



Facultad de Ciencias Empresariales

EL BIG DATA EN EL SECTOR ASEGURADOR

Clave: 201605760

EL BIG DATA EN EL SECTOR ASEGURADOR

ÍNDICE DE CONTENIDOS

1. INTRODUCCIÓN (Pág-4)

1.1 ¿Qué es el Big Data? (Pág-5)

1.2 Introducción al sector asegurador (Pág-8)

2. EL BIG DATA EN EL SECTOR ASEGURADOR (Pág-11)

3. TIPOS DE TÉCNICAS MULTIVARIANTES (Pág-16)

3.1 Técnicas de aprendizaje supervisado (Pág-16)

3.2 Técnicas de aprendizaje no supervisado (Pág-19)

4. APLICACIONES DEL BIG DATA AL SECTOR ASEGURADOR (Pág-21)

4.1 Aplicaciones a los seguros de automóviles (Pág-23)

4.2 Aplicaciones a los seguros de salud (Pág-28)

5. LOS LÍMITES DEL BIG DATA (Pág-32)

6. CONCLUSIONES (Pág-34)

7. REFERENCIAS BIBLIOGRÁFICAS (Pág-36)

RESUMEN

El presente trabajo pretende explicar las diferentes iniciativas de Big Data que se están implementando en el sector asegurador, tratando de explicar las diferentes técnicas que existen de análisis multivariante, diferenciando entre las técnicas de aprendizaje supervisado y las de aprendizaje no supervisado. Ambos tipos se explican brevemente, recordando los objetivos y metodología de cada técnica en concreto para poder ver más adelante las diferentes aplicaciones que tienen las mismas en los seguros de automóvil y en los seguros de salud. Lo que se pretende es crear un marco básico tanto como teórico como práctico para que el lector pueda entender las diferentes metodologías. Por último, se tratan los límites que presenta el uso del Big Data y cuáles son los extremos que pueden alcanzar las compañías aseguradoras en lo respectivo a estas herramientas y tratamiento de datos.

Palabras clave: Big Data, sector asegurador, análisis multivariante, aprendizaje supervisado, aprendizaje no supervisado, seguros de automóvil, seguros de salud.

ABSTRACT

The following project aims to explain the different Big Data initiatives that are being implemented in the insurance sector, trying to explain the different techniques that exist for multivariate analysis, differentiating between supervised learning and unsupervised learning techniques. Both types are explained briefly, remembering the objectives and methodology of each specific technique to be able to see later the different applications that they have in car insurance and health insurance. What is intended is to create a basic framework, both theoretical and practical, so that the reader can understand the different methodologies. Lastly, the limits presented by the use of Big Data are discussed and what are the extremes that insurance companies can reach regarding these tools and data processing.

Key words: Big Data, insurance sector, multivariate analysis, supervised learning, unsupervised learning, car insurance, health insurance.

1. Introducción

Lo que pretendo a través de la realización del trabajo es aumentar tanto mis conocimientos como los de las personas que puedan llegar a leerlo. Nunca antes he dado una asignatura relacionada con Big Data y con el sector asegurado por lo que ha supuesto un reto ya que he intentado aumentar mis conocimientos en estos campos de manera considerable. Se presentan conceptos sobre todo teóricos, sin entrar en la aplicación matemática o estadística de los diferentes técnicas de análisis multivariante que se exponen.

El trabajo se divide principalmente en cuatro grandes bloques. Primeramente, se hace una pequeña introducción, en la que se hace un repaso a los conceptos de los que va a tratar el trabajo como son el Big Data (sección 1.1) o el sector asegurador (sección 1.2). En el siguiente apartado se pretende hacer una síntesis y explicar lo que supone el Big Data pero ya dentro del sector asegurado (Capítulo 2). Seguidamente se procede a introducir las diferentes técnicas de análisis multivariante (Capítulo 3), dónde se hace distinción entre las técnicas de aprendizaje supervisado (Sección 3.1) y las de aprendizaje no supervisado (Sección 3.2). En el siguiente bloque se pretende explicar las diferentes aplicaciones que tiene tanto el Big Data como las técnicas de análisis multivariante al sector asegurador (Capítulo 4). Además, se pretende profundizar explicación la aplicación a sectores específicos como el seguro de automóvil (Sección 4.1) o a los seguros de salud (Sección 4.2). Seguidamente, se encuentra un apartado en el que se tratan los límites que puede alcanzar el Big Data (Capítulo 5). Para terminar, se presentan las conclusiones correspondientes (Capítulo 6) y las referencias bibliográficas utilizadas en la realización del trabajo (Capítulo 7).

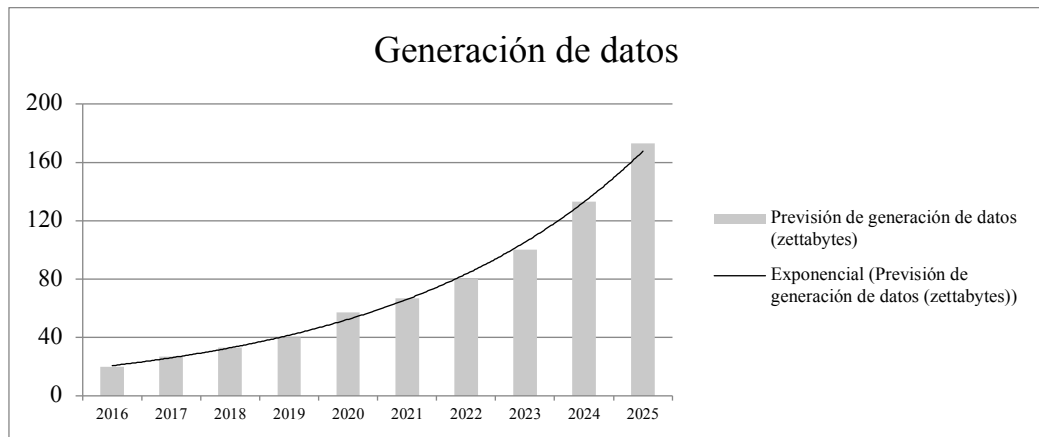
1.1 ¿Qué es el Big Data?

El concepto podemos decir que es dinámico porque los métodos y los sistemas utilizados para el análisis de datos van evolucionando y adoptando cada vez herramientas mucho más innovadoras. El término Big Data describe una nueva generación de tecnología y arquitectura diseñadas para extraer valor de manera económica de volúmenes muy grandes de una amplia variedad de datos, permitiendo la captura, el descubrimiento o el análisis a alta velocidad (Gantz y Reinsel, 2011). La primera referencia que se conoce de este concepto fue hace 23 años a través de Cox y Ellsworth (1997) cuando comenzaron a ver que la cantidad de información generada estaba comenzando a ser difícil de manejar con las herramientas tradicionales.

En el año 2001, D. Laney desarrolla el concepto y define las características del Big Data, realizando un informe técnico y creando el modelo de las tres “V’s”: Volumen, Variedad, Velocidad.

- Volumen: Como ya he indicado con anterioridad, lo principal del término son las enormes cantidades de datos que manejan ya que los volúmenes de datos almacenados crece de forma exponencial. La gran mayoría de las empresas en Estados Unidos tienen como mínimo cien mil Gigabytes de datos almacenados (IBM, 2018). Por ejemplo, en una hora de vuelo se generan 20TB de datos, Google procesa cada minuto un total de 2.400.000 búsquedas e intenta dar respuesta en 0,2 segundos o en Facebook que en 1 minuto, se llegan a publicar 500.000 comentarios o a subir 140.000 fotos. Cada individuo somos una fuente inagotable de datos, los generamos de manera inconsciente y los cedemos a empresas o instituciones. Según un artículo de “The Economist” (Cukier, 2010), la cantidad de información digital se multiplica por 10 cada 5 años, mientras que la Ley de Moore (Moore, 1965) indica que la capacidad de procesamiento se duplica cada 18 meses. Los datos crecen a un ritmo mayor que la capacidad que disponemos actualmente de procesarlos. Según la consultora Deloitte, desde el comienzo de la humanidad hasta 2003, se generaron 2 exabytes de información mientras que en 2011, se creó el mismo volumen en tan solo 2 días y en 2020, se tardará menos de 10 minutos. Según Anchen y Dave (2019), en el año 2025, los datos habrán

crecido globalmente un 27% respecto a los niveles actuales, hasta 175 Zettabytes, un tercio de los cuales serán en tiempo real (Gantz y Reinsel, 2018).



Elaboración propia. Fuente: Swiss Re Institute (Anchen y Dave, 2019)

- Variedad: la definición de dato, la encontramos en Davenport y Prusak (1998), quienes lo entienden como una representación simbólica (numérica, alfabética, algorítmica, etc...) de un atributo o variable cuantitativa o cualitativa. Dentro de los datos que se manejan cabe la distinción entre los datos estructurados y los no estructurados, entendiendo los primeros como aquellos que se encuentran organizados y pueden ser procesados y ordenados de una forma sencilla. Por otro lado, los no estructurados son aquellos que no están organizados y con una estructura difícil de identificar (generados por usuarios). El 80% de todos los datos son no estructurados, siendo el 20% restante los que provienen de sistemas tradicionales, correspondiéndose con los estructurados. Esta es la que se conoce como la “regla del ochenta por ciento”, siendo una regla comúnmente aceptada en el entorno del Big Data. La aparición de las redes sociales, y la nueva gran variedad y cantidad de datos no estructurados significa que las empresas se encuentran ante una nueva oportunidad de conseguir información extra acerca de aspectos tan importantes como sus productos, sus empleados o sus clientes entre otros (Mills y Forder, 2012).
- Velocidad: las empresas obtienen datos de forma cada vez más rápida, llegando a obtener datos en tiempo real gracias a los sistemas de conectividad y

geolocalización. Esta característica hace referencia también a la rapidez con la que se están creando los datos y la transmisión y agregación de los mismos.

