Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Toma de decisiones en la gestión de siniestros: Ensayo sobre técnicas de *machine learning* en seguros

AUTOR: JUAN GABRIEL STRICCOLI

TUTOR: RODRIGO DEL ROSSO

JUNIO 2024

Resumen

El mercado de seguros transita una paulatina transformación tanto en sus procesos, su tecnología, así como en su cultura, producto del cambio en otros mercados generados por el advenimiento de la cuarta revolución industrial, la digitalización global y el foco en la voz del cliente. No obstante, dicha transformación dista de asentarse, situándose la gestión de siniestros de dicho mercado en un contexto donde las decisiones aún continúan basándose principalmente en la realización de análisis descriptivos y en la experiencia de sus colaboradores.

El presente trabajo tiene por objeto la búsqueda de modelos que ayuden a predecir comportamientos inherentes a la gestión de siniestros, mediante la explotación y exploración de bases de datos de compañías de seguros obtenidas gracias a las nuevas tecnologías digitales implementadas durante los últimos años, con el propósito de realizar los cambios necesarios en sus procesos a modo de mantener un costo medio competitivo, una gestión de riesgos adecuada y elevados estándares de servicio al asegurado.

Palabras clave

Seguros; gestión de siniestros; toma de decisiones; asegurado y tercero.

Índice

Introducción

La irrupción de la pandemia por coronavirus en todo el mundo a fines de 2019 acentuó en los mercados la necesidad de acelerar su transformación digital, que se venía gestándose paulatinamente desde comienzos de la década del 2000. El mercado de seguros y su gestión de siniestros no fue ajeno a dicha necesidad, en donde se manifestó un nuevo escenario en el que para poder continuar con su propuesta de valor había que comunicarse y brindar su servicio de una forma diferente, con los medios digitales como su principal canal. La adopción de nuevas tecnologías digitales, la revisión y mejora continua de sus procesos, el aumento en la capacidad de almacenamiento de información producto de la disminución de sus costos, así como también la necesidad de escuchar y documentar la voz del cliente en cada interacción con las compañías de seguro del mercado, abren un nuevo abanico de posibilidades en términos de análisis de la información para la mejora en la gestión, muchas de ellas aún no exploradas.

Actualmente, una de las problemáticas existentes dentro del mercado, al menos en la República Argentina, consiste en comprender y controlar el *ratio* de traspasos de los reclamos extrajudiciales de terceros a instancias judiciales (mediación o juicio), dado que los tiempos y costos de estos se multiplican hasta por diez veces en algunos casos. Otro aspecto usual en el mercado es el de establecer procesos de atención al asegurado en función del tipo de riesgo contratado y el tipo y grado de respuesta que brinda su red de proveedores, sin cuestionarse quizá si no es necesario aperturar el servicio en función de la agrupación de algunas características clave de dichos asegurados.

En función a dichas problemáticas existentes, el presente trabajo tiene como objetivo la realización de ensayos para medir el potencial impacto de la implementación de técnicas multivariadas y de *machine learning* en la predicción de comportamientos en la gestión y toma de decisiones de siniestros. Por tal motivo, se parte de la hipótesis que el modelo de datos a analizar es útil para aplicar modelos predictivos o de *machine learning* que mejorarán el proceso decisorio en la gestión de siniestros.

En línea con la hipótesis general establecida, se acudirá inicialmente a bases de datos confidenciales de una compañía aseguradora del mercado argentino de la cual se logró obtener los datos necesarios, a modo de realizar la exploración correspondiente para evaluar y poner a prueba los diversos ensayos en línea con las problemáticas planteadas. Cabe destacar que dicha compañía es una organización multinacional, la cual se dedica al ofrecer tanto seguros patrimoniales como de vida y retiro y que se encuentra entre las primeras 20 compañías de

seguros acorde al ranking de compañías del mercado asegurador a diciembre de 2023 (Todo Riesgo, 2024).

Ahora bien, una vez seleccionadas, las bases pasarán por un proceso de anonimización, transformación y normalización de los datos para asegurar la confidencialidad de los asegurados y todas las partes intervinientes en la gestión del siniestro, con la intención asimismo de no perder valor en el análisis. Las herramientas que se utilizarán serán los lenguajes R y Python, al igual que la herramienta de bajo código *Azure Machine Learning*, y las metodologías serán tanto multivariadas, como ser la técnica de clústers mediante la implementación del algoritmo de *k-means* o el método componentes principales (PCA), así como también algunas técnicas clasificación de *machine learning*.

Seguros y siniestros: Evolución hacia un nuevo paradigma

> Históricamente en un contrato de seguros el asegurador y el asegurado tenían conjuntos de información diferentes. […] El asegurador intentaba obtener la máxima información a través de cuestionarios, observaciones y estadísticas para inferir cómo se iba a comportar el asegurado; y por su lado el asegurado se esforzaba en minimizar el riesgo, maximizar el valor de los siniestros y "manipular" a su favor el precio del seguro. […]
>
> Los desarrollos en IA y la disponibilidad de datos masivos han alterado completamente esta asimetría. Así como antes la información era incompleta, estática, fragmentada y no estaba actualizada, con la llegada del *big data* la información ha pasado a ser comprensible, accesible desde múltiples dispositivos y actualizada en tiempo real. El asegurado deja de comportarse como un sujeto pasivo en relación con la aseguradora, para posicionarse como alguien que reclama el data *ownership*, en un intento de equilibrar la asimetría anteriormente expuesta. Además, los costes de la obtención de la información se han reducido a la mínima expresión. Los datos se han convertido en una mercancía y la IA nos va a ayudar a explotarla. (Torres Ayala, 2020)

Se cree pertinente que, para el inicio del presente trabajo, se debe contextualizar al lector respecto del camino que transitó el mercado asegurador y, sobre todo, el sector de siniestros, para dimensionar el por qué se cree relevante el desarrollo de las investigaciones a desarrollar en el presente trabajo. A tal fin, en un primer subapartado, se presenta una descripción de la evolución y cambios recientes dentro del mercado de seguros. En un segundo subapartado, se hace foco en las contribuciones más recientes por la inteligencia artificial (IA) en el mercado de seguros a la fecha. Y en un tercer subapartado, por último, se introducirá en la conceptualización de las problemáticas propuestas a investigar.

1.1. Análisis de contexto del mercado de seguros

Para hablar del mercado de seguros y su actualidad, se debe primero hablar obligadamente de un suceso global, un cambio de paradigma que ha alcanzado a todas las industrias del Siglo XXI y que influye no solo a las organizaciones sino también a las personas y la forma en que se relacionan: la cuarta revolución industrial. En palabras del autor Blanco (2019) este nuevo suceso representa "un salto cualitativo en la manera en que las tecnologías digitales, físicas y biológicas se fusionan, generando impactos significativos en todos los aspectos de nuestra sociedad" (p. 134).

Se puede decir que la Cuarta Revolución Industrial tiene su origen definido en función de los avances tecnológicos que han convergido en los últimos años, como por ejemplo la inteligencia artificial, el Internet de las cosas, la computación en la nube y la robótica avanzada, entre otros. Estas tecnologías han generado una fuerte interconexión entre sí sin precedentes, las cuales junto a la generación masiva de datos han dado lugar a una nueva era de innovación y transformación, y abren nuevas posibilidades en términos de eficiencia, productividad y personalización de los productos y servicios. En otras palabras, se puede decir que "la Cuarta Revolución Industrial surge de un modo parecido a las anteriores. Aparecen tecnologías nuevas que permiten fabricar productos y prestar servicios en formas y lugares completamente nuevos, siendo la conectividad la característica principal que […] las hace auténticamente disruptivas" (Blanco, 2019, p. 11).

Es precisamente este suceso el que ha despertado a una industria histórica que estuvo por varios siglos a gusto o, por lo menos, acostumbrada a una dinámica de servicio que en la actualidad se había vuelto obsoleta gracias a lo manifestado en otras industrias, como el transporte de pasajeros, venta de mercaderías o *delivery* de alimentos, por mencionar algunos. En estos negocios, la digitalización, la omnicanalidad, la transparencia en la trazabilidad del producto y la inmediatez en la respuesta de servicio se han convertido en el *core* de estos. En consonancia con lo anteriormente dicho, Martín Ferrari, fundador de 123Seguro, menciona que "La industria del seguro lleva cientos de años haciendo las cosas igual y se fue corriendo de las necesidades de los clientes " (citado en Álvarez Plá, 2018, p. 80), quien posteriormente complementa con una interesante afirmación al establecer que "la principal barrera ante la innovación es el miedo a cambiar" (p. 80).

El resultado de este nuevo paradigma en la industria del seguro fue el surgimiento en los primeros años de la década de 2010 de una nueva clase de competidor o agente en la industria: la *Insurtech*, denominación que "surge de la conjunción de las palabras inglesas *insurance* (seguros) y *technology* (tecnología) y está inspirada en el término *Fintech* (unión de *finance* y *technology*) que alude a un fenómeno que se anticipó en muchos años al de *Insurtech*" (Zapiola Guerrico, 2020, sección 2). Y aunque si bien las *Fintech* fueron las precursoras, Muñoz Pardo (2018), destaca una fortaleza de las *Insurtech* respecto de su hermano mayor al mencionar que "[…] el *fintech* tardó siete años (2006 – 2013) en conseguir el volumen de inversiones de unos 2.500 millones de dólares, mientras que las *Insurtech* lo han conseguido en solo 4 años (2011 – 2015)" (p. 24). Las *Insurtech* han traído al mercado una nueva visión disruptiva que va en

línea con la demanda de las nuevas generaciones, con un fuerte sustento en el desarrollo tecnológico y digital, donde tanto la inmediatez en el servicio y la comunicación como la facilidad de acceso desde diversos canales mediante el uso de dispositivos tecnológicos y el Internet, representan el *slogan* de estos nuevos agentes. Álvarez Plá (2018) lo resume muy bien de la siguiente forma:

> […] De eso hablamos cuando hablamos de Insurtech. De esas startups de base tecnológica que tienen por objetivo aplicar al seguro las más diversas herramientas tecnológicas, ampliando así su cobertura y generando valor añadido, pero también obligándolas a pensar nuevos productos y servicios que se adapten a las necesidades de la sociedad digital. (p. 77).

