033	0,8546	0,8707	0,8734	0,8546
20	0,8792	0,7928	0,7971	0,8606	0,8776	0,8750	0,8606
30	0,8679	0,8081	0,8005	0,8621	0,8753	0,8734	0,8621
40	0,8584	0,8137	0,8127	0,8559	0,8710	0,8710	0,8559
50	0,8477	0,8264	0,8246	0,8541	0,8680	0,8648	0,8541

Tabla 23- Tabla de con los valores de AUC de todas los métodos de selección con la dirección de búsqueda hacia delante y todos los tamaños de conjuntos de atributos considerados.

	Backward						
NºAtributos	MIM	MIFS	MRMR	CIFE	CONDRED	CMI	JMI
10	0,8793	0,6413	0,8627	0,8694	0,8694	0,8694	0,8671
20	0,8792	0,7006	0,8488	0,8767	0,8767	0,8767	0,8739
30	0,8679	0,7146	0,8384	0,8729	0,8729	0,8729	0,8695
40	0,8584	0,7620	0,8382	0,8705	0,8705	0,8705	0,8705
50	0,8477	0,8295	0,8436	0,8697	0,8697	0,8697	0,8697

Tabla 24- Tabla de con los valores de AUC de todas los métodos de selección con la dirección de búsqueda hacia atrás y todos los tamaños de conjuntos de atributos considerados.

En las siguientes dos gráficas se ven las tendencias de los métodos de selección. De nuevo los métodos que utilizan únicamente la redundancia obtienen los peores resultados.

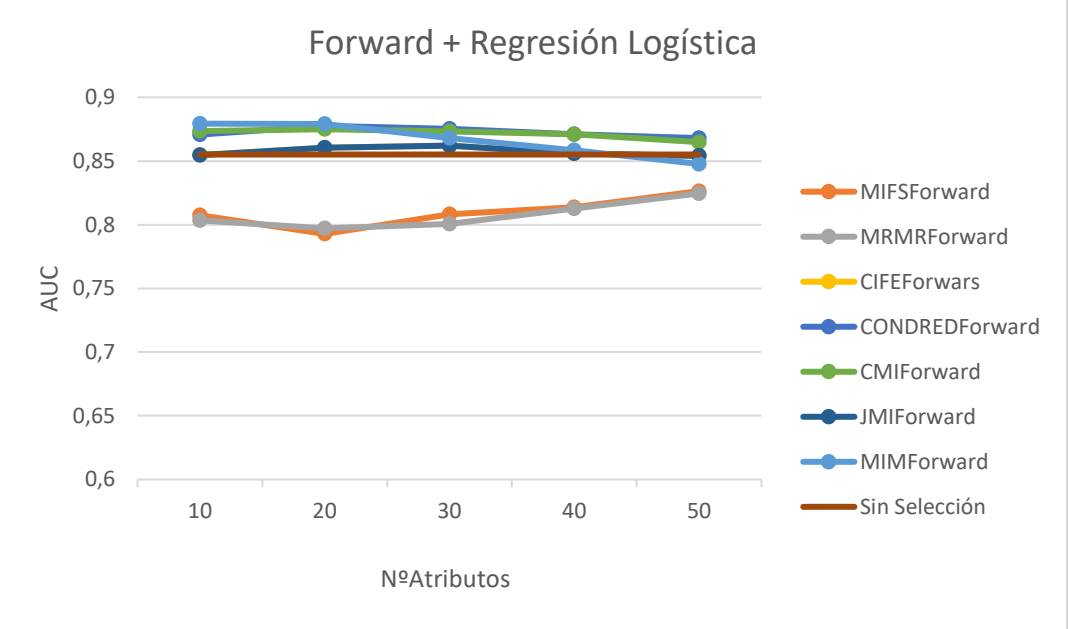

Figura 22- Gráfica que muestra la tendencia del valor de AUC con los métodos de selección y la dirección de búsqueda hacia delante con la Regresión Logística.