El Big Data por tanto, se centra en la creación y en la extracción de información de grandes cantidades de datos que se encuentran disponibles, y en la aplicación de enfoques y métodos sofisticados en el análisis de los datos cuyo fin es ampliar el alcance del conocimiento y la obtención de diferentes puntos de vista, englobando tanto las arquitecturas de sistemas buscando nuevas formas de organizar la información y los algoritmos y herramientas que explotarán la información.

El Big Data sigue un proceso que suele ser muy general, basado principalmente en tres fases. Se centran primero en la identificación y el análisis de las fuentes internas y externas. Después, capturan y recopilan los datos que han adquirido a través de las fuentes y finalmente, se centran en el estudio y tratamiento de los datos. Como aspectos a destacar, el Big Data ayuda a entender y segmentar a los clientes que disponen las compañías que hacen uso de esta tecnología, ayudando a comprender sus comportamientos y preferencias con el objetivo principal de aumentar las ventas y la creación de modelos productivos. También se utiliza para entender y ayudar en la optimización de los procesos de negocio basándose en diferentes predicciones, destacando el estudio de la competencia que genera mucho valor para la empresa ya que se centra en el conocimiento avanzado de sus principales competidores. Finalmente, puede ayudar a cuantificar y optimizar no sólo a las empresas, sino al rendimiento personal de la sociedad. Más adelante se explicarán con detalle las aplicaciones directas que se pueden realizar con esta herramienta en el sector asegurador. Gracias al Big Data podemos medir, y por tanto, conocer profundamente mucho más sobre cualquier aspecto de la sociedad, los negocios y por supuesto, de los seguros que es en lo que voy a centrar mi estudio (McAfee y Brynjolfsson, 2012). La posibilidad de obtener datos relacionados con un evento en concreto está relacionada con la disponibilidad del dato, la capacidad para su almacenamiento y finalmente su procesado.

Respecto a los límites de la Inteligencia Artificial, la tecnología debe estar fundamentada en unos principios y valores. Según una entrevista reciente (Martínez, 2017), hay que destacar tres principios fundamentales:

- El propósito debe ser aumentar la inteligencia humana, pero siempre bajo nuestro control.
- La transparencia en cuanto a su propósito y el origen de sus datos.
- El compromiso por desarrollar el talento; es decir, las capacidades y el conocimiento humanos que se requerirán para trabajar con estos sistemas.

El Big Data para que pueda consolidarse y aportar valor a las empresas debe en primer lugar, realizar un gran esfuerzo de integración de los sistemas, dotarlos de flexibilidad y escalabilidad para una fácil adaptación a futuros requisitos. Por otro lado, deberán centrarse en facilitar la identificación de datos que aporten valor y sean correctos ya que acumular y procesar datos erróneos supondrá un tremendo desperdicio de recursos.

En la actualidad, ante el aumento del volumen y la variedad de los datos, se produce la posibilidad de realizar los análisis en tiempo real junto con la toma de decisiones. En un futuro cercano, los sistemas buscarán la eficacia mediante la aplicación de un análisis predictivo con el fin de dirigir las acciones que se encuentran derivadas de los datos.

1.2 Introducción al sector asegurador

Una compañía aseguradora es una empresa que se encarga de proteger de ciertos riesgos a terceros, de manera que se produce una protección de los bienes materiales ante los riesgos a los que los clientes se encuentran expuestos. También se encargan de resguardar los bienes intangibles como ocurre en los seguros de vida. A cambio de esta protección, las compañías exigen pagos periódicos (cuotas o primas) para que las posibles coberturas puedan tener lugar y cuyo importe viene determinado por el nivel de riesgo que calcule la compañía en función del bien y de los posibles daños o pérdidas que puedan sufrir. Las correspondientes cláusulas y condiciones entre el asegurado y el asegurador se pactan a través de un contrato que se conoce como póliza de seguros. La misión principal de las aseguradoras consiste en proteger desde la perspectiva económica aquellos bienes que pueden sufrir daños y que se encuentran bajo algún tipo de riesgo aunque este sea mínimo,

de modo que si se produce algún perjuicio sobre los bienes protegidos, será la compañía quien incurra en reparar o subsanar los daños provocados.

Según Fundación Mapfre (s.f.-c), el seguro es un medio para la cobertura de los riesgos al transferirlos a una aseguradora que se va a encargar de garantizar o indemnizar todo o parte del perjuicio producido por la aparición de determinadas situaciones accidentales. La industria aseguradora tiene una curiosa particularidad ya que desde sus inicios, se ha centrado en la venta de un mismo producto lo que podría llevarnos a pensar que es un sector que no ha avanzado ni se ha transformado. Sin embargo, la industria aseguradora está en constante cambio y muchas medidas y herramientas se están implementando a nivel mundial. La ciencia actuarial, que es una disciplina que para la evaluación de riesgos, siempre se ha centrado y ha requerido de la acumulación y procesamiento de datos para su correcto análisis, pudiendo realizar probabilidades de eventos o hechos que permitían la evolución de los riesgos. La calidad de un modelo se encuentra directamente relacionado con la cantidad de datos disponibles, por lo que el objetivo principal de cualquier compañía de este sector es la disposición de grandes cantidades de datos de calidad para la realización de sus modelos y generación de valor.

Las entidades ofrecen una amplia gama de seguros a sus clientes, tan distintos como los diferentes riesgos que cubren. Según Fundación Mapfre (s.f.-c), se puede hablar de tres tipos de seguros que son los más frecuentes en la práctica y se clasifican en función del objeto asegurado:

- Seguros personales: el objeto asegurado es la persona física, protegiendo al individuo de cualquier riesgo que le afecte directamente como puede ser el fallecimiento o su salud. Dentro de esta categoría existen los seguros de vida, de salud, de dependencia y de accidentes personales.
- Seguros de daños o patrimoniales: el objeto asegurado es el patrimonio de las personas y tienen como finalidad principal el reparar dichos daños, distinguiendo entre seguros de cosas (seguros de automóvil o multirriesgo) y los seguros de responsabilidad (seguros de robo, incendios y responsabilidad civil).

- Seguros de prestación de servicios: la obligación del asegurador en este caso consiste en la prestación de servicios al asegurado. Entre este tipo de seguros cabe destacar los seguros de defensa jurídica, asistencia en viajes y de decesos.

El sector asegurador se encuentra actualmente, como muchos otros, desarrollando una importante transformación en su negocio para así poder adaptarse y mantenerse en las tendencias actuales como son la digitalización, atención y servicio a clientes cada vez más exigentes y que demandan servicios más personalizados, o la adaptación a fenómenos de moda como la economía colaborativa o las nuevas compañías competidoras como las *Insurtech*. Hoy en día, la variable del precio es imprescindible para que ciertas ramas de seguros acaben con éxito como es el caso de los seguros de automóviles, hogar y salud. Según Flores (2017) de Pricing LAB, más del 85% de los españoles que contrataron en los últimos meses un seguro de autos, tenían varias ofertas, comparables según su criterio, antes de decidir, y que más del 56% de ellos acabó eligiendo la más barata de las que encontró. Aunque también es cierto, que más del 40% de dichos asegurados no eligieron la entidad más barata y por lo tanto, tomaron decisiones en base a algún tipo de diferencia percibida. La gestión de precios es de extrema complejidad ya que busca el equilibrio entre la competitividad y la solvencia técnica.

Actualmente, se puede ver como hay cambios clave de la era digital que están influyendo directamente y transformando a las compañías aseguradoras. El uso del móvil supone un cambio en el proceso de la cadena de valor de los seguros, siendo un canal ideal para ofrecer nuevos servicios y posible reclamo de siniestros. Por otro lado, los clientes cada vez se centran más en valorar a la hora de contratar un seguro en la opinión de las personas que tienen a su alrededor. Además las personas jóvenes, que se encuentran integradas en la era digital, demandan servicios cada vez más digitales lo que obliga a las compañías a reinventarse o perder clientes. A través del rastro digital que vamos dejando, se está produciendo el retorno de las pólizas a medida junto con un ajuste en los precios. La inmersión del sector en la utilización de grandes volúmenes de datos, el Big Data Analysis y las *Insurtech* obliga a no quedarse ajenas a esta nueva realidad y necesitan adaptarse rápidamente para poder sobrevivir a estas herramientas y tener éxito. Aquí sólo estamos hablando de factores que se encuentran íntimamente relacionados con lo digital, pero los aspectos socioculturales, el cambio de hábitos y el envejecimiento de la población también están configurando la industria aseguradora del futuro.

2. El Big Data en el sector asegurador

En los dos primeros apartados se ha realizado un primer acercamiento a los conceptos que se van a tratar. El principal objetivo del estudio es poder ver cuáles son los avances que está llevando a cabo el sector asegurador en el proceso de transformación digital, en concreto con el Big Data Analytics. También se presentan algunas reflexiones sobre lo que supone la incorporación de esta herramienta a las compañías aseguradoras.