Este punto de vista es realmente significativo, dado que comienza a manifestarse un cambio no sólo sobre los procesos internos de las aseguradoras, sino también sobre su forma de relacionarse con los asegurados, proveedores y productores, así como la forma en que idean sus productos. Esto, por supuesto, crea nuevas oportunidades, pero también nuevos competidores, que, si bien en ocasiones puede que inicialmente no sean expertos en seguros, sí lo son en tecnología, datos e inteligencia artificial. En palabras de Torres Ayala (2020):

> Las aseguradoras tradicionales están empezando a afrontar la competencia tanto de las Insurtech como la de las grandes compañías tecnológicas, que se están apalancando en tecnologías avanzadas para presentar productos innovadores, por ejemplo, ofreciendo productos con coberturas totalmente personalizadas (hypercustomization), microseguros en modalidad *switch on/off* u ofreciendo negocios 100% nativos digitales. Este tipo de empresas, además, está abriendo nuevos canales de venta y nuevas formas de relación inspiradas en otros sectores donde la relación con el cliente es más continua, cercana y recurrente. (p. 30)

Francisco Astelarra, secretario general la Federación Interamericana de Empresas de Seguros (FIDES), reafirma lo antedicho al establecer que “la tecnología es la variable estratégica que va a determinar la evolución del negocio asegurador” (citado en Álvarez Plá, 2018, p. 77). Esto es sobre todo cierto si tenemos presente lo que ha transcurrido en estos últimos años, en especial luego del comienzo de la pandemia por COVID-19 en 2020, donde hemos podido observar una aceleración de este proceso impulsado por el incremento en el uso personal de dispositivos tecnológicos producto del aislamiento que hemos transitado como sociedad, lo cual surtió un efecto acelerador de este proceso, ya que no solo las nuevas generaciones saben hoy en día

utilizar las nuevas tecnologías. Tanto la pandemia, como el movimiento cada vez más creciente en el cuidado de la salud física y mental y el uso de dispositivos digitales, han despertado el hecho de que "los nuevos asegurados ya no sólo quieren asegurar objetos, sino también su estilo de vida por lo que las compañías tendrán que ofrecer servicios personalizados para poder dar una respuesta satisfactoria a las nuevas necesidades de sus clientes" (Álvarez Plá, 2018, p. 84).

Afortunadamente, no todo el mercado vio inicialmente con temor al advenimiento de este nuevo rumbo, ya que, en opinión de Eduardo Iglesias, CEO de Colón Seguros y Fundador de eColón –la primera aseguradora completamente digital de la Argentina– "la tecnología no es un riesgo para el mercado asegurador, son herramientas que nos hacen ser más eficientes, bajar nuestros costos y llegar a más asegurados" (citado en Álvarez Plá, 2018, p. 78). Otra persona que ve al desafío con otros ojos es Mauricio Monroy (2023), VP para LATAM y España en Equisoft, quien afirma que "el objetivo es atraer a las nuevas generaciones que son nativas digitales y para los cuales los tradicionales canales de venta y servicio no son atractivos ni cumplen con sus expectativas" (citado en Perazo, p. 90). Iglesias (2018) también hace hincapié en el cliente, ubicándolo en el eje de esta transformación, producto de sus nuevas costumbres digitales:

> 'Gran parte de los asegurados son nativos digitales y no quieren una experiencia analógica, y hacia ellos debemos reenfocar el negocio', y resuelve que 'si vemos al cliente como alguien inteligente, digital, y nos centramos en él, los cambios se darán solos'. (citado en Álvarez Plá, p. 80)

Otro que va en igual línea que Iglesias y Monroy es Lionel Moure, socio de la consultora Deloitte, quien establece que "las oportunidades son muchas y en lo que respecta al mercado tradicional las Insurtech 'son tanto competidoras como facilitadoras'" (citado en Álvarez Plá, 2018, p. 82). Moure (2018), a la vez establece que hay dos grandes modelos de este nuevo desarrollo tecnológico en seguros:

> 'Uno sería la reconversión de las aseguradoras al modelo de negocio digital, asociándose o adquiriendo compañías tecnológicas'. […] Bajo este supuesto, 'las aseguradoras tradicionales se adaptan al nuevo entorno y continúan liderando el negocio'. El otro modelo es el de las compañías tecnológicas entrando al negocio asegurador y compitiendo con las aseguradoras tradicionales. 'En este escenario –indica Moure–, las compañías tecnológicas podrían relegar a las tradicionales'. (citado en Álvarez Plá, p. 82)

Desde la consultora Deloitte creen que ambos modelos van a estar "conviviendo entre sí" (Álvarez Plá, 2018, pag. 84). Moure (2018), a la vez, indica que "el predominio de uno u otro vendrá dado por la habilidad de las aseguradoras tradicionales para reconvertir su negocio, así como por la habilidad de las *Insurtech* en desarrollarlo" (citado en Álvarez Plá, p. 84). En esta convivencia, Gonzalo Delger (2018), cofundador y CEO del comparador de seguros de autos *SnapCar*, "también es optimista y afirma que 'las aseguradoras tienen el capital y las *startups* tienen el foco. Cada uno tiene lo que el otro necesita. Estas alianzas son un *win-win*.'" (citado en Álvarez Plá, p. 84).

A criterio del presente autor, esta distinción de modelos es algo que se disipará con el correr de los años para converger en un nuevo modelo único, apoyándose en lo establecido por Zapiola Guerrico (2020):

> […] existe una interacción creciente entre las Insurtech y las aseguradoras ya establecidas: en algunos casos las Insurtech son creadas por las propias aseguradoras "establecidas" y en otros casos estas nuevas compañías proveen servicios a las ya establecidas a efectos de mejorar y "digitalizar" aspectos o sectores específicos de su operatoria. (sección 2)

Que el cambio es un hecho y está en marcha, no cabe duda, pero, ahora bien, el asunto ahora radica en entender cómo se manifestará y qué áreas serán afectadas en primer lugar. En este sentido, Moure (2018) cree que ese manifiesto se presentará de la siguiente manera:

> La evolución hacia el modelo digital en seguros está comenzando por la cadena de valor ligada al marketing y a la comercialización. Luego se profundizará en materia de suscripción de riesgos y determinación de tarifas y, por último, en la gestión de siniestros y procesos de administración. (citado en Álvarez Plá, p. 78)

Según Moure (2018), podremos ver "cada vez productos más personalizados (…), ya que gracias a la gran cantidad de información y velocidad de procesamiento podremos conocer a los asegurados y tarifar mejor sus seguros. 'Tarifas […] que podrán adecuarse al riesgo asociado a su comportamiento'" (citado en Álvarez Plá, p. 78). Y no solo surgió la personalización de productos, sino también, y en especial, la creación de nuevos. Álvarez Plá (2018) nos brinda algunos interesantes ejemplos al respecto:

> […] están apareciendo productos que hasta ahora no imaginábamos. Sirva como ejemplo Pensumo, en España, un sistema de pensiones alternativo en el que los

> comercios adscritos realizan una micro aportación a un producto de ahorro cada vez que un cliente hace una transacción comercial en alguno de dichos establecimientos, o Testamenta, una empresa, también española, que ofrece la tramitación de testamentos online y que ya colabora con compañías como Zurich o Axa, que han decidido ofrecer los servicios de la startup de forma gratuita a quienes contraten su seguro de Vida. Por no hablar de las decenas de nuevos productos relacionados con la ciberseguridad que se comercializan en todo el mundo. (p. 84)

Ahora bien, es de suma importancia mencionar que, a diferencia de otras industrias, y debido a la asimetría en la información mencionada inicialmente entre el asegurado y las aseguradoras, la industria aseguradora incorporó hace décadas la presencia de agentes para que intermedien entre ambos, denominados productores o brokers de seguros, los cuales, al ser un eslabón más del servicio, han complejizado y enlentecido de cierta forma este camino hacia la digitalización y la conectividad directa entre la aseguradora y el asegurado, más aún si se lo compara contra el resto de los mercados. Precisamente por esto, según palabras de Álvarez Plá (2018), es lógico cuando vemos que "algunos productores se muestran preocupados por la incursión de los nuevos jugadores que podrían, piensan algunos, reducir el volumen de sus negocios" (p. 80). Para Ferrari (2018), "'si algunos productores tienen este temor es porque ya identificaron que la necesidad está cerca' y, en su opinión, 'deberían transformar el temor en acción, sin olvidar el gran valor que le brindan a la industria'" (citado en Álvarez Plá, p. 80).

Lo cierto es que, sin importar las voluntades de los agentes individuales del mercado, el cambio está en pleno desarrollo y las inversiones en tecnología continúan en aumento, y con el correr del tiempo se afianza -aunque lentamente- este nuevo orden. Para contextualizar lo antedicho, según CIO España (2017) y Álvarez Plá (2018), Tecnocom informó se han invertido más de USD 160 mil millones en transformación digital e Insurtech a nivel Global entre 2013 y 2017, aproximadamente. Y si nos trasladamos a la actualidad, Jiménez (2023), vicepresidente regional para América Latina en Mambu, establece que todo "el sector financiero ya hace uso de IA y se prevé que puede generar un valor agregado hasta de U$S 1 trillón anual, según McKinsey & Company" (p. 27).

No obstante, que las inversiones estén de manifiesto y que el mercado haya reaccionado con el objetivo de, entre otros, no quedar fuera de la competencia, no implica que el camino sea sencillo y no haya tenido sus obstáculos. Los proyectos en digitalización e IA han tenido sus vaivenes producto de la irrupción de la pandemia y el contexto económico global, en donde por

ejemplo y en palabras de Gustavo Bresba (2023), director de DC Sistemas y Servicios, en la actualidad "hubo una reactivación y una vuelta a una situación pre-pandemia. Esto hizo que varios proyectos que habían quedado truncos se reactivaran y otros nuevos nacieran. El mayor foco lo notamos en la actualización de los sistemas de gestión" (citado en Perazo, p. 88). Este limitante no ha sido lo suficientemente fuerte para frenar el crecimiento tecnológico en la industria, ya que en palabras de Fernando López Orlandi (2023), Gerente Comercial para Latinoamérica en Charles Taylor Insurtech, "la coyuntura económica ha impactado el sector; no obstante, el ecosistema *Insurtech* ha venido creciendo un 18% anual por su significativo aporte en la innovación en modelos de negocio y a distancia" (citado en Perazo, 2023, p. 82).

