TRABAJO FINAL DE MÁSTER

TIPOLOGÍA DE CLIENTE SEGÚN EL USO DE LOS SEGUROS DE SALUD

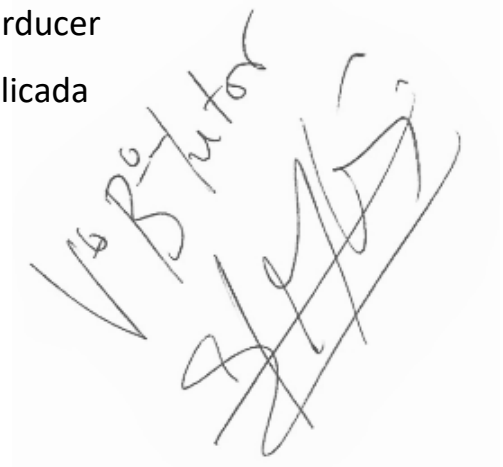
Alumno: Daniel Ferrer Lázaro

Tutor: Ignacio Martínez de Lejarza Esparducer

Departamento del tutor: Economía aplicada

Curso académico: 2018/2019

Fecha de depósito: 16/09/2019



TIPOLOGÍA DE CLIENTE SEGÚN EL USO DE LOS SEGUROS DE SALUD

Alumno:
Ferrer Lázaro, Daniel
Máster en Ciencias Actuariales y Financieras
dafela@alumni.uv.es

Tutor:
Martínez de Lejarza Esparducer, Ignacio
Departamento de Economía Aplicada
ignacio.martinez-lejarza@uv.es

Abstract

Este trabajo se centra en la aplicación de procedimientos de Machine Learning a un conjunto de datos relacionado con los asegurados de un seguro de Salud.

Es por esto por lo que se plantean dos objetivos: segmentación de los asegurados en función del comportamiento con respecto al uso del seguro de Salud y selección de un algoritmo que permita clasificar a estos en función de su rentabilidad. Como metodología se ha escogido la aplicación de técnicas de Aprendizaje Supervisado (J48 y Naïve Bayes) y No Supervisado (CLARA y K-means).

Por un lado, los resultados de la segmentación de asegurados plantean la existencia de 5 grupos de comportamiento en función de la frecuencia de uso del seguro de Salud, aunque la distribución de asegurados en cada método varía significativamente.

Por otro lado, la implementación de las técnicas de Aprendizaje Supervisado señala al método J48 como el más eficaz para la clasificación al cometer menores errores de clasificación y, por tanto, generar un menor coste derivado de este error.

Palabras clave: *Seguros de Salud, Segmentación, Árboles de decisión, J48, Naïve Bayes, K-means, CLARA*

Tabla de contenido

1. Introducción ... 1

1.1. Objetivo, metodología y marco teórico ... 3

2. Análisis actuarial de los seguros de Salud ... 6

2.1. Tipos de Seguro de Salud ... 7

2.2. La prima del seguro de Salud y sus determinantes ... 8

2.3. Coberturas del Seguro de Salud ... 10

2.4. Cuestionario de Salud ... 11

3. Base de Datos ... 12

3.1. Preparación de los datos ... 15

3.2. Análisis exploratorio de la base de datos ... 16

3.3. Análisis de la correlación entre las variables ... 28

4. Métodos de Aprendizaje No Supervisado ... 29

4.1. Variables de interés ... 29

4.2. Algoritmo CLARA ... 31

4.2.1. Determinación del número de clústeres ... 32

4.2.1.1. Implementación en la cartera ... 33

4.2.1.2. Implementación en el Ramo 29 ... 34

4.2.1.3. Implementación en el Ramo 54 ... 35

4.2.1.4. Implementación en el Ramo 82 ... 36

4.2.2. Comparación de los resultados ... 38

4.3. Algoritmo K-means ... 39

4.3.1. Implementación en la cartera ... 39

4.3.2. Implementación en el Ramo 29 ... 40

4.3.3. Implementación en el Ramo 54 ... 41

4.3.4. Implementación en el Ramo 82 ... 42

4.3.5. Comparación de los resultados ... 44

4.4. Algoritmo CLARA vs algoritmo K-means ... 44

5. Métodos de Aprendizaje Supervisado ... 46

5.1. Árbol de decisión J48 ... 47

5.2. Algoritmo clasificador “Naïve Bayes” ... 51

5.3. Comparación de métodos ... 54

6. ANOVA la siniestralidad ... 54

7. Conclusión ... 56

Referencias ... 58

Apéndice 1. Tablas del grado de correlación ... 61

Listado de tablas

Tabla 1. Peso del Sector Asegurador en España en 2017 (datos en millones de euros) 1
Tabla 2. Variables contenidas en la BBDD 12
Tabla 3. Composición de la cartera por Ramos 17
Tabla 4. Distribución de los ramos por sexo del asegurado/a 17
Tabla 5. Estadísticos sobre la Edad de los asegurados 18
Tabla 6. Proporción de asegurados por tramos de edad 19
Tabla 7. Resumen estadístico sobre el índice de siniestralidad (%) 21
Tabla 8. Estadísticos descriptivos para la frecuencia de los servicios médicos estudiados 21
Tabla 9. Frecuencia de la variable "Visitas" 22
Tabla 10. Frecuencia de la variable "Radiodiagnosis" 23
Tabla 11. Frecuencia Análisis Clínicos 23
Tabla 12. Frecuencia de Otros Diagnósticos 24
Tabla 13. Frecuencia de Actos Profesionales 24
Tabla 14. Frecuencia de Anestesia 25
Tabla 15. Frecuencia de Rehabilitación 25
Tabla 16. Frecuencia de Prótesis 26
Tabla 17. Tabla de frecuencias de Otros Conceptos Facturables 26
Tabla 18. Tabla de frecuencias de la variable "Otros" 27
Tabla 19. Estadísticos descriptivos para los costes de los servicios médicos 27
Tabla 20. Estadísticos descriptivos para la duración de la póliza 28
Tabla 21. Resumen de las variables de interés 31
Tabla 22. Silueta media por ramo y diferente 'k' clústeres 32
Tabla 23. Media de las variables de interés por clúster para la cartera (CLARA) 33
Tabla 24. Media de las variables de interés por clúster para el ramo 29 (CLARA) 34
Tabla 25. Media de las variables de interés por clúster para el ramo 54 (CLARA) 36
Tabla 26. Media de las variables de interés por clúster para el ramo 82 (CLARA) 37
Tabla 27. Distribución de asegurados en función de la tipología de uso (CLARA) 38
Tabla 28. Medias de las variables de interés para cada clúster de la cartera (K-means) 39
Tabla 29. Medias de las variables de interés por clúster para el ramo 29 (K-means) 41
Tabla 30. Medias de las variables de interés por clúster para el ramo 54 (K-means) 42
Tabla 31. Medias de las variables de interés por clúster para el ramo 82 (K-means) 43
Tabla 32. Distribución de asegurados por en función de la tipología de uso (K-means) 44
Tabla 33. Tabla de contingencia para los clústeres de los algoritmos CLARA y K-means 45
Tabla 34. Variables de interés para los métodos de Aprendizaje Supervisado 46
Tabla 35. Matriz de confusión del algoritmo J48 para el total de la cartera 48
Tabla 36. Matriz de confusión del algoritmo J48 por ramo 49
Tabla 37. Matriz de confusión del método Naïve Bayes 52
Tabla 38. Comparativa de la precisión entre el método J48 y Naïve Bayes 54
Tabla 39. Análisis del Coste de la clasificación 54
Tabla 40. Prueba ANOVA sobre la siniestralidad 55
Tabla 41. Siniestralidad media por clúster 55
Tabla 42. Resumen de clústeres 57
Tabla 43. Estadístico η2 para la correlación entre variables numéricas y nominales 61
Tabla 44. Estadístico de Cramer para la correlación entre variables nominales 61

Listado de gráficos

Gráfico 1. Respuesta a “Si se presentara actualmente una de estas situaciones, ¿qué elegiría para que le atendiesen la sanidad pública o su seguro privado?” 2
Gráfico 2. Composición de la cartera por ramos 17
Gráfico 3. Distribución por edad de los asegurados 17
Gráfico 4. Distribución de asegurados por provincia de residencia (I) 19
Gráfico 5. Distribución de asegurados por provincia de residencia (II) 20
Gráfico 6. Proporción de asegurados rentables y no rentables por ramo 20
Gráfico 7. Matriz de correlación de las variables cuantitativas 28

Listado de ecuaciones

Ecuación 1. Prima de tarifa mensual por edad de un seguro de Salud 8
Ecuación 2. Prima pura diaria de un seguro de salud 8
Ecuación 3. Fórmula probabilidad condicionada 51

1. Introducción

La actividad aseguradora tiene un papel relevante tanto en la economía como en la estabilidad financiera dada la naturaleza de su actividad. De hecho, las entidades de seguros actúan en el sector financiero como proveedores de servicios básicos en la gestión de riesgos a la vez que realizan el papel de inversores institucionales (González Martínez & Marqués Sevillano, 2013).

En España, el sector asegurador goza de gran relevancia dentro de la economía. Tal es la magnitud que el peso del sector asegurador en relación con el PIB a cierre del año 2017 fue del 5,54% aproximadamente, de acuerdo con (DGSFP, 2018).

Tabla 1. Peso del Sector Asegurador en España en 2017 (datos en millones de euros)

	2014	2015	2016	2017
Primas brutas / PIB (%)	5,40%	5,31%	5,80%	5,54%
PIB a p.m.	1.037.025	1.075.639	1.118.522	1.163.662
Primas devengadas brutas	56.263	56.016	57.073	64.514

Fuente: Elaboración propia a partir de datos del Informe 2017 de Seguros y Fondos de Pensiones de la DGSFP

Como se aprecia en la Tabla 1, se ha experimentado un ligero descenso con respecto a 2016. Este decremento vino causado por un descenso de las primas brutas devengadas y un leve crecimiento del PIB.

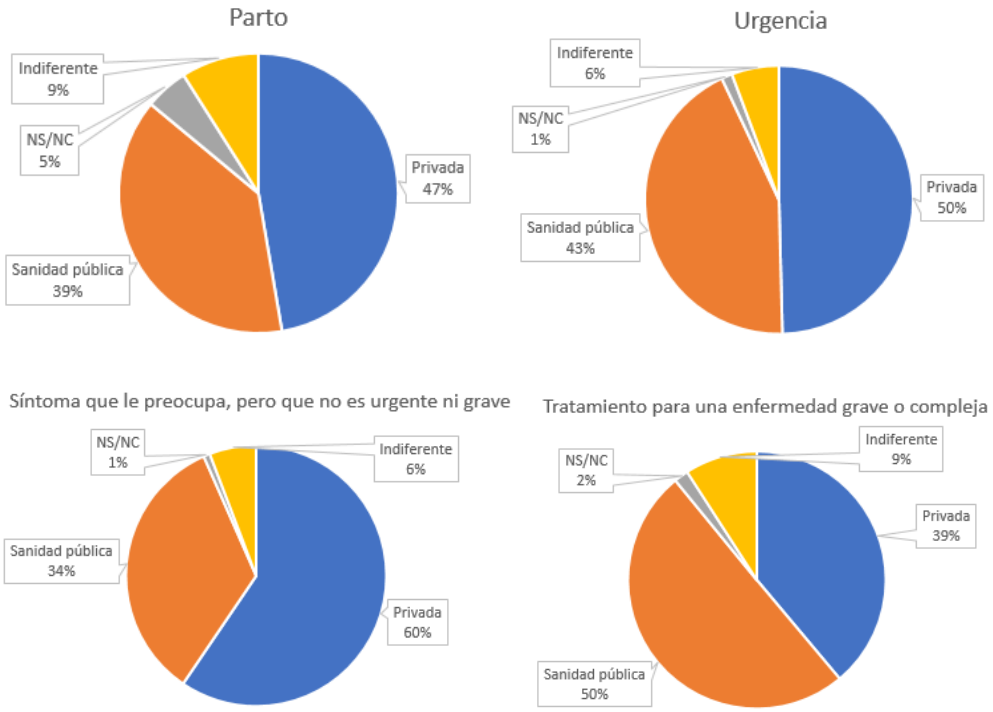
El ámbito del seguro tiene dos grandes manifestaciones en la sociedad: la Seguridad Social, sistema obligatorio de cobertura administrado por el Estado, y los Seguros Privados, pudiendo estos últimos ser de suscripción voluntaria u obligatoria (por ejemplo, seguro de autos).

En el campo de la salud, hoy en día, la masificación de la Asistencia Primaria, los recortes sanitarios, la precariedad de empleo del personal sanitario, la ineficiencia de prestación de servicios y las largas listas de espera han afectado significativamente a la Sanidad Pública (EUROPA PRESS, 2010), favoreciendo que el mercado de los seguros de Salud se haya mantenido en un continuo auge.

De acuerdo con el Barómetro sobre el estado del Sistema Sanitario Español realizado por (CIS, 2019), el 26.2% de los encuestados afirmó que el sistema sanitario público

funcionaba bastante bien; el 47.1% de los encuestados consideró que éste funcionaba bien, aunque era necesarios algunos cambios; y un 26.2% opinó que necesita cambios fundamentales. El porcentaje restante de los encuestados eligió "No sabe/No contesta" (NS/NC) o "Indiferente". Aun así, según este mismo Barómetro del 2018, el Sistema Sanitario Público Español obtuvo de valoración una media de 6.57 puntos, ligeramente inferior a la puntuación obtenida en el Barómetro de 2017, que alcanzó los 6.68 puntos (CIS, 2018), siendo 1 "muy insatisfecho" y 10 "muy satisfecho".

Gráfico 1. Respuesta a "Si se presentara actualmente una de estas situaciones, ¿qué elegiría para que le atendiesen la sanidad pública o su seguro privado?"



Fuente: elaboración propia a partir de datos del Barómetro Sanitario 2019

En este mismo estudio se realizó la siguiente pregunta a los encuestados: "Si se presentara actualmente una de estas situaciones, ¿qué elegiría para que le atendiesen la sanidad pública o su seguro privado?". Contemplando el Gráfico 1, destaca que, para un parto, síntoma no grave o urgencia, la mayor parte de los encuestados elegiría una asistencia mediante seguro privado. No obstante, si se tratase de un tratamiento para una enfermedad grave o compleja, la opción preferida entonces es la Sanidad Pública.

Del informe (Fundación IDIS, 2017) se desprende que, dentro de una escala del 1 al 10, el grado de satisfacción medio por los servicios ofrecidos por la sanidad privada en

España es de 7.6, mientras que los encuestados otorgan a la sanidad pública un 6.1, suponiendo más del 75% de los encuestados consumidores de la sanidad pública y privada simultáneamente.

Esta breve introducción nos sirve para dirigir nuestra mirada hacia el sector asegurador español en materia de seguros de salud, ya que este trabajo se centra en dicho ramo y, más concretamente, en la cartera de asegurados de una entidad aseguradora del mercado español.

1.1. Objetivo, metodología y marco teórico

Dentro del mundo empresarial, algunas entidades aseguradoras, como compañías dentro del sector privado, presentan interés por la búsqueda de la maximización de los beneficios, intentando además minimizar los posibles costes.

Parte de las acciones para la maximización de beneficios y reducción de costes se apoyan en las decisiones tomadas por la Dirección de la entidad y la planificación de las estrategias comerciales. Esto hace que resulte imprescindible conocer y clasificar si los clientes son rentables y saber segmentarlos adecuadamente para estudiar los diversos comportamientos y características de estos (Alcaide, 2015).

Centrando nuestra atención en la clasificación de clientes, de acuerdo con lo mencionado en (Rosales Estrada & Guadarrama Tavira, 2015), estos autores recomiendan clasificar a los clientes según su valor para la empresa y no necesariamente por sus necesidades.

Con respecto a la segmentación de clientes, (Alcaide, 2015) afirma que “la empresa debe intentar aproximar con criterios segmentados y objetivos el interés que, en términos económicos, tienen los clientes de cada segmento para ella” y que “en caso de no hacerlo, la discriminación de clientes daría un resultado «extraño», quizá erróneo”. De igual forma, este autor apoya la idea de que la empresa ha de especializarse en la parte más rentable de los clientes, asumiendo el riesgo de abandono de otros clientes que también son rentables, aunque en menor medida.

Teniendo en cuenta lo anterior, el presente trabajo presenta doble objetivo. Por un lado, busca establecer una segmentación de los asegurados en función de su comportamiento

con respecto al uso de los servicios médicos prestados por el seguro de Salud. Por otro lado, busca seleccionar un algoritmo que permita clasificar a los asegurados de la entidad estudiada en función de su rentabilidad.

De este trabajo cabe destacar el uso de una base de datos correspondiente a la cartera de asegurados de un seguro de Salud, teniendo en cuenta la escasa literatura relacionada con la aplicación del Machine Learning en este campo de los seguros.

Dirigiendo nuestra atención a la metodología, para llevar a cabo nuestros objetivos nos centraremos las técnicas del llamado *Machine Learning* o Aprendizaje automático. A través de la implementación de estos métodos, se intentan descubrir patrones en los datos, a la vez que se crean sistemas para incorporar nuevos patrones tal y como se va introduciendo nuevos volúmenes de información.

De este modo, las técnicas de Machine Learning deben dinamizar los procesos de segmentación y clasificación de los asegurados con la finalidad de ofrecerles una oferta especializada que permita al asegurado percibir un mayor valor a su seguro y que las compañías conozcan mejor a sus clientes. Como bien explica (Recuerdo de los Santos, 2017), existen dos tipos de técnicas en Machine Learning:

- Aprendizaje Supervisado: el algoritmo se entrena con un histórico de datos y así aprende a asignar la etiqueta de salida adecuada las nuevas observaciones. Esto, es, el objetivo final es predecir el valor de salida y, por tanto, se suele emplear en problemas de clasificación. Ejemplo: Identificar grupos de clientes con una alta probabilidad de solicitar la baja del servicio en cuanto finalice su contrato. El objetivo es predecir si se dará de baja.