En los últimos años, una gran cantidad de sectores se encuentran evolucionando y estamos presenciando una revolución tecnológica que está cambiando muchos aspectos de la vida como la forma de trabajar y relacionarnos, alterando también los modelos de negocio de las empresas tradicionales. El sector asegurador apenas había evolucionado lo que promete grandes cambios en un corto periodo de tiempo, dejando atrás la estandarización de la oferta actual. Nos encontramos en un periodo en el que ya no estamos ante una opción de adaptarse o no al Big Data sino que las empresas que mejor implementen esta herramienta tendrán un elemento diferenciador y por tanto, una ventaja competitiva. El ritmo al que están ocurriendo estos acontecimientos es cada vez mayor, lo que está llevando a plantear retos incesantes y se está produciendo un cambio en los paradigmas que se encontraban existentes hasta hoy. Este crecimiento ha obligado a que se tengan que integrar la ciencia de los datos en los nuevos modelos que se encuentra centrado en estudiar el sector asegurador junto con grandes fuentes de información disponibles. En un estudio reciente (Deloitte, 2016), para el año 2020 la cuota de mercado europea de las compañías aseguradoras que ofrezcan soluciones digitales alcanzará el 17%, y superará los 15 billones de euros de facturación.

Las compañías del sector asegurador siempre han intentado mantener una relación cercana con sus clientes y corredores como una forma de mantener su clientela frente a terceros por lo que ahora tienen en su mano, medios atractivos y beneficiosos que van a ayudar a lograr dicho cometido. Los clientes cuentan con un acceso mucho más fácil tanto a la tecnología como a la información, haciéndolos cada vez más digitales y permitiendo una mejor y eficaz toma de decisiones. Mediante las herramientas de Big Data se tendrá la posibilidad de acceder a seguros personalizados, a un costo razonable como nunca se han tenido. Además la clientela y la sociedad en general ha cambiado su mentalidad, sus hábitos y sus comportamientos por lo que se centran en los canales digitales para poder

llevar a cabo las tareas de una forma rápida, eficaz y poder ser en el largo plazo mucho más rentables. Las compañías deben aprovechar estas tendencias para no caer en la obsolescencia y no “matar a la innovación”. La implementación de estas medidas tiene que ser rápida ya que el mundo avanza cada vez más rápido y deben responder adecuadamente para así cumplir con las expectativas de los clientes.

El sector asegurador tiene muchos puntos en común con el sector financiero, y al igual que este, se encuentran bajo amenaza de entrada de nuevos competidores que se centran en la innovación y en los cambios culturales como está ocurriendo con las *Insurtech* y las *Fintech*. Estos dos términos se refieren al fenómeno de “*startups*” que están innovando utilizando la tecnología para mejorar tanto el modelo de negocio del sector asegurador como por otro lado, el del sector financiero respectivamente. Se están observando movimientos interesantes por parte de las aseguradoras como la creación de apps para los clientes o la venta directa de seguros pero esto puede cambiar rápidamente ya que el mundo avanza continuamente y aparecen nuevas herramientas y tendencias. Por ejemplo, Lemonade y Progressive son dos compañías que avanzan mediante la implementación de las nuevas tendencias y encuentran en la innovación la forma de subsanar sus procesos críticos. Hacen uso de algoritmos en vez de evaluadores de riesgos, utilizan *bots* conversacionales para la venta de seguros y hacen uso de la inteligencia artificial para el manejo de las reclamaciones.

Como hemos visto, la transformación digital es posible pero los modelos de negocio se tienen que adaptar al nuevo entorno. Las oportunidades son muchas pero la decisión de desarrollar estas herramientas puede ser difícil de llevar a cabo porque el avance tecnológico implica que las empresas adopten alteraciones y saquen ventajas de las nuevas tecnologías que ayuden a la expansión del negocio.

Nos encontramos en una era en la que la mayoría de las fuentes de información disponibles se encuentran digitalizadas. El mundo cambia constantemente y las organizaciones necesitan sacar ventajas de la información de la que disponen lo que supone descubrir nuevos aspectos que hasta ahora se pasaban desapercibidos. Las ventajas se focalizan en las compañías que se adapten rápidamente a los nuevos desafíos y sean capaces de saber manejar la gran cantidad de datos y la modelización predictiva ya que son imprescindibles y una oportunidad para canalizar la teoría del riesgo hacia la

predicción de posibles riesgos. La era del Big Data ha llegado para quedarse y el sector asegurador se ha visto involucrado en estos cambios radicales que demandan transformaciones profundas también a nivel organizacional.

Las aseguradoras están invirtiendo cada vez más tiempo y recursos en estas iniciativas tecnológicas aunque su inicio fue lento. IBM recomienda tres pasos que cualquier aseguradora debería tener en cuenta a la hora de incorporar Big Data a su negocio. El primer paso sería identificar los posibles casos de uso del Big Data en su negocio. El segundo, clasificarlos por la importancia y el valor que tendrían en la empresa, y por último, la creación de un “laboratorio de datos”, donde estudiar estos casos y validarlos antes de incorporarlos al negocio (Mills y Forder, 2012).

Según un estudio reciente (Anchen y Dave, 2019), la perspectiva a futuro es prometedora, ya que los analistas prevén que el gasto en datos y analítica en todas las industrias aumentará a una tasa de crecimiento anual compuesto del 13% en los próximos cuatro años, y animan a las aseguradoras a que sigan el ritmo. Sin embargo, la paciencia es fundamental debido a la complejidad inherente a la cadena de valor del seguro. Creen que a medida que un mayor número de aseguradoras busquen capacidades diferenciadoras, el actual desarrollo de infraestructuras, recursos y conocimientos específicos del sector ayudará a liberar todo el potencial de la analítica en el sector del seguro.

El aumento exponencial de las fuentes disponibles de datos ofrece una gran cantidad de oportunidades a este sector, que se centra en el análisis y el uso intensivo de la información. Actualmente, existen tecnologías como IBM Watson, que es una plataforma de inteligencia digital mediante la cual pueden analizar tanto los datos estructurados como los no estructurados, siendo estos últimos mucho más complejos (imágenes, textos o información de redes sociales). Estas herramientas son imprescindibles porque en las compañías aseguradoras se están acumulando gran cantidad de datos y ayudan a segmentar y adaptarlos a cada departamento. La forma en la que se analizan los datos hoy en día ha sufrido cambios como también ha ocurrido en la estructura interna del negocio. Ahora se invierten gran cantidad de recursos en el desarrollo de un plan estratégico basado en la cultura de los datos o más conocida en el sector como “data driven”. También se ha realizado una actualización de los sistemas informáticos en las compañías que estaban

quedando obsoletos, sustituyéndolos por los equipos de Big Data que son cada vez más comunes y utilizados en el sector asegurador.

La mayoría de aseguradoras de P&C (92% según una encuesta realizada en EE.UU.) ha previsto iniciativas sobre Big Data y analítica avanzada (Breeding y Pauli, 2018). Con este dato tan relevante podemos decir que muchas se encuentran familiarizadas con este término pero hay que puntualizar que muchas de ellas se encuentran en las primeras etapas de implementación de las bases de las iniciativas analíticas. Aunque el gasto en TI del sector se ha mantenido constante en los últimos años, los analistas prevén un reajuste dentro de los estáticos presupuestos conforme las aseguradoras completen las actualizaciones de los sistemas principales y asignen más fondos a iniciativas novedosas como las digitales y analíticas (Harris-Ferrante, 2018). Según un informe de McKinsey (Balasubramanian et al. 2017), los responsables de datos y analítica de aseguradoras globales dijeron que estaban invirtiendo en analítica de datos unos 80 millones de USD al año y la mayoría comentó que tenía previsto el gasto por lo que la envergadura de la inversión seguramente se irá incrementando a corto plazo. IDC prevé que el gasto en soluciones analíticas y Big Data en todas las industrias crecerá a una tasa de crecimiento anual compuesto del 13,2% en el periodo de 2018 a 2022, y alientan a las compañías aseguradoras a que sigan este ritmo y no se queden estancadas (IDC, 2019). Como es normal, las compañías aseguradoras más grandes y presentes en muchos lugares del mundo son las que más inversión realizan en este ámbito aunque siempre realizan inversión en proyectos pequeños ya que la gestión, probabilidad de éxito y obtención de beneficios en proyectos de escala muy grande puede ser incierta. El gasto de las aseguradoras en analítica podría ser superior a 20.000 millones de USD al año en todo el sector, representando alrededor del 15% del gasto en TI de las aseguradoras.

Finalmente, como ya he indicado con anterioridad, la digitalización y entender a la nueva clientela están creando nuevas oportunidades en el sector asegurador aunque son necesarios una serie de cambios significativos. En primer lugar, se requerirá hacer cambios en la estructura interna de la organización y adoptar nuevas competencias ya que el nuevo mundo en el que vivimos que está totalmente digitalizado, desafía a los modelos de negocio tradicionales. Por otro lado, situar al cliente como el eje central del negocio supondrá una ventaja competitiva para aquellas compañías que consigan identificar a aquellos que les aportan un mayor valor y les ayudará a satisfacer la demanda que se

encuentra basada en las interacciones electrónicas, mejorar la gestión de la experiencia del usuario e implementar nuevos y novedosos productos. Por último, la digitalización en su conjunto, siempre y cuando se implemente adecuadamente en la estrategia de las compañías, supondrá una ventaja competitiva a largo plazo.