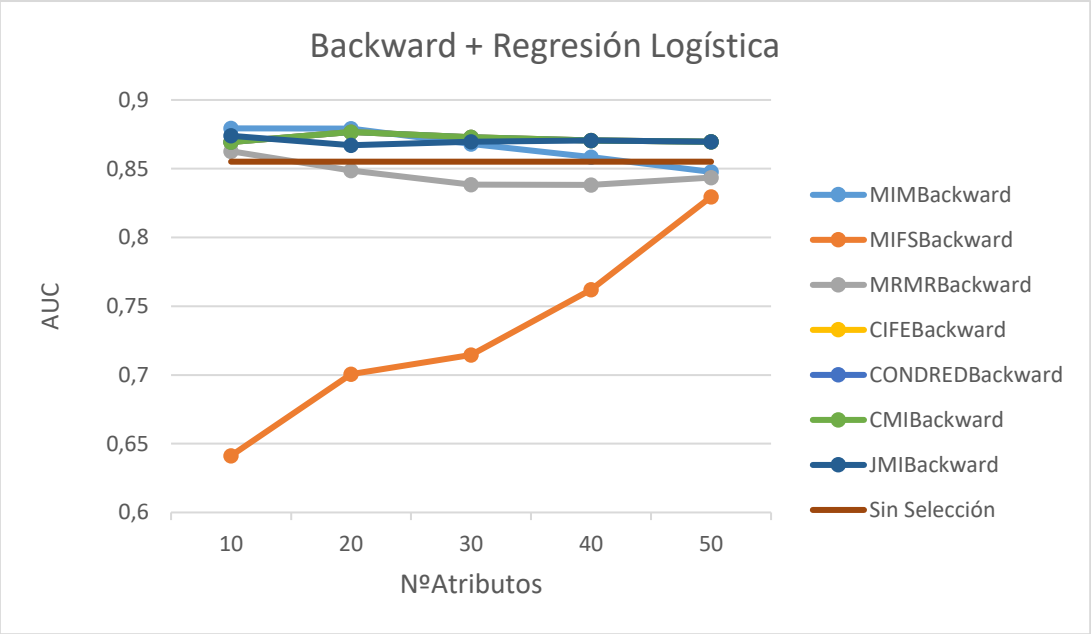

Figura 23- Gráfica que muestra la tendencia del valor de AUC con los métodos de selección y la dirección de búsqueda hacia atrás con la Regresión Logística.

El mejor resultado lo obtenemos con el ranking MIM obteniendo un 0.8793 mejorando más de un 2% en el valor de AUC con respecto a la clasificación sin la selección.En las siguientes dos tablas vienen dados los mejores métodos de selección de características según el número de atributos seleccionado y en ambas direcciones de búsqueda teniendo en cuenta también el método MIM.

Forward		
Método	NºAtr	AUC
MIM	10	0,8793
MIM	20	0,8792
CONDRED	30	0,8753
CMI	40	0,8710
CONDRED	50	0,8680
Sin Selección	todos	0,8551

Tabla 25- Tabla con los mejores valores de AUC en la dirección hacia delante obtenidos en cada tamaño del conjunto de atributos con la Regresión Logística.

Backward		
Método	NºAtr	AUC
MIM	10	0,8793
MIM	20	0,8792
CIFE	30	0,8729
CIFE	40	0,8705
CIFE	50	0,8697
Sin Selección	todos	0,8551

Tabla 26- Tabla con los mejores valores de AUC en la dirección hacia atrás obtenidos en cada tamaño del conjunto de atributos con la Regresión Logística.

En la siguiente gráfica se ven los 3 mejores métodos comparándolos con el valor de AUC sin realizar selección. Se consideran los métodos CONDRED con dirección hacia delante, el método CIFE con dirección hacia atrás y el raking MIM que es el método que mejor resultado ha obtenido en este caso.