- Aprendizaje No Supervisado: este aprendizaje tiene lugar cuando no se dispone de datos inicialmente agrupados. El fin es describir la estructura de los datos a través de agrupamientos basados en similitudes. Ejemplo: ¿Se agrupan mis clientes de alguna manera de forma? No existe ninguna variable objetivo para la segmentación.

Por tanto, en este trabajo se acude al Aprendizaje Supervisado con el fin de identificar y clasificar si los clientes de un seguro de salud de la entidad aseguradora estudiada son rentables, mientras que se implementan algoritmos del Aprendizaje No Supervisado con

la finalidad de analizar si los asegurados presentan alguna agrupación innata en cuanto a su comportamiento en el uso de servicios médicos prestados por dicho seguro.

En lo que concierne a la literatura, en la actualidad existe una extensa cantidad de artículos y publicaciones académicas referidas al uso del Machine Learning para la clasificación y segmentación. No obstante, aquella referida a la aplicación del Aprendizaje Automático a los seguros y, más específicamente, a los de Salud es bastante escasa. Es por esto por lo que la literatura que se va a incluir en este apartado no está relacionada directamente con los seguros de Salud.

En primer lugar, nos centramos en la literatura relativa a la clasificación de objetos (Aprendizaje Supervisado). (Patil & Sherekar, 2013) proponen la implementación de los algoritmos Naïves y J48 sobre su conjunto de datos, concluyendo que este último representa una técnica de generación de árboles de decisión más simple y eficiente.

El trabajo de (Imran, Ali, Khan, Ahmad, & Maqsood, 2012) implementa los dos métodos comentados anteriormente y, además, los algoritmos REPTree, Random Tree, y Random Forests. Los resultados muestran que este último clasificador es el que obtiene una mayor precisión en su implementación a los datos propuestos y consigue superar el problema del sobreajuste por la presencia de valores perdidos. En cambio, según (Kalmegh, 2015), de entre los diversos métodos implementados, el algoritmo Random Tree es aquél que consigue mayor precisión a la hora de clasificar las observaciones.

En segundo lugar, dirigimos nuestra atención hacia la literatura centrada en la segmentación de objetos, esto es, Aprendizaje No Supervisado. (Jin & Han, 2017) y (Deepali, Arora, & Varshney, 2015) comparan el método K-Means con diversos métodos K-Medoids (PAM, CLARA, CLARANS, etc.), destacando que estos últimos proporcionan agrupaciones más robustas ante la presencia de *outliers* en comparación con el primer método. No obstante, el resultado contrario se obtiene en (Balabantaray, Sarma, & Jha, 2013), en el que K-Means se presenta como una alternativa más eficiente que K-Medoids, al poder emplear en el primero la distancia Manhattan.

Siguiendo la comparación entre K-Means y K-Medoids, (Velmurugan & Santhaman, 2010) añaden que la ventaja del algoritmo K-Means se encuentra en la rapidez del

tiempo de ejecución, pero que ve reducida su eficiencia ante grandes conjuntos de datos, situación en la que los algoritmos K-Medoids presentan mayor rendimiento.

La investigación de (Steinbach, Karypis, & Kumar, 2000) sobre la comparación del algoritmo K-means, su variación "bisecting" K-means y el método jerárquico concluye que los primeros presentan un mejor comportamiento y rendimiento a la hora de segmentar que el método jerárquico.

Concluyendo, tanto para segmentar como clasificar, no existe un consenso general en la literatura al seleccionar qué algoritmo predomina y proporciona mejor rendimiento, ya que depende en gran cantidad del conjunto de datos sobre los que se trabaja.

Para finalizar, este trabajo presenta la siguiente estructura: en primer lugar, se incluye una introducción y el marco teórico que enmarca este trabajo; en segundo lugar, se realiza un análisis actuarial de los seguros de salud con el objetivo de acercar más al lector a esta línea de negocio; en la tercera sección, se presenta la base de datos empleada , así como un estudio pormenorizado de la cartera y su composición con el fin de conocer mejor con qué datos se están trabajando; las secciones cuarta, quinta y sexta recogen los resultados de la implementación del aprendizaje Supervisado y No Supervisado; y, para finalizar, se presentan las conclusiones.

2. Análisis actuarial de los seguros de Salud

En primer lugar, se debe concretar qué se entiende por un seguro de salud. Por un lado, si nos referimos al término "Seguro", el art. 1 de la Ley 50/1980, de 8 de octubre, de Contrato de Seguro lo define como "aquel por el que el asegurador se obliga, mediante el cobro de una prima y para el caso de que se produzca el evento cuyo riesgo es objeto de cobertura a indemnizar, dentro de los límites pactados, el daño producido al asegurado o satisfacer un capital una renta u otras prestaciones convenidas".

Una vez conocida la definición de seguro, se puede hacer la extensión al seguro de salud: es el contrato mediante el cual la compañía asume, tras el pago de una prima y en los límites establecidos, el compromiso de proporcionar, a través de los servicios concertados, la asistencia médica, quirúrgica y hospitalaria que proceda en los supuestos de enfermedad o lesión.

El riesgo que cubre un seguro de Salud tiene doble naturaleza:

1. Evento futuro e incierto en su realización, como podría ser una enfermedad;
2. Eventos ciertos en su realización como es una visita programada a un profesional médico.

A continuación, se explican diferentes elementos relevantes dentro del ámbito de los seguros de Salud: los tipos, la prima y sus determinantes, coberturas y el Cuestionario de Salud.

2.1. Tipos de Seguro de Salud

Los Seguros de Salud no tienen una única modalidad, sino que se dividen en diferentes categorías. Por lo general, los principales tipos de seguro que podemos encontrar son los siguientes:

- Seguros de cuadro médico: ésta es la modalidad más demandada por los usuarios de los Seguros de Salud. En este tipo de seguros, la entidad aseguradora proporciona al asegurado un listado con los facultativos y centros médicos concertados a los que puede asistir sin coste adicional.

- Seguros de reembolso de gastos: tal y como hace honor su nombre, el usuario es el encargado de hace frente al coste de la prestación de los servicios médicos a los que acuda para que, posteriormente, la entidad aseguradora le reembolse un porcentaje establecido del importe desembolsado inicialmente. Este porcentaje de reembolso suele situarse entre el 80% y el 100% del importe, siempre dentro del límite establecido en la póliza.

- Seguros de Salud mixtos: este tipo de seguros médicos presenta una dualidad en su funcionamiento. Por una parte, el asegurado dispone de un listado de médicos y centros sanitarios donde acudir sin un coste adicional. Por otra parte, el asegurado dispone también de la opción de acudir a un centro o médico fuera del cuadro médico realizando el desembolso inicial y solicitando el reembolso posteriormente.

2.2. La prima del seguro de Salud y sus determinantes

Para realizar este apartado, nos hemos basado en el trabajo de (Pérez Torres, 2011), (Planells Jalón, 2010) y (Peña Sánchez, 2010). Teóricamente, la prima de tarifa mensual (p_x^{cm}) para cada edad o grupo de edad (x) es la siguiente:

Ecuación 1. Prima de tarifa mensual por edad de un seguro de Salud

$$p_x^{cm} = \frac{p_x^{pd} \cdot 365/12}{(1 - \alpha - \beta - \gamma - \delta - \lambda)}$$

Donde p_x^{pd} es la prima pura diaria, α es el porcentaje de gastos de adquisición sobre la prima de tarifa, β es el porcentaje de gastos de administración sobre la prima de tarifa, γ es el porcentaje de otros gastos técnicos sobre la prima de tarifa, δ es el margen de beneficios sobre la prima de tarifa y, por último, λ es el porcentaje de otros recargos sobre la prima de tarifa, tales como tasas o impuestos.

A su vez, la prima pura diaria (p_x^{pd}) se puede expresar como:

Ecuación 2. Prima pura diaria de un seguro de salud

$$p_x^{pd} = f_x \cdot \overline{I_x}$$

Donde f_x es la tasa de morbilidad, que indica el número de personas enfermas en un lugar y tiempo determinado, y el elemento $\overline{I_x}$ denota el coste medio diario.

No obstante, en la práctica, existen diferentes tipos de prima a la hora de tarificar:

- Prima nivelada: se calcula una cuota fija independiente de la edad y que permanece inalterada con el paso del tiempo. Para las edades más jóvenes ésta es más elevada que la prima natural y, para edades más avanzadas, la prima nivelada será inferior.
- Prima natural en función de la edad: se trata de una tarifa variable, ya que cambia con la edad el asegurado.
- Prima nivelada por tramos de edad: es una combinación de las dos anteriores. Trata de calcular una prima nivelada para cada tramo de edad.

Por último, no puede existir diferencias en las primas y prestaciones en los seguros por razones de sexo. Y ello en virtud de la Directiva (UE) 2004/113/CE, de diciembre de 2004, por la que se aplica el principio de igualdad de trato entre hombres y mujeres al acceso a bienes y servicios y su suministro, y el Real Decreto 1361/2007, de 19 de octubre, por el que se modifica el Reglamento de ordenación y supervisión de los seguros privados, aprobado por el Real Decreto 2486/1998, de 20 de noviembre, en materia de supervisión del reaseguro, y de desarrollo de la Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva de mujeres y hombres, en materia de factores actuariales. Como consecuencia, ni sexo ni la consideración de los costes asociados al embarazo y parto podrán ser un criterio en la determinación de la tarifa.

En los Seguros de Salud, existen una serie de condicionantes que pueden determinar el importe de la prima en el momento de la contratación. A continuación, se nombrarán y se detallará brevemente cómo pueden afectar algunos de estos factores.

En primer lugar, cabe destacar uno de los factores más determinantes de la prima: la edad del individuo. En general, cuando mayor es la edad el asegurado en el momento de la contratación el importe de la prima aumenta debido a la asunción de un mayor riesgo por la entidad aseguradora. Además, algunos seguros de Salud establecen un límite de edad para la contratación, normalmente situado entre los 65 y 70 años.

En segundo lugar, se encuentra que la provincia o ciudad de residencia del asegurado (zona geográfica) puede condicionar el precio del seguro de Salud. Esto se debe a que el precio de los proveedores en las diferentes localizaciones geográficas diverge, variando éste en gran medida por la mayor o menor oferta/demanda de proveedores sanitarios.

El precio del seguro también depende de las coberturas contratadas. Los seguros de Salud en su precio base cubren prestaciones de servicios sanitarios básicos. Así, si se desea mayor amplitud de cobertura sanitaria y servicios o contratar garantías adicionales, mayor será también el precio.

En cuarto lugar, encontramos como determinante de la tarifa la introducción del llamado copago. Éste consiste en que el usuario deberá pagar un importe determinado cada vez que haga uso del seguro médico, a parte de la prima abonada. De este modo,

las entidades aseguradoras ofrecen la opción de copago con el objetivo de reducir la prima, cuando el asegurado no tiene previsión de hacer un uso intensivo del seguro.

Otros factores determinantes en el precio de la prima son los recargos o descuentos. Por lo general, las entidades aseguradoras aplican recargos a las pólizas si éstas están compuestas por un único asegurado, haciendo incrementar el precio del seguro. No obstante, conforme aumenta el número de asegurados se aplican descuentos, que reducen dicho importe.

2.3. Coberturas del Seguro de Salud

Las coberturas incluidas en este tipo de seguros varían de una compañía aseguradora a otra, no teniendo por qué coincidir en el paquete de coberturas ofrecidas. Además, las coberturas ofrecidas no son las mismas en un seguro de Salud básico que en uno completo o de reembolso. No obstante, se van a comentar las coberturas frecuentemente incluidas.

Entre las especialidades más demandadas por los usuarios podemos nombrar pediatría, medicina interna, gineco-obstetricia y cirugía general. No obstante, también ofrecen coberturas adicionales y/o opcionales como las siguientes:

- Servicio de ambulancia
- Atención domiciliaria
- Apoyos diagnósticos especializados
- Fisioterapia
- Prótesis
- Ortesis y endoprótesis
- Prótesis
- Cirugía especializada
- Seguro dental preventivo y curativo
- Hematología e inmunología

Por otro lado, las entidades aseguradoras suelen excluir una serie de coberturas del seguro de Salud. Entre las principales encontramos: el parto y servicios médicos de prevención y control cuando no haya transcurrido el periodo de carencia; tratamientos de esterilidad o fertilidad (en ocasiones se incluye en los seguros de Reembolso); tratamientos estéticos, de calvicie u obesidad; estudios y tratamiento para las alteraciones del sueño; tratamientos psicológicos o psiquiátricos (no en vano, cada vez

es más frecuente su inclusión, aunque con muchas limitaciones); cualquier servicio médico fuera del territorio nacional; y las enfermedades preexistentes.

2.4. Cuestionario de Salud

En el momento de la contratación, la entidad aseguradora debe de tener la información adecuada. Es por ello por lo que, como norma general, la entidad aseguradora solicita al interesado la cumplimentación del Cuestionario de Salud.

Este documento, que debe ser cumplimentado por el asegurado y refleja sus datos médicos, tiene la finalidad de ayudar a la entidad a realizar un diagnóstico sobre el posible asegurado a la vez que valora y acota correctamente el riesgo a cubrir. De acuerdo con los apuntes de (Vidal Meliá, 2018), las preguntas contenidas en este documento deben versar sobre las enfermedades, dolencias, lesiones e intervenciones quirúrgicas, y se ha de conseguir que las respuestas del asegurado sean claras y concisas.

Por tanto, el modelo de preguntas debe reunir las siguientes cualidades: debe ser manejable, comprometedor, fiable, completo y económicamente viable. En caso de ocultación, omisión o mentir en algún dato sobre el estado de salud en el cuestionario, se podría excluir al asegurado de ciertas coberturas o bien impugnar el contrato.

En ocasiones, el Cuestionario de Salud puede ir acompañado de un Informe Médico, si así lo requiriera la aseguradora, con el fin de tener un conocimiento más exacto del riesgo. La negativa a someterse al examen médico liberaría al asegurador de cualquier compromiso para la celebración del contrato.

Este documento permite a la empresa determinar las preexistencias y periodos de carencia. El primer término hace referencia a las patologías que el asegurado tiene antes de formalizar el contrato. En cuanto al periodo de carencia, es el lapso de tiempo que debe transcurrir desde el alta de la póliza hasta el momento en el que se permiten emplear ciertas coberturas.

Por último, resulta interesante comentar la protección de datos recolectados relativos al asegurado. En 2018, con la introducción de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, se limita la obtención de datos de manera que se prohíbe a las entidades aseguradoras la

recolección de datos que no tengan como objetivo valorar un riesgo concreto y determinado o que sean incompatibles con dicha finalidad.

3. Base de Datos

Para este trabajo se emplea una Base de Datos, llamada BBDD a partir de ahora, cedida por una entidad aseguradora. Más concretamente, dentro de las diferentes áreas de negocio, esta BBDD proviene del departamento de seguros de Salud.

A la hora de obtener los datos ha sido preciso acotar los ramos de seguros de Salud ofrecidos por esta empresa. Así, con el objetivo de homogeneizar las primas y el tipo de coberturas ofrecidas, este trabajo se ha centrado en los llamados ramos 29, 54 y 82.

Estos ramos pertenecen a un producto que ofrece cobertura completa para los asegurados que lo contratan. La diferencia entre estas modalidades reside en la manera en la que se determina la prima. Mientras que el ramo 54 recoge una prima nivelada, el ramo 29 establece en su nota técnica una prima por tramos no nivelada y, finalmente, el ramo 82 se ingenió con la idea de incluir una prima nivelada con por tramos de edad.

Los datos incluidos en la BBDD van referidos al año natural 2018, esto es, entre 01/01/2018 y 31/12/2018. La BBDD está compuesta por 193,852 asegurados, previa a realizar modificación alguna. No obstante, por motivos explicados más adelante, el número de observaciones se verá disminuido.

Las variables recogidas en esta BBDD pueden dividirse en tres categorías: variables relacionadas con aspectos sociodemográficos, aspectos relacionados con la póliza y aspectos de la siniestralidad del asegurado. En la tabla siguiente se resumen y detallan brevemente las variables incluidas en la BBDD.

Tabla 2. Variables contenidas en la BBDD

VARIABLES RELACIONADAS CON LA PÓLIZA	
Ramo (Ramo)	Indica a qué ramo corresponde el producto contratado.
Número de asegurado (numero)	Código único que identifica a cada asegurado.

Fecha de Alta (falta)	Fecha en la que el asegurado se dio de alta en la póliza.
Fecha de Baja (fbaja)	Fecha en la que el asegurado se dio de baja de la póliza.
Duración (Antigüedad)	Años de antigüedad del asegurado en la póliza.
VARIABLES SOCIODEMOGRÁFICAS	
Fecha de nacimiento (FechaNac)	Variable que recoge el día, mes y año de nacimiento del asegurado.
Edad (Edad)	Edad del asegurado a fecha 31/12/2018.
Sexo (Sexo)	Variable binaria que indica si el asegurado es mujer (M) u hombre (V).
Relación de Parentesco (Parentesco)	Relación del asegurado con el titular de la póliza. A: Abuelo/a — G: Hermano/a B: Nieto/a — H: Hijo/a D: Otros (empleado, tío, primo…) — M: Padre/Madre E: Esposo/a o compañero/a — T: Titular
Provincia (Provincia)	Provincia de residencia del asegurado en cuestión.
Tramo de edad (Tramo)	Variable creada a partir de la variable "Edad" que clasifica al asegurado en función de intervalos de edad: ▪ [0,20] ▪ [46,50] ▪ [56,60] ▪ [71,75] ▪ [21,45] ▪ [50,55] ▪ [61,70] ▪ [76, ∞)
VARIABLES RELACIONADAS CON LA SINIESTRALIDAD	
Número de visitas ambulatorias (Visitas)	Número de visitas ambulatorias que ha realizado el asegurado. Engloba urgencias, primeras visitas, visitas sucesivas, urgencias a domicilios…
Coste de las visitas médicas (CteVisitas)	Importe en euros (€) del coste de las visitas médicas realizadas por el asegurado.