3. Tipos de técnicas multivariantes para el Análisis del Big Data

La importancia del Big Data ya ha quedado clara en el anterior punto por lo que en esta sección se van a describir de una forma breve las técnicas de análisis multivariante que se utilizan en el ámbito asegurador. Se entiende el análisis multivariante como el conjunto de técnicas de análisis de datos que se utilizan en la actualidad y se encuentran en expansión. Estas técnicas nacen de lo que se conoce como Machine Learning, que se puede definir como aquellos métodos que se basan en la experiencia para la mejora de las predicciones. El uso de las herramientas informáticas colaboró a iniciar la era del análisis multivariante, con un número de técnicas que se pueden aplicar a multitud de situaciones en la actualidad

Conviene matizar primero la diferenciación entre qué es el aprendizaje supervisado (*supervised learning*) y el aprendizaje no supervisado (*unsupervised learning*). Se entiende el aprendizaje supervisado como aquel algoritmo capaz de aprender la relación entre unas variables de entrada (*inputs, features, X*) y unas de salida (*output, target, Y*). A partir de unos ejemplos etiquetados (conocemos input y output correctos), el algoritmo es capaz de predecir el valor del output para casos nuevos, no usados en el proceso de aprendizaje o entrenamiento. Por otro lado, en el aprendizaje no supervisado el algoritmo extrae conocimientos de los datos de entrada (*input*) sin decirle qué es lo que tiene que aprender. Se centra en encontrar regularidades en los datos de entrada y se encarga de descubrir patrones sin contar con objetivo.

3.1 Técnicas de aprendizaje supervisado

Estas técnicas hacen uso de datos históricos que se encuentran de una forma etiquetados y que son utilizados para realizar las predicciones. Suelen trabajar primero con una muestra de entrenamiento para así realizar un ajuste en el modelo. Con este modelo, se realizan predicciones en la otra información que se denomina muestra de prueba (datos no incluidos en la muestra de entrenamiento). Para realizar el aprendizaje de forma correcta, ambas muestras, han de tener la misma distribución de probabilidad y el principal objetivo es predecir en la muestra de prueba los patrones o comportamientos observados con la muestra de entrenamiento. Entre las diferentes técnicas de aprendizaje supervisado voy a tratar las redes neuronales, las máquinas de vectores de soporte (SVM),

la regresión logística, el modelo regresión múltiple, el análisis discriminante múltiple, los modelos de probabilidad, los árboles de decisión condicionales y el análisis multivariante de la varianza y covarianza.

Las redes neuronales pueden aplicarse tanto a tareas de clasificación como de regresión, consiguiendo predicciones muy buenas pero puede llegar a ser un algoritmo no interpretable. Nacen intentando “imitar” la estructura y el funcionamiento que siguen las neuronas y las conexiones del sistema nervioso. A partir de un conjunto de datos compuesto por patrones de entrada y salida, se busca que la red construya un modelo a partir del proceso generado por estas entradas y salidas. Estos modelos pueden ser de tres tipos: por corrección de error, por refuerzo y estocástico (Casanovas, Merigó y Torres, 2014). Los resultados que se obtienen son poco intuitivos y generan poca información aunque destacan por su capacidad para encontrar relaciones no lineales y su capacidad de aprender de dichas relaciones (Padilla-Barreto et al. 2017). Esta técnica es muy diferente a las técnicas que se van a detallar y que son más tradicionales tanto en la formulación del modelo como en los tipos de relaciones más complejos que se pueden formular con esta técnica.

Las máquinas de vector soporte (SVM) son un método de predicción alternativo y comúnmente utilizado en problemas de clasificación, análisis de regresión y detección de valores extremos (Padilla-Barreto et al. 2017). Sus principales limitaciones son el tiempo y recursos que consume entre la fase de entrenamiento y la de prueba (Karatzoglou, Meyer y Hornik, 2006). Estas máquinas aplican un método lineal simple al conjunto de datos pero en un espacio de alta dimensión que no está relacionado linealmente con el espacio de entrada. La popularidad de las máquinas de vector soporte se debe a la simplicidad junto con su rendimiento.

El modelo de regresión logística es un modelo lineal generalizado que se encuentra muy generalizado en la ciencia actuarial y cuyo estudio se centra en variables dicotómicas o de respuesta binaria. Se trata de un modelo muy conocido y resulta fácil de entender y explicar. Así mismo, resulta bastante informativo dado que a-priori aporta pistas sobre la influencia de las variables en la modelización, lo cual es ideal al momento de decidir la configuración final del modelo (Padilla-Barreto et al. 2017). Su forma funcional y la dificultad de extrapolar los resultados en caso de sobreajuste, son dos de las principales limitaciones de este modelo.

El modelo de regresión múltiple es el método utilizado cuando en el problema del investigador se incluye una única variable métrica dependiente suponiendo que se encuentra relacionada con una o más variables métricas independientes. Siempre y cuando el investigador esté interesado en predecir la cantidad de la variable dependiente, este método será útil. El objetivo de este análisis se consigue a través de la regla estadística de los mínimos cuadrados y consiste en predecir los cambios que sufrirá la variable dependiente si se producen cambios en las variables independientes.

El análisis discriminante múltiple es el método de análisis apropiado si la única variable dependiente es dicotómica, multidicotómica y no métrica. Este análisis resulta útil cuando la muestra puede subdividirse en grupos basados en una variable dependiente. Los objetivos de este análisis se centran en entender las diferencias que ofrecen los grupos y poder predecir la verosimilitud de que alguna persona u objeto pertenezca a uno en particular basándose en variables métricas independientes.

Los modelos de probabilidad lineal, más conocidos como *análisis logit*, consisten en una combinación entre el análisis discriminante múltiple y el modelo de regresión múltiple. Se trata de una mezcla ya que la variable dependiente en este caso es no métrica, como ocurre con el análisis discriminante y al contrario del modelo de regresión múltiple aunque puede llegar a ser similar ya que se utilizan varias variables independientes para predecir una única variable dependiente binaria. Estos modelos se diferencian del análisis discriminante en que se acomodan variables independientes métricas y no métricas y no se requiere de la suposición de normalidad multivariante. Sin embargo, cuando se pretende analizar una variable dependiente con más de dos niveles, el análisis discriminante se posiciona como la técnica más apropiada.

Los árboles condicionales (CTREE) son una clase de árboles de decisión caracterizados por ser poderosas herramientas de clasificación y visualización, además son útiles en situaciones para agrupar a segmentos de clientes, toma de decisiones o la identificación de características de grupos. En este tipo de árbol de decisión, se seleccionan variables en dos fases. Primero se estudia la dependencia entre la variable objetivo y las variables independientes planteando una hipótesis global de independencia en términos de m hipótesis parciales (Padilla-Barreto et al. 2017). Si no se puede rechazar la hipótesis planteada, se detiene el proceso recursivo. En contraposición, si la hipótesis global de independencia es rechazada, el siguiente paso es medir el nivel de asociación entre la

variable dependiente y cada una de las variables explicativas, lo cual permite generar nuevas divisiones del árbol de manera secuencial (Padilla-Barreto et al. 2017). Son aplicables a todo tipo de problemas de regresión incluidas las variables de respuesta nominal, ordinal, numérica, así como variables de respuesta multivariante (Hothorn et al. 2006).

El análisis multivariante de la varianza (MANOVA) es una técnica estadística que representa una extensión del análisis univariante de la varianza (ANOVA) y es usada para explorar similitudes y relaciones entre categorías de variables independientes con dos o más variables métricas dependientes. Por otro lado, el análisis multivariante de la covarianza (MANCOVA) resulta de utilidad cuando se diseña una situación experimental y permitir la comprobación de hipótesis correspondientes con la varianza de respuesta de grupos. Este análisis se puede usar conjuntamente con el análisis de la varianza para eliminar el efecto de cualquier variable independiente que no se encuentre controlada sobre las variables dependientes.

3.2 Técnicas de aprendizaje no supervisado

Como he explicado con anterioridad, estos métodos son utilizados para agrupar o clasificar elementos de datos históricos. Se suelen utilizar en situaciones en las que se dispone de mucha información, con gran cantidad de variables que están correlacionadas entre sí en mayor o menos grado (Aldás y Uriel, 2017). Entre las diferentes técnicas de aprendizaje no supervisado voy a tratar el análisis *cluster*, el análisis de componentes principales, el análisis factorial común, el análisis multidimensional y el análisis de correspondencias

El análisis *cluster* es una técnica estadística multivariante que se utiliza para desarrollar subgrupos significativos de individuos u objetos y cuyo principal objetivo es clasificar una muestra en un número pequeño de grupos mutuamente excluyentes pero basados en similitudes. En este análisis los grupos no se encuentran predefinidos como ocurría con el análisis discriminante por lo que se usa la técnica para la identificación de los grupos. Para llevar a cabo este análisis hay que realizar al menos dos etapas. Una primera, dónde se trata de determinar cuántos grupos existen en la muestra en base a similitudes o asociaciones. La segunda etapa consiste en describir a las personas para determinar su

composición, pudiéndose llevarse a cabo mediante la aplicación del análisis discriminante a los grupos que se han identificado con el análisis *cluster*.

El análisis factorial, que engloba tanto al análisis de componentes principales como al análisis factorial común, es un método estadístico utilizado para estudiar las similitudes y relaciones que existen entre un gran número de variables y explicar dichas variables en función de sus dimensiones comunes (factores). El objetivo final de este análisis es un modo de resumir y condensar la información contenida en las variables originales y convertirlas en un conjunto más pequeño de variables (factores) con una pérdida poco significativa de información en las mismas. Son técnicas utilizadas de una manera general para simplificar bases de datos.

El análisis multidimensional es una técnica que se centra en la reducción de variables y de dimensiones cuya finalidad es transformar los juicios de los consumidores de similitud en distancias representadas en un espacio multidimensional. Aquí los factores que denominábamos en el análisis factorial se entienden como dimensiones y los datos provienen de la comparación entre pares de variables.