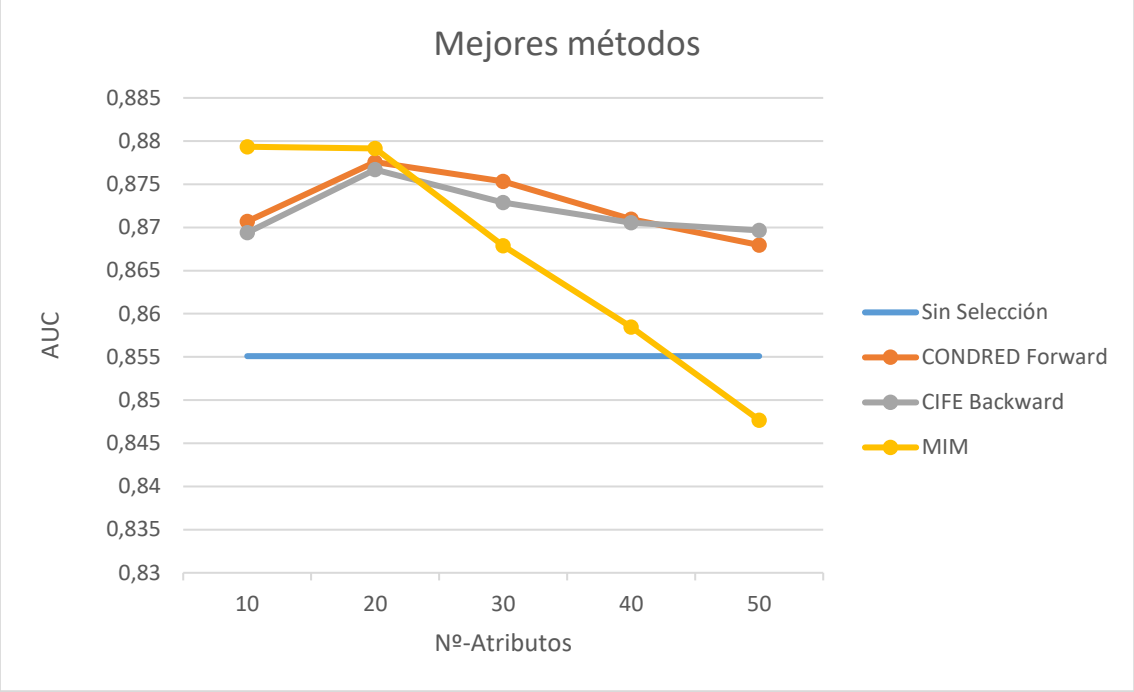

Figura 24- Gráfica comparativa del valor de AUC obtenido para cada tamaño de subconjunto de atributos en los mejores métodos de selección para la Regresión Logística.

7.2.2 Naïve Bayes

Sin realizar una selección de características el clasificador Naïve Bayes obtiene un valor de AUC ligeramente superior al valor de AUC de la Regresión Logística. Sin embargo las mejoras con los diferentes métodos de selección son mínimas. En las siguientes dos tablas vemos los mejores resultados.

Forward		
Método	NºAtr	AUC
CIFE	10	0,8234
CMI	20	0,8469
CMI	30	0,8552
CMI	40	0,8578
CMI	50	0,8602
Sin Selección	todos	0,8589

Tabla 27- Tabla con los mejores valores de AUC en la dirección hacia delante obtenidos en cada tamaño del conjunto de atributos y Naïve Bayes.

Backward		
Método	NºAtr	AUC
MRMR	10	0,8319
MRMR	20	0,8619
MRMR	30	0,8543
MRMR	40	0,8503
MIM	50	0,8560
Sin Selección	todos	0,8589

Tabla 28- Tabla con los mejores valores de AUC en la dirección hacia atrás obtenidos en cada tamaño del conjunto de atributos y Naïve Bayes.

La mejor configuración viene dada por el método MRMR hacia atrás y 20 atributos dando un valor de 0,8619 mejorando un 0,3% el valor de AUC. En las siguientes dos gráficas representan los valores de AUC para cada método de selección. Se observa que con este clasificador no han funcionado los métodos de selección propuestas. Esta tendencia se repite también en el problema de clasificación de eventos cardiovasculares. Podemos concluir que Naïve Bayes es menos sensible a la selección de características que la regresión logística en esta aplicación.