Número de radiografías (Radiodiagnosis)	Número de radiografías que se le han realizado al asegurado. Esta categoría engloba las radiologías simples, radiologías con contrastes, ecografías, resonancias, densitometría y los TAC o escáneres.
Coste de las radiografías (CteRadio)	Importe en euros (€) del coste de las radiografías realizadas al asegurado.
Número de análisis clínicos (AnClinicos)	Número de análisis clínicos que le han realizado el asegurado.
Coste de los análisis clínicos (CteAnClinico)	Importe en euros (€) del coste de los análisis clínicos realizados al asegurado.
Número de otras pruebas diagnósticas (OtrosDiagnosticos)	Número de otras pruebas diagnósticas diferentes de las anteriores que se le han realizado el asegurado. Engloba las pruebas de medicina nuclear, así como sus tratamientos, pruebas de neurofisiología y pruebas de anatomía patológica.
Coste de otras pruebas diagnósticas (CteOtrosDiag)	Importe en euros (€) del coste de otras pruebas diagnósticas realizadas al asegurado.
Número de actos profesionales (ActosProfesionales)	Número de actos profesionales que le han realizado el asegurado. Como actos profesionales se consideran los actos quirúrgicos, partos, cesáreas, comadronas, quimioterapia…
Coste de actos profesionales (CteActos)	Importe en euros (€) del coste de los actos profesionales realizados al asegurado.
Número de anestesias (Anestesia)	Número de anestesias que ha necesitado el asegurado.
Coste de anestesias (CteAnest)	Importe en euros (€) del coste de las anestesias necesitadas por el asegurado.
Número de rehabilitaciones (Rehab)	Sesiones de rehabilitación realizadas por el asegurado.
Coste de las rehabilitaciones (CteRehab)	Importe en euros (€) del coste de las sesiones de rehabilitación realizadas por el asegurado.
Número de prótesis (Prótesis)	Número de actos relacionados con prótesis que ha necesitado el asegurado. Engloba aquellas prótesis relacionadas con traumatología, cirugías vasculares, cirugías cardiacas y otras.

Coste de las prótesis (CteProtesis)	Importe en euros (€) del coste de las prótesis necesitadas por el asegurado.
Número de "Otros conceptos facturables" (OtrosCptosFac)	Número de actos médicos diferentes a los anteriores clasificados como "Otros conceptos facturables" que ha solicitado el asegurado. Incluye actos diversos como los relacionados con quirófanos ambulatorios, preoperatorios, material sanitario y otros tratamientos ambulatorios.
Coste de "Otros conceptos facturables" (CteOtrosCptoFac)	Importe en euros (€) del coste por asegurado de actos médicos clasificados como "Otros conceptos facturables".
Número de "Otros" (Otros)	Número de actos médicos clasificados como "Otros" que ha realizado el asegurado. Incluyen los recobros, capitativos, rappels, prestaciones especiales, asistencia en el extranjero y demás.
Coste de "Otros" (CteOtros)	Importe en euros (€) del coste por asegurado de los actos médicos clasificados como "Otros".
Uso total (Total)	Número total de usos médicos que ha realizado el asegurado.
Coste total (Ctetotal)	Importe total en euros (€) de todos los conceptos médicos expuestos anteriormente.
Siniestralidad% (GTA)	Índice de siniestralidad del asegurado. Esta ha sido creada como el cociente entre los gastos e ingresos del asegurado.
Rentabilidad (Rentable)	Variable dicotómica que toma valor "Si" y "No" si el asegurado es rentable para la entidad aseguradora o no, respectivamente. Es decir, si la Siniestralidad% es menor que 100%, tomará valor "Si". En caso contrario, su valor será "No".

Fuente: Elaboración propia

3.1. Preparación de los datos

La preparación de datos consiste en tratar los datos previamente al análisis. Dentro de esta etapa del trabajo entrarían actividades como, por ejemplo, combinar un conjunto de datos provenientes archivos distintos, seleccionar subconjunto de datos, división de la BBDD en varias partes, transformación de variables y agregación o eliminación de variables, entre otras operaciones.

Para la obtención de la BBDD empleada en este trabajo se han tenido que realizar diferentes operaciones. En primer lugar, para la extracción de los datos se tuvo que acudir al software Cognos de IBM. Concretamente, a IBM Cognos Report Studio. Este programa permite crear informe con estructuras de datos relaciones o dimensionales.

No obstante, este software tiene una forma peculiar de generar las BBDD. Así, en el momento que se introduce una variable en el informe de diferente categoría de clasificación, sólo introduce aquellas observaciones que tengan un valor estrictamente positivos o negativos, eliminado aquellas con valor cero.

Esta incidencia ha obligado a generar tres BBDD diferentes para que no eliminara todas las observaciones (asegurados). La unión de los diferentes archivos de CSV se realizó mediante la función BUSCAR.V de Excel.

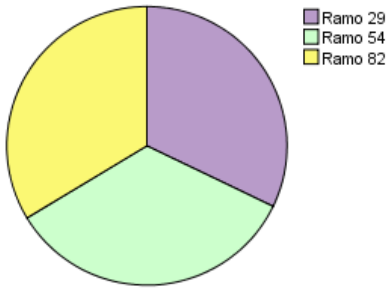
El número inicial de observaciones ronda la cifra de 193,852 asegurados. No obstante, la BBDD presenta valores NA, es decir, ausencia de valores para ciertas variables de interés, y ciertas observaciones sospechosas de error. Una vez comprobado que no existe error en la carga de datos, se procede a eliminar dichas observaciones, resultado la BBDD final empleada en este trabajo en un total de 168,009 asegurados.

3.2. Análisis exploratorio de la base de datos

El Análisis Exploratorio de Datos consiste en examinar los datos antes de implementar cualquier método estadístico u algoritmo con el fin de extraer la máxima información posible acerca de la BBDD con la que se trabaja.

Como se ha comentado más arriba, la BBDD estudiada está compuesta por asegurados que pertenecen a tres modalidades o ramos diversos: los ramos 29, 54 y 82. En primer lugar, vamos a verificar si se trata de una BBDD equilibrada en cuanto a ramos. Para ello acudimos al Gráfico 2. Éste muestra que estamos trabajando con una BBDD en la que aproximadamente el número de asegurados de cada ramo es similar. Concretamente, de acuerdo con la Tabla 3, el 32.15% de los asegurados pertenece al ramo 29, el 34.20% al ramo 54 y, por último, en el ramo 82 encontramos el 33.65% de asegurados.

Gráfico 2. Composición de la cartera por ramos



Fuente: elaboración propia

Tabla 3. Composición de la cartera por Ramos

Ramo	Porcentaje (%)
29	32.15%
54	34.20%
82	33.65%
Total	100.00%

Fuente: elaboración propia

Si nos fijamos en la distribución por sexo de los asegurados, encontramos que el 57.08% de los asegurados son mujeres frente al 42.92% de los cuales corresponden a hombres. Si analizamos esta distribución por ramos (Tabla 4) nos encontramos con que la composición en función del sexo del asegurado/a es similar.

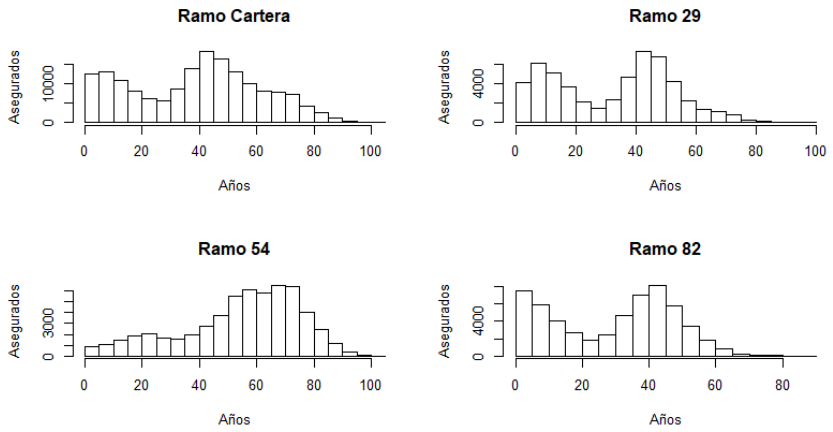
Tabla 4. Distribución de los ramos por sexo del asegurado/a

Ramo	**Mujer (%)**	**Hombre (%)**
29	56.91%	43.09%
54	57.57%	42.43%
82	56.73%	43.27%

Fuente: elaboración propia

La edad de los asegurados también es una cuestión relevante que analizar, ya que una cartera podría estar envejecida o, por el contrario, rejuvenecida. Este hecho podría tener importantes consecuencias para la entidad aseguradora.

Gráfico 3. Distribución por edad de los asegurados



Fuente: elaboración propia

El Gráfico 3 muestra para la cartera en general y para cada uno de los ramos la frecuencia de asegurados en función de la edad. En el primer gráfico en la esquina superior izquierda, se observa como la cartera tiene el grueso de asegurados repartidos entre los de mediana edad (35-55 años) e infantes y adolescentes (0-15 años).

Si el análisis se realiza para cada uno de los ramos, los resultados muestran, por un lado, como la cartera para el ramo 54 es la más envejecida, teniendo la mayoría de los asegurados más de 55 años. Por otro lado, el ramo 29 y el ramo 82 son los que más asegurados jóvenes tienen y, por tanto, más rejuvenecido está. De estos dos últimos, destaca cómo a partir de los 60 años disminuye drásticamente el número de asegurados.

Para poder analizar mejor la distribución de edad, analizaremos algunos estadísticos presentes en la Tabla 5. Se observa con claridad como la edad media es más elevada en el ramo 54 alcanzando una edad media de 55.15 años, frente a los 33 y 29.85 años de los ramos 29 y 82, respectivamente. No obstante, cuando se analiza la cartera en su conjunto, la edad media se sitúa en los 39.51 años.

Estudiando los cuartiles, destaca que el 25% de los asegurados del ramo 54 tienen 44 años o menos, en comparación con los 13 y 11 años, de los ramos 29 y 82, respectivamente. Estos estadísticos ya nos indican cierto envejecimiento del ramo 54, que se acaba de confirmar cuando se mira el tercer cuartil: el 75% de los asegurados del ramo 54 tiene 71 años o menos. No obstante, si se atiende a la cartera en su conjunto, la edad media es ligeramente inferior a los 40 años, por lo que no se puede hablar de cartera envejecida.

Tabla 5. Estadísticos sobre la Edad de los asegurados

Ramo	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
29	0	13	38	33	48	99
54	0	44	59	55.15	71	103
82	0	11	35	29.85	44	87
Cartera	0	19	42	39.51	55	103

Fuente: elaboración propia

Por último, se va a acudir al análisis por tramos de edad, cuya información se encuentra en la Tabla 6. Esto acaba de confirmarnos que el ramo 29 y 82 están rejuvenecidos, pero

el 54 presenta un problema de elevada edad de sus asegurados. Aun así, la cartera en su conjunto presenta niveles adecuados de edad, predominando asegurados jóvenes.

Tabla 6. Proporción de asegurados por tramos de edad

Ramo	[0,20]	[21,45]	[46,50]	[51,55]	[56,60]	[61,65]	[66,70]	[71,75]	[76,)
29	35.39%	33.44%	12.57%	7.79%	4.08%	2.53%	2.02%	1.43%	0.75%
54	9.10%	17.32%	6.58%	9.65%	10.68%	10.12%	11.37%	11.19%	13.99%
82	35.79%	42.79%	10.15%	6.08%	3.15%	1.47%	0.43%	0.11%	0.03%
Cartera	26.53%	31.08%	9.71%	7.85%	6.02%	4.77%	4.68%	4.32%	5.04%

Fuente: elaboración propia

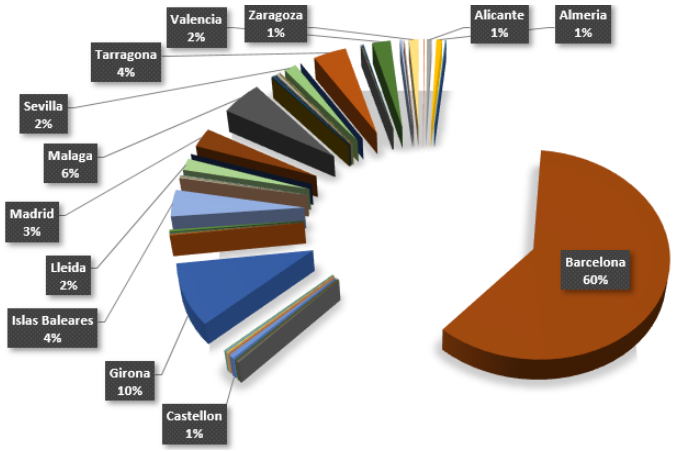
Otro punto de interés sería analizar las provincias donde esta entidad aseguradora tiene presencia. Esta información viene recogida en el Gráfico 4, el cual recoge un mapa coroplético de España por provincias. Como se puede observar, el grueso de asegurados de esta entidad aseguradora se encuentra en la CCAA de Cataluña, donde el 76% aproximadamente de los asegurados tienen su residencia principal.

Gráfico 4. Distribución de asegurados por provincia de residencia (I)

Fuente: elaboración propia

El siguiente gráfico recoge el porcentaje de asegurados que residen en cada en cada provincia. Así, se verifica que la mayoría, el 60%, de los asegurados tienen su residencia en Barcelona. La siguiente provincia española con mayor número de asegurados es Girona con un 10% del total, seguido de Málaga (6%) e Islas Baleares junto con Tarragona, empatadas con un 4% de asegurados cada una. Por otro lado, encontramos que en Madrid sólo residen el 3% de los asegurados.

Gráfico 5. Distribución de asegurados por provincia de residencia (II)

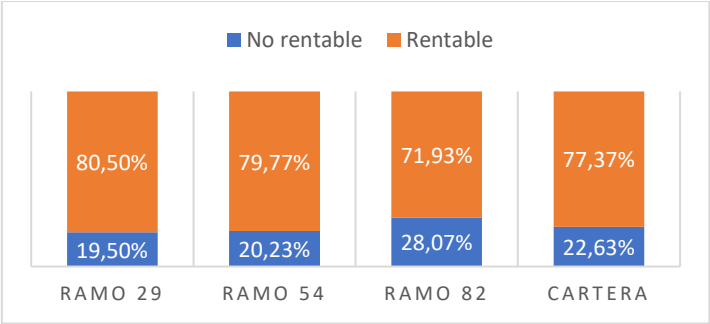


Fuente: elaboración propia

Una vez analizadas las características socioeconómicas de los asegurados, se procede a analizar las variables relacionadas con la rentabilidad de las pólizas, es decir, costes, frecuencia de uso médico, primas e índices de rentabilidad.

Para empezar con el análisis de las variables relacionadas con la rentabilidad, cabe profundizar sobre cuantos asegurados estudiados son rentables para la entidad aseguradora. Para ello acudimos a dos medidas: a la variable rentabilidad (Gráfico 6) y la Siniestralidad (Tabla 7).

Gráfico 6. Proporción de asegurados rentables y no rentables por ramo



Fuente: elaboración propia

En cuanto al Gráfico 6, éste muestra la proporción de asegurados rentables y no rentables para cada ramo y la cartera. Se observa claramente como en todos los casos la proporción de asegurados rentables es muy superior al de no rentables. Destaca el ramo 82, el cual tiene la mayor cantidad de asegurados no rentables (28.97%). Esto es contrario a lo esperado, ya que la siniestralidad se concentra más en el ramo más

rejuvenecido (ramo 82), que en el envejecido (ramo 54). Esto puede deberse a la mayor prima nivelada. El ramo 29 es el que tiene menor siniestralidad media (19.50%).

Tabla 7. Resumen estadístico sobre el índice de siniestralidad (%)

Ramo	1º Cuartil	Mediana	Media	3º cuartil
29	18.54	40.57	81.49	81.74
54	17.61	39.57	85.36	83.07
82	24.17	53.79	110.33	110.86
Cartera	19.80	44.10	92.52	91.92

Fuente: elaboración propia

Los estadísticos situados en la Tabla 7 permiten estudiar en mayor profundidad la siniestralidad de los asegurados de la cartera. En primer lugar, si comparamos la siniestralidad media, destaca el hecho de que los ramos 29 y 54 tengan en promedio una siniestralidad alta pero inferior al 100%, hecho que implica que las primas son superiores a los gastos incurridos por el asegurado. No obstante, la siniestralidad media de los asegurados del ramo 82 se sitúa en 110.33%, posibles síntomas de problemas de insuficiencia de prima. En general, la cartera tiene de media un 92.52% de siniestralidad.

Al principio del trabajo, se ha desarrollado la composición de la variable siniestralidad como cociente entre los coste incurridos por el asegurado y la prima de éste. Concentrándonos en el coste, la BBDD incluye información desglosada acerca del tipo de servicio médico y el coste incurrido. A continuación, se muestran estadísticos para cada categoría médica, como apoyo para el análisis de las tablas de frecuencia.