Finalmente, el análisis de correspondencias es una técnica que se ha desarrollado recientemente y ayuda a facilitar la reducción dimensional de una clasificación de objetos sobre un conjunto de atributos y el mapa perceptual de objetos relativos a estos atributos. Esta técnica destaca por su capacidad para acomodar tanto datos no métricos como relaciones no lineales, proporcionando una representación multivariante de datos no métricos que no es posible realizar con otra metodología. El análisis emplea una tabla de contingencia con las dos variables categóricas para después transformar dichos datos no métricos en un nivel métrico y realizar una reducción de los mismos, junto con un mapa perceptual.

4. Aplicaciones del Big Data en el sector asegurador

El Big Data es una herramienta que se está implantando muy rápidamente en compañías de multitud de sectores. En este apartado se explicarán las aplicaciones directas de dichas herramientas pero al sector asegurador en particular, diferenciando de los usos generales que se puede hacer de estas herramientas que fue explicado en la introducción al Big Data.

Como primera acción, se puede destacar el análisis predictivo que consiste en captar diferentes flujos de datos disponibles y realizar un análisis mediante el cruce de los datos. A través de este análisis, las compañías obtienen una visión mucho más amplia y definida del consumidor, permitiendo la mejora de las acciones publicitarias y de marketing automatizado dirigidas a segmentos determinados de la población. El objetivo que se pretende conseguir mediante este cruce de datos es llegar a conocer de alguna forma el comportamiento del consumidor, lo que generará ventajas para la compañía como anticiparse ante sus necesidades y enviarles estímulos en momentos indicados y correctos.

Hoy, los clientes se han vuelto mucho más exigentes y conocen mucho más en profundidad el sector asegurador lo que provoca que las compañías han de ser innovadoras en sus productos junto con precios competitivos. Para ayudar ante esta circunstancia, el Big Data, se centra en ofrecer a los clientes de las compañías experiencias personalizadas en función de la información que poseen de los mismos como poner sugerencias en función de sus búsquedas. Muy relacionado con esto se encuentra el real-time marketing que permite también una personalización hacia el cliente con la aparición de anuncios íntimamente relacionados con los gustos de los usuarios que navegan en sus páginas web. Por otro lado, el Big Data puede ayudar también a crear valor para el cliente ya que las aseguradoras quieren mejorar la relación con sus clientes personalizándolo al máximo lo que permite detectar las debilidades en su trato y así poder actuar en consecuencia.

Otro de los usos más importantes es ayudar a las compañías aseguradoras a mejorar en la detección del fraude, que supone una gran amenaza debido a su naturaleza cambiante. Esto produce mucho gasto innecesario por parte del sector por lo que el análisis de datos

y el uso de modelos predictivos será fundamental para reducir costes. Para ayudar en dicho cometido, una de las opciones consiste en cruzar bases de datos de antiguas reclamaciones con las actuales para ver si existen coincidencias. Como con cualquier modelo predictivo, cuántas más variables se manejen, más opciones de éxito y probabilidad de detección del fraude. Tampoco hay que centrar únicamente dicho análisis en las pólizas o reclamaciones, sino también en las redes sociales, su entorno y sus comportamientos.

La gestión de riesgos también es considerada como uno de los principales ámbitos dónde el Big Data podría aportar un mayor valor a las compañías, siendo trascendental en una aseguradora debido a la entrada en vigor en 2016 de la normativa Solvencia II y ayuda en la determinación de las primas que pagan los clientes. Muchas compañías están beneficiándose de los datos telemáticos y de internet de las cosas para predecir y calcular los riesgos mediante el uso de modelos predictivos y utilizando tanto datos de comportamiento de los clientes como fuentes de datos exógenas.

La tarificación y comercialización de los seguros toma gran importancia, mediante la intención de retener a los clientes de menor riesgo junto con la asignación de precios competitivos para ellos. Todas las compañías tienen contratados a especialistas en tarificación ya que actualmente es muy difícil conseguir nuevos clientes y el objetivo es intentar retenerlos en la compañía. El objetivo de toda aseguradora consiste en la aplicación de precios bajos que resulten interesantes para el mercado pero a la vez lo suficientemente altos para garantizar alcanzar los capitales mínimos exigidos y con ello el beneficio pretendido. Según un estudio reciente (Deloitte, 2016), en los últimos años han nacido varias aseguradoras digitales que distribuyen su aún estandarizada oferta a través de medios online, siendo su principal elemento diferenciador el precio. Esta estrategia adoptada por las pequeñas compañías está llevando a las grandes a tener que ir ajustando sus ofertas.

A continuación, se realizará una explicación de las aplicaciones que tiene el Big Data tanto al seguro del automóvil como al seguro de salud de las personas.

4.1 Aplicaciones a los seguros de automóviles

El seguro de automóvil se encuentra clasificado dentro de los seguros de daños o patrimoniales, ya que el asegurado es un objeto patrimonio de las personas por lo que se enmarca dentro de los seguros de cosas. Este seguro tiene como finalidad principal el reparar los daños que se puedan causar en el vehículo.

Como se busca en toda la industria, se debe romper con la oferta homogénea pero requiriendo un cambio diferenciador en los servicios junto con un cambio en el canal de comercialización. Hoy en día, los avances tecnológicos son tangibles y han permitido que se puedan conectar los dispositivos móviles al vehículo lo que ayuda a la obtención de datos de conducción en tiempo real. A través de esta tecnología inteligente o Bluetooth se puede controlar la velocidad que adoptan los clientes, si realizan muchos acelerones o frenazos bruscos, etc...

Actualmente las primas de los seguros de automóvil son calculadas en base a parámetros como la edad del tomador, siendo más altas las primas si se encuentran entre los 17 y 25 años. También se tiene en cuenta el lugar de residencia, la ocupación del tomador, el uso y tipo de vehículo a asegurar, el tipo de cobertura que se pretende contratar, si va a haber más conductores del mismo. Por otro lado, el género del conductor habitual del vehículo no se tiene en consideración para el cálculo de la prima. Si las compañías tuvieran acceso a los datos que se obtienen gracias a la innovación tecnológica, se acabaría con las pólizas genéricas y de los factores que se utilizan hoy en día y se pasaría a calcular las primas de las pólizas mediante la personalización y la realización a medida de las mismas. En estas pólizas a medida se trataría de recompensar a los conductores que tienen unos hábitos y un estilo óptimo de conducción mientras que se centrarían en poner precios más altos y penalizaciones a las personas que tengan comportamientos inadecuados o violentos al volante.

Por otro lado, está el dilema ético de si los clientes serían capaces de ceder su historial de conducción a las compañías. Según un estudio de la consultora Deloitte (Oliva y Flores, 2017), ante la posibilidad de ceder los datos y hábitos de conducción a las compañías aseguradoras, son los “jóvenes innovadores” y los “adultos exploradores” los que más dispuestos estarían a compartir sus datos y cambiar de modelo asegurador. Lo que puede

llegar a generar más incertidumbre a los clientes es si los datos serían utilizados para controlar o para adecuar las tarifas a la forma de conducción del cliente. No obstante, aún existen ciertas dudas por parte de los clientes que se muestran reticentes ante la cesión de los datos a externos.

Los clientes cada vez se encuentran más concienciados de que nos encontramos en una era digital y que los datos suponen un gran valor, por lo que esperan poder obtener a cambio de su cesión ciertos beneficios. Según Deloitte (2016), cabe destacar que los ciudadanos españoles somos los que más dispuestos estaríamos a ceder los datos relativos a la conducción, junto con Bélgica, Irlanda y Polonia. En base a esta misma encuesta, un 58% del total de los encuestados ha declarado que la principal razón del temor a compartir la información es la privacidad (el 4% no la compartiría por miedo a ser facilitada a la policía). Esto es un aspecto de suma importancia que será tratado de forma individual para ver los límites que puede alcanzar el Big Data en este sector porque los datos podrían perderse, ser robados, ser vendidos a terceros o utilizada de forma distinta a lo acordado con el cliente. El intercambio de información del que se habla plantea numerosas cuestiones éticas que deben ser garantizadas por parte de la compañía. Ante un fallo en este sentido, los clientes podrían no confiar en el modelo y como consecuencia, podría no funcionar esta prometedora medida.

Otro de los avances que se han realizado en los seguros de automóvil viene muy ligado a lo anterior, teniendo que ver con la generación de datos telemáticos pero en este caso mediante la implantación de un GPS que registra las distancias recorridas y formas de conducir, al igual que la anterior. Su utilización aporta elementos de prevención de accidentes, así como identificación de segmentos de conductores con comportamientos potencialmente peligrosos (Guillén, 2016). Se obtienen diversidad de datos como velocidades máximas, medias, los periodos de descanso que realiza en los viajes, horario de conducción, día de la semana o tipo de vía por el que circula. Aquí aparece el seguro de automóvil que se encuentra basado en el uso (*“pay as you drive”*), en el que los conductores más activos suelen pagar pólizas mayores por lo que esto se puede controlar gracias a los datos telemáticos. A parte de estos datos que se obtienen con estas herramientas telemáticas, las compañías ya conocen la edad de los conductores, su sexo, la antigüedad de su permiso de conducir, la marca y modelo del vehículo del que forma parte el seguro y el número de siniestros en los que se ha encontrado inmerso. A parte del

tiempo que pasen al volante, los hábitos generales de conducción también se tienen en cuenta. Todos estos factores que he explicado suponen una posible explicación de la probabilidad de que el conductor se vea inmerso en un accidente, lo que generará daños en el vehículo. También se puede hacer uso de modelos predictivos para identificar si hay mucha probabilidad de que sus clientes sufran un accidente comprobando sus datos personales con datos exógenos. Muchas de las compañías aseguradoras ofrecen cantidad de ventajas como seguros extra o descuentos para sus clientes si estos ceden información personal ya que necesitan profundizar en la tecnología portátil y posteriormente, aplicarlo en su negocio. También existen aplicaciones para el móvil que son propiedad de las aseguradoras que evalúan la conducción del cliente y le aconsejan sobre los malos hábitos (Exastax, 2017). Estas aplicaciones móviles pueden ayudar también a la mejora de la relación con los clientes y la retención y fidelización de los mismos en la compañía.