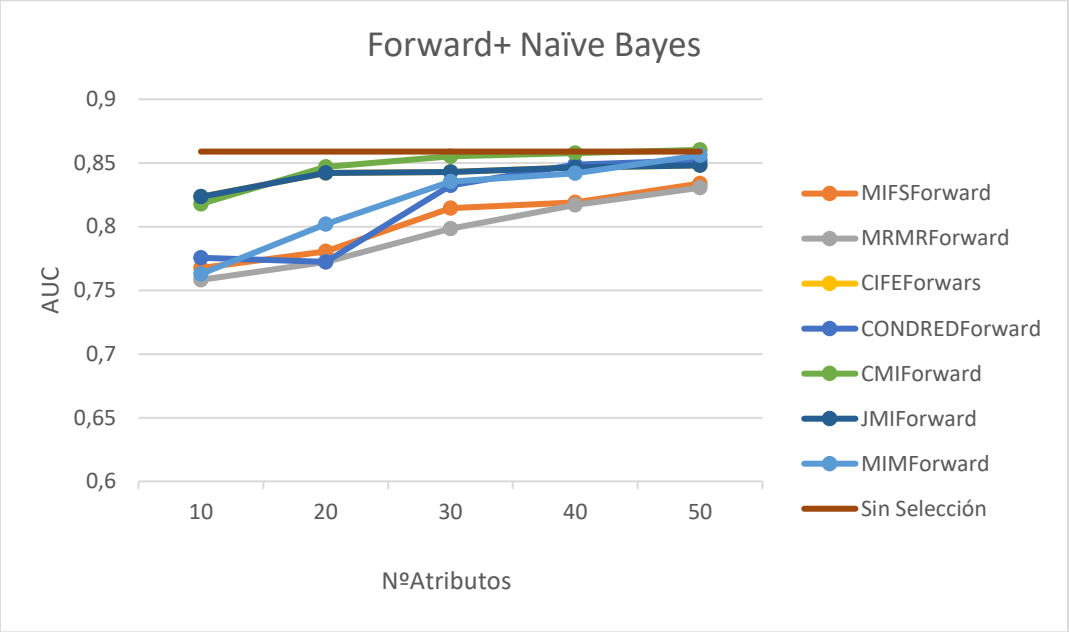

Figura 25- Gráfica comparativa de todos los métodos de selección de características con la dirección de búsqueda hacia delante. Se muestra la evolución del AUC con respecto al número de atributos seleccionado.