Si bien la pandemia representó un desafío para aquellos agentes que querían invertir en el futuro de la industria del seguro, momento en el cual tuvieron que entender el nuevo comportamiento del mercado y replanificar sus proyectos, también fue el disparador para que varias compañías lo entendieran como una oportunidad para dar ese salto que anteriormente le temían. De hecho, ese tiempo les ha servido para comprender los nuevos hábitos de consumo digital por parte de los asegurados y empezar a delinear las bases de un nuevo modelo de negocios. En línea con esto, Martín Centeno, Cofundador de Autoreclamo, indica que "el impulso que generó la pandemia en cuanto a la demanda de soluciones digitales aceleró en muchos casos la adopción de tecnología y afortunadamente eso se ha reflejado en respuestas acordes a las nuevas exigencias del mercado" (citado en Perazo, 2023, p. 80). Y es por esto que, tal como indica Gonzalo Delgado Zemborain (2023), director para Argentina y Chile en ARM Services, "en 2023 se reordenaron muchas decisiones vinculadas a desarrollos que se precipitaron en el contexto de la cuarentena, y se priorizaron aquellos que generen ahorros por sobre aquellos de crecimiento" (citado en Perazo, 2023, p. 80). En otras palabras, la inversión no se frenó, sino que se priorizó.

En el caso de Latinoamérica, algunos especialistas indican que la inversión fue aún mayor en estos últimos años, y es que "según datos de la consultora IDC hay una creciente inversión en tecnología en todo el mundo, pero en particular de Latinoamérica" (Perazo, 2023, p. 76). Quien coincide en esto es Sebastián Schuartzman (2023), CEO de Adminse, el cual opina que:

> […] en los últimos años ha habido un impulso significativo hacia la adopción de tecnología en el sector asegurador. Las compañías de seguros han reconocido la necesidad de adaptarse a los cambios tecnológicos y han estado invirtiendo en soluciones digitales para mejorar la eficiencia operativa. Se han adoptado y desarrollado

> nuevas tecnologías; sin embargo, creo que o mejor está aún por venir y por eso […] considero que se avanzó mucho, pero estamos a mitad de camino. (citado en Perazo, 2023, p. 78)

Asimismo, si observamos no sólo Latinoamérica sino también el mercado argentino, Orlandi (2023) indica lo siguiente:

> Evidenciamos una aceleración en los planes de inversión de las organizaciones, sin embargo, la dirección de estas inversiones, en Latinoamérica, es bastante diversa. En la Argentina la transformación digital ya sucedió, por lo que las compañías de seguro están dirigiendo sus esfuerzos en mejorar procesos, reducir costos, obtener mayores resultados, fortalecer los canales de comunicación y encontrar vías de monetización de sus datos. (citado en Perazo, p. 82)

Para algunos, motivos para continuar por este rumbo sobran, pero si se es consecuente con todo lo que se ha mencionado en párrafos anteriores, podemos estar de acuerdo con Schuartzman (2023) en sus dichos respecto a que "esta aceleración se debe a la creciente conciencia de la importancia de la transformación digital para mantenerse competitivo en el mercado actual" (citado en Perazo, p. 78). Es evidente que el miedo al cambio comienza a dejarse de lado, no solo porque otros mercados nos muestran cómo es el resultado final de las experiencias digitales con sus productos y servicios, sino porque ya es una realidad y quien no abrace el cambio sabe muy bien que quedará marginado de la competencia. Pero esto no es algo que se presente en la totalidad de los casos, ya que tal como indica Daniel Gabas (2023), CEO de la compañía Fraudkeeper, "algunas compañías de seguros han abrazado plenamente la transformación, mientras que otras están en proceso de adaptación" (citado en Perazo, p. 92).

Ahora bien, amén de que los proyectos varían en función de la estrategia de cada competidor, los proveedores de tecnología del mercado han mencionado algunos puntos en común sobre el rumbo que la industria optó seguir respecto al ordenamiento de sus inversiones. En primer lugar, varios expertos coinciden en que la digitalización llegó a todo el mercado argentino, y así lo afirma Perazo (2023):

> El índice de Maduración Digital 2023, realizado por la Asociación Argentina de Compañías de Seguros y la Cámara de la Industria Argentina del Software, evidenció que la transformación digital comienza a consolidarse en las compañías de seguros. Este índice destaca que el 100% de las empresas participantes utilizan pólizas digitales. (p. 76)

Otros aspectos en los que la mayoría del mercado coincide en que se ha avanzado son la renovación de sus sistemas core, la implementación de chatBot con IA, sistemas de detección y prevención de fraude e integración de datos en la nube, entre los principales. Por ejemplo, tanto Sebastián Anselmi (2023), co-founder de Leverbox, como Gastón Ramos Adot (2023), gerente general de la Unidad de Negocio Cono Sur de Sistran, opinan que tanto desde 2022 el foco de las compañías de seguro estuvo en sus sistemas Core, ya sea para actualizarlos o para directamente cambiarlos por otros más modernos (citados en Perazo, pp. 76, 95). Otro que opina en igual sentido y explaya los motivos, es Mauricio Monroy (2023), VP para LATAM y España en Equisoft, al establecer que las aseguradoras "[…] necesitan sistemas ágiles y sistemas core flexibles. Asimismo, un objetivo claro de las aseguradoras es potencializar la transformación digital y optimizar el *customer journey*; para esto es necesario invertir en plataformas core modernas que soporten, permitan y aceleren dicha transformación" (citado en Perazo, p. 90).

No obstante, varios analistas y proveedores del mercado coinciden que las principales inversiones fueron dedicadas a la seguridad de los sistemas y la información, dada la creciente amenaza de hackeos cibernéticos mediante *ramsomware* a organizaciones grandes y medianas. Uno de ellos es Nahuel Ramírez (2023), desarrollador de negocios en Colinet Trotta, quien comenta que "observamos la activación de proyectos de inversión en tecnología […], con particular interés en las cuestiones relacionadas con ciberseguridad y protección de infraestructura, que sin dudas se transformaron en temas de agenda para todos los actores del mercado" (citado en Perazo, p. 86). Este autor menciona asimismo su parecer respecto de la IA en la industria aseguradora: "Si bien la IA aparece como tema de agenda obligado en foros y convenciones del sector, creemos que aún lo hace con carácter aspiracional" (citado en Perazo, 2023, p. 86). Fabián Aquino (2023), CEO de Codeicus, se alinea con la visión de su colega, al declarar que "se nota la aceleración en la inversión principalmente en seguridad informática. Los servicios de hacking ético y demás procesos que buscan prevenir ataques y blindar los sistemas de las empresas de seguro se encuentran con importante demanda" (citado en Perazo, p. 86).

Pero hay otros agentes de mercado que creen que, si bien la ciberseguridad es un aspecto crítico para poder prosperar en este nuevo paradigma, no fue la única ocupación del mercado en estos últimos dos años, ya que según ellos también la IA comienza a tener mucho más protagonismo y no se limita a ser solo un aspiracional o promesa de la industria. Orlandi (2023), por ejemplo, afirma que "este ha sido el año de la Inteligencia Artificial y la ciberseguridad, pues los usuarios

demandan hoy mayor rapidez, autenticidad y personalización" (citado en Perazo, p. 84). Y quien se explaya un poco más en este aspecto del mercado es Gabas (2023), quien destaca lo siguiente:

> En el sector asegurador, el 2023 es el año de la IA porque revolucionó la forma en que las aseguradoras analizan datos, automatizan procesos y mejoran la precisión en la detección de fraudes. El segundo gran punto es la seguridad; aquí el panorama se ha vuelto crítico debido al aumento de las amenazas cibernéticas y la creciente sofisticación de los ataques. Las compañías de seguros están tomando medidas para protegerse contra estas amenazas y garantizar la seguridad de los datos confidenciales de sus clientes, y en ese punto las *insurtechs* somos una parte importante en la colaboración. (citado en Perazo, p. 92)

Al margen de las diferencias que cada agente del mercado tiene respecto de las prioridades de inversión y la evolución de este, lo que si resulta claro es que el avance es tangible, aunque a un ritmo distinto al de otras industrias. En su artículo publicado, Perazo (2023) presentó la opinión de 22 proveedores de tecnología especialistas en seguros, de los cuales 20 de ellos han respondido, con un puntaje de 1 a 10, a una pregunta referida a su parecer respecto del grado de adopción de tecnología por parte del mercado asegurador. El puntaje promedió en una calificación de 6,5 puntos, donde la mayoría coincide en que es visible el avance en tecnología del mercado pero que aún están a mitad de camino (pp. 78-98). Adot (2023), uno de los encuestados, explica que se observa que "para 2026, el 40 por ciento de la infraestructura de TI adoptará Inteligencia Artificial y Analítica de Datos a la toma de decisiones corporativas" (citado en Perazo, p. 76).

Algunos ponen en tela de juicio que el avance sea por iniciativa y convencimiento propio, sino más bien por el mero hecho de no querer quedar fuera de la competencia. En otros casos, incluso, creen que el avance no es tal como el que se predica en las conferencias abiertas del mercado, y que el carácter conservador prevalece en este contexto. Una explicación a esto la puede brindar la posición de Emilio Moses (2023), Cofundador y responsable comercial de Kurt, el cual indica que "el mercado asegurador en general es lento en la adopción de nuevas tecnologías en comparación con la industria financiera y además cuenta con una muy alta cantidad de procesos manuales o realizados por humanos" (citado en Perazo, p. 94). Agostina Fernández (2023), Sales Manager para Argentina y Cono Sur en Hexagon, entiende que "el mercado […] aún es muy conservador a la hora de invertir en herramientas que vayan más allá

de particularidades o necesidades concretas" (citado en Perazo, p. 94). Bresba (2023), por su parte, tiene una visión algo más cruda de la realidad:

> El principal desafío que yo veo en las compañías de seguros argentinas es cambiar la forma en que ven el dinero que destinan en tecnología. Hay muchas que, reconociéndolo o no, lo siguen viendo como un costo y no como una inversión. Parece absurdo, pero hay casos donde prefieren no asumir el costo de una hora de trabajo a cambio de no aplicar un cambio, que probablemente luego tenga muchísimo más costo en recursos para la compañía. Ese enfoque erróneo es el que pone un freno enorme a los saltos de calidad que la tecnología puede ofrecer. (citado en Perazo, pp. 88-90)

La conclusión en este aspecto es que lo hecho hasta ahora no deja de ser positivo, pero dista aún de lo que muchos especialistas entienden que se puede lograr con las soluciones que varias *Insurtech,* o mismo otros mercados, ya han desarrollado o están en proceso de desarrollo. La industria avanza, pero de forma más homogénea y gradual, aún sin ningún referente que tome las riendas de esta transformación, por lo que, según Santiago Urrizola (2023), CEO y Co-Founder de Flux IT, "el desafío es que la industria acelere sus planes. La empresa que acelere hoy a nivel tecnológico va a ganar un gran terreno en el mediano plazo" (citado en Perazo, p. 91). Bresba (2023), nuevamente, deja una visión realista en consonancia con esta visión:

> No veo que haya ninguna tecnología en particular picando en punta. Las cosas realmente novedosas todavía no se filtraron a gran escala en el mercado asegurador nacional. Mi análisis es que las compañías se encuentran todavía en una situación más precaria en sus procesos de digitalización y no podrían absorber cambios demasiado más profundos en el paradigma en el que se mueven. (citado en Perazo, p. 90)

Lo cierto es que la industria afronta (y afrontará por muchos años) varios desafíos, por lo que tiene aún un largo camino por recorrer. Estos desafíos son variados y difieren según la óptica de cada participante, pero se compartirán los más destacados según el criterio de este autor. Por ejemplo, Schuartzman (2023), entiende que "las empresas tecnológicas enfocadas al seguro deben encontrar formas de diferenciarse, cumplir con las regulaciones, garantizar la seguridad de los datos y superar los desafíos técnicos y de adopción por parte de los consumidores para prosperar en este entorno […]" (citado en Perazo, p. 76). En un sentido similar, Anselmi (2023), considera que el gran reto del mercado a futuro "[…] va a ser la diferenciación. Es un mercado lleno de necesidades y por momentos siento que hay muchos enfocándose en lo mismo. Creo que hay que ampliar la mirada en este plano para poder generar mayores oportunidades de

negocios" (citado en Perazo, p. 95). Adot (2023) presenta una óptica diferente, preocupado por la capacidad de obtener y retener talento, al reconocer que "el principal desafío de las compañías de software será enfrentar las consecuencias de la escasez de recursos altamente capacitados y la alta rotación de personal que según LinkedIn ronda el 22 por ciento anual" (citado en Perazo, p. 76). Finalmente, Delgado Zemborain (2023), brinda una visión holística en consonancia con la opinión del presente autor:

> Las *insurtech* son un socio para que la industria del seguro acelere la transformación; el desafío lo tienen las aseguradoras y *brókers*, quienes deben aceptar que cada día pueden controlar menos todas las verticales y necesitan generar alianzas por proyectos; los ecosistemas sólo son viables si las partes que los conforman logran articularse; hay que dejar de hablar de cliente-proveedor donde uno impone las condiciones, y empezar a entenderse como socios de negocios donde todos ganen. (citado en Perazo, p. 80)

1.2. Aplicación de la inteligencia artificial en seguros y siniestros

Como bien se ha indicado en párrafos anteriores, varias compañías empiezan gradualmente a implementar inteligencia artificial (IA) para algunas aplicaciones de uso que generen ahorros, ya sea para automatizar tareas como para aplicar algunos análisis de riesgos que mitiguen pérdidas operativas. Esto se debe a que, para poder aplicar *machine learning* o IA generativa, es necesario antes tener sistemas Core que provean información correcta y eficiente, así como también una correcta integración de datos en la nube para poder tener una visión unificada de estos, con procesos de gobierno y calidad de datos para mejores y más prontas decisiones. Varios de estos proyectos aún están en desarrollo, pero algunas compañías ya han compartido con el mercado algunas implementaciones y sus logros. A modo de adelanto, y tal como indica Iván Ballón (2023), Gerente de Desarrollo para América Latina e Iberia en Friss, "las aseguradoras ya experimentan por ejemplo con ChatGPT[1] para sus procesos, también en reconocimiento de voz e imagen, o en la virtualización de situaciones para los procesos de suscripción" (citado en Perazo, p. 93).

Uno de los primeros sectores que han comenzado con la explotación de la IA es el área de *marketing* o producto. Esto es así porque suelen obtener no sólo información recolectada de los

[1] ChatGPT es un modelo de lenguaje capaz de producir texto de forma autónoma, que puede ser utilizado para tareas de generación de texto, traducción automática, resúmenes, entre otras. En pocas palabras, es una herramienta de inteligencia artificial que se utiliza para generar respuestas a preguntas y realizar otras tareas naturales del lenguaje (Infobae, 2023).

asegurados una vez contratada la póliza, sino también con potenciales clientes al interactuar con la compañía ya sea en internet como en las redes sociales, y esto brinda un volumen de información para aplicar diversos análisis con *machine learning* y *deep learning* envidiables por otros sectores. En la actualidad este sector puso de moda el término de marketing digital, término que revolucionó la forma en que las compañías de todas las industrias llegan de forma más efectiva y personalizada a los clientes con sus productos y servicios. "[…] El marketing digital supone una serie de acciones con el objetivo de atraer nuevas oportunidades y desarrollar una identidad de marca. Pero, en realidad, es la mejor manera de lograr interactividad con el público objetivo y afianzar relaciones" (Christian Blousson, citado en Fiorentino, 2023, p. 30). Aplicaciones en dicha área son muchas, casi en su totalidad centradas en comprender y acercar al cliente, e Inés Rodríguez (2023), gerente de *Marketing* y *Analytics* de Hipotecario Seguros, describe algunas que aplica actualmente en dicha compañía:

> El marketing digital es imprescindible para que las empresas puedan llegar de manera efectiva a su público objetivo y ofrecerle soluciones adaptadas a sus intereses. En nuestro caso, utilizamos técnicas de audiencias *look-alike* basadas en algoritmos y modelos estadísticos para buscar usuarios similares a nuestros clientes actuales en términos de comportamiento e intereses. También usamos *customer audiences* para segmentar y personalizar la comunicación y las ofertas, así como mapas de calor en nuestro sitio *web* para conocer los contenidos más clickeados. (Fiorentino, p. 30)

Lovagnini (2023) resume los principales usos de la IA en *marketing* digital:

- **Atención al cliente automatizada**: Las empresas están utilizando chatbots basados en inteligencia artificial para brindar respuestas rápidas y precisas a las consultas de los clientes. Estos chatbots pueden resolver problemas comunes, proporcionar información sobre productos o servicios, o ayudar a los clientes a tomar decisiones de compra. (p. 16)
- **Segmentación y personalización**: La IA se puede utilizar para analizar grandes cantidades de datos de clientes y generar perfiles detallados, lo que mejora la relevancia y efectividad de las campañas de marketing. (p. 16)
- **Recomendaciones de productos**: Los algoritmos de IA pueden analizar el comportamiento de compra y navegación de los clientes para hacer recomendaciones personalizadas de productos o servicios. Esto se utiliza

ampliamente en plataformas de comercio electrónico y sitios web para mejorar la experiencia de compra y aumentar las ventas. (p. 16)

- **Optimización de anuncios**: La IA se puede utilizar para optimizar y personalizar anuncios. Los algoritmos pueden analizar datos en tiempo real sobre el comportamiento de los usuarios y ajustar automáticamente los anuncios para mejorar el rendimiento y aumentar la tasa de conversión. (p. 16)
- **Análisis de sentimientos y opiniones**: La IA permite también analizar las opiniones de los clientes en las redes sociales, foros y otros canales en línea. Esto implica para las empresas comprender mejor el sentimiento de sus clientes, identificar problemas o tendencias emergentes, y responder de manera adecuada. (p. 18)

Tal como lo mencionaba Moure (2018), una vez avanzado con soluciones para el sector comercial o de *marketing*, el manual indicaba que el mercado debiera tener que continuar en el avance de asuntos referidos a suscripción de riesgos y determinación de tarifas, para luego finalmente atacar cuestiones referidas a la gestión de siniestros y procesos de administración. Y como se ha dejado en claro en el subapartado anterior, dada la irrupción de la pandemia y con la necesidad de focalizar en proyectos que generen ahorros, el orden en algunos casos ha cambiado un poco, con la priorización en la retención de cartera y mejora en la experiencia de cliente.

En ese sentido, el departamento de siniestros es uno de los que más ha tenido actividad en estos últimos años, dado que es el otro gran departamento en las compañías de seguros que tiene contacto directo con el cliente a la hora de la verdad, es decir, al momento de brindar el servicio el cual el asegurado contrató y pagó. Porque el seguro es un producto que se contrata con el propósito de obtener un servicio en caso de -desafortunadamente- sufrir un incidente o eventual pérdida económica o material (denominada siniestro), por lo que no necesariamente todos los clientes de una aseguradora van a recibir ese servicio y esto difiere de lo que sucede en otros mercados. Por ejemplo, acorde al último informe estadístico del ramo automotor emitido por la Superintendencia de Seguros de la Nación (SSN, 2022) en la sección "Siniestros de casco – Todos los hechos generadores", subtítulo "Frecuencia Siniestral por Hecho Generador y Año de Ocurrencia", la frecuencia del ramo fue de un 19,4% para dicho año, lo que implica que, de cada 100 autos asegurados en el mercado, se siniestraron casi 20 de ellos.

Dicho de otra forma, la ocurrencia de un siniestro es un momento clave para cualquier aseguradora, en ese momento su relación con el asegurado está en juego ya que deberá mostrarle

valor durante la gestión del caso, dado que el asegurado necesita contención y una respuesta rápida dado que sufrió un evento tanto inesperado como indeseado. Por otra parte, durante esta gestión la aseguradora debe aplicar todos los procedimientos necesarios para no solo verificar que el riesgo asegurado corresponda ser indemnizado, sino también eficientizar los costos y tiempos de gestión mientras vela por que no se cometa fraude en la denuncia de siniestro recibida. La finalidad aquí es, en definitiva, gestionar el caso en un tiempo acorde a la expectativa del asegurado e indemnizar el monto exacto que corresponda, ni más, ni menos. Y precisamente es aquí donde la IA puede y ofrece algunas soluciones muy útiles, las cuales se detallarán a en los próximos párrafos.