Hoy en día, muchas compañías de seguros del mundo ya comercializan contratos basados en la telemetría, pero dada la introducción tan reciente de este sistema, este es todavía escasamente conocido y su fundamento técnico no ha sido investigado como merece (Guillén, 2016). En definitiva, el avance de la telemetría constituye una verdadera revolución en la manera de abordar el precio de los seguros, de forma que en el futuro en lugar de valorar el tiempo como factor de riesgo, se considerará la distancia recorrida como la verdadera variable de exposición al riesgo y, además, se tendrán en cuenta los hábitos de conducción, mantenimiento del vehículo y prevenciones particulares de cada asegurado (Guillén, 2016).

En cuanto al uso de las técnicas de análisis multivariante de aprendizaje supervisado en los seguros de automóvil, los modelos de regresión lineales sirven para analizar si va a ocurrir o no un siniestro pudiendo ver la probabilidad del cliente de verse involucrado en un accidente, modelizar el número de siniestros que puede llegar a haber y su coste asociado, y también para analizar aspectos de fidelización tratando de conocer si un cliente tiene riesgo de ir a otra compañía o no. El modelo de regresión múltiple sirve para analizar las variables que influyen en la siniestralidad de las diferentes pólizas y predecir el precio de la misma para cada asegurado y compañía. El análisis discriminante múltiple sirve para clasificar la cartera de clientes en diferentes clases y así poder hacer estudios de siniestralidad. Según Lozano (2015), las redes neuronales son utilizadas como modelos de predicción de insolvencia, análisis de detección de fraude, modelos de clasificación de

riesgo y asegurados, análisis de la demanda, riesgo de suscripción de pólizas para coches y como modelo de predicción de daños. Las máquinas de vector soporte (SVM) sirven para la detección de sucesos raros como fingir el daño al vehículo y también para solucionar problemas de clasificación de nuevos clientes y su distribución de riesgo subyacente (Lozano, 2015). La clasificación de nuevos clientes también se realiza con los árboles de decisión condicionales (CTREE), considerándose poderosas herramientas de clasificación (Guelman et al., 2014). Para terminar las técnicas de aprendizaje supervisado, la regresión logística es empleada para la ocurrencia o no de un evento como si se va a producir o no una reclamación, si el asegurado va a cambiar de compañía y también para la detección de posibles fraudes.

A continuación, profundizaré en explicar los usos que se puede hacer de las técnicas de aprendizaje no supervisado. El análisis *cluster* se utiliza para realizar clasificaciones dentro de una muestra por lo que es útil en la identificación de factores que influyen en el comportamiento del asegurado y ayuda a las compañías a ver su perfil. También es utilizado para poder ver todas las tipologías de impacto que pueden darse en una colisión (González, 2015). El análisis de componentes principales funcionales sirve para asignar primas dentro de este sector. El análisis factorial común sirve como técnica de reducción de datos y para su agrupación en grupos homogéneos de variables como puede ser clasificar la gente que tiene permiso de conducción entre si es de ciclomotor o de coche. El análisis multidimensional permite a las aseguradoras integrar los datos que provienen de los sensores que se encuentran conectados en los vehículos. Como última herramienta de aprendizaje no supervisado, el análisis de correspondencias puede ayudar a las compañías a ver determinados riesgos o posibles accidentes estableciendo relaciones entre los accidentes que se producen y los comportamientos al volante de los conductores.

Finalmente, destacar el proceso de tarificación en los seguros del automóvil. El principal objetivo de los sistemas de tarificación es la obtención y fijación de primas, ajustándose a los principios de invariabilidad, suficiencia, equidad y indivisibilidad siempre de acuerdo con el marco legal. Existen los sistemas de tarificación a priori y a posteriori pero en este caso, las herramientas de Big Data sólo sirven para el primer sistema de tarificación, también conocido como “*class-rating*”. Este sistema permite asignar una prima a un nuevo riesgo basándose en unas características o datos que se disponen de dicho riesgo, sin tener experiencia sobre la siniestralidad de dicho riesgo. Únicamente

conociendo determinadas características de la póliza determinaremos su prima asignándole una siniestralidad esperada (Boj, Claramunt y Costa, 2017). Cumpliendo con el principio de equidad, cada asegurado debe hacer frente a pagar un importe de prima que sea proporcional al riesgo en el que incurre para la compañía de seguros.

Según Boj, Claramunt y Costa (2017), en el proceso de tarificación se pueden distinguir tres fases diferenciadas, más una inicial:

- En esta fase inicial, se busca obtener los datos iniciales y realizar un proceso de homogeneización y depuración de los datos ya que la aplicación de un método estadístico-matemático a datos de poca calidad supondrá una pérdida de tiempo.
- En la primera fase se realiza una selección de las variables tarificadoras y de las clases de tarifa pudiendo hacer referencia tanto a las características del objeto asegurado como a características del asegurado, del tomador, etc...
- La segunda fase se centra en la obtención de los grupos de tarifa junto con la formación de grupos homogéneos de riesgo (exclusivos y exhaustivos), a partir de las clases de tarifa.
- Finalmente y como tercera fase, se procederá a la estimación de las primas.

Para la tarificación a priori se emplean técnicas de análisis multivariante, pudiendo servir alguna para todas las fases ya que estas fases no son independientes. En la fase 1, se utilizan los Análisis *cluster*, análisis discriminante y técnicas de segmentación. Adicionalmente, se utiliza la regresión lineal y modelos lineales generales. Para la fase 3 también se utilizan los modelos lineales generalizados como el método de Bailey y Simon o el método de los mínimos cuadrados. A la hora de estimar las primas pueden establecerse estructuras aditivas o multiplicativas, de forma que la prima correspondiente a un grupo se calcule como suma de los efectos de pertenecer a cada una de las clases o como producto (Boj, Claramunt y Costa, 2017). Algunos de los modelos que se acaban de exponer se aplican sobre datos con clasificación cruzada.

4.2 Aplicaciones a los seguros de salud

Los seguros de salud para personas son clasificados dentro de los seguros personales, siendo el objeto asegurado la persona física, protegiendo al individuo en este caso de cualquier enfermedad o problema de salud que pueda sufrir. El campo de la biomedicina es uno de los sectores dónde se han observado mayores oportunidades del uso del Big Data, pasando de costar 1.000€ la secuencia genética junto con la lectura del genoma a 100€ en un periodo de cinco años.

Uno de los objetivos principales de las compañías consiste en conseguir métodos efectivos de tarificación y una de las aportaciones del Big Data puede ser aplicada sobre este objetivo en el contexto de los seguros de salud. El objetivo es la posibilidad de mejorar los precios de las coberturas privadas para las personas mayores mediante el uso del principio de diversificación. Según Guillén (2016), en este caso se trata de lograr una evaluación de la longevidad conjuntamente con la predicción del número de servicios médicos demandados por el paciente, en función de las características personales y del histórico de uso de los servicios registrados. En la mayoría de los casos para el cálculo de la prima se utilizan variables como la edad de los clientes, el género y su estado de salud en el momento de contratación de la póliza.

Cuando una persona acude al médico porque presenta síntomas o se enferma ya se está produciendo una reclamación por su parte para que la póliza les ofrezca cobertura. Ante la presencia de una enfermedad, la probabilidad de supervivencia disminuye ya que supone una reducción de las condiciones de salud y mediante la aplicación de un tratamiento médico vuelven a aumentarse las posibilidades de sobrevivir. Por tanto, se produce un efecto de disminución de las condiciones de salud y de aumento de las perspectivas de supervivencia sin que se sepa cuál de los dos efectos es mayor (Guillén, 2016).

Si la evaluación de riesgo para calcular la prima es proporcional a la demanda de servicios médicos multiplicado por la probabilidad de sobrevivir todo el año, la aproximación de modelo conjunto puede proporcionar una herramienta dinámica para generar una evaluación de riesgo dinámica y personalizada, es decir continua e individualizada, con la que fijar precios competitivos (Guillén, 2016). Lo que saco en claro es que la

transformación que se está produciendo requiere de un cambio de enfoque de los métodos mediante la incorporación de un mayor número de variables para que la tarificación sea más precisa. En síntesis, la abundancia de datos y la modelización predictiva constituyen una oportunidad para canalizar la teoría del riesgo hasta la predicción de pérdidas (Guillén, 2016).

La consultora Capgemini (2020), ha realizado un informe en el que identifica las 8 tendencias que se van a desarrollar en este 2020 relacionadas con los seguros de salud:

- Presión por parte de los mercados para reestructurar la cadena de valor debido a las nuevas tecnologías, los crecientes costes sanitarios, las preocupaciones y expectativas de consumidores y la entrada de nuevos actores.
- Aumento de la transparencia para ganar la confianza de los clientes. Los clientes analizan cada vez con más exhaustividad las diferentes opciones con las que cuentan y ayudaría a la satisfacción de los mismos.
- Experiencias 360° para los clientes con servicios que aportan valor añadido. En este sentido, el canal móvil resultará vital para las aseguradoras.
- Inclusión de *chatbots* para ayudar y guiar a los clientes ya que se trata de un programa que simula el mantenimiento de una conversación con el cliente al proveer respuestas automáticas.
- Utilización de plataformas digitales que impulsan la comunicación usuario-cliente y ayudar a medir diversos parámetros de salud para así ofrecer un servicio más completo y personalizado.
- Análisis predictivo para identificar y elaborar perfiles de riesgo gracias a la mayor predisposición de los consumidores a compartir sus datos personales.
- Uso de la automatización para optimizar las gestiones y realizar el pago de indemnizaciones en tiempo real para la mejora de la experiencia del usuario.
- Aumento de la interoperabilidad gracias a la gestión de las APIs, que son interfaces de programación de aplicaciones que permiten comunicación entre diferentes productos y servicios.