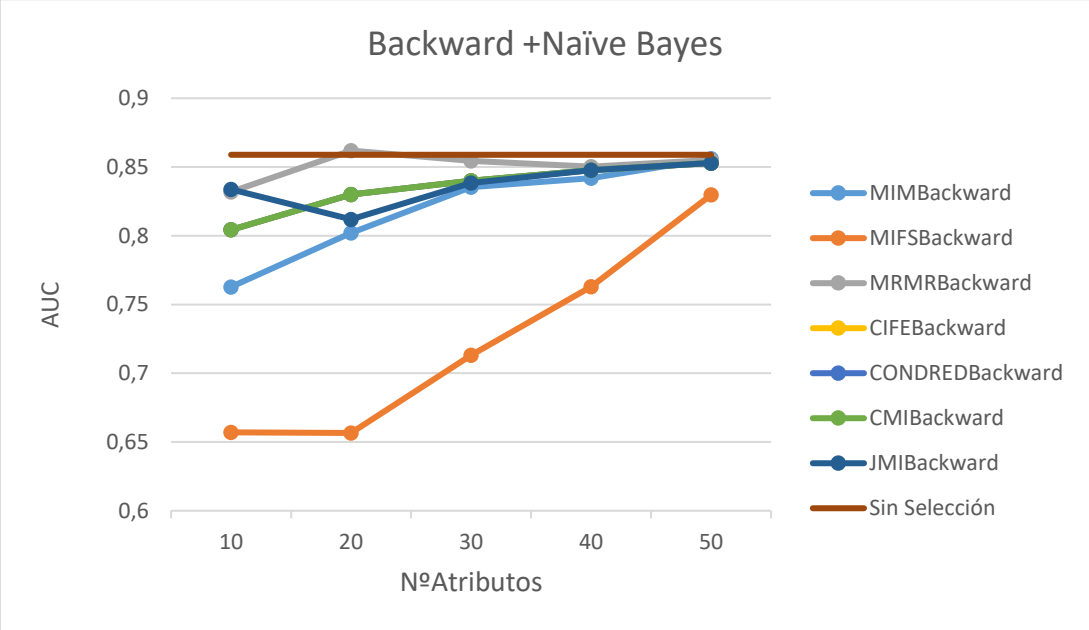

Figura 26- Gráfica comparativa de todos los métodos de selección de características con la dirección de búsqueda hacia atrás. Se muestra la evolución del AUC con respecto al número de atributos seleccionado.

7.2.3 Algoritmo Genético

De la misma forma que para el problema de clasificación para eventos cardiovasculares se ha ejecutado el algoritmo genético con 20 individuos. Utilizando solamente como algoritmos de aprendizaje la Regresión Logística y Naïve Bayes.

Se ha lanzado un genético para cada dataset y se han obtenido los siguientes resultados. En la primera columna se ven las mejores configuraciones para cada conjunto, o lo que es lo mismo los mejores individuos en cada genético. Siendo RL la Regresión Logística y NB Naïve Bayes.

	Configuración	MG	TNR	TPR	ACC	AUC
Dataset1	MRMR Backward 31-Atributos NB	0,8809	0,8084	0,8814	0,7414	0,8797
Dataset2	MRMR Backward 11-Atributos RL	0,8889	0,7930	0,8896	0,7069	0,8652
Dataset3	CIFE Backward 16-Atributos RL	0,8749	0,8312	0,8752	0,7895	0,8967
	CMI 0,74 0,82 Backward 21-Atributos RL	0,8740	0,8308	0,8744	0,7895	0,8869
	CONDRED Backward 21-Atributos RL	0,8740	0,8308	0,8744	0,7895	0,8869
Dataset4	CIFE Backward 30-Atributos RL	0,9524	0,8148	0,9534	0,6964	0,9364
Dataset5	MIM 44-Atributos RL	0,9061	0,7636	0,9071	0,6429	0,8578

Tabla 29- Tabla de con las mejores configuraciones para realizar la clasificación para cada dataset según el algoritmo genético.

En la siguiente tabla se muestran los resultados de cada configuración de la tabla anterior evaluada con todos los conjuntos de datos obteniendo la media de las medidas de precisión.

Configuración	MG	TNR	TPR	ACC	AUC
MRMR Backward 31-Atributos NB	0,7742	0,8880	0,6804	0,8873	0,8554
MRMR Backward 11-Atributos RL	0,8015	0,9041	0,7161	0,9034	0,8617
CIFE Backward 16-Atributos RL	0,7840	0,9169	0,6761	0,9160	0,8817
CMI 0,7437 0,8250 Backward 21-Atributos RL	0,7721	0,9341	0,6386	0,9330	0,8764
CONDRED Backward 21-Atributos RL	0,7721	0,9341	0,6386	0,9330	0,8764
CIFE Backward 30-Atributos RL	0,7812	0,9159	0,6746	0,9149	0,8679
MIM 44-Atributos RL	0,7565	0,9409	0,6105	0,9397	0,8547

Tabla 30- Valores de las medidas de precisión de las mejores configuraciones obtenidas para cada dataset evaluadas para todos los dataset calculando su media aritmética.

7.3 Resumen experimentos

Para resumir la experimentación realizada en las siguientes tablas se presentan los mejores resultados en cada problema de clasificación y los métodos y algoritmos de clasificación para obrenerlos.