Antes de avanzar, he aquí los principales beneficios que, según Muñoz Pardo (2018), el departamento de siniestros espera obtener con el avance de la digitalización y la IA en sus operaciones:

- Mejorar la experiencia del cliente
- Mejorar la eficiencia en procesos
- Mejorar el tiempo de tramitación del siniestro
- Mejorar la comunicación con la red de colaboradores
- Aumentar la retención de clientes
- Reducir costos (p. 64)

Con esto en mente, la primera aplicación de IA en el sector, y que ha sido la estrella del sector estos últimos años, es la detección de intento de fraude en siniestros, con el objetivo de evitar erogar pagos de indemnización en caso donde la descripción del hecho ha sido adulterada o inventada con la intención de recibir un pago no legítimo. Y es que se tornó un asunto serio en el mercado argentino, ya que estos "[…] registraron un incremento en los últimos años, ocasionando pérdidas al sector en torno a los US$ 100 millones anuales, según datos de la Asociación Argentina de Compañías de Seguros (AACS), una situación que genera preocupación en cámaras y empresas […]" (Díaz Zettelmann, 2023). Gustavo Trías, presidente de la AACS, detalla que "en pólizas de hogar, por ejemplo, se analizan en profundidad entre el cuatro y cinco por ciento de los siniestros denunciados, y se determina que aproximadamente el 25% son fraudulentos, según los estudios de la AACS" (citado en Díaz Zettelmann, 2023). El resto de las regiones, por su lado, sufren este tipo de eventos de igual manera, aunque cada uno con sus peculiaridades:

> En Estados Unidos, por ejemplo, la *Coalition Against Insurance Fraud* considera que el fraude supone unos 80.000 millones dólares en pérdidas al año. En Australia, el *Insurance Council of Australia*, reporta que entre el 10% y el 15% de los siniestros tienen signos de fraude. Y en Europa, los valores son ligeramente inferiores, estimándose que el fraude representa el 10% de todos los gastos en reclamaciones. […]
>
> A estos preocupantes datos, se añaden los resultados del estudio realizado por AXA para su informe VII Mapa del fraude, que indican que la tasa de fraude (detectado) se ha duplicado en la última década pasando del 0,85% al 1,94%. (Torres Ayala, 2020, p. 39)

La buena noticia es que se cuenta actualmente con la tecnología para poder mejorar las detecciones, y sólo es cuestión de tiempo y de cultura de datos y antifraude para que se asiente y brinde sus frutos. En este sentido, Muñoz Pardo (2018), afirma lo siguiente:

> Las tecnologías de *Big Data* e Inteligencia Artificial proporcionarán unas herramientas para poder explotar toda la información y datos que disponen las aseguradoras. Estos datos pueden proceder mediante fuentes internas como el historial de siniestralidad de las pólizas, el CRM o por fuentes externas como las redes sociales. (p. 61)
>
> La llegada de los dispositivos conectados de *Internet of Things* y de sus sensores al mismo tiempo que proporcionarán nuevas fuentes de información que podrán ser añadidas a las bases de datos ya disponibles, ayudarán a la detección de nuevos fraudes. (p. 61)
>
> […] Para una correcta detección de fraudes, será imprescindible que exista una comunicación e intercambio de información no solo entre entidades aseguradoras, sino también con proveedores del ecosistema de *Internet of Things*. (p. 62)

Lo cierto es que, ya sea que se apliquen reglas de detección predeterminadas, o algoritmos de *machine learning* como ser árboles de decisión, *random forest*, regresión logística o análisis de clústeres, entre otros, la detección de fraude se torna un arte en sí. Esto se debe a que, si las reglas de detección implementadas son muy exigentes, es probable que se eleve no solo la cantidad de siniestros de fraude, sino de aquellos siniestros que finalmente son verídicos y en donde su gestión es demorada por el proceso de verificación de los hechos que lleva adelante el sector de lucha contra el fraude. Ahora bien, si por el contrario las reglas de detección son más benevolentes, la compañía se perdería la detección de una porción de casos de fraude que

le ingresa. El impacto de ambos escenarios es negativo para la organización, ya sea por la potencial pérdida de clientes debido a la demora en la gestión de su siniestro, como por la disminución en los ahorros producto de la menor detección de casos de fraude.

En segundo lugar, debemos mencionar que la función principal del departamento de siniestros es la de gestionar los reclamos de los asegurados, con el afán de brindar el mejor servicio posible a esa porción de asegurados que sufren un daño material o físico. Los cambios que la tecnología y la internet produjeron en la sociedad hicieron que las aseguradoras también replantearan la forma en que gestionaban estos casos, ya que entendieron que las necesidades han cambiado dado que los asegurados se han puesto más exigentes. En palabras de Muñoz Pardo (2018):

> Años atrás la gestión de un siniestro hasta su cierre se podía demorar y no se recibían quejas, hoy en día una lenta gestión de un siniestro puede conllevar además de las correspondientes quejas a una pérdida de la póliza y a una publicidad negativa.
>
> […] Bajar los tiempos de cierre de un siniestro es importante y también será un objetivo para las aseguradoras, pero se hace más transcendental la inmediatez en la respuesta a una consulta del asegurado y la información actualizada del estado del siniestro dado que acaba creando una mejor experiencia con el cliente. (p. 52)

Es por ello que otra de las aplicaciones de la IA que dicho sector comienza a aplicar es, y de acuerdo con su búsqueda de automatizar tareas, el incremento en la eficiencia y efectividad tanto en el análisis de riesgos como en la evaluación de reclamos. Y es que "las aseguradoras tienen una elevada cantidad de documentación con la que han de trabajar. Destinar recursos físicos para gestionar esta documentación provoca que se consuman muchos recursos y se aumenten los costes" (Muñoz Pardo, 2018, p. 52). Motivos para el cambio sobran, y lo confirma la consultora McKinsey & Company, la cual informa que "la digitalización de unos 20/30 procesos clave ayudaría a poder reducir en un 50% los gastos totales de una compañía" (Muñoz Pardo, 2018, p. 52). Es por tal motivo que, Lovagnini (2023) comenta, gracias a el soporte de ChatGPT, que "[…] la IA puede analizar datos históricos y en tiempo real para evaluar los riesgos asociados con diferentes perfiles de asegurados. Esto ayuda a las compañías de seguros a tomar decisiones más informadas al emitir pólizas y calcular primas" (p. 18). En ese aspecto, considero que no solamente se limita a la emisión y *pricing* de la póliza, sino también a facilitar el análisis de riesgo de los reclamos recibidos. Y en relación con este punto, precisamente Lovagnini (2023) amplía que "[…] la IA puede ayudar en la automatización y agilización del

proceso de evaluación de reclamaciones, mejorando la eficiencia y reduciendo los costos operativos" (p. 18).

Algunas de estas automatizaciones se han nombrado y ya son una realidad en el mercado, y otras por lo menos comenzaron a idearse desde la irrupción de las *Insurtech*. Precisamente en ese aspecto, Muñoz Pardo (2018), ha realizado en dicho momento algunas presunciones de las áreas o aspectos a automatizar:

1. **Utilización de aplicaciones de mensajería instantánea para la gestión documental**: La utilización de una aplicación de mensajería instantánea como podría ser *whatsapp* podría ayudar al proceso de la inmediatez de respuesta para las compañías aseguradoras. La recepción de comunicaciones o de documentación o fotografías a través de esta aplicación podría mejorar los tiempos de respuesta en gran medida. Además […], la compañía facilitaría el contacto al cliente a través de una de las aplicaciones más utilizadas por los clientes. Facilitando así la comunicación con la compañía y adaptándose a los nuevos hábitos de los consumidores.

   […] Las compañías deberán adaptar sus programas y dirigirse hacia una estrategia de omnicanalidad, en la cual todos los canales estarán interrelacionados y los clientes puedan utilizar el que más se adapte a ellos. (pp. 53-54)
2. **Utilización de una aplicación móvil para el seguimiento e interacción del siniestro:** […] Otro de los procesos de mejora en las que deben incurrir las aseguradoras, es la utilización de esta aplicación móvil para que los clientes puedan no solo contactar con la aseguradora sino llegar a interactuar de manera que también puedan aperturar un nuevo siniestro, realizar consultas y seguimiento del expediente, adjuntar documentación o incluso llegar a cambiar citas con perito o reparadores. (p. 54)
3. **Utilización de *chatbots* para la apertura de siniestros:** Los chatbots pueden ser un gran aliado para la gestión de siniestros para ayudar al cliente en la apertura de un nuevo siniestro o para realizar una consulta sobre garantías.

   […] la utilización de un *chatbot* en el proceso de apertura de un siniestro puede ser la solución. El *chatbot* guiara al asegurado durante todo el proceso ayudándolo a describir correctamente lo que ha ocurrido y recogiendo los datos que la compañía considera necesarios.

   […] la utilización de un *chatbot*, […] podría comunicar al asegurado si podría tener o no cobertura el siniestro.

> La utilización de *bots* ayudará a reducir el porcentaje de errores, a mejorar la velocidad de respuesta y por lo tanto mejorará la calidad del servicio que se ofrece al cliente. (pp. 54-56)

Lo cierto es que la atención personalizada a través del uso de *chatbots* no se limita específicamente a la atención de consultas y recomendación de productos, y lo todo lo mencionado en la página anterior es una realidad en muchas compañías en la actualidad. Y es que la digitalización, omnicanalidad e implementación de IA vienen a resolver todo lo que se ha mencionado anteriormente, en línea con la necesidad de eliminar la asimetría, acercarse al asegurado, lograr una conexión con este y brindarle seguridad y transparencia en el servicio que les proveen. Un buen caso de uso que se puede presentar es el mencionado por Rodríguez (2023), quien con orgullo establece que:

> En Hipotecario Seguros implementamos un *chatbot* que resuelve el 30 por ciento de las consultas de nuestros asegurados y los resultados han sido satisfactorios. Además, recientemente hemos introducido la opción de realizar la denuncia de siniestros de forma digital e inteligente y ya un 25 por ciento de nuestros asegurados ha optado por utilizarla. La ventaja diferencial radica en que estamos en constante comunicación con los clientes siniestrados para mantenerlos informados sobre los avances en la resolución de sus casos a través de canales digitales, como email marketing. (citado en Fiorentino, p. 34)

Finalmente, es importante mencionar que las *Insurtech* del mercado comienzan a desarrollar soluciones específicas para necesidades igual de específicas acorde a la coyuntura del mercado de cada país. A su vez, comienzan a idear futuras soluciones a medida que no solo la tecnología de las compañías evolucione, sino también que la sociedad adopte nuevos dispositivos que acompañen a esa transformación.

En relación con el primer punto, una aplicación de éxito y necesaria en el mercado argentino es el provisto por *Claims Services*, uno de sus proveedores de tecnología, y que ya brinda su servicio a varias compañías aseguradoras en la Argentina. El servicio de mención consiste en desplegar un sistema de licitación de repuestos para que cada compañía seleccione, al mejor precio, el repuesto necesario en una reparación de un automóvil en función de la oferta provista por una red de proveedores de repuestos (o más comúnmente denominados repuesteros). En palabras de su CEO, Leonardo Valseche (2023):

> A raíz de las trabas a la importación y los cambios del dólar, entre muchos otros factores, nuestra plataforma *ClaimsServices* Repuestos ha crecido mucho. En un contexto desfavorable ha logrado hacer frente a estas dificultades, solucionando efectivamente uno de los grandes problemas del mercado: conseguir el repuesto al mejor precio y garantizar la entrega en tiempo y forma a través del seguimiento. (citado en Perazo, p. 84)

Y en relación con el segundo punto, Muñoz Pardo (2018) mencionó algunas aplicaciones que el mercado posiblemente hará uso en un futuro cercano:

- **Utilización de dispositivos de *Internet of Things* (IoT)**: Se trata de dispositivos o sensores que permiten a la compañía captar y recoger una gran cantidad de información.