Tabla 8. Estadísticos descriptivos para la frecuencia de los servicios médicos estudiados

	Mínimo	Máximo	Media	Desviación estándar
Visitas	0	77	8.63	8.632
Radiodiagnosis	0	56	2.50	3.587
Análisis Clínicos	0	30	1.39	2.010
Otros Diagnósticos	0	24	0.47	1.110
Actos Profesionales	0	60	2.30	3.906
Anestesia	0	13	0.14	0.550
Rehabilitación	0	154	2.22	7.615
Prótesis	0	5	0.02	0.136
Otros Cptos Facturables	0	40	0.04	0.382
Otros	0	44	0.21	0.816

Fuente: elaboración propia

En primer lugar, tenemos las visitas que se podrían enmarcar dentro de actos ambulatorios. Los estadísticos nos informan de que un asegurado promedio hace 8.63 visitas ambulatorias a lo largo del año, mientras que el número máximo de visitas de este carácter ha sido de 77. En la Tabla 9, se muestra la frecuencia con la que los asegurados han realizado el uso del seguro privado para visitas médicas de carácter ambulatorio. Encontramos que lo más frecuente es que los aseguradas hagan 1 o 2 visitas, con un 10.2% y 10.6%, respectivamente. Los resultados pueden llevar a afirmar que es bastante frecuente para los asegurados de la entidad aseguradora estudiada realizar visitas de este tipo.

Tabla 9. Frecuencia de la variable "Visitas"

	Frecuencia	Porcentaje	Porcentaje acumulado	(cont.)	Frecuencia	Porcentaje	Porcentaje acumulado
0	3125	1.9	1.9	10	6350	3.8	72.0
1	17139	10.2	12.1	11	5296	3.2	75.1
2	17767	10.6	22.6	12	4849	2.9	78.0
3	14802	8.8	31.4	13	4186	2.5	80.5
4	14260	8.5	39.9	14	3613	2.2	82.7
5	12145	7.2	47.2	15	3157	1.9	84.5
6	10812	6.4	53.6	16	2928	1.7	86.3
7	9150	5.4	59.0	17	2403	1.4	87.7
8	8454	5.0	64.1	18	2246	1.3	89.0
9	6916	4.1	68.2	19 ≤	18411	11.0	100.0

Fuente: elaboración propia

Si fijamos nuestra atención al coste, las visitas médicas ambulatorias suponen de media un coste de 184.48€ por asegurado a la entidad aseguradora, siendo el coste máximo alcanzado por un asegurado de 6,166.66€

En cuanto a las radiodiagnosis, la media de actos de este carácter ronda las 2.5 visitas. Si centramos nuestra atención en la tabla de frecuencias situada abajo, la información muestra que el 40% de los asegurados no ha necesitado ninguna visita de esta categoría médica, el 13.5% ha realizado 2 pruebas diagnósticas, el 11.9% ha necesitado 3 visitas y que el porcentaje de asegurados va disminuyendo conforme aumenta la frecuencia, siendo el porcentaje marginal cuando se supera la décimo visita.

Tabla 10. Frecuencia de la variable "Radiodiagnosis"

	Frecuencia	Porcentaje	Porcentaje acumulado	(cont.)	Frecuencia	Porcentaje	Porcentaje acumulado
0	67330	40.1	40.1	**6**	6023	3.6	89.0
1	22692	13.5	53.6	**7**	4292	2.6	91.5
2	19959	11.9	65.5	**8**	3296	2.0	93.5
3	14231	8.5	73.9	**9**	2483	1.5	95.0
4	11347	6.8	80.7	**10≤**	8420	5.0	100.0
5	7936	4.7	85.4				

Fuente: elaboración propia

En este caso, en promedio, este servicio médico genera un coste de 88.90€/asegurado a la entidad aseguradora. No obstante, existen asegurados que han alcanzado la cifra de 2,701.53€ en esta categoría. El bajo importe de la media puede deberse al exceso de ceros que presenta esta variable.

En tercer lugar, encontramos la frecuencia con la que se han realizado Análisis Clínicos. La Tabla 11 indica que casi el 50% de la muestra no necesitó ningún análisis clínico. Por el contrario, vemos que el 15.9% y 11.76% de los asegurados realizó 1 y 2 análisis, respectivamente. Destaca el hecho de que el 8,420 asegurados necesitaron 10 o más análisis, mientras que el 30% de los clientes realizaron entre 3 y 9 actos de esta categoría. No obstante, aunque la media ronda un análisis por asegurado, en la muestra de la que disponemos se encuentran asegurados que han realizado hasta 30 análisis clínicos. Este servicio médico supone, en promedio, un coste de 43.81€ por asegurado.

Tabla 11. Frecuencia Análisis Clínicos

	Frecuencia	Porcentaje	Porcentaje acumulado	(cont.)	Frecuencia	Porcentaje	Porcentaje acumulado
0	82540	49.1	49.1	**4**	11906	7.1	92.2
1	26697	15.9	65.0	**5**	10778	6.4	98.6
2	19669	11.7	76.7	**6≤**	2276	1.4	100.0
3	14143	8.4	85.1				

Fuente: elaboración propia

Otro tipo de acto médico del que se dispone información es el clasificado como otras pruebas diagnósticas. Aproximadamente el 7 de cada 10 asegurados no necesitó realizar este tipo de acto. De hecho, en promedio, el número de visitas de esta categoría es de

0.47. Lo siguiente más frecuente es la realización de una visita para otras pruebas diagnósticas, suponiendo 33,926 asegurados o, lo que es lo mismo, un 20.2% del total.

Tabla 12. Frecuencia de Otros Diagnósticos

	Frecuencia	Porcentaje	Porcentaje acumulado
0	120384	71.7	71.7
1	33926	20.2	91.8
2	7124	4.2	96.1
3	2627	1.6	97.7
4≤	3948	2.3	100.0

Fuente: elaboración propia

El coste medio por asegurado asociado a "Otros diagnósticos" es de 13.73€. Esta cifra tan baja es consecuencia de la gran cantidad de asegurados que no han necesitado estas pruebas médicas y que, consecuentemente, están distorsionando la media.

En lo que a los Actos Profesionales se refiere, un asegurado realiza 2.30 visitas de esta categoría de promedio. En cuanto a la frecuencia, por orden de relevancia destaca los que no han necesitado ningún acto profesional, los que han realizado una visita y los que han realizado 2, respectivamente. Tal y como es de esperar, conforme va aumentado el número de visitas, la frecuencia disminuye.

Tabla 13. Frecuencia de Actos Profesionales

	Frecuencia	Porcentaje	Porcentaje acumulado	(cont.)	Frecuencia	Porcentaje	Porcentaje acumulado
0	72154	42.9	42.9	**6**	4935	2.9	90.2
1	28712	17.1	60.0	**7**	3552	2.1	92.3
2	19015	11.3	71.4	**8**	2722	1.6	93.9
3	11618	6.9	78.3	**9**	1969	1.2	95.1
4	9011	5.4	83.6	**10≤**	8278	4.9	100.0
5	6043	3.6	87.2				

Fuente: elaboración propia

En relación con el coste asociado, el gasto máximo es de 17,953.98€. Sin embargo, el coste medio de este tipo de servicios médicos se sitúa en 111.30€ por asegurado.

Otro tipo de acto a analizar es el categorizado como "Anestesia". De acuerdo con la Tabla 8, un asegurado medio ha necesitado 0.14 anestesias. Este dato se puede explicar por

el hecho de que el 91.5% de los asegurados no ha necesitado ninguna anestesia, llegando suponer 153,800 asegurados de los 168,009 estudiados.

Tabla 14. Frecuencia de Anestesia

	Frecuencia	Porcentaje	Porcentaje acumulado
0	153800	91.5	91.5
1	9008	5.4	96.9
2	2286	1.4	98.3
3	2336	1.4	99.7
4≤	579	0.3	100.0

Fuente: elaboración propia

Analizando las sesiones de rehabilitación, el grupo más numeroso es la ausencia de este tipo de acto médico, llegando suponer 146,509 asegurados de los 168,009 estudiados. Es más, resalta los que han realizado 10 sesiones, que suponen un 3.2% del total.

Tabla 15. Frecuencia de Rehabilitación

	Frecuencia	Porcentaje	Porcentaje acumulado	(cont.)	Frecuencia	Porcentaje	Porcentaje acumulado	(cont.)	Frecuencia	Porcentaje	Porcentaje acumulado
0	146509	87.2	87.2	**7**	488	0.3	89.2	**14**	333	0.2	94.2
1	461	0.3	87.5	**8**	693	0.4	89.6	**15**	1005	0.6	94.8
2	437	0.3	87.7	**9**	1094	0.7	90.3	**16**	267	0.2	95.0
3	494	0.3	88.0	**10**	5396	3.2	93.5	**17**	245	0.1	95.1
4	449	0.3	88.3	**11**	289	0.2	93.7	**18**	358	0.2	95.3
5	517	0.3	88.6	**12**	313	0.2	93.8	**19**	518	0.3	95.6
6	517	0.3	88.9	**13**	278	0.2	94.0	**20≤**	2084	1.2	96.9

Fuente: elaboración propia

De acuerdo con los resultados obtenidos, las sesiones de rehabilitación suponen, en promedio, un coste de 18.86€ por cada asegurado. Debido a que casi el 90% de los asegurados no han necesitado servicios médicos, esta media está distorsionada.

En cuanto a las prótesis, un usuario medio realiza 0,02 visitas, siendo 5 el número máximo de visitas alcanzado por los asegurados. De manera marginal, encontramos que el 1.3% de la muestra, esto es, 2201 asegurados han precisado una visita de este tipo de servicios, mientras que sólo 175 usuarios han precisado de dos visitas

Tabla 16. Frecuencia de Prótesis

	Frecuencia	Porcentaje	Porcentaje acumulado
0	165610	98.6	98.6
1	2201	1.3	99.9
2	175	0.1	100.0
3	19	0.0	100.0
4 ≤	4	.0	100.0

Fuente: elaboración propia

La categoría "Prótesis" es la segunda con el mayor coste máximo, que ronda los 14,500€. En lo que se refiere al coste medio por cada cliente, éste es de 21.05€, distorsionado por el hecho de que sólo 2,399 clientes han generado coste en esta categoría médica.

Dejando atrás las prótesis, también se ha hablado de otros conceptos facturables, en el que se engloban actos diversos como los relacionados con quirófanos ambulatorios, preoperatorios, material sanitario y otros tratamientos ambulatorios. Esta categoría ha generado de media un coste de 5.26€ por asegurado a la entidad aseguradora. Los estadísticos descriptivos nos indican que prácticamente, de media, los asegurados no hacen uso de estos servicios médicos. Esto queda confirmado con tal sólo observar Tabla 17, en la que se muestra que al 98.6% de los asegurados no hizo uso de este acto médico.

Tabla 17. Tabla de frecuencias de Otros Conceptos Facturables

	Frecuencia	Porcentaje	Porcentaje acumulado
0	164072	97.7	97.7
1	3109	1.9	99.5
2	504	0.3	99.8
3≤	324	.2	100.0

Fuente: elaboración propia

Para finalizar con el análisis de los costes, cabe estudiar la variable "Otros". Al igual que en el caso anterior, la media indica que prácticamente el usuario medio no hace uso de estos servicios médicos. Concretamente, 144,208 asegurados no hicieron uso de estos servicios, mientras que 18,260 usuarios lo requirieron una vez y tan solo 5,541 clientes tuvieron necesidad de acudir 2 o más veces. Así pues, el coste promedio por asegurado es de 157.83€ para los actos médicos que engloba esta categoría médica.

Tabla 18. Tabla de frecuencias de la variable "Otros"

	Frecuencia	Porcentaje	Porcentaje acumulado
0	144208	85.8	85.8
1	18260	10.9	96.7
2	3742	2.2	98.9
3	955	0.6	99.5
4≤	844	.5	100.0

Fuente: elaboración propia

Una vez analizado el coste desglosado, debe analizarse el coste total y la prima. La Tabla 19 recoge los estadísticos descriptivos que nos informa de que, en promedio, un usuario de seguro de salud genera un coste de 663.55€. Si comparamos, este dato con la prima media, que es de 763.63€, puede afirmarse que en promedio los asegurados generan rentabilidad a la entidad aseguradora, ya que los costes generados por los usuarios no superan los ingresos percibidos por éstas.

Tabla 19. Estadísticos descriptivos para los costes de los servicios médicos

	Mínimo	Máximo	Media	Desviación estándar
Ctevisitas	0	6166.66	184.4812	185.98863
CteRadio	0	2701.53	88.9052	138.30159
CteAnClinico	0	3217.46	43.8058	78.20541
CteOtrosDiag	0	2641.91	13.7328	66.33506
CteActos	0	17953.98	111.2980	337.28539
CteAnest	0	2863.84	18.3143	70.80510
CteRehab	0	3576.20	18.8639	67.98009
CteProtesis	0	14478.78	21.0544	271.78247
CteOtrosCptoFac	0	8560.10	5.2613	80.00524
CteOtros	0	9993.75	157.8314	648.69216
Ctetotal	0	29838.06	663.5485	1256.68339

Fuente: elaboración propia

Para finalizar con el análisis exploratorio, se debe analizar la duración de los asegurados en póliza. Como expone la tabla situada debajo, los asegurados permanecen en esta entidad aseguradora una media de 10 años. No obstante, la más frecuente es permanecer 1 año, ya que es el valor que más se repite. Si analizamos los cuartiles, el 25% de los asegurados permanece 3 o menos años en póliza, mientras que el 75% de estos tiene una duración de 15 años o menos.

Tabla 20. Estadísticos descriptivos para la duración de la póliza

	Media	Mediana	Moda	Percentil 25	Percentil 75
Antigüedad	10	6	1	3	15

Fuente: elaboración propia

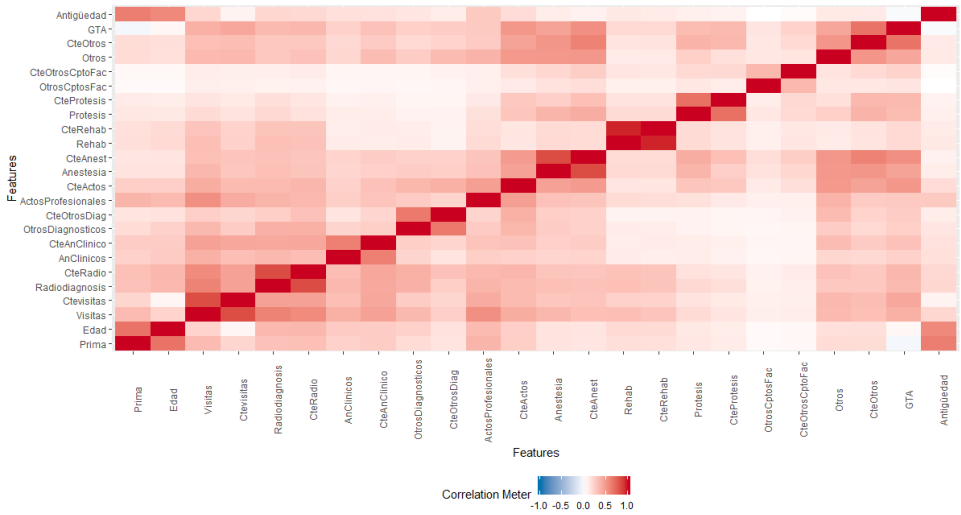
Una vez se ha analizado y conocido en profundidad el conjunto de datos con el que se va a trabajar, se puede proceder con el siguiente apartado.

3.3. Análisis de la correlación entre las variables

El objeto de esta etapa es realizar un examen de las relaciones entre las variables analizadas y su grado de interrelación. Es decir, estudiar la existencia de relación entre ellas y detectar la presencia de multicolinealidad.

Para simplificar en el análisis, se ha procedido a incluir únicamente la matriz de correlación reducida, es decir, aquella que únicamente incluye las variables de interés.

Gráfico 7. Matriz de correlación de las variables cuantitativas



Fuente: elaboración propia

Tras el análisis también se detecta un elevado grado de correlación positivo entre la frecuencia de los diversos servicios médicos y sus respectivos costes. Es por ello por lo que posiblemente haya que recurrir a la eliminación de alguna de dichas variables, sin pérdida de capacidad explicativa dada la redundancia que ocasionan. El resto de las variables presentan cierta correlación positiva entre ellas, pero no supone un grado de asociación tan elevado como para generar multicolinealidad.

En la matriz de correlación anterior sólo se incluyen variables del tipo cuantitativas, pero no categóricas (nominales u ordinales). Es por ello por lo que debemos acudir a otros métodos para medir la correlación entre variables cuantitativas y categóricas, y entre las variables categóricas entre sí. Para el primer caso de estudio se acudirá al coeficiente de correlación η^2, mientras que para estudiar el grado de asociación entre variables categóricas se analiza el estadístico de Cramer.

Los estadístico η^2 y los estadísticos de Cramer para estudiar el grado de asociación se encuentran recogidas en el Apéndice 1 en la Tabla 43 y Tabla 44. Por un lado, en lo que se refiere al grado de asociación entre variables cuantitativas y categóricas, el estadístico η^2 indica que no existe grado de asociación suficiente entre las variables analizadas, a excepción de la mayoría de las variables con la rentabilidad. Por otro lado, analizando los resultados del estadístico de Cramer para estudiar la correlación entre factores, se puede concluir que existe una relación débil entre ciertas variables categóricas entre las variables ramo y sexo con parentesco, y entre el ramo y provincia. No obstante, no generará ningún problema de multicolinealidad.

4. Métodos de Aprendizaje No Supervisado

En este apartado, nos centramos en el llamado "Aprendizaje No Supervisado" con el objetivo de detectar grupos (clústeres) de asegurados en función de la tipología de uso a través de variables relevantes que permitan descubrir comportamientos o características de cada grupo que ya se han manifestado previamente en los datos.

Esta segmentación debe implementarse de manera que las observaciones sean lo más homogéneas entre sí dentro de un clúster, pero heterogéneas respecto al resto de clústeres. Si ocurriera lo contrario, se estaría aplicando un criterio de agrupación que no aportaría valor al no permitir de manera adecuada establecer las estrategias comerciales ni ayudar a la toma de decisiones.