	Eventos Cardiovasculares	
	Configuración	AUC
	Sin Selección-**Regresión Logística**	0,7618
Experimentación 1	CONDRED Forward 20-Atributos -**Regresión Logística**	0,7719
Experimentación Genético	MIM 31-Atrbutos - **Regresión Logística**	0,7721

Tabla 31- Tabla resumen de los mejores valores de AUC en el problema de los eventos cardiovasculares.

	HPE	
	Configuración	AUC
	Sin Selección- **Naïve Bayes**	0,8589
Experimentación 1	MIM 20-Artributos - **Regresión Logística**	0,8792
Experimentación Genético	CIFE Backward 16-Atributos - **Regresión Logística**	0,8817

Tabla 32- Tabla resumen de los mejores valores de AUC en el problema de las hospitalizaciones potencialmente evitables.

8 Conclusiones y líneas futuras

Después de la realización de los experimentos y del estudio de los resultados se han llegado a diversas conclusiones. En primer lugar cabe destacar que este proyecto se ha empezado, como ya se ha mencionado anteriormente, a partir de los resultados de otro trabajo. Se partía del conocimiento de que con los conjuntos de datos proporcionados y los problemas a resolver los mejores clasificadores son la Regresión Logística y Naïve Bayes. A pesar de ello se han realizado pruebas con otros clasificadores para observar la influencia que tienen sobre estos los distintos métodos implementados.

Con respecto al problema que trata de clasificar eventos cardiovasculares se observa que es un problema extremadamente difícil. Se ha conseguido mejorar en un 1% el AUC, o área bajo la curva ROC con uno de los métodos de selección y la Regresión Logística. En general los métodos que han dado peores resultados han resultado ser los que utilizan la redundancia entre atributos sin utilizar la información compartida dada la clase y el método qué mejores resultados ha dado es el CONDRED curiosamente el único que no utiliza la redundancia entre atributos y sí la información compartida. Esto puede deberse a que los datos que se han utilizado ya están muy optimizados y apenas son redundantes, de forma que si utilizamos un método que penaliza mucho la redundancia puede que estemos penalizando información compartida entre atributos, que de alguna forma es una redundancia “buena” o necesaria.

Por otro lado el problema de clasificación de las hospitalizaciones potencialmente evitables se ha mejorado un 2% el AUC con respecto a la clasificación sin selección con la Regresión Logística. En general los resultados han sido mejores, pero el problema atacado también es más sencillo. Naïve Bayes a pesar de obtener un resultado ligeramente mejor en el conjunto original, la Regresión tras las selecciones ha obtenido unos resultados mejores.

En ambos casos Naïve Bayes no ha mejorado apenas los resultados junto con las selecciones. Parece que no es un clasificador que sea sensible a la clasificación en este problema. Sin embargo la Regresión ha mejorado en ambos casos notablemente con los subconjuntos seleccionados.

Como líneas futuras en primer lugar se podrían investigar métodos de selección de características basados en la ganancia de información pero que no sean una combinación lineal de esta. Yendo más allá sería interesante considerar técnicas que no sean filtros, estudiar los resultados aplicando wrappers o métodos embebidos. En relación al genético implementado se podría añadir como gen los parámetros configurables de los clasificadores utilizados, para así optimizar al máximo el proceso de clasificación.

Por otro lado, alejándonos un poco del aspecto técnico del proyecto y poniendo el foco de atención en el ámbito de la salud se podrían estudiar otras clases objetivo parecidas, como eventos cerebrovasculares, fármacos que puedan provocar partos prematuros, etc.

9 Bibliografía

Data Mining. Practical Machine Learning Tools and Techniques. Ian H. Witten, Eibe Frank, Mark A. Hall. 2011

Data Preprocessing in Data Mining. Salvador García, Julián Luengo, Francisco Herrera. 2015

[1] ”Elaboración de un sistema inteligente para la predicción de eventos adversos relacionados con la polimedicación en atención primaria” M. X. Uriz, M. Galar. Trabajo fin de Máster. 24 FEBRERO 2017

[2] “Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection” G. Brown, A. Pocock, M. Zhao, M. Luján. Journal of Machine Learning Research. Vol.13, Issue 1. Pages:27-66. ENERO 2012