  […] El objetivo es que todos estos dispositivos inteligentes que se pueden instalar en la vivienda, comunidad u oficina, se puedan comunicar entre sí, sean más inteligentes y que se puedan controlar a través de smartphone o *smartwaches*.

  […] La aplicación de sensores permitirá poder automatizar el riesgo, pudiendo conectar a internet estos sistemas y controlarlos desde cualquier lugar a través de un smartphone, por ejemplo.

  […] Al utilizar esta tecnología se recibirá una gran cantidad de información de la vida de los clientes. Una vez analizada esta información, se podrá segmentar a los clientes y ofrecer servicios y productos más personalizados.

  […] La utilización de sensores ayudará a bajar la siniestralidad y por tanto a reducir los costes medios de los siniestros […]. (pp. 58-60)
- **Análisis de datos y prevención de siniestros:** Las aseguradoras ya no ofrecerán una rápida indemnización o reparación, sino que el futuro está en la prevención del siniestro.

  A medida que avance la domótica y automatización de los riesgos y la instalación de sensores por parte de las compañías se podrán analizar los siniestros |que hayan ocurrido para analizarlos y determinar si existe algún nexo o relación.

  […] A partir de los resultados de los análisis realizados, en el caso de encontrar relaciones o nexos en la ocurrencia de un siniestro, se podrá facilitar a los asegurados dicha información para que reparen o modifiquen sus instalaciones antes de la

ocurrencia de un siniestro, o se podrán ofrecer reparación o servicios preventivos […].

El objetivo será el de analizar y detectar que servicios preventivos o sensores pueden ser de utilidad para reducir la siniestralidad, y siempre teniendo en cuenta el coste que conlleva llevar a cabo o instalar estos servicios o sensores para que no sobrepase los beneficios que repercutan los mismos. (pp. 62-63)

- **Nuevos servicios a ofrecer gracias a la utilización de sensores:** Aparte de la prevención de siniestros, la utilización de estos dispositivos que están conectados a la red permitirá a las compañías ofrecer nuevos servicios al asegurado como el de avisarle de la ocurrencia de un siniestro antes de que lo descubra, al poder detectar la compañía la presencia de un siniestro gracias a los sensores.

  Al detectar el siniestro antes que el asegurado se le podrá informar a través de la aplicación de la compañía en su smartphone de la ocurrencia del mismo, ofreciéndole la posibilidad de enviarle a un reparador o de realizarle la indemnización en el caso de que tuviéramos cuantificados los daños o únicamente se hubiese producido una fuga. (p. 63)

Si bien el internet de las cosas o IoT es una tecnología que tarde o temprano brindará un gran soporte a las aseguradoras, lo cierto es que las soluciones que enumeramos basadas en ella aún distan de ser una realidad en muchas regiones del mundo. Esto se debe principalmente a dos factores: en primer lugar, por el elevado costo asociado a la compra e instalación de algunos de estos dispositivos basados en sensores; en segundo lugar, porque para que las compañías de seguros puedan utilizar dicha información es necesario previamente tener una infraestructura moderna de datos para dar soporte a todo el proceso de recolección, transmisión, almacenamiento y análisis de riesgos de los datos provistos por los dispositivos de mención.

1.3. Riesgo judicial en la actualidad y abordaje del servicio a asegurados

Una vez provista tanto la situación actual del mercado asegurador, como las aplicaciones de la IA que se han implementado en este, se profundizará en las dos temáticas que el presente trabajo final integrador pretende investigar: el entendimiento de la gestión del servicio post siniestro y como las compañías lo analizan, y la gestión del riesgo judicial en la actualidad.

En cuanto a la primera temática, ya hemos enunciado en los subapartados anteriores que el interés en la segmentación y personalización de los productos de cara a los asegurados es parte de las prioridades de la industria. Pero a la vez, se ha mencionado que ese foco se manifestó en

las áreas de producto o marketing, impulsado por el interés comercial tanto de retener como de capturar nuevos clientes, así como también ampliar la oferta de servicios a los asegurados actuales y atender las consultas diarias de estos. Si a eso le sumamos el hecho de que el servicio que brinda una compañía de seguros en un siniestro alcanza a una fracción del total de su cartera, pensar en la segmentación de clientes para brindar el servicio puede que no se entienda como algo demasiado obvio y, por ende, no se considere como prioritario.

Esto no significa que las compañías de seguros no se interesen en entender cómo brindan su servicio con el afán de intentar mejorarlo, todo lo contrario, sino que el foco es distinto. Una de las principales herramientas que utilizan para tal fin es el envío de encuestas a los asegurados post gestión de un siniestro, con el propósito de entender la efectividad y eficiencia de sus procesos, a la vez que intenta retener, mediante una estrategia basada en incentivos, a la cartera de asegurados que han tenido malas experiencias. La medición de dichas encuestas se suele realizar mediante la utilización de un método o índice denominado NPS (*Net Promoter Score*), el cual brinda un sistema de puntuación y medición específico (que se detallará más adelante) basado en las calificaciones de los asegurados luego de contestar dichas encuestas. Este análisis ex post no deja de ser un análisis descriptivo, el cual tiene dos propósitos: mitigar los efectos de las malas experiencias en asegurados cuando el proceso no tiene éxito, y entender las maneras de mejorar los procesos existentes a modo de evitar que vuelva a suceder. Según Lotko (2015), la filosofía y el origen de este índice se remonta a 2003:

> *F. Reichheld introduced Net Promoter Score index (NPS index) based on customers willingness to recommend a company to a friend (Reichheld, 2003).*
>
> *[…] The idea behind NPS index is very simple and is based on asking customers just one question: „How likely is it that you would recommend us to a friend or colleague?" on a 10-point scale. Subsequently, customers are clustered into 3 groups: promoters, passively satisfied and detractors. Customers who will rate 9 or 10 are considered as promoters, 7-8 as passively satisfied (neutrals, fence sitters) and the rest are classified as detractors. Each group is linked with expected customer behavior. Promoters are likely to stay with a company in case of emergence of competitors. Moreover, they are more likely to repeat purchases. Finally, they may have positive impact on other potential customers. Therefore promoters are expected to contribute to growth of company performance. On the other hand, detractors have negative impact on business performance expectations. They are likely to create negative opinions or switch to*

> *competitors. NPS is calculated as subtraction between share of promoters and share of detractors (NPS = P – D). NPS index above 0 is considered as positive and the value above 50 means excellent situation (Reichheld 2003).*

Esta explicación indica que el índice busca también agrupar a los clientes en grupos, a modo de poder entender qué se espera más adelante sobre el accionar de cada uno de los tres grupos de clientes (encuestados) que menciona el autor.

Ante esto, Lotko (2015) en su artículo nos muestra un estudio en el cual buscó determinar si los clientes de un centro de contacto (o *call center*) pueden agruparse en clústeres homogéneos basados en su evaluación de la calidad del servicio, si estos clústeres coinciden con la clasificación del índice NPS (promotores, neutrales y detractores), y cuáles son las características de cada clúster. La investigación busca tanto explicar la estructura de los grupos (clústeres) que unen a los clientes del centro de contacto en función de su percepción de la calidad de los servicios, como también buscar si las características en común de cada clúster pueden ser utilizadas para diseñar servicios específicos, con especial atención a la calidad, para satisfacer las expectativas de estos clientes. Para develar el misterio, los resultados mostraron que la calidad del servicio en los centros de contacto puede evaluarse eficazmente mediante el análisis de clúster, y que estos clústeres coincidieron con la clasificación del índice NPS. Por ende, este enfoque permite comprender mejor el comportamiento de los clientes y sus percepciones de calidad, y el autor abre la puerta para utilizarlo y así ayudar a diseñar servicios que satisfagan mejor sus expectativas (pp. 6-26).

El ensayo que se compartirá más adelante tiene un objetivo similar, pero obviamente aplicado al servicio de siniestros, y con la diferencia de que no haremos foco en la hipótesis sobre el entendimiento de si los clústers resultantes coinciden o no con la agrupación del NPS. De la investigación realizada para el presente trabajo final integrador no se ha encontrado algo similar aplicado en siniestros, por lo que se avanzará con el ensayo para realizar un primer entendimiento en esta área.

La segunda temática abordará un problema de negocio que, desde el punto de vista del presente autor, es una realidad y prioridad de negocio, pero en donde paradójicamente no se prioriza la inversión en modelos predictivos para mitigarlo. Hablamos, nada más y nada menos, del riesgo judicial que tienen los reclamos de terceros, riesgo que se encuentra latente en algunas partes del mundo pero que ha penetrado con fuerza en la Argentina.

Precisamente, la redacción de Clarín (2006/2017) nos menciona en su artículo que en este país se implementó en el año 1996, y a través de la Ley 26094, un sistema de mediación obligatoria previo al juicio a modo de intentar mitigar la judicialización. Y este sistema ha tenido éxito, dado que menciona que, en los 10 años que transcurrieron entre su implementación y la fecha del artículo original (2006), el 65% de los reclamos extrajudiciales que pasaron a mediación han sido resueltos en esa instancia y, por ende, no han traspasado a juicio. El artículo continúa y dice que lo que busca esta instancia es intentar acercar a las dos partes del conflicto (llamados partes requirente y requerido), en la cual intermedia un tercero neutral (denominado mediador), con la intención de intentar resolver su disputa, acelerar los tiempos, abaratar los costos y aliviar la congestión del sistema judicial. Esta obligatoriedad rige para distintas áreas o ramas del derecho, como el patrimonial, de familia, sucesiones, daños y perjuicios, entre otros. Esto quiere decir que alcanza a los reclamos en seguros, aunque en ese sentido el artículo aclara que, según la doctora Silvana Greco de la Fundación Libra, a diferencia del promedio de todas las ramas del derecho, en los accidentes de tránsito (aplicables a las coberturas de automotores y accidentes personales) la cantidad de juicios se reduce a más de un 50%. Finalmente, y no menos importante, se enuncia en el artículo que, en la Argentina y en promedio, un juicio dura aproximadamente cuatro años debido a los plazos extensos de la burocracia judicial, mientras que un acuerdo en mediación necesita de algunas reuniones para que las partes definan un acuerdo.