Cabe destacar que el fin último de este análisis no es predecir, sino encontrar patrones en los datos tal y como se ha mencionado en el párrafo anterior para poder diferenciar en cuanto a comportamientos de uso de los servicios médicos en los seguros de Salud.

4.1. Variables de interés

Antes de aplicar los diferentes algoritmos, hay que considerar que es posible incurrir en ciertos problemas: problema del tamaño de las variables, problema de multicolinealidad (redundancia) y problema de "*missing values*" o valores perdidos. Como anteriormente se ha descrito, la BBDD empleada en este análisis está libre de valores perdidos, por lo que nos permite obviar esta cuestión y centrarnos en las dos primeras.

En cuanto al tamaño de las variables, se debería estudiar si las distintas variables difieren en la dimensión de sus valores. El sentido común indica que se podría estar incurriendo en este problema ya que las variables de interés están asociadas a la frecuencia de los diferentes actos médicos y su correspondiente coste. Por tanto, hay que evitar un método sin distancia euclídea al cuadrado, ya que asigna a todas las variables el mismo peso.

Para estudiar el segundo problema, el relativo a la existencia de una posible multicolinealidad, acudimos a la matriz de correlaciones presentada en el apartado anterior. Ésta hace patente la correlación existente entre las diferentes variables, en especial entre la frecuencia de cada servicio médico y su coste. Esta correlación es positiva y significativa a un nivel de significancia del 1% y, por tanto, confirma la presencia de redundancia.

Considerando lo anterior, nos enfrentamos un problema con respecto al tamaño de las variables y con respecto a la redundancia. ¿Qué método debemos aplicar entonces? Para poder solucionar este problema encontramos dos alternativas:

1. Tipificación o estandarización de las variables de interés: esta alternativa solucionaría el problema del tamaño de las variables, pero no el de multicolinealidad.
2. Aplicar el método de Ward con la distancia de Mahalanobis: esto equivaldría a realizar el método empleando la distancia euclídea y como predictores los factores obtenidos por el método de los Componentes Principales. Esta opción solucionaría ambas incidencias.
3. Eliminar aquellas variables de interés que generen la multicolinealidad.

De entre las presentes alternativas, se escoge a tercera opción: la supresión de las variables redundantes. Como nuestro objetivo es agrupar en clústeres según tipología

de uso de los servicios médicos, las variables de las que se prescinde serán aquellas relacionadas con los costes. Como consecuencia, solucionamos el problema de la multicolinealidad a la vez que el de tamaño de las variables.

Tabla 21. Resumen de las variables de interés

Número	Variable	Número	Variable
1	Visitas	**6**	Anestesia
2	Radiodiagnosis	**7**	Rehabilitación
3	Análisis Clínicos	**8**	Prótesis
4	Otros Diagnósticos	**9**	Otros conceptos facturables
5	Actos profesionales	**10**	Otros

Fuente: Elaboración propia

En este apartado, se aplican dos métodos de Análisis de Clústeres, algoritmo CLARA y el algoritmo K-means, para estudiar cuál alcanza a realizar una mejor agrupación de los asegurados en función de la tipología de uso del seguro de Salud.

4.2. Algoritmo CLARA

En esta subsección, se implementa el algoritmo CLARA (Clustering Large Application), una extensión del método de agrupación PAM (Partioning Around Medoids), destinado a muestras grandes que reduce el tiempo de cálculo[1]. Según los autores (Kaufman & Rousseeuw, 1990), el algoritmo CLARA, en vez de emplear el conjunto de datos entero para encontrar la distancia media o medoid, considera una pequeña muestra con un tamaño fijo (*sampsize*) e implementa el algoritmo para generar el conjunto óptimo de medoides.

A continuación, se mencionan las etapas que conforman este algoritmo:

1. Crear de manera aleatoria múltiples muestras de tamaño fijo (*sampsize*), partiendo del conjunto de datos original;

[1] El tiempo de cálculo aumenta en gran medida por la cantidad de observaciones del conjunto de datos y la potencia de cálculo del ordenador disponible tanto en memoria RAM como velocidad.

2. Implementar el algoritmo PAM en cada una de las submuestras y elegir el número óptimo de medoides;
3. Asignar cada observación de la muestral al medoid más próximo;
4. Calcular la media o la suma de las desemejanzas de las observaciones hasta el medoid más cercano, con el fin de obtener una medida de bondad de la agrupación;
5. Conservar la submuestra cuya media o suma obtenida en el cuarto paso se minimice.

4.2.1. Determinación del número de clústeres

Para determinar el número de clústeres, podemos acudir a las siluetas de cada grupo (S_i). Éstas miden la similitud de un objeto "i" con los otros objetos de su propio clúster frente a los del clúster vecino. La silueta S_i toma valores dentro del rango $[-1,1]$, indicando 1 que el objeto está bien agrupado y lo contrario cuando se acerca a -1.

La tabla siguiente muestra los valores medios de la silueta para diferentes números de clústeres (k) en cada uno de los ramos. El ramo 29 tiene la mayor silueta media para 6 clústeres, mientras que la cartera y los ramos 54 y 82 alcanzan la mayor media con 5.

Tabla 22. Silueta media por ramo y diferente 'k' clústeres

Ramo	$k = 4$	$k = 5$	$k = 6$
Cartera	0.29	0.36	0.35
29	0.28	0.35	0.42
54	0.32	0.40	0.26
82	0.32	0.33	0.27

Fuente: Elaboración propia

Nos encontramos ante un dilema, ya que no coinciden en cuanto al número de clústeres idóneos. No obstante, ya que dos de los tres ramos y la cartera consiguen una mayor agrupación con $k = 5$, se elegirá éste como numero de clústeres para todos los ramos.

Una vez decidido y generado el número óptimo de los clústeres o segmentos de asegurados, recurrimos a un test de comparación de medias (ANOVA), que nos confirma que las medias de las variables de interés empleadas, junto con la edad, son estadísticamente significativas entre los grupos y, por tanto, son de utilidad para detectar diferencias en el comportamiento de los grupos de asegurados.

4.2.1.1. Implementación en la cartera

La tabla siguiente nos informa acerca de la media de las variables de interés, junto con la edad y la duración de la póliza, por clúster con el objetivo de identificar las diferentes tipologías de uso de seguro médico encontrados para la cartera de asegurados.

Tabla 23. Media de las variables de interés por clúster para la cartera (CLARA)

Variables	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5	Cartera
Visitas	33.46	17.39	16.56	7.36	2.23	8.63
Radiodiagnosis	8.28	6.44	4.39	2.31	0.57	2.50
AnClinicos	3.41	2.05	2.44	1.53	0.47	1.39
OtrosDiagnosticos	1.59	0.81	0.85	0.44	0.15	0.47
ActosProfesionales	8.66	5.07	4.12	2.11	0.51	2.30
Anestesia	0.68	0.46	0.28	0.09	0.01	0.14
Rehab	1.95	26.69	0.76	0.67	0.27	2.22
Protesis	0.07	0.09	0.02	0.01	0.00	0.02
OtrosCptosFac	0.09	0.13	0.06	0.03	0.01	0.04
Otros	1.12	0.46	0.40	0.15	0.03	0.21
Edad	52.16	54.19	43.46	39.26	34.14	39.51

Fuente: Elaboración propia

El primer clúster va asociado a una elevada frecuencia media de usos médicos, excepto para las sesiones de rehabilitación y otros conceptos facturables. De entre las diferentes categorías médicas destacan las visitas ambulatorias, con una media de 33 visitas aproximadamente. La edad media de un individuo de este grupo es de 52.16 años.

El Clúster 2 es muy similar al primero, diferenciándose en que éste tiene un uso medio mayor para las prótesis, otros conceptos facturables y, sobre todo, las sesiones de rehabilitación, pero inferior en el resto de los actos médicos. En cuanto a la edad promedio, es el clúster más envejecido.

Por el contrario, el Clúster 5 presenta el uso promedio de las categorías médicas más bajo de todos los clústeres, señalando así el poco uso que estos asegurados hacen del seguro médico. Cabe destacar que este clúster es el más rejuvenecido de todos.

El Clúster 4 presenta un comportamiento parecido al Clúster 5. La diferencia radica en que la frecuencia media de uso de los servicios prestados por el seguro es ligeramente superior en todas las categorías de actos médicos.

En cuanto al tercer grupo, atendiendo a las medias de las variables de interés, se caracteriza por un comportamiento intermedio entre el de los clústeres 2 y 4: presenta una frecuencia media mayor que el tercer grupo, pero inferior al del cuarto grupo.

Por tanto, a las diferentes tipologías de uso se les puede nombrar como sigue: Clientes Hipocondriacos (Clúster 1), Clientes Habituales (Clúster 2), Usuario Medio (Clúster 3), Clientes Precavidos (Clúster 4) y Clientes Pasivos (Clúster 5).

4.2.1.2. Implementación en el Ramo 29

Al igual que lo realizado en la sección anterior, acudiremos a la media de las variables de interés y a la variable edad para identificar los clústeres en el ramo 29, aplicando la mismas nomenclaturas anteriores para facilitar la futura comparación.

Tabla 24. Media de las variables de interés por clúster para el ramo 29 (CLARA)

Variables	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5	Ramo 29
Visitas	32.16	6.6	14.06	14.94	2.22	7.81
Radiodiagnosis	8.46	1.8	3.86	5.62	0.67	2.21
AnClinicos	3.11	1.12	2.33	1.83	0.54	1.26
OtrosDiagnosticos	1.47	0.37	0.73	0.74	0.16	0.42
ActosProfesionales	6.55	1.48	3.61	3.6	0.55	1.85
Anestesia	0.62	0.06	0.25	0.42	0.01	0.12
Rehab	2.67	0.7	1.08	29.43	0.27	1.86
Protesis	0.06	0	0.02	0.07	0	0.01
OtrosCptosFac	0.07	0.02	0.05	0.13	0.01	0.03
Otros	0.93	0.1	0.34	0.36	0.03	0.17
Edad	38.85	31.13	36.02	45.9	30.93	33

Fuente: Elaboración propia

En primer lugar, el Clúster 1 destaca por una elevada frecuencia media de uso en relación con el resto de las agrupaciones prácticamente en todas las categorías, destacando Visitas, Radiodiagnosis, Análisis Clínicos, Actos Profesionales y Otros. En cuanto a la edad media de los clientes de este grupo, se sitúa casi en los 39 años. Este grupo podría clasificarse como de uso muy frecuente del seguro de Salud: “Clientes Hipocondriacos”.

En segundo lugar, el segundo segmento de asegurados destaca por un bajo número medio de servicios asociados a la categoría “Anestesia”, mientras que el resto de los servicios se encuentran en la media de frecuencia de uso del ramo. Así, los asegurados

de este grupo se consideran como clientes con un perfil de uso ocasional. Los podemos englobar dentro del nombre de “Clientes Precavidos”. La edad media de este grupo es ligeramente superior a los 31 años.

En lo que respecta al tercer clúster, un asegurado medio perteneciente a este conglomerado se caracteriza por una frecuencia relativamente alta en análisis clínicos y servicios médicos clasificados como otros. Es por ello por lo que, al situarse el resto de las categorías cerca de la frecuencia media del ramo, estos asegurados pueden nombrarse como “Usuario Medio”. La edad media de este clúster es de 36 años, una de las más altas.

En el cuarto lugar, el siguiente clúster destaca por una elevada media de sesiones de rehabilitación en su haber. Además, destaca por su elevado número medio de asistencias a servicios incluidos dentro de “Otros conceptos” y “Prótesis”. Los asegurados de este grupo, en relación a la tipología de uso del seguro, se pueden considerar como de uso frecuente y le damos el nombre de “Clientes Habituales”. Este clúster es el que presenta el mayor envejecimiento, con una edad media de casi 46 años.

En último lugar, encontramos el clúster que podrían considerarse como usuarios con prácticamente una nula frecuencia media de uso, presentado un mayor número medio de las visitas ambulatorias. Así, estos clientes pueden denominarse “Clientes Pasivos”. La edad media de este grupo es la menor de todos los clústeres: 30.93 años.

4.2.1.3. Implementación en el Ramo 54

Volvemos a repetir el proceso con el fin de segmentar los asegurados de este ramo en función del uso del seguro de Salud.

Se observa el parecido existente entre el Clúster 1 y 2. Principalmente las diferencias radican en que los asegurados del Clúster 1 tienen un uso medio muy superior de actos ambulatorios y de actos profesionales. En cambio, el Clúster 2 presenta una frecuencia media muy superior en sesiones de rehabilitación. Es por ello por lo que al primer clúster lo podríamos nombrar como “Clientes Hipocondriacos” y a los del segundo “Clientes Habituales”. Estos primeros alcanzan una edad media de 68.02 años, mientras que para los segundos es de 62.73 años. Los primeros se agrupan el clúster más envejecido.

Tabla 25. Media de las variables de interés por clúster para el ramo 54 (CLARA)

Variables	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5	Ramo 54
Visitas	36.97	18.67	2.63	8.32	18.26	10.57
Radiodiagnosis	9.46	6.77	0.91	2.85	5.45	3.31
AnClinicos	3.35	2.16	0.69	2	2.72	1.74
OtrosDiagnosticos	1.9	0.81	0.2	0.53	1.06	0.61
ActosProfesionales	13.5	5.11	0.94	3.23	5.01	3.48
Anestesia	0.85	0.5	0.02	0.13	0.32	0.20
Rehab	4.4	31.23	0.35	0.9	1.43	3.22
Protesis	0.11	0.11	0	0.02	0.04	0.27
OtrosCptosFac	0.11	0.12	0.01	0.03	0.05	0.04
Otros	1.53	0.46	0.05	0.22	0.5	0.31
Edad	68.02	62.73	48.24	55.75	60.76	55.15

Fuente: Elaboración propia

Por otro lado, podemos relacionar también al tercer y cuarto clúster, siendo estos los que menor frecuencia media presentan en general para las diferentes categorías de servicio médicos. Centrando nuestra atención a las medias, se puede afirmar que los "Clientes Pasivos" serían aquellos pertenecientes al Clúster 3, mientras que a los del Clúster 4 se les podría definir como "Clientes Precavidos", al presentar estos últimos medias mayores en algunas categorías médicas como Visitas, Actos Profesionales o Análisis Clínicos. En cuanto a la edad, los "Clientes Pasivos" serían los más jóvenes, con una edad media de 48.24 años frente a los 56 años de los "Clientes Precavidos".

Para finalizar, el Clúster 5, por descarte, lo denominaremos como "Usuario Medio". Observando la Tabla 26, se puede confirmar dicho sobrenombre dado que prácticamente su frecuencia en los diferentes servicios se sitúa sobre la media del ramo, excepto para Análisis Clínicos y Otros diagnósticos con un uso medio relativamente elevado. La edad promedio de este grupo ronda los 61 años.

4.2.1.4. Implementación en el Ramo 82

Para finalizar, vamos a analizar los clústeres en el ramo 82 siguiendo el mismo procedimiento que en el resto de los ramos.

Tabla 26. Media de las variables de interés por clúster para el ramo 82 (CLARA)

Variables	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5	Ramo 82
Visitas	11.64	22.55	6.3	11.85	2.19	7.45
Radiodiagnosis	1.37	5.56	2.44	5.47	0.56	1.97
AnClinicos	1.06	2.93	1.64	1.77	0.3	1.16
OtrosDiagnosticos	0.32	0.94	0.46	0.7	0.12	0.37
ActosProfesionales	1.67	4.48	1.67	3.06	0.45	1.54
Anestesia	0.11	0.4	0.08	0.34	0.01	0.10
Rehab	0.37	0.97	0.51	22.93	0.20	1.55
Protesis	0.01	0.03	0.01	0.06	0	0.01
OtrosCptosFac	0.08	0.09	0.03	0.15	0.02	0.05
Otros	0.18	0.53	0.13	0.34	0.03	0.15
Edad	21.2	30.51	33.46	42.39	28.3	29.85

Fuente: Elaboración propia

En primer lugar, el Clúster 1 engloba a los asegurados denominados como “Usuario Medio”. En comparación con el resto de los clústeres no destaca en ninguna de las variables, ya que éstas tienen una media similar a la del ramo. No obstante, podemos resaltar el uso medio de las visitas ambulatorias, situado casi en 12 visitas, y la baja frecuencia de pruebas de radiodiagnosis y de análisis clínicos. Este segmento de asegurados tiene la edad media más joven, ligeramente superior a los 21 años.

En segundo lugar, encontramos al Clúster 2, que presenta unas características muy similares a las del cuarto clúster. Sin embargo, las propias de los “Clientes Hipocondriacos” vienen representadas mejor por el Clúster 2 al tener una media superior en todas las categorías médicas excepto para las sesiones de rehabilitación. Así pues, el Clúster 4 englobará a los “Clientes Habituales”. Destaca que estos últimos son los que mayor edad media presentan.

En tercer lugar, encontramos a los “Clientes Precavidos” enmarcados dentro del tercer segmento de asegurados. Estos no destacan por ninguna característica en particular y su edad media está en los 33.46 años.

Para finalizar, el Clúster 5 muestra signos de ser los “Clientes Pasivos” ya que presentan un uso medio muy poco frecuente para casi todos los servicios médicos. La edad media se centra en los 28.30 años.

4.2.2. Comparación de los resultados

Para finalizar con este algoritmo, vamos a comparar los clústeres obtenidos para cada ramo. Para ello acudimos a la distribución de asegurados por cada tipología de uso del seguro médico obtenido: Clientes Hipocondriacos, Clientes Habituales, Usuario Medio, Clientes Precavidos y Clientes Pasivos.