Otra aplicación que se puede hacer a este tipo de seguros es la utilización de datos telemáticos o la tecnología portátil ya que a través de los rastreadores de actividad se puede llegar a controlar los hábitos, actividad física y comportamiento de los asegurados.

Estas herramientas pueden ayudar a las compañías tanto a determinar el precio de las primas de la póliza y ayudar en la gestión de riesgos.

Por otro lado, en los seguros de salud se están implementando estos dispositivos también para mejorar la relación con los clientes. Por ejemplo, se están utilizando dispositivos para la prevención de la diabetes lo que significa ofrecer un servicio al cliente.

En cuanto al uso de las técnicas de análisis multivariante de aprendizaje supervisado en los seguros de salud, los modelos de regresión lineales permiten construir sistemas de tarificación sostenibles en el tiempo que incorporan un conjunto de variables que pueden predecir el riesgo biométrico con suficiente fortaleza estadística (Rodríguez-Pardo, 2012). También permite calcular los efectos económicos que supondrá una determinada consulta o problema a la entidad aseguradora. El modelo de regresión múltiple sirve para analizar las variables que influyen en la salud y en la probabilidad de fallecimiento de los individuos, sirviendo así mismo para el cálculo de las primas al identificar las diferentes variables que influyen. El análisis discriminante múltiple al igual que en los seguros de auto, sirve para clasificar en diferentes grupos a la cartera de clientes y así poder realizar estudios en función de personas con características homogéneas. Las redes neuronales permiten analizar los posibles fraudes, modelos de clasificación en función de la salud de los clientes y como modelo de predicción de gastos en los que pueda incurrir el asegurado. Las máquinas de vector soporte (SVM) son utilizadas para detectar cosas extrañas que sean difíciles de clasificar como puede ser un cliente con muchas patologías o una enfermedad rara. También pueden ser utilizadas para ver si un paciente tiene posibilidades de sobrevivir ante una situación clínica complicada. Los árboles de decisión condicionales también sirven para clasificación de nuevos clientes como se ha hecho ver en los seguros de automóvil. Para finalizar con estas técnicas, la regresión logística es empleada para hacer estimaciones de qué individuos con determinadas patologías sobrevivirán a una determinada fecha junto con una estimación de su posible fallecimiento. También se utiliza para ver si se va a producir un determinado servicio de salud o no y si el asegurado va a cambiar o no seguro de salud.

Finalmente, procedo a explicar los usos que se puede hacer de las técnicas de aprendizaje no supervisado en los seguros de salud. El análisis *cluster* se utiliza para ver y estudiar la totalidad de gastos en los que incurren las compañías en este tipo de seguros. Así mismo,

ayuda a identificar casos extraños a las compañías y las clasificaciones dentro de una muestra sean más precisas. El análisis de componentes principales al igual que el análisis factorial común sirven para realizar una reducción de los datos médicos que manejan las compañías y así hacer más fácil su análisis para agruparlos en grupos homogéneos. Estos grupos homogéneos pueden diferenciarse entre si tienen un determinado nivel arterial u otro o el peso de los clientes. El análisis multidimensional permite a las aseguradoras la integración de los datos que reciben a través de los *chatboxs* o dispositivos digitales utilizados para medir determinadas variables médicas en pacientes. Por último, el análisis de correspondencias sirve para analizar las relaciones que existen entre unos determinados médicos con las enfermedades, pudiendo ayudar a establecer de forma más precisa la probabilidad de sufrir problemas de salud o no.

5. Los límites del Big Data

Hoy en día, la sociedad y el mundo en general tiene muy presente la normativa sobre la protección de datos y los límites que tienen las compañías a la hora del tratamiento de la información. El uso intensivo de redes sociales y de dispositivos electrónicos está provocando que las personas dejen una huella digital que puede llegar a ser perjudicial para su futuro ya que son los datos utilizados por los algoritmos para crear modelos y detectar patrones y comportamientos. Se han producido numerosos escándalos por el uso y tratamiento de la información disponible de las personas de forma ilegal por lo que en el año 2016, entró en vigor el Reglamento General de protección de datos. Al igual que con las nuevas herramientas y procedimientos para la gestión de riesgos, las aseguradoras deben mantenerse al día e informados sobre todo con el impacto que puede hacer sobre los consumidores y el uso de sus datos personales. Las aseguradoras deberán hacer un seguimiento y adaptarse a cambios regulatorios en lo relacionado con el uso de nuevas fuentes de datos y analítica, especialmente en suscripción y siniestros (Anchen y Dave, 2019).

La normativa de protección de datos que existe puede afectar al tratamiento de los datos en el sector asegurador porque lo principal que busca dicha normativa es proteger a las personas de las que disponen de información. Esta información servirá para crear valor para las empresas pero se debe buscar un marco de legalidad y ética ya que, por ejemplo, en el sector de seguros de la salud puede llegar a ser bastante peligroso al tratarse de datos sobre la salud de las personas. Para evitar todos estos problemas, las aseguradoras se han centrado en la creación de procesos totalmente estructurados y salvaguardados para la evaluación de los datos. Según Anchen y Dave (2019), en sus entrevistas con directivos de diferentes compañías del sector, han conocido que muchas aseguradoras han establecido a propósito salvaguardas para garantizar la privacidad.

Las aseguradoras a parte del tema que acabo de tratar, tendrán que tener en cuenta otros aspectos regulatorios como es la legislación sobre las compañías competidoras del sector y las restricciones que se hayan podido aceptar sobre el tratamiento de los datos de los clientes.

La seguridad, la privacidad y la confidencialidad son tres aspectos fundamentales para el sector asegurador en el ámbito del tratamiento de datos. Chema Alonso (2017), miembro del Comité Ejecutivo de Telefónica y experto en ciberseguridad, a través de la revista “Actuarios”, no se decanta sobre si estos tres aspectos son un peligro o oportunidades para el sector. Lo que sí que tiene claro es que las tablas de generación de las cuotas de aseguradoras cambiarán, indicando que estos formularios han pasado a ser tratados mediante algoritmos de Big Data con Machine Learning y en el futuro, se aplicarán nuevos sistemas. Adicionalmente, dice que los usuarios reclamarán cada vez más transparencia sobre esos mecanismos para gestionar el riesgo y para que puedan ser auditados por los clientes para así entender mejor los mecanismos de clasificación de los mismos y saber porque sus datos dicen si son clientes de riesgo o no.

6. Conclusiones

En el pasado, en las compañías, se aplicaba análisis descriptivo a los datos para realizar informes que servían de ayuda a la toma de decisión (Berlangua, 2017). Como bien se ha podido comprobar a lo largo del trabajo, el Big Data está siendo utilizado actualmente en el sector asegurador de forma activa mediante la inclusión de elementos que ayudan a la captación de un mayor número de datos y a su posterior análisis a través de las técnicas de análisis multivariante. El diseño de una buena estrategia de Big Data parte como fundamental si se quiere conseguir resultados en la transformación digital que están emprendiendo las compañías ya que tienen que adaptarse a las nuevas demandas y necesidades de los clientes. Dichos clientes son cada vez mucho más exigentes ya que conocen y tienen un acceso mucho más fácil tanto a la tecnología como a la información, pudiendo demandar servicios y productos a las compañías cada vez más personalizados. La respuesta por parte de las compañías a estas tendencias debe ser rápida para no caer en la obsolescencia y poder responder adecuadamente para así cumplir con las expectativas que prevén los clientes. Los usuarios, en especial los suscriptores, requieren una explicación clara de los resultados para generar confianza: necesitan transparencia para que los resultados sean más interpretables (Anchen y Dave, 2019).

La correcta aplicación del Big Data es crucial ya que trata de preparar y conseguir nuevas fuentes de datos. Posteriormente, aparece la analítica avanzada para que se ejecuten los diferentes modelos y técnicas de análisis multivariante, que son las encargadas del análisis de los datos y cuyas conclusiones pueden aportar conclusiones contrastadas e ideas al negocio de las compañías. En estas conclusiones es de suma importancia que los datos que se hayan utilizado sean veraces ya que sino las decisiones empresariales que se hayan podido tomar no serán correctas y no servirán para que la empresa se adapte mejor a las diferentes necesidades que pueda contar el consumidor. Como bien he explicado con anterioridad, estas técnicas son aplicadas a la tarificación de los diferentes productos, predicción de ventas, gestión de riesgos o segmentación de la muestra de clientes en diferentes grupos con características homogéneas. Tal como afirman Caballero y Martín (2015), el fin último del Big Data no es acumular datos, sino extraer información útil a partir de los datos.