Si bien no se ha encontrado información más actual del estado de situación de estas instancias, lo cierto es que, en función de la experiencia del presente autor, no ha variado demasiado este número a pesar de la irrupción de la pandemia, por lo que nos sirve de referencia para el argumento de este ensayo. De hecho, la SSN provee información pública con estadísticas del mercado asegurador, en las cuales informa los estados contables y resultados operativos del mercado, tanto a nivel general como por compañía. El inconveniente radica en el hecho de que, en lo que refiere exclusivamente a los reclamos judiciales, los informes sumarizan o integran en sus indicadores tanto a los juicios como las mediaciones, salvo en los pasivos. Precisamente, gracias a esta investigación, hemos podido deducir algunos indicadores que muestran una comparación económica aproximada entre la instancia de mediación y la de juicio. Por ejemplo, la SSN presentó en su publicación de indicadores del mercado asegurador a diciembre 2023, la cantidad de juicios pendientes (recordar que integra juicios más mediaciones) del ramo patrimoniales y mixtos, equivalente a 216.818 casos (SSN, 2023b). Por otra parte, en otro de sus informes denominado situación del mercado asegurador a diciembre 2023, informa en su

cuadro número 5 a los pasivos[2] de las mediaciones (AR$ 75.004.603.249) y de los juicios (AR$ 469.955.253.502), esta vez sí de manera desglosada, y también para el ramo de patrimoniales y mixtos (SSN, 2023a). Si aplicamos la relación 35 / 65 que afirma Clarín (2006/2017) existe entre mediaciones que se acuerdan y mediaciones que traspasan a juicio, podemos deducir que, de los 216.811 casos, unos 75.884 corresponden a mediaciones y los restantes 140.927 a juicios, aproximadamente. Si estos números los aplicamos sobre los pasivos ya desglosados, podemos deducir que las reservas promedio por cada caso son de 532.223 para las mediaciones y de 6.193.074 para los juicios, lo que nos muestra que el costo medio de un juicio equivale a 10,6 veces más que el de una mediación, aproximadamente. Y decimos que es aproximado porque la deducción se basa en las reservas de los casos (estimaciones de lo que se pagará) y no en sus pagos (lo que realmente se erogó por su indemnización), los cuales, si bien se informan en los reportes de la SSN, se hacen de forma agregada ya que agrupa toda la cartera (pagos extrajudiciales, en mediación y en juicio).

Podemos resumir entonces que, atento a lo expuesto en los últimos dos párrafos, permitir que un caso traspase a juicio equivale a aceptar que se pagará, en promedio, alrededor de 10 veces más lo que se pagaría en mediación, y se demoraría también alrededor de 10 veces más, si se tiene presente que una mediación tiene aproximadamente una demora entre 3 y 6 meses (promedio de 4,5 meses), y un juicio demora en promedio 4 años. Si a esto le sumamos el factor inflacionario que se trasladan a las tasas judiciales, este riesgo de traspaso judicial es clave para cualquier compañía de seguros, y el hecho de poder predecirlo con alguna metodología de machine learning debiera al menos contemplarse en su estrategia de mediano plazo.

Ahora bien, como veremos más adelante en detalle durante el desarrollo del ensayo que se aplicará a esta temática, más precisamente en el subapartado 2.1, es importante mencionar que la debilidad en general de este proceso es la pobre calidad de los datos que usualmente se tiene al recibir un reclamo, sea en instancia extrajudicial como en mediación. Y si estos datos son de mala calidad o no están completos, la aplicación de un modelo que prediga el riesgo judicial de traspaso a juicio al recibir el reclamo del tercero no será del todo efectivo y puede conducir a interpretaciones erróneas. Debido a esto, las compañías deberán mejorar sus estrategias de recolección de datos en la notificación de los siniestros, ya que mismo hoy en día se presenta

[2] En seguros, el pasivo refiere al monto reservado que tiene la cartera de juicios y mediaciones poder afrontar el pago futuro de un siniestro.

como un desafío para el mercado cuando debe determinar la responsabilidad del asegurado en el hecho. Así lo afirma Carlos Vitagliano (2012), al mencionar que:

> En muchos casos la mala evaluación de la responsabilidad por falta de elementos objetivos o por un análisis reducido a los elementos básicos recogidos durante la denuncia, hace que no se termine comprendiendo la verdadera magnitud del siniestro y se concluya dictaminando la inexistencia de responsabilidad u otras circunstancias que luego no pueden ser validadas en el curso de las diferentes instancias pre o judiciales.

Finalmente, y de forma similar a lo sucedido con la primera temática, no se han encontrado ensayos o trabajos orientados a la predicción de traspasos de instancia judicial. Lo más similar que se ha encontrado es una investigación realizada por Morselli, Masias, Crespo, y Laengle (2013), la cual tenía el propósito de predecir los resultados de las sentencias en juicios penales de una red de narcotráfico de la Red Caviar, basándose en información extraída en redes sociales. Esta investigación utilizó medidas de centralidad para predecir el veredicto (inocente o culpable), y la duración de la sentencia en años. En el estudio se aplicaron modelos de regresión logística y modelos lineales para predecir estos resultados, y aplicaron las medidas de centralidad. Más allá de los resultados específicos que obtuvo, lo importante aquí es que nos muestra que modelos como la regresión logística son efectivos para estos modelos de clasificación (dato binario: inocente o culpable), el cual es uno de los dos modelos que se utilizará en el correspondiente ensayo. También corrobora que la predicción del comportamiento de cierre de las mediaciones no es en la actualidad el foco de las industrias, donde se incluye la industria de seguros.

Reclamos de terceros: Predicción del riesgo de judicialización

> Los conflictos jurídicos que hoy enfrentan las aseguradoras siguen siendo los altos porcentajes de incapacidad que fijan los peritos médicos y/o legistas, las elevadas tasas de interés, los altos montos de condena y la aplicación de la ley de Defensa del Consumidor a los contratos de seguros.
>
> En Automotores, el descontrol que implica la actuación de los peritos de oficio en la justicia y la disparidad de criterios en distintas sedes judiciales implican mayores dificultades en la negociación de los siniestros en etapa extrajudicial y mediación. Así es como crece la litigiosidad. (Carelli, 2020).

Una vez puesto en contexto al lector sobre la actualidad del negocio y las áreas en las cuáles el presente trabajo final integrador (TFI) pretende profundizar, se avanzará en el presente capítulo en el primer ensayo de los dos propuestos. El ensayo busca identificar el riesgo judicial que posee la notificación de una mediación presentada por un tercero cuando ingresa a una compañía o, en otras palabras, el propósito de este radica en encontrar un modelo lo suficientemente bueno como para arrojar a un analista, durante el ingreso del ingreso de la mediación, un indicador porcentual con el riesgo potencial que posee un reclamo de que traspase a la etapa de juicio. El objetivo final es que, gracias a la implementación de dicho modelo, se evite que las mediaciones futuras traspasen a dicha instancia. Esto permitirá evitar los costos excesivos que conlleva la gestión y pago de una indemnización en dicha etapa, como asimismo lograr eficiencia operativa al buscar su cierre en mediación con la aplicación de alguna estrategia de cierre más agresiva. Por ende, lo que se busca resolver es un problema de clasificación (Traspaso SI o NO).

A tal fin, en un primer subapartado, se presenta el problema de negocio que se busca resolver con más detalle, junto con la descripción del *dataset* utilizado para realizar la prueba y desarrollar el modelo. En un segundo subapartado, se profundizará en el análisis del *dataset* y las variables que contiene luego de realizar su correspondiente análisis exploratorio. En un tercer subapartado, se muestran los pasos realizados para preparar los datos antes de comenzar con el entrenamiento del modelo. El cuarto subapartado detalla cómo ha sido el diseño y evaluación de los dos modelos predictivos seleccionados. Y en un quinto subapartado, por último, se procede a evaluar los resultados para seleccionar el modelo más propicio para predecir un traspaso.

2.1. Descripción del problema de negocio y de la base de datos seleccionada

Inicialmente, y antes del comienzo del ensayo, se había ideado la posibilidad de dividir el análisis en dos partes, correspondientes al traspaso de cada instancia, es decir: el traspaso de un reclamo extrajudicial a una mediación, y el traspaso de un reclamo en mediación a juicio. El inconveniente radica en que las bases de datos utilizadas para predecir cada traspaso de instancia son diferentes, y el análisis se debe hacer sobre dos instancias de captura de información también totalmente diferentes. Como se mencionaba anteriormente, para poder tener éxito en el ensayo, este debe aplicarse sobre la información que se obtiene al inicio de la gestión de cada reclamo, y en el caso de un reclamo extrajudicial la información que se obtiene es escasa, nula, o sujeta a verificación por parte de un estudio liquidador. Esto se debe a que, o

bien no se recibe el reclamo por parte directa del tercero sino por parte del asegurado, donde en ocasiones este último omite información clave, o anota mal la información provista por el tercero, o directamente no registra los datos del tercero (ya que el interés de la denuncia se focaliza en los daños propios), o bien el tercero presenta el reclamo con información incompleta por desconocimiento de cómo completar la denuncia. Ni mencionar el hecho de que, en varias ocasiones, ni el asegurado ni el tercero presentan el reclamo extrajudicial y llega, directamente, la notificación de la mediación sin aviso previo y por recomendación de un abogado allegado.

Todos los escenarios descriptos anteriormente plantean un real desafío al intentar predecir los traspasos de un reclamo extrajudicial a una mediación, y es por ello que se descartó su análisis en el presente TFI. Es por dicho motivo que finalmente se seleccionó la segunda base de mención: la de *notificaciones de mediación*, a modo de predecir los traspasos de mediación a juicio, la cual es más completa y fiable en cuanto a la calidad de sus datos, ya que provienen de cartas documento enviadas por los terceros reclamantes.

Por otra parte, se recuerda al lector que el modelo a resolver plantea un *problema de clasificación*, por lo que para el presente experimento se seleccionaron finalmente dos modelos que se consideran adecuados para resolver este tipo de problema: *Regresión Logística* y *Decision Forest*. Los motivos se enunciarán en el subapartado 2.4.