Tabla 27. Distribución de asegurados en función de la tipología de uso (CLARA)

	Ramo 29	Ramo 54	Ramo 82	Cartera
Clientes Hipocondriacos	4.08%	6.20%	10.23%	4.86%
Clientes Habituales	4.07%	7.29%	5.06%	6.23%
Usuario Medio	21.07%	18.48%	13.86%	13.84%
Clientes Precavidos	31.06%	30.83%	30.91%	38.21%
Clientes Pasivos	39.73%	37.20%	39.93%	36.86%

Fuente: Elaboración propia

Destaca que, a diferencia de los ramos, cuando se considera los clústeres para el total de la cartera, el porcentaje de Clientes Precavidos aumenta en detracción de los llamados Clientes Medios. No obstante, la distribución de los asegurados derivado del algoritmo CLARA es bastante similar para los diferentes ramos y la cartera. De hecho, las tipologías de mayor uso son aquellas que contienen un menor número de asegurados, a excepción del ramo 82 donde el porcentaje de Clientes Hipocondriacos supera al de Habituales, al suponer el 10.23% del total.

Los Usuarios Medios, Clientes Precavidos y Clientes Pasivos son los segmentos de asegurados donde se encuentra mayor concentración de asegurados, siendo las agrupaciones de clientes que menos uso de media hacen de su seguro de Salud. Suponen aproximadamente el 80% y el 90% del total de asegurados en los tres ramos.

Este resultado nos indica que la entidad aseguradora estudiada tiene su grueso de clientes en aquellos asegurados con una frecuencia de uso del seguro moderada o prácticamente nula y, por el contrario, los asegurados que hacen un uso intensivo del seguro suponen una pequeña proporción del total de clientes.

4.3. Algoritmo K-means

Para medir la homogeneidad entre los datos, el algoritmo K-means utiliza la distancia entre los objetos. Las observaciones que se parecen tendrán una menor distancia entre ellas. En general, como medida se utiliza la distancia euclídea, aunque también se pueden utilizar otras funciones como la distancia Manhattan.

Para este método, vamos a seguir el mismo proceso que en el punto anterior. Con el objetivo de poder comparar los dos algoritmos, vamos a implementar el método K-Means con 5 clústeres, al igual que con el algoritmo CLARA. De hecho, emplearemos las tipologías utilizadas en el algoritmo CLARA para denominar los grupos obtenidos.

Una vez generados los clústeres o segmentos de asegurados, debemos analizar qué características ayudan a distinguir el comportamiento de cada uno de los grupos haciéndolos más heterogéneos entre sí y homogéneos entre sus propias observaciones. Para ello volvemos a recurrir a la prueba ANOVA. Los resultados obtenidos muestran que todas las variables de interés empleadas, incluso la edad, son de utilidad para detectar diferencias de comportamiento entre los grupos.

4.3.1. Implementación en la cartera

Tal y como se ha realizado en el algoritmo CLARA, segregaremos a los asegurados de la cartera según el uso del seguro de Salud ateniendo a sus características en promedio.

Tabla 28. Medias de las variables de interés para cada clúster de la cartera (K-means)

Variables	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5	Cartera
Visitas	3.89	13.99	33.55	23.99	11.43	8.63
Radiodiagnosis	1.14	3.78	8.67	8.07	4.63	2.50
AnClinicos	0.84	2.30	3.24	2.40	1.70	1.39
OtrosDiagnosticos	0.25	0.73	1.63	1.03	0.60	0.47
ActosProfesionales	1.01	3.77	9.33	6.17	2.97	2.30
Anestesia	0.03	0.23	0.71	0.74	0.21	0.14
Rehab	0.18	0.20	3.33	43.51	15.81	2.22
Protesis	0.00	0.02	0.08	0.15	0.04	0.02
OtrosCptosFac	0.02	0.05	0.10	0.19	0.07	0.04
Otros	0.07	0.34	1.15	0.65	0.24	0.21
Edad	35.64	42.37	53.75	57.04	50.07	39.51

Fuente: Elaboración propia

El Clúster 1 presenta el uso promedio de las categorías médicas más bajo de todos los clústeres, confirmando el poco uso del seguro médico por parte de estos. Esto nos indica que es el segmento de asegurados llamado “Clientes Pasivos”. Cabe destacar que este clúster es el más rejuvenecido de todos

Por el contrario, los clústeres 3 y 4 exponen comportamientos similares, mostrando un alto uso medio de los servicios médicos prestados por el seguro de Salud. El primero va asociado a una elevada frecuencia media de usos médicos superior a la del Clúster 4, excepto para las sesiones de rehabilitación, anestesia, prótesis y otros conceptos facturables. En lo referente a la edad media, destaca que el Clúster 4 es el más envejecido con 57 años de media, frente a los casi 54 años del Clúster 3. Atendiendo a los resultados, el Clúster 4 serán los “Clientes Habituales” y el tercer grupo los “Clientes Hipocondriacos”.

El segundo grupo presenta unas características que lo sitúan entre los clústeres 1 y 5. Esto es, una frecuencia media inferior al Clúster 5, pero superior a la del Clúster 1. Este comportamiento les otorga el sobrenombre de “Clientes Precavidos”.

Para finalizar, el Clúster 5 se caracteriza por no destacar en ninguno de los servicios médicos prestados, a excepción de las sesiones de rehabilitación con una media de casi 16 visitas. Este clúster engloba a los llamados “Usuario Medio”.

4.3.2. Implementación en el Ramo 29

En segundo lugar, vamos a analizar las características de los clústeres del ramo 29 usando para ello la media de las variables de interés. De este modo, segregaremos a los asegurados según el uso del seguro de Salud.

En cuanto al Clúster 1, destaca por la baja frecuencia media en relación con el resto de las agrupaciones. Es muy similar al Clúster 5, pero su frecuencia media de uso está por encima de la de los asegurados de dicho grupo. Este segmento equivaldría a los “Clientes Precavidos”, cuya edad media es de 34 años aproximadamente.

Por tanto, en el Clúster 5, un cliente medio no hace de media uso de los servicios médicos disponibles, presentando la frecuencia media más baja en todas las categorías.

Por tanto, este grupo englobaría a los “Clientes Pasivos”. La edad media de estos clientes es de prácticamente 30.71 años.

Por otro lado, el segundo clúster destaca por tener la frecuencia media de usos más alta en las categorías de Rehabilitación, Radiodiagnosis, Anestesia y Prótesis en comparación con el clúster anterior. Dados los resultados, los podemos enmarcar dentro de “Clientes Habituales”, siendo éste el grupo más envejecido en promedio.

Tabla 29. Medias de las variables de interés por clúster para el ramo 29 (K-means)

Variables	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5	Ramo 29
Visitas	12.98	19.68	9.85	31.10	3.70	7.81
Radiodiagnosis	3.39	6.96	4.01	7.98	1.05	2.21
AnClinicos	2.11	2.23	1.42	3.16	0.76	1.26
OtrosDiagnosticos	0.65	0.93	0.54	1.45	0.24	0.42
ActosProfesionales	3.05	4.74	2.27	7.48	0.88	1.85
Anestesia	0.20	0.65	0.18	0.63	0.03	0.12
Rehab	0.16	41.31	14.16	2.74	0.09	1.86
Protesis	0.01	0.12	0.03	0.05	0.00	0.01
OtrosCptosFac	0.04	0.21	0.05	0.07	0.01	0.03
Otros	0.28	0.54	0.19	0.95	0.06	0.17
Edad	34.14	47.57	42.80	39.36	30.71	33.00

Fuente: Elaboración propia

El Clúster 3 destaca por su elevada frecuencia media de sesiones de rehabilitación, de aproximadamente 14 veces por asegurado, y por el poco uso en promedio de Actos Profesionales y de servicios médicos categorizados como Otros. Coincide con la tipología de “Usuario Medio” del algoritmo CLARA. La edad media del asegurado de este grupo ronda casi los 43 años.

Por último, el Clúster 4 es muy similar al Clúster 2, pero despuntan al presentar, en promedio, el mayor uso de todas las categorías médicas, a excepción de rehabilitación, prótesis y otros conceptos facturables. Este grupo se puede definir como “Clientes Hipocondriacos” y presentan una edad media de 39 años.

4.3.3. Implementación en el Ramo 54

Repetimos el mismo proceso que para el ramo 29, pero aplicado al ramo 54, para segmentar en función del uso de los servicios médicos prestados por el seguro de Salud.

Los clústeres 1 (Clientes Hipocondriacos) y 5 (Clientes Habituales) destacan por su elevada frecuencia media de visitas médicas por asegurado en casi todas las categorías médicas, a excepción de rehabilitación y visitas ambulatorias. Es aquí donde radica la principal diferencia entre estos, teniendo el Cliente Hipocondriaco un número promedio de rehabilitaciones muy superior al de un Cliente Habitual promedio, pero inferior en visitas ambulatorias. La edad media es 67.27 para el Clúster 1 y de 65.44 años para el Clúster 5.

Tabla 30. Medias de las variables de interés por clúster para el ramo 54 (K-means)

Variables	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5	Ramo 54
Visitas	36.68	4.41	14.21	15.68	29.74	10.57
Radiodiagnosis	9.46	1.52	5.44	4.82	9.68	3.31
AnClinicos	3.36	1.14	1.94	2.61	2.64	1.74
OtrosDiagnosticos	1.88	0.31	0.64	0.93	1.25	0.61
ActosProfesionales	11.63	1.50	4.17	5.42	8.07	3.48
Anestesia	0.83	0.05	0.33	0.29	0.90	0.20
Rehab	4.23	0.42	22.92	1.13	50.11	3.22
Protesis	0.11	0.01	0.07	0.03	0.20	0.27
OtrosCptosFac	0.10	0.01	0.08	0.05	0.22	0.04
Otros	1.43	0.10	0.34	0.46	0.82	0.31
Edad	67.27	50.44	61.43	60.61	65.44	55.15

Fuente: Elaboración propia

Con respecto al Clúster 2 (Clientes Pasivos), un asegurado medio presenta la frecuencia media más baja en todas las categorías de servicios médicos. Estos resultados indican que este tipo de asegurados no hacen un uso activo del seguro. Cabe destacar que la edad media de esta agrupación es la más baja: 50.44 años de media por asegurado.

Los asegurados del Clúster 3 se caracterizan por una elevada frecuencia media en servicios de rehabilitación, aunque inferior a la de los “Clientes Habituales”, sin destacar en el resto de las categorías médicas, otorgándoles así el nombre de “Usuario Medio”.

Para finalizar, encontramos al Clúster 4 (Clientes Precavidos). Esta agrupación es muy similar al Clúster 2 en cuanto a su comportamiento, diferenciándose en que presenta una media de uso ligeramente mayor en todos los actos médicos.

4.3.4. Implementación en el Ramo 82

En esta subsección, se procede a realizar el mismo análisis que en los puntos anteriores, pero para la cartera de asegurados del ramo 82.

En el Clúster 1, llamado "Clientes Precavidos", los asegurados destacan por tener una frecuencia media de usos ligeramente por encima de la media en la categoría "Visitas", siendo poco frecuente el uso de otras categorías médicas. La edad media de estos clientes es de la menor de entre todos los clústeres: 28.39 años.

En referencia al Clúster 2, es decir, los "Clientes Pasivos", al contrario del resto de grupos, destaca por la baja frecuencia media en todos los servicios médicos prestados. La edad media de estos clientes es de prácticamente 29 años.

La tipología del Clúster 3, "Clientes Habituales", es la más envejecida, con un promedio de 43.45 años. Destaca por su elevado uso medio de los servicios relacionados con Visitas, Radiodiagnosis y, en especial, Rehabilitación.

Tabla 31. Medias de las variables de interés por clúster para el ramo 82 (K-means)

Variables	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5	Ramo 82
Visitas	11.80	3.40	18.04	9.27	27.48	7.45
Radiodiagnosis	2.84	0.90	6.60	3.79	6.62	1.97
AnClinicos	1.90	0.66	1.93	1.46	2.89	1.16
OtrosDiagnosticos	0.54	0.21	0.77	0.51	1.10	0.37
ActosProfesionales	2.42	0.72	3.83	1.97	5.43	1.54
Anestesia	0.17	0.02	0.57	0.14	0.49	0.10
Rehab	0.12	0.07	36.41	13.44	1.91	1.55
Protesis	0.01	0.00	0.11	0.02	0.04	0.01
OtrosCptosFac	0.07	0.02	0.21	0.08	0.11	0.45
Otros	0.24	0.05	0.51	0.18	0.65	0.15
Edad	28.39	28.98	43.45	40.91	30.60	29.85

Fuente: Elaboración propia

En el Clúster 4, también llamado "Usuario medio", los clientes realizan un uso medio del seguro, destacando tan sólo la frecuencia media de las sesiones de rehabilitación, siendo la segunda más alta sólo por detrás del cuarto clúster. Su edad media ronda los 41 años.

Por último, los "Clientes Hipocondriacos" se encuentran en el Clúster 5. Corresponde con los asegurados que hacen de media un uso elevado de las diferentes categorías médicas. No obstante, un cliente medio de este clúster sólo hace dos visitas a rehabilitación. La edad media de estos asegurados ronda los 30.60 años.

4.3.5. Comparación de los resultados

Una parte de este trabajo se centra en comparar posteriormente ambos algoritmos de segmentación. Así pies, tal y como se ha realizado con el primer método, antes debemos analizar la proporción de asegurados que esta técnica agrupa en cada tipología de uso.

Tabla 32. Distribución de asegurados por en función de la tipología de uso (K-means)

	Ramo 29	Ramo 54	Ramo 82	Cartera
Clientes Hipocondriacos	4.56%	6.66%	5.69%	5.08%
Clientes Habituales	1.71%	2.02%	1.62%	1.87%
Usuario Medio	6.58%	6.05%	5.77%	6.83%
Clientes Precavidos	23.54%	25.82%	25.02%	23.27%
Clientes Pasivos	63.60%	59.46%	61.89%	62.96%

Fuente: Elaboración propia

Se observa claramente como este método de agrupación concentra la mayoría de los asegurados en dos tipologías de uso: Clientes Precavidos y Clientes Pasivos. Atendiendo a los resultados, se observa que el grueso de asegurados (en torno al 60%) está clasificado en todos los casos como "Clientes Pasivos", es decir, aquellos con menor uso medio de los actos médicos. Después, encontramos a los "Clientes Precavidos" como siguiente grupo más numeroso.

Si comparamos a los llamados "Clientes Hipocondriacos" y "Usuarios Medios", los resultados indican que suponen una proporción muy similar, suponiendo aproximadamente un 5% o 6% del total de asegurados cada clúster.

Por último, si continuamos analizando la tabla, se puede afirmar que este método apenas considera a los asegurados como "Clientes Habituales", con proporción que alcanza a lo sumo un 2% de los asegurados.

4.4. Algoritmo CLARA vs algoritmo K-means

En este apartado, se realiza con la finalidad de resumir los resultados derivados de la segmentación por ambos algoritmos y ayudar a la selección de uno de ellos como el más adecuado para esta cartera de asegurados.

En primer lugar, cabe destacar que ambos algoritmos coinciden a la hora de agrupar comportamientos similares, pudiendo así emplear los mismos sobrenombres a los

clústeres de asegurados generados por ambos métodos. No obstante, si analizamos la agrupación conseguida por cada método, se observa que el algoritmo CLARA realiza una segmentación más pareja entre los clústeres, a diferencia del K-means que tiende a la concentración de asegurados en aquellos clústeres con menor frecuencia media de uso, dejando al resto de los segmentos con poca representación.

Tabla 33. Tabla de contingencia para los clústeres de los algoritmos CLARA y K-means

		K-means				
	Cross Table	Hipocondriacos	Habituales	Medios	Precavidos	Pasivos
CLARA	Hipocondriacos	7,676	2	0	493	0
	Habituales	722	3,132	6,590	30	0
	Medios	135	0	1,247	21,864	0
	Precavidos	0	0	3,343	16,701	44,147
	Pasivos	0	0	293	0	61,634

Fuente: Elaboración propia

Para continuar con la comparación, acudimos a una tabla cruzada o de contingencia (Tabla 33). Ésta se ha implementado sólo para la cartera de asegurados. Si centramos nuestra atención en los Clientes Hipocondriacos, ambos métodos coinciden en gran medida en su clasificación ya que aproximadamente el 94% de los casos son clasificados bajo el mismo clúster. También coinciden prácticamente con los Clientes Pasivos, en los que sólo el 0.47% de los casos K-means difiere en su clasificación.

Continuando con el análisis, existe una clara divergencia en lo que se refiere a los Clientes Habituales según CLARA, ya que el algoritmo K-means clasifica aproximadamente el 29.90% como tal, pero el 62.92% como Usuario Medio. También encontramos que ambos algoritmos coinciden en un 26.02% en los Clientes Pasivos, pero el 41.73% K-means los clasifica como Clientes Precavidos.

En cuanto al "Usuario Medio", encontramos la mayor diferencia al tan sólo coincidir ambos métodos en el 5.36% de los casos. Es más, K-means clasifica el 94.05% de los asegurados clasificados por CLARA como Usuario Medio en Clientes Precavidos. Esto confirma nuestros resultados, ya que K-means concentra a los asegurados en los Clientes Precavidos y Clientes Pasivos, en detracción del resto de grupos.

Para finalizar el análisis, cabría realizar un test de χ^2 para estudiar la independencia entre los diferentes clústeres obtenidos por ambos algoritmos. Los resultados muestran un p-valor próximo a cero (p-value < 2.2e-16), rechazando la hipótesis nula y concluyendo la independencia de los clústeres obtenidos por ambos métodos.

En resumen, aunque se ha conseguido segmentar los clústeres en las mismas tipologías de uso del seguro de Salud por CLARA y K-means, se considera que este primero nos proporciona mejores resultados, dada su mejor distribución entre los diferentes segmentos, sobre todo al tener una mayor proporción de Clientes Habituales e Hipocondriacos.

5. Métodos de Aprendizaje Supervisado

En el apartado anterior se han analizado distintos métodos del llamado Aprendizaje No Supervisado, el cual se basa en el conocimiento de un conjunto de datos sin existir una variable respuesta asociada. En cambio, en los métodos de Aprendizaje Supervisado, el objetivo se centra en predecir, en nuestro si el asegurado será rentable o no, haciendo uso para ello de las distintas variables de interés.