Las necesidades de las compañías de digitalizar todos sus procesos, está obligando a aumentar su inversión en herramientas análisis y máquinas de procesamiento de datos. La inclusión de estos dispositivos, está provocando que las empresas adopten nuevas medidas y mecanismos en lo que respecta al tratamiento, almacenamiento y su análisis. En la actualidad, el incremento en el volumen y variedad de datos disponibles, permite realizar análisis en tiempo real, teniendo la capacidad de tomar decisiones también en tiempo real (Berlanga, 2017). Para el futuro próximo, las compañías deberán anticiparse y buscar la eficacia, siempre adaptándose a las nuevas tecnologías, procesos y marcos regulatorios ya que hay que tener cuidado a los límites que puede tener su aplicación. Para este sector creo que será mucho más sencillo en comparación con otros debido a que la ciencia actuarial siempre ha estado familiarizada con la acumulación y el procesamiento de datos. Mirando al futuro, los modelos analíticos necesitarán ajustes para incorporar nuevos datos de tecnologías en desarrollo, haciendo esenciales los marcos de gobernanza y la calidad de los datos para sacar el máximo provecho de las ventajas de la digitalización (Anchen y Dave, 2019).

7. Referencias bibliográficas

- Aldás, J., y Uriel, E. (2017). Análisis multivariante aplicado con R (2o edición). Paraninfo. Recuperado 26 de Marzo de 2020, de <https://www.paraninfo.es/catalogo/9788428329699/analisis-multivariante-aplicado-con-r--2%C2%AA-ed->
- Alonso, C. (2017). Entrevista Chema Alonso. Big-Data Revista Anales del Instituto de Actuarios Españoles (40), 26–28. Recuperado 10 de Marzo de 2020 de, de <https://actuarios.org/wp-content/uploads/2017/07/actuarios-40.pdf>
- Anchen, J., y Dave, A. (2019). Analítica avanzada: ampliando los límites del seguro de P&C (N.º 4/2019). Recuperado 5 de Marzo de 2020, de <https://www.swissre.com/institute/research/sigma-research/sigma-2019-04.html>
- Balasubramanian, R., Kaur, K., Libarikian, A. y Narula, N. (2017). Raising returns on analytics investments in insurance, McKinsey. Recuperado 30 de Marzo de 2020, de <https://www.mckinsey.com/industries/financial-services/our-insights/raising-returns-on-analytics-investments-in-insurance>
- Berlanga, A. (2017). Entrevista Chema Alonso. Big-Data Revista Anales del Instituto de Actuarios Españoles (40), 7–9. Recuperado 10 de Marzo de 2020, de <https://actuarios.org/wp-content/uploads/2017/07/actuarios-40.pdf>
- Boj, E., Claramunt, M., y Costa, T. (2017). Tarificación y provisiones. Recuperado 16 de Marzo de, de <http://diposit.ub.edu/dspace/bitstream/2445/107124/6/MA-OMADO.pdf>
- Breading, M. y Pauli, K. (2018). AI is changing the game in workers' comp, Strategy Meets Action. Recuperado 16 de Marzo de 2020, de https://info.claraanalytics.com/white-paper_sma_april2018
- Caballero, R., y Martín, E. (2015). Las bases de Big Data. Recuperado 5 de Abril de 2020, de <https://reubenbooks.co.uk/324159-VNMJKUHQXL/>
- Capgemini. (2020, 8 enero). Las 8 tendencias en 2020 en el seguro de Salud. Seguros News. Recuperado 25 de Marzo de 2020, de <https://segurosnews.com/news/las-8-tendencias-en-2020-en-el-seguro-de-salud-y-las-10-tendencias-en-multirriesgos?cn-reloaded=1>
- Casanovas, M., Merigó, J. M., y Torres, A. (2014). Inteligencia computacional en la gestión del riesgo asegurador: operadores de agregación OWA en procesos de tarificación. Madrid: Fundación MAPFRE.
- Cox, M., y Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. Proceedings. Visualization '97 (Cat. No. 97CB36155), 235-244,. Recuperado 26 de Febrero de 2020, de <https://doi.org/10.1109/VISUAL.1997.663888>
- Cukier, K. (2010). Data, data everywhere. The Economist. Recuperado 15 de Marzo de 2020, de <https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf>

Davenport, T. y Prusak, L. (1998). *Working Knowledge: How organisations manage what they know*. USA: Harvard Business School Press, 224p.

Deloitte. (2016, 23 noviembre). El futuro del sector asegurador. Recuperado 30 de Marzo de 2020, de <https://www2.deloitte.com/es/es/pages/financial-services/articles/el-futuro-del-sector-asegurador-europa.html>

Exastax. (2017). Top 7 Big Data Use Cases in Insurance Industry. Recuperado 30 de Marzo de 2020, de <https://www.exastax.com/big-data/top-7-big-data-use-cases-in-insurance-industry/>

Flores, A. (2017). ¿Quiénes son de verdad tus competidores? Big-Data Revista Anales del Instituto de Actuarios Españoles (40), 34–37. Recuperado 10 de Marzo de 2020, de <https://actuarios.org/wp-content/uploads/2017/07/actuarios-40.pdf>

Fundación Mapfre. (s.f.-b). ¿Qué es el seguro? Recuperado 30 de Marzo de 2020, de https://www.fundacionmapfre.org/fundacion/es_es/

Fundación Mapfre. (s.f.-c). Tipos de seguros. Recuperado 25 de Marzo de 2020, de https://www.fundacionmapfre.org/fundacion/es_es/

Gantz, B. J., y Reinsel, D. (2011). *Extracting Value from Chaos State of the Universe: An Executive Summary*, IDC iView, 1-12.

Gantz, J., Rydning, J., y Reinsel, D. (2018). *The Digitization of the World From Edge to Core (#US44413318)*. Recuperado 10 de Marzo de 2020, de <https://resources.moredirect.com/white-papers/idc-report-the-digitization-of-the-world-from-edge-to-core>

González, E. (2015). *Análisis de datos aplicado a siniestros de automóviles*. Universidad Carlos III de Madrid. Recuperado 3 de Abril de 2020, de https://e-archivo.uc3m.es/bitstream/handle/10016/23045/TFG_Eduardo_Gonzalez_Gonzalez.pdf?sequence=1&isAllowed=y

Guelman, L., Guillén, M., y Pérez Marín, A. M. (2014). *A survey of personalized treatment models for pricing strategies in insurance*. *Insurance: Mathematics and Economics*. Recuperado 27 de Marzo de 2020, de <https://doi.org/10.1016/j.insmatheco.2014.06.009>

Guillén, M. (2016). *Big Data en seguros*. Recuperado 16 de Febrero de 2020, de <http://www.revistaindice.com/numero67/p28.pdf>

Harris-Ferrante, K. (2018). *CIO Agenda: Insurance Industry Insights*, Gartner. Recuperado 10 de Marzo de 2020, de <https://www.gartner.com/en/documents/3891568/2019-cio-agenda-insurance-industry-insights>

Hothorn, T., Hornik, K., y Zeileis, A. (2006). Ctree: Conditional Inference Trees. *Journal of Computational and Graphical Statistics*. Recuperado 30 de Marzo de 2020, de <https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf>

IBM. (2018). The Four Vs of Big Data. Recuperado 10 de Febrero de 2020, de http://www.ibmdatahub.com/sites/default/files/infographic_file/4-Vs-of-bigdata.jpg

IDC. (2019, 4 abril). IDC Forecasts Revenues for Big Data and Business Analytics Solutions Will Reach USD 189.1 Billion This Year with Double-Digit Annual Growth Through 2022. Recuperado 10 de Febrero de 2020, de <https://www.idc.com/getdoc.jsp?containerId=prUS44998419>

Karatzoglou, A., Meyer, D., y Hornik, K. (2006). Support Vector Algorithm in R. *Journal of Statistical Software*, 15(9), 1-28. Recuperado 10 de Marzo de 2020, de <https://doi.org/10.18637/jss.v081.b02>

Laney, D. (2001). 3D management: Controlling data volume, velocity, and variety. *Application Delivery Strategies*, META Group, Inc. Recuperado 16 de Febrero de 2020, de <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>

Lozano, C. (2015). Modelo de predicción de daños, una aplicación de redes neuronales artificiales a riesgos en el seguro de naranja. Fundación Mapfre. Recuperado 10 de Marzo de 2020, de https://www.fundacionmapfre.org/documentacion/publico/i18n/catalogo_imagenes/imagen_id.cmd?idImagen=1098471

Martínez, M. (2017). Entrevista Marta Martínez IBM. *Big-Data Revista Anales del Instituto de Actuarios Españoles* (40), 4–6. Recuperado 10 de Marzo de 2020, de <https://actuarios.org/wp-content/uploads/2017/07/actuarios-40.pdf>

McAfee, A., y Brynjolfsson, E. (2012). Big Data. The management revolution. *Harvard Business Review*, 90(10), 61-68. Recuperado 26 de Febrero de 2020, de <https://doi.org/10.1007/s12599-013-0249-5>

Mills, S., y Forder, S. (2012). IBM Smarter Analytics: Big Data and Insurance. IBM Smarter Analytics.

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8), 114-117. Recuperado 25 de Febrero de 2020, de <https://www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf>

Oliva, F., y Flores, M. (2017). La transformación de las compañías de seguros en la era digital. Recuperado 30 de Marzo de 2020, de <https://www2.deloitte.com/uy/es/pages/strategy-operations/articles/La-transformacion-de-las-companias-de-seguros-en-la-era-digital.html>

Padilla-Barreto, A., Guillén, M., y Bolancé, C. (2017). Big-Data Analytics en Seguros. *Revista Anales del Instituto de Actuarios Españoles*, Recuperado 10 de Marzo de 2020, de https://www.actuarios.org/wp-content/uploads/2017/11/1_19_ANALES_1_Final.pdf

Rodríguez-Pardo, J. M. (2012). Modelos predictivos aplicados al seguro de vida.
Recuperado 3 de Abril de 2020, de
<https://app.mapfre.com/fundacion/html/revistas/gerencia/n114/docs/Estudio1.pdf>