En los procedimientos enmarcados como "Aprendizaje No Supervisado" se entrena un conjunto de datos históricos conocidos a priori. Así, al conjunto de datos se le realiza una partición obteniendo un conjunto de entrenamiento (Train) y un conjunto de prueba (Test). Para realizar este apartado se han seleccionado las siguientes variables:

Tabla 34. Variables de interés para los métodos de Aprendizaje Supervisado

Número	Variable	Número	Variable
1	Sexo	9	Actos Profesionales
2	Edad	10	Otros diagnósticos
3	Provincia	11	Anestesia
4	Duración	12	Rehabilitación
5	Parentesco	13	Prótesis
6	Visitas	14	Otros conceptos facturables
7	Radiodiagnosis	15	Otros
8	Análisis Clínicos	16	Rentabilidad (predicción)

Fuente: Elaboración propia

Para este trabajo se han seleccionado dos algoritmos en concreto: el algoritmo de clasificación J48 y el algoritmo clasificador bayesiano ingenuo, también llamado Naïve Bayes. Mediante estos algoritmos se tratará de ajustar un modelo que permita relacionar la respuesta (rentabilidad) a los predictores (variables de interés).

5.1. Árbol de decisión J48

Los árboles de decisión se componen de una estructura que parte de los nodos y de los cuales surgen las ramas, que conectan con el resto de los nodos. Cada uno de estos nodos representa cada uno de los atributos o característica, mientras que las ramas representan el rango de valores (Zhao & Zhang, 2008).

El algoritmo C4.5 fue introducido por primera vez por (Quinlan, 1986) con el objetivo de generar árboles de decisión. Posteriormente, con la necesidad de aplicar dicho algoritmos a otros lenguajes de programación surgió el algoritmo J48. Éste no es más que una adaptación al lenguaje de programación Java del algoritmo C4.5 publicado en 1986 (Salzberg, 1994).

Este método crea un árbol para modelizar el proceso de clasificación. Una vez generado árbol, se aplica a cada una de las tuplas de la base de datos y, en consecuencia, se muestran los resultados para cada una de éstas (Patil & Sherekar, 2013).

Este método presenta la ventaja que puede ser aplicado a cualquier tipo de variables predictoras, tanto variables continuas como categóricas. A su vez, la rapidez de cálculo y la facilidad de entendimiento e interpretación de este algoritmo también lo enmarcan dentro uno de los métodos más empleados para generar árboles de decisión.

En este trabajo, el árbol de decisión tiene persigue el fin de clasificar a los asegurados como clientes “rentables” o “no rentables”. Esta metodología se aplica al conjunto de datos correspondientes a la cartera en su totalidad y para cada uno de los ramos.

A la hora de generar el árbol de decisión, nos hemos decantado por realizar una validación cruzada de 10 iteraciones (10-fold cross validation). Una vez implementado el algoritmo J48, de entre los outputs resultantes, hay que prestar especial atención a la matriz de confusión. Ésta muestra la precisión del problema de clasificación, conteniendo información sobre la clasificación actual y estimada por un sistema de

clasificación. En otras palabras, informa acerca del número de casos que fueron asignados a una clase $C_{i,j}$, pero que de hecho su clasificación correcta sería en la clase C_i (Dunham & Sridhar, 2006). La solución perfecta sería aquella en la que se encontraran ceros en la contradiagonal de la matriz de confusión, indicando así que no se ha clasificado incorrectamente ninguna observación.

A continuación, la Tabla 35 muestra la matriz de confusión resultante de la validación cruzada del algoritmo J48. En esta matriz "b" indica que el asegurado es rentable, mientras que "a" indica lo contrario. En la tabla se muestra los casos de "a" y "b" clasificados por el algoritmo tanto en número de asegurados como en porcentaje.

Tabla 35. Matriz de confusión del algoritmo J48 para el total de la cartera

CARTERA		Predicción a = "NO"	Predicción b = "SI"	Predicción a = "NO"	Predicción b = "SI"	Precisión
Original	a = "NO"	27,812	10,214	73.14%	26.86%	90.04%
	b = "SI"	6,523	123,460	5.02%	94.98%	

Fuente: Elaboración propia

En la tabla de arriba, la diagonal muestra los llamados True Positive o Verdaderos Positivos (TP), mientras que en la contradiagonal se localizan los False Positive o Falsos Positivos (FP). Los resultados indican lo siguiente:

- $\boldsymbol{C(a = No\,; a = No)}$: esta predicción se refiere al número o porcentaje de asegurados "No rentables" que están predichos como tal. El algoritmo aplicado consigue que esta cifra alcance el 73.14% de aciertos.
- $\boldsymbol{C(a = No; b = Si)}$: en esta ocasión encontramos que el 26.86% de los asegurados clasificados inicialmente como "No rentables" son estimados como "Rentables".
- $\boldsymbol{C(b = Si; a = No)}$: indica que tan sólo el 5.02% de las veces un asegurado realmente "Rentable" será predicho como "No rentable" con este modelo.
- $\boldsymbol{C(b = Si; b = Si)}$: el 94.98% de los asegurados rentables son estimados correctamente como tal.

A modo de resumen, cuando se aplica el método generados de árboles de decisión J48 a la cartera en su totalidad, compuesta por 168,009 asegurados, se obtiene que las predicciones correctas realizadas por este algoritmo constituyen del 90.04%, es decir,

este método consigue predecir correctamente como rentables o no rentables 151,272 asegurados mientras que el número de asegurados incorrectamente clasificados se concentra en 16,737 clientes.

Tabla 36. Matriz de confusión del algoritmo J48 por ramo

RAMO 29		Predicción a = "NO"	Predicción b = "SI"	Predicción a = "NO"	Predicción b = "SI"	Precisión
Original	a = "NO"	7,474	3,059	70.96%	29.04%	91.05%
	b = "SI"	1,775	41,714	4.08%	95.92%	
RAMO 54		**Predicción a = "NO"**	**Predicción b = "SI"**	**Predicción a = "NO"**	**Predicción b = "SI"**	**Precisión**
Original	a = "NO"	8,451	3,172	72.71%	27.29%	90.59%
	b = "SI"	2,233	43,595	4.87%	95.13%	
RAMO 82		**Predicción a = "NO"**	**Predicción b = "SI"**	**Predicción a = "NO"**	**Predicción b = "SI"**	**Precisión**
Original	a = "NO"	11,873	3,997	74.81%	25.19%	88.69%
	b = "SI"	2,400	38,267	5.90%	94.10%	

Fuente: Elaboración propia

La Tabla 36 recoge los resultados desglosados por ramos. Tal y como se ha realizado anteriormente, se procede a realizar un breve análisis de las matrices de confusión:

- $\boldsymbol{C(\ a = No\ ;a = No)}$: el 70.96%, 72.71% y el 74.81% de los asegurados no rentables de los ramos 29, 54 y 82, respectivamente, son clasificados correctamente.
- $\boldsymbol{C(a = No\ ;b = Si)}$: este árbol de decisión es incapaz de predecir correctamente el 29.04%, 27.29% y el 25.19% de los asegurados no rentables de los ramos 29, 54 y 82, respectivamente. Estos casos se engloban dentro de los llamado False Positive.
- $\boldsymbol{C(b = Si; a = No)}$: centrándonos en el resto de los False Positive, el 4.08%, 4.87% y 5.90% de los casos, para los ramos 29, 54 y 82 respectivamente, un asegurado "Rentable" se ha predicho como "No rentable".
- $\boldsymbol{C(b = Si; b = Si)}$: por último, el 95.92%, 95.13% y el 94.10% de los asegurados rentables de los ramos 29, 54 y 82, respectivamente, son predichos correctamente.

Atendiendo a los resultados, la precisión global de cada modelo se sitúa en torno al 90% en los tres ramos, alcanzando el ramo 29 la precisión más alta con un 91.05% y la más

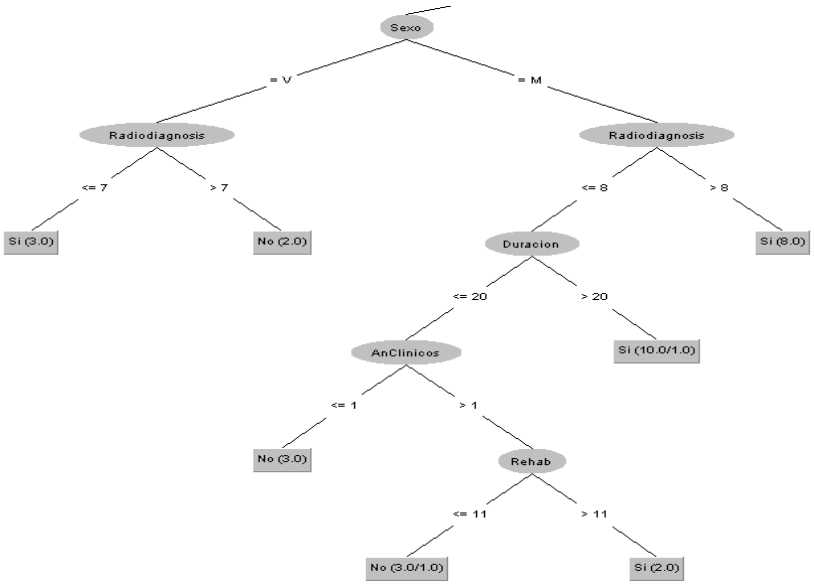
baja el ramo 82, con un 88.69%. No obstante, destaca el hecho de que, para el ramo 82, la precisión general del modelo es inferior al del resto de ramos, pero es la que comete menos errores a la hora de clasificar a asegurados no rentables como si lo fueran.

Este algoritmo no predice adecuadamente el grupo para prácticamente entre el 25% y 30% de los asegurados "no rentables", depende del ramo. Es necesario considerar que, como consecuencia de los errores de clasificación $C(No; Si)$, se podría incurrir en un coste para la entidad aseguradora al poder ofrecer descuentos o estrategias comerciales al suponer cierta rentabilidad en ellos.

Lo contrario ocurre con $C(Si; No)$, es decir, el error al clasificar a un asegurado "rentable" como "no rentable". El coste asociado al cometer este error está más relacionado con la posibilidad de "expulsión" del cliente de la empresa, al no premiar o bonificar a éste por su buen comportamiento siniestral.

Este algoritmo genera un árbol de decisión a partir del cual se clasifican los asegurados en función de su rentabilidad. Debido a la gran extensión de los árboles, se va a exponer una pequeña partición del árbol generado para la cartera. Cabe mencionar que el método J48 ya proporciona el llamado árbol podado, eliminando los atributos no relevantes, aunque en este tipo de algoritmos la presencia de dichos atributos no suele afectar significativamente a la precisión del modelo.

Ilustración 1. Output WEKA: parte del árbol de decisión del algoritmo J48



Fuente: elaboración propia a partir del software WEKA

Para finalizar este apartado, podemos concluir que el algoritmo J48 es adecuado para clasificar a los asegurados como rentables, pero que aproximadamente 3 de cada 10 asegurados "no rentables" sean predichos como "rentables". Esto denota cierto error de predicción por parte del modelo. No obstante, como la proporción de asegurados "no rentables" es bastante inferior a la de "rentables", no se ve reflejado en la precisión general del modelo.

5.2. Algoritmo clasificador "Naïve Bayes"

El método Naïve Bayes se puede definir como el paradigma clasificatorio en el que se emplea el teorema de Bayes junto con la hipótesis de independencia condicional de las variables explicativas dada la clase. El nombre de este clasificador deriva de la inclusión de la hipótesis simplificadora sobre las que se basa. Este método clasificador aparece por primera vez en la literatura en (Cestnik, Kononenko, & Bratko, 1987), pero no recibió este nombre hasta la publicación de (Kononenko, 1990).

Los clasificadores bayesianos asignan cada uno de los casos, descritos por un vector de características, a la clase más probable. Este algoritmo simplifica significativamente el aprendizaje utilizando el teorema de Bayes y asumiendo que la independencia de todas las características dado el valor de la variable clase (Rish, 2001). Esto es:

Ecuación 3. Fórmula probabilidad condicionada

$$P(X|C) = \prod_{i=1}^{n} P(X_i|C)$$

donde $X = (X_1, \dots, X_n)$ es el vector de características y C es la clase.

Pese a que este supuesto es poco realista, la clasificación resultante de la aplicación del algoritmo Naïve Bayes es considerablemente exitosa y, la mayoría de las veces, compite con los clasificadores más sofisticados. Según (Domingos & Pazzani, 1997), el éxito de este algoritmo en presencia de independencia entre características puede explicarse por el hecho de que la optimalidad en términos de error en la clasificación no está necesariamente relacionada con la calidad del ajuste de la distribución de probabilidad. Más bien, se tiene una clasificación óptima siempre y cuando la distribución actual y la predicha coincidan en la case más probable.

Al igual que con el algoritmo J48, a la hora de aplicar este método al conjunto de datos, la clasificación se ha generado junto con una validación cruzada de 10 iteraciones (10-fold cross validation). En la Tabla 37 encontramos la matriz de confusión resultante.

Tabla 37. Matriz de confusión del método Naïve Bayes

CARTERA		Predicción a = "NO"	Predicción b = "SI"	Predicción a = "NO"	Predicción b = "SI"	Precisión
Original	a = "NO"	23,886	14,140	62.81%	37.19%	86.06%
	b = "SI"	9,275	120,708	7.14%	92.86%	
RAMO 29		**Predicción a = "NO"**	**Predicción b = "SI"**	**Predicción a = "NO"**	**Predicción b = "SI"**	**Precisión**
Original	a = "NO"	7,451	3,082	70.74%	29.26%	89.24%
	b = "SI"	2,732	40,756	6.28%	93.72%	
RAMO 54		**Predicción a = "NO"**	**Predicción b = "SI"**	**Predicción a = "NO"**	**Predicción b = "SI"**	**Precisión**
Original	a = "NO"	7,933	3,690	68.25%	31.75%	87.75%
	b = "SI"	3,349	42,479	7.31%	92.69%	
RAMO 82		**Predicción a = "NO"**	**Predicción b = "SI"**	**Predicción a = "NO"**	**Predicción b = "SI"**	**Precisión**
Original	a = "NO"	10,535	5,335	66.38%	33.62%	86.17%
	b = "SI"	2,485	38,182	6.11%	93.89%	

Fuente: Elaboración propia

Atendemos a la tabla anterior con el fin de estudiar la precisión del modelo generado a través del método clasificador Naïve Bayes. Por tanto, la información que aporta la tabla se puede resumen como sigue:

- $\boldsymbol{C(a = No\ ;\ a = No)}$: el 62.81%, 70.74%, 68.25 y 66.38% de los asegurados no rentables de la cartera y ramos 29, 54 y 82, respectivamente, son predichos correctamente en su clase por el algoritmo.
- $\boldsymbol{C(a = No\ ;\ b = Si)}$: el algoritmo es incapaz de predecir correctamente la clase del 37.19%, 29.26%, 31.75% y el 33.62% de los asegurados no rentables de la cartera y de los ramos 29, 54 y 82, respectivamente.
- $\boldsymbol{C(b = Si\ ;\ a = No)}$: el porcentaje de asegurados "rentables" que son incorrectamente predichos por el algoritmo sigue situándose a un nivel aceptable para todos los casos.

- $\boldsymbol{C(b = Si\ ;\ b = Si)}$: el número de asegurados "rentables" cuya clase es correctamente estimada supera el 90% en todos los casos.

Volvemos a enfrentarnos a un elevado coste asociado al error de cometer falsos positivos por predecir asegurados "no rentables" como "rentables" y, por tanto, ofrecer mejores estrategias comerciales o descuentos al suponer cierta rentabilidad en estos. Este coste ha aumentado al aplicar esta metodología en comparación con el árbol de decisión J48. A su vez, el coste asociado al error de clasificar o predecir clientes "rentables" como si no lo fueran también ha aumentado, pudiendo desincentivar en cierta medida el buen comportamiento siniestral.

Si analizamos la precisión general del método Naïve Bayes, para la cartera y cada uno de los ramos es ligeramente inferior al 90%. Este porcentaje es elevado, pero al igual que con el método anterior, no refleja fielmente la realidad. Al existir un desequilibrio en el número de asegurados "rentables" y "no rentables", el porcentaje de precisión no refleja que, para la cartera en su conjunto, por ejemplo, casi el 40% de los asegurados "no rentables" serán predichos erróneamente como "rentables". Este es un porcentaje superior al obtenido con el método J48.

Al contrario que en J48, el método clasificador Naïve Bayes no proporciona durante su implementación ningún árbol de decisión. Es por esto por lo no se incluye ningún apartado para su análisis a diferencia del método J48. No obstante, cabe destacar que este método reduce la calidad de su precisión si existen atributos no relevantes. Así, efectuamos un filtrado de atributos, eligiendo un método de evaluación de atributos de la familia Wrapper y un método de búsqueda llamado "Best First". De acuerdo con estos métodos, se deben emplear las siguientes variables:

- Ramo 29: Sexo, Visitas, Prótesis, Otros y Duración
- Ramo 54: Sexo, Edad y Otros
- Ramo 82: Visitas, Radiodiagnosis, Prótesis y Otros.

Pese a la ligera mejoría en la precisión general que este filtrado genera para los diversos ramos, no se van a mostrar los resultados, ya que siguen siendo menos precisos a los que se obtienen por el algoritmo J48.

5.3. Comparación de métodos

Una vez implementados los dos algoritmos clasificadores, J48 y Naïve Bayes, debemos realizar una comparativa entre estos dos métodos para determinar cuál se adecua mejor a nuestros datos y, por ende, genera una clasificación más precisa.

Tabla 38. Comparativa de la precisión entre el método J48 y Naïve Bayes

Precisión	CARTERA	RAMO 29	RAMO 54	RAMO 82
J48	90.04%	91.05%	90.59%	88.69%
Naïve Bayes	86.06%	89.24%	87.75%	86.17%

Fuente: Elaboración propia

En la tabla anterior, se observa de manera clara que el error en la predicción a la hora de clasificar a los asegurados según su rentabilidad ha aumentado con respecto al del método J48 generador de árboles de decisión, viéndose reducida la precisión general del modelo.

Tabla 39. Análisis del Coste de la clasificación

	CARTERA		RAMO 29		RAMO 54		RAMO 82	
Rentabilidad	J48	Naïve Bayes	J48	Naïve Bayes	J48	Naïve Bayes	J48	Naïve Bayes
Sí	6,523	9,275	1,175	2,732	2,233	3,349	2,400	2,485
No	10,214	14,140	3,059	3,082	3,172	3,690	3,397	5,335

Fuente: Elaboración propia

Además, se realiza un estudio del rendimiento de ambos métodos a través del Análisis de Costes. Los resultados demuestran que el algoritmo J48 genera menores costes o errores de clasificación que el método Naïve Bayes. De esta manera, queda probado que para este conjunto de datos el algoritmo J48 es más eficiente en términos de costes en comparación con el otro método empleado.

Concluyendo, los resultados confirman que el método J48 es una técnica más adecuada para clasificar a los asegurados de esta BBDD al proporcionar una mayor precisión a la hora de predecir la clase a la que pertenece el asegurado.

6. ANOVA la siniestralidad

Antes de concluir este trabajo, se plantea un apartado que sirva de unión entre la parte del trabajo dedicada al Aprendizaje Supervisado (clasificación) y el No Supervisado (segmentación). Para ello, se realizan pruebas ANOVA de la siniestralidad sobre los clústeres generados por la implementación de CLARA y K-means con el objetivo de analizar si realmente estos clústeres difieren significativamente en su siniestralidad.

Tabla 40. Prueba ANOVA sobre la siniestralidad

	K-means			CLARA		
	g.l.	F-stat	Sig.	g.l.	F-stat	Sig.
Clústeres Cartera	4	6,327	<2e-16 ***	4	6,552	<2e-16 ***
Clústeres Ramo 29	4	2,889	<2e-16 ***	4	3,037	<2e-16 ***
Clústeres Ramo 54	4	1,936	<2e-16 ***	4	1,941	<2e-16 ***
Clústeres Ramo 82	4	2,666	<2e-16 ***	4	2,605	<2e-16 ***

Fuente: Elaboración propia

Los resultados del test ANOVA nos informan de que la siniestralidad nos puede ayudar a diferenciar entre los clústeres, tanto para los derivados de K-means como CLARA. En la siguiente tabla se presentan la siniestralidad media:

Tabla 41. Siniestralidad media por clúster

CLARA	Hipocondriacos	Habituales	Medios	Precavidos	Pasivos
Cartera	283.35	209.19	171.87	82.48	28.11
Ramo 29	297.99	203.79	147.07	64.56	25.14
Ramo 54	240.53	172.66	137.24	74.63	25.27
Ramo 82	321.88	264.93	138.08	99.44	35.30
K-MEANS	Hipocondriacos	Habituales	Medios	Precavidos	Pasivos
Cartera	285.50	281.61	149.37	142.18	44.89
Ramo 29	294.71	271.75	131.28	123.58	38.25
Ramo 54	260.31	235.09	137.77	125.23	39.91
Ramo 82	376.77	376.44	167.63	160.67	50.96

Fuente: Elaboración propia

La tabla anterior nos proporciona bastante información. En primer lugar, nos indica que, además de clasificar según la tipología de uso de los servicios médicos del seguro de Salud, cada uno de los clústeres se caracteriza por una determinada siniestralidad media que difieren estadística y significativamente entre sí. De este modo, cuanto menor uso medio del seguro, menor siniestralidad media se esperará que tenga el asegurado. Lo contrario ocurrirá si el cliente tiene una frecuencia elevada de uso de los servicios.

Otro dato que destacar es que, por un lado, el método K-means implementa la segmentación de manera que el único grupo resultante rentable para la entidad es el de los Clientes Pasivos, ya que el resto de los clústeres tiene una rentabilidad media superior al 100%. Por otro lado, con el algoritmo CLARA, los asegurados clasificados como Precavidos y Pasivos son, en promedio, rentables para la entidad aseguradora. Llama la atención como para K-means la diferencia de siniestralidad media entre los Clientes Hipocondriacos y Habituales es mínima, pudiendo generar confusión a la hora de clasificarlos en un grupo u otro.

Resumiendo, los clústeres generados por ambos algoritmos se caracterizan por cierta siniestralidad, que los diferencia entre ellos. Esta información sobre su rentabilidad complementa los resultados estudiados acerca del comportamiento de los grupos.

7. Conclusión

Para finalizar, se presentan las conclusiones obtenidas en este trabajo. Antes de detallarlas, se realiza un resumen del trabajo con la finalidad de recapitular lo realizado.

Para llevar a cabo este trabajo, se ha empleado una BBDD consistente en una cartera de asegurados de un seguro de Salud, cedida por una entidad aseguradora. A modo de recordatorio, el presente trabajo presenta doble objetivo:

- Establecer una segmentación de los asegurados en función de su comportamiento con respecto al uso de los servicios médicos prestados por el seguro de Salud.
- Seleccionar un algoritmo que permita clasificar a los asegurados de la entidad estudiada en función de su rentabilidad.

En cuanto a la metodología implementada, se ha recurrido al Aprendizaje No Supervisado y Supervisado. Por un lado, en lo que a Aprendizaje No Supervisado se refiere, hemos acudido a los algoritmos CLARA y K-means con la finalidad de detectar agrupaciones de asegurados en función del uso del seguro de Salud. Por otro lado, en lo que respecta al Aprendizaje Supervisado, se han ejecutado el árbol de decisión J48 y el algoritmo clasificador Naïve Bayes para clasificar si los clientes de un seguro de salud de la entidad aseguradora estudiada son rentables.

Dirigiendo nuestra atención hacia los resultados, en lo que respecta a los derivados de la segmentación de asegurados, la implementación de los algoritmos CLARA y K-means evidencia la existencia de cinco clústeres, resumidos en la siguiente tabla.

Tabla 42. Resumen de clústeres

Clústeres	Cliente Hipocondriaco	Cliente Habitual	Usuario Medio	Cliente Precavido	Cliente Pasivo
Frecuencia	Alta	Media - alta	Media	Media - baja	Baja

Fuente: Elaboración propia

De la comparación de ambos métodos, se deduce que ambos proporcionan una clasificación parecida en cuanto al comportamiento de los usuarios, pero no distribuyen de la misma manera los asegurados. Así, K-means concentra a los asegurados como precavidos y pasivos, en detracción del resto de grupos. Mientras tanto, CLARA reparte de manera más equitativa los asegurados entre los usuarios anteriores y los clientes medios. Esta redistribución de los asegurados es la que hace que nos decantemos por el algoritmo CLARA frente al K-means.

Los resultados relacionados con el Aprendizaje Supervisado se centran en la precisión de los métodos empleados: J48 y Naïve Bayes. De esta manera, este segundo método se ve superado por el árbol de decisión J48 en cuanto a precisión se refiere para clasificar a los clientes de nuestra BBDD en función de su rentabilidad. Por tanto, para este conjunto de datos, éste es más eficaz que el método Naïve Bayes al permitir un menor coste relacionado con la clasificación de asegurados no rentables como si lo fuesen y un menor coste de errar en la clasificación de usuarios rentables como si no lo fueran, en términos de estrategias comerciales y decisiones empresariales.

Concluyendo, hay que hacer referencia a posibles líneas de investigación para este trabajo. Una extensión interesante consistiría en conseguir aumentar la BBDD incluyendo variables relacionadas con el estado de salud del asegurado, tales como presión arterial, peso, si fuma o no, si realiza actividad física, etc. Estas nuevas variables nos permitirían cambiar el enfoque de este trabajo y dirigirlo hacia un nuevo tema: la búsqueda de factores que permitan explicar la siniestralidad de los asegurados.

Referencias

Alcaide, J. C. (2015). Fidelización de clientes. Madrid: ESIC Editorial.

Balabantaray, R. C., Sarma, C., & Jha, M. (junio de 2013). Document Clustering using K-Means. *International Journal of Knowledge Based Computer System , 1*(1), pp. 5-7.

Cestnik, B., Kononenko, I., & Bratko, I. (1987). Assistant 86 : A Knowledge-Elicitation Tool for Sophisticated Users. *Progress in Machine Learning, 62*, pp. 31-45.

CIS. (2018). *Estudio nº 8817: Barómetro Sanitario 2017.* MINISTERIO DE SANIDAD, CONSUMO Y BIENESTAR SOCIAL, DIRECCIÓN GENERAL DE SALUD PÚBLICA, CALIDAD E INNOVACIÓN, SUBDIRECCIÓN GENERAL DE INFORMACIÓN SANITARIA. Recuperado el 29 de julio de 2019, de https://www.mscbs.gob.es/estadEstudios/estadisticas/BarometroSanitario/Barom_Sanit_2017/BS2017_ma.pdf

CIS. (2019). *Estudio nº 8188: Barómetro Sanitario 2018.* MINISTERIO DE SANIDAD, CONSUMO Y BIENESTAR SOCIAL, DIRECCIÓN GENERAL DE SALUD PÚBLICA, CALIDAD E INNOVACIÓN, SUBDIRECCIÓN GENERAL DE INFORMACIÓN SANITARIA. Recuperado el 28 de julio de 2019, de https://www.mscbs.gob.es/estadEstudios/estadisticas/BarometroSanitario/Barom_Sanit_2018/BS2018_mar.pdf

Deepali, D., Arora, P., & Varshney, S. (2015). Analysis of K-Means and K-Medoids Algorithm For Big Dat. *Procedia Computer Science, 78*, pp. 507-512.

DGSFP. (2018). *Informe 2017. Seguros y Fondo de Pensiones.* Ministerio de Economía, Industria y Competitividad. Recuperado el 24 de junio de 2019, de http://www.dgsfp.mineco.es/es/Publicaciones/DocumentosPublicaciones/INFORME%20SECTOR%202017.pdf

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning, 29*(2-3), pp. 103-130.

Dunham, M. H., & Sridhar, S. (2006). *Data mining. Introductory and Advanced Topics* (1 ed.). Nueva Jersey: Pearson Education.

EUROPA PRESS. (24 de abril de 2010). Los problemas de la sanidad pública no están incluidos en los programas electorales. *El Confidencial*. Recuperado el 14 de julio de 2019, de https://www.elconfidencial.com/elecciones-generales/2019-04-24/programas-electorales-sanidad-publica_1961038/

Fundación IDIS. (2017). *Barómetro de la Sanidad Privada 2017.* Madrid. Recuperado el 11 de agosto de 2019, de https://www.fundacionidis.com/wp-content/informes/informe_barometro_idis2017_05.pdf

González Martínez, C. I., & Marqués Sevillano, J. M. (13 de noviembre de 2013). Las entidades de seguros ante el nuevo entorno financiero. *Revista de Estabilidad Económica del Banco de España*(25), pp. 127-138.

Imran, M., Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues, 9*(3), pp. 272-278.

Jin, X., & Han, J. (2017). K-Medoids -clustering. (C. Sammut, & G. I. Webb, Edits.) *Encyclopedia of Machine Learning and Data Mining*.

Kalmegh, S. (2015). Analysis of WEKA Data Mining Algorithm REPTree, Simple CART and Random Tree for Classification of Indian News. *International Journal of Innovative Science, Engineering and Technology, 2*(2), pp. 438-445.

Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New Jersey: Wiley-Intersciencie.

Kononenko, I. (1990). Comparison of inductive and naive Bayesian learning approaches to automatic knowledge adquisition. *Current trends in Knowledge Adquisition*.

Patil, T., & Sherekar, S. (2013). Performance Analysis of Naive Bayes and J48: Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications, 6*(2), pp. 256-261.

Peña Sánchez, I. (2010). Aplicación práctica de modelos de credibilidad en la tarificación de Seguros de Salud. *Gerencia de Riesgos y Seguros*(108), pp. 37-51. Recuperado el 25 de julio de 2019, de https://www.fundacionmapfre.org/documentacion/publico/i18n/catalogo_imagenes/grupo.cmd?path=1062014

Pérez Torres, J. L. (2011). Teoría General del Seguro. En J. L. Pérez Torres, *Teoría General del Seguro* (págs. pp. 369-386). Barcelona.

Planells Jalón, F. (2010). El seguro de Salud. Col·legi d'Actuaris de Catalunya. Barcelona: Umeser.

Quinlan, J. R. (1986). Induction of decision trees. *Kluwer Academic Publisher, 1*(1), pp. 81-106.

Recuerdo de los Santos, P. (16 de noviembre de 2017). Los 2 tipos de aprendizaje en Machine Learning: supervisado y no supervisado. *LUCA Telefonica data unit*. Recuperado el 13 de agosto de 2019, de https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence, 3*(22), pp. 41-46.

Rosales Estrada, E. M., & Guadarrama Tavira, E. (2015). Marketing relacional: valor, satisfacción, lealtad y retención del cliente. Análisis y reflexión teórica. *Ciencia y Sociedad, 40*(2), pp. 307-340.

Salzberg, S. (1994). Book Review: C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Kluwer Academic Publishers, 16*(3), pp. 235-240.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A Comparison of Document Clustering Techniques. *KDD workshop on text mining*.

Velmurugan, T., & Santhaman, T. (2010). Computational Complexity between K-Means and K-Medoids Clustering Algorithms. *Journal of Computer Science, 6*(3), pp. 363-368.

Vidal Meliá, C. (28 de noviembre de 2018). Apuntes Tema 4: Salud, de la asignatura Prestaciones y Seguros de Salud y Dependencia.

Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research, 41*(12), pp. 1955-1959.

Legislación

Directiva 2004/113/CE. (13 de diciembre de 2004). Diario Oficial de la Unión Europea. *Directiva 2004/113/CE, de 13 de diciembre, por la que se aplica el principio de igualdad de trato entre hombres y mujeres al acceso a bienes y servicios y su suministro*. Bruselas, Bélgica.

Ley 50/1980. (8 de octubre 1980). Boletín Oficial del Estado*. Ley 50/1980, de 17 de octubre, de Contrato de Seguro*. Madrid, España.

Ley Orgánica 03/2007. (22 de marzo de 2007). Boletín Oficial del Estado. *Ley Orgánica 03/2007, de 23 de marzo, para la igualdad efectiva de mujeres y hombres en materia de factores actuariales*. Madrid, España.

Ley Orgánica 03/2008. (5 de diciembre de 2018). Boletín Oficial del Estado. *Ley Orgánica 03/2008, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.* Madrid, España.

Real Decreto 1361/2007. (19 de octubre de 2007). Boletín Oficial del Estado. *Real Decreto 1361/2007, de 23 de octubre, por el que se modifica el Reglamento de ordenación y supervisión de los seguros privados en materia de supervisión del reaseguro, aprobado por el Real Decreto 2486/1998, de 20 de noviembre, en materia de supervisión del reaseguro, y de desarrollo de la Ley Orgánica 3/2007, de 22 de marzo, para la igualdad efectiva de mujeres y hombres, en materia de factores actuariales*. Madrid, España.

Apéndice 1. Tablas del grado de correlación

Tabla 43. Estadístico η2 para la correlación entre variables numéricas y nominales

η^2	Rentabilidad	Ramo	Parentesco	Sexo	Provincia
Prima	0.001	0.3502	0.1365	0.0009	0.008
Edad	0.002	0.2535	0.4204	0.0031	0.005
Visitas	0.274	0.0266	0.0189	0.0064	0.028
Ctevisitas	0.308	0.0027	0.0018	0.0070	0.006
Radiodiagnosis	0.162	0.0266	0.0876	0.0416	0.012
CteRadio	0.182	0.0270	0.0873	0.0211	0.005
AnClinicos	0.079	0.0163	0.0435	0.0094	0.008
OtrosDiagnosticos	0.076	0.0092	0.0417	0.0346	0.009
CteOtrosDiag	0.045	0.0059	0.0116	0.0020	0.001
ActosProfesionales	0.100	0.0484	0.0433	0.0010	0.008
CteActos	0.147	0.0203	0.0262	0.0001	0.004
Anestesia	0.175	0.0056	0.0114	0.0000	0.002
CteAnest	0.195	0.0034	0.0110	0.0002	0.002
Rehab	0.058	0.0093	0.0197	0.0011	0.002
CteRehab	0.056	0.0082	0.0180	0.0008	0.004
Protesis	0.043	0.0038	0.0049	0.0002	0.000
CteProtesis	0.020	0.0032	0.0026	0.0000	0.000
OtrosCptosFac	0.013	0.0004	0.0005	0.0000	0.002
CteOtrosCptoFac	0.012	0.0003	0.0006	0.0000	0.000
Otros	0.152	0.0081	0.0110	0.0000	0.001
CteOtros	0.180	0.0080	0.0115	0.0000	0.001
GTA	0.281	0.0044	0.0053	0.0014	0.001
Antigüedad	0.003	0.3334	0.0535	0.0002	0.027

Fuente: Elaboración propia

Tabla 44. Estadístico de Cramer para la correlación entre variables nominales

Estadístico de Cramer	Ramo	Sexo	Provincia	Rentabilidad	Parentesco
Ramo	1.000	0.0070	0.1330	0.0930	0.153
Sexo	0.007	1.0000	0.0260	0.0540	0.132
Provincia	0.133	0.0260	1.0000	0.0330	0.041
Rentabilidad	0.093	0.0540	0.0330	1.0000	0.093
Parentesco	0.153	0.1320	0.0410	0.0930	1.000

Fuente: Elaboración propia