El indicador de negocio por excelencia que mide la performance del modelo a posteriori es el *transfer ratio*, una tasa que mide la cantidad de traspasos (en cada instancia) sobre el total de casos ingresados en la instancia anterior. Por ejemplo, el transfer ratio en mediaciones representa la cantidad de mediaciones ingresadas del mes dividida por la cantidad de denuncias extrajudiciales de ese mes, es decir, se compara contra los ingresos de la instancia anterior. Lo mismo sucede para las mediaciones que traspasaron a instancia Judicial. Es importante mencionar dicho indicador porque es la forma en que la compañía podrá comprobar si, una vez aplicado el modelo, hubo éxito o no en la implementación y nuevo proceso de gestión. Pero es igual de importante dejar en claro que queda por fuera del alcance de este ensayo su cálculo, en primer lugar, porque es necesario que el modelo sea exitoso y se despliegue en producción en una compañía y, en segundo lugar, porque implica obtener y cruzar otras bases que están fuera del objetivo del ensayo.

Como se adelantó inicialmente, al resultar que no existe actualmente en el negocio un modelo predictivo vigente que permita mejorar la toma de decisiones al momento de negociar la indemnización del reclamo con el tercero, la posición que se espera de una gerencia en una

primera iteración es la de implementar un modelo que permita identificar al menos un porcentaje de potenciales traspasos, a modo de realizar una negociación más agresiva con el tercero y evitar el correspondiente traspaso. Este punto es relevante ya que *no se espera tener una alta tasa de verdaderos positivos a costa de aumentar los falsos positivos*, ya que caso contrario podría aumentar el costo medio de la instancia actual más allá de lo pretendido (se cerrarían casos por encima del promedio de indemnización cuando quizá no tenían riesgo de traspasarse a mediación o juicio), y ya con el hecho de identificar y evitar un porcentaje de traspasos (aunque no supere el 50%) es mejor que no identificar ninguno.

Asimismo, tampoco se espera una alta tasa de verdaderos positivos dado que varios de los predictores que se necesitarían para mejorar el *accuracy* se capturan durante la gestión del siniestro (a posteriori), y no al momento del ingreso del reclamo, y a priori se espera que el modelo se ejecute con el ingreso de la notificación de la mediación, momento en el cual se espera dar una definición del tipo de estrategia a utilizar para darle las indicaciones correspondientes al estudio de abogados al momento de derivarle el caso (cada día que pasa, aumenta el riesgo de traspaso).

En relación al *dataset*, este se conformó por un *joineo* o cruce de tablas de la base de datos de la compañía, en donde mediante la utilización de SQL se cruzó información de las notificaciones de mediación durante todo el año 2019, junto con otra base con información histórica de los reclamos judiciales pero con información actualizada a Septiembre de 2022, la cual se utilizó para traer algunos campos adicionales y entender cuáles de todas esas notificaciones se traspasó a Juicio en ese lapso de menos de tres años. La tabla resultante inicial contenía 49 columnas, con un total de 2.671 registros que corresponden a cada mediación ingresada en dicho año. Posteriormente se realizó un análisis para reducir la cantidad de variables a analizar, en donde se quitaron aquellas que eran nulas, redundantes o irrelevantes para el análisis, por lo que terminó utilizándose un *dataset* de 17 columnas para el desarrollo del modelo.

Las variables o predictores seleccionados de la base son las siguientes:

1. **Número de Juicio (NRO_JUICIO)**: Código ID que identifica de forma unívoca al reclamo judicial, ya sea que se encuentre en instancia de mediación o juicio, por lo que es un mismo código para ambas instancias.

2. **Fecha de notificación del reclamo a la compañía (FEC_NOTIF_CIA)**: Fecha en que la compañía recibe la carta documento de la mediación y toma conocimiento del reclamo.
3. **Abogado que representa a la compañía (ABOGADO_CIA)**: Abogado o estudio de abogados que selecciona la compañía para que la represente, defienda y negocie durante la gestión de la mediación.
4. **Abogado que representa al tercero (ABOGADO_3ERO)**: Abogado o estudio de abogados que representa al tercero en la negociación y que, entre otras cosas, envía la carta documento a modo de presentar el reclamo formal.
5. **Posición de la Compañía (POSICION_CIA)**: Posición de la compañía en el reclamo de mediación o juicio. Los códigos son G (en Garantía) o S (Sindicada), donde la primera es un caso donde la aseguradora proporciona una garantía o aval al asegurado a través del contrato del seguro, y la segunda, se presenta cuando la compañía es acusada o implicada en el proceso judicial.
6. **Tipo de reclamo (TIPO_RC)**: Código que tipifica el reclamo del tercero y se divide en 3: RCC (Responsabilidad civil cosas), RCL (Responsabilidad civil lesiones) o RCHO (Responsabilidad civil homicidios).
7. **Indeterminado:** Indica si la demanda en mediación es determinada, es decir, si el tercero deja explícito o no un monto de demanda en el reclamo por el cual solicita ser resarcido.
8. **Cobertura**: Análisis preliminar que decreta si la póliza asociada perteneciente al asegurado, al cual el tercero reclama, posee cobertura o no. Los códigos son N (NO) o S (SI). Esto puede determinar un posible rechazo inicial, pero si el caso prospera a Juicio el Juez asignado puede llegar a determinar que, por más de no poseer cobertura, la aseguradora debe pagar por la responsabilidad del hecho de todas maneras.
9. **Fecha de Siniestro (FEC_STRO)**: Fecha en que ocurrió el siniestro o evento por el cual el tercero reclama una indemnización por daños y perjuicios.
10. **Código de Jurisdicción (COD_JURISDICCION)**: Código que representa la jurisdicción donde se presentó el reclamo, correspondiente en general a un municipio.
11. **Código de Fuero (COD_FUERO)**: Un fuero refiere a la competencia que tiene un tribunal o juez para conocer y resolver un determinado caso o tipo de asuntos. Los

códigos existentes en la base son CI (Civil), CC (Civil y Comercial) y NN (No informado).

12. **Código de Juzgado (COD_JUZGADO)**: Código correspondiente con el Juzgado dentro de la jurisdicción.
13. **Productor**: Intermediario entre el Asegurado y la compañía de seguros.
14. **Fecha de Demanda (FECHA_DEMANDA)**: Fecha en que ingresó la demanda de mediación.
15. **Tipo de Siniestro (TIPO_DE_SINIESTRO)**: Listado con las tipificaciones con las distintas clases de eventos que describen sintéticamente la causa del siniestro.
16. **Demanda**: Monto total que el tercero reclama en la carta documento de la mediación, en caso de haber.
17. **Traspaso de Juicio**: Indica si el caso finalmente traspasó o no a instancia de Juicio, por lo que resulta ser la variable a predecir.

A modo de mantener una absoluta confidencialidad, la base fue anonimizada y transformada para conservar la privacidad de los asegurados, terceros, proveedores y agentes que trabajan con la compañía de seguros seleccionada, así como la compañía en sí.

2.2. Análisis exploratorio de las variables predictoras y a predecir

En el presente subapartado se realizará un Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) de la base finalmente seleccionada, el cual es un enfoque analítico que se utiliza para resumir las principales características de un conjunto de datos. Su objetivo principal es comprender la estructura subyacente de los datos, identificar patrones, detectar anomalías, probar hipótesis y verificar supuestos a través del uso de estadísticas descriptivas y visualizaciones gráficas.

En esta oportunidad, el perfilado de datos para realizar el análisis exploratorio se realizó en un notebook con *python*, en el cual se aplicó la función *ProfileReport* de la librería pandas. El script se adjunta en el apartado Apéndices para consultar con mayor detalle. El *profiling report* obtenido arrojó una serie de estadísticas y gráficos relevantes para comprender el *dataset* en detalle, los cuales se compartirán y describirán a continuación.

En primer lugar, se presenta un *overview* o visión general el cual muestra cuestiones genéricas, donde algunas ya fueron mencionadas anteriormente. Por ejemplo, indica que el *dataset* contiene 17 variables, 2.671 observaciones, 216 celdas vacías (equivalente a un 0,5% del total

de campos), no tiene registros duplicados, y finalmente, de dichas variables: cuatro son campos de texto (*text*), dos son campos de fecha (*datetime*), ocho son variables categóricas (*categorical*) y tres son numéricas (*numeric*).

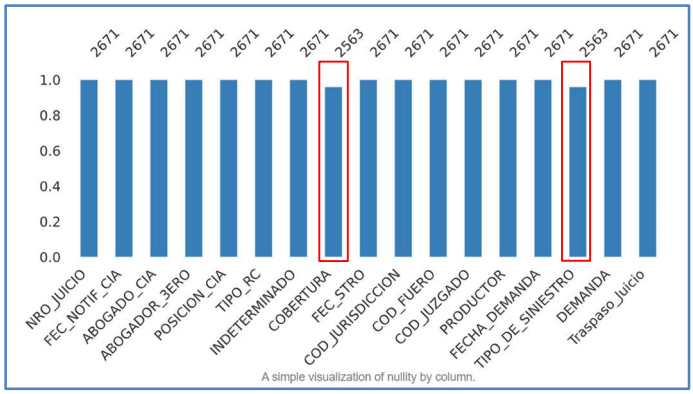
Posteriormente, *pandas profiling report* brinda un interesante sistema de alertas, donde indica cuestiones relevantes a tener presente en el *dataset*, entre las que se encuentra las variables que tienen una alta correlación (que se abordará posteriormente en detalle dentro de este subapartado), aquellos datos que están desbalanceados (existencia de más registros de una clase o categoría que otra/s), variables con valores faltantes y, por último, variables con valores únicos. Se comparte el detalle del análisis del reporte:

Figura 1 - Alertas del Profiling Report

En cuanto al desbalanceo en las variables predictoras, este no siempre perjudica el análisis o la construcción de modelos de predicción, más aún si el desbalance es una característica natural del negocio. En dicho caso, puede no ser necesario balancear las variables predictoras, donde forzar un balance podría introducir ruido y hacer que el modelo aprenda patrones que no son representativos del negocio real. Este es el caso del *dataset* presentado donde, por ejemplo, la mayoría de las demandas que recibe una compañía de seguros es para que se presente "en garantía" (valor G en campo POSICION_CIA), o la mayoría de las demandas recibidas se dirigen a una póliza que posee cobertura de responsabilidad civil (en la cartera automotor, por ejemplo, es obligatorio legalmente y es la cartera de mayor frecuencia), a la vez que hay desbalanceos en el fuero dado que la mayoría de las demandas en mediación no lo indican. Asimismo, modelos como el *decision forest* pueden manejar variables desbalanceadas sin demasiados problemas.

Respecto a los valores faltantes, anteriormente se indicó que representan el 0,5% del total del *dataset*, y solo se presentan en dos variables: COBERTURA y TIPO_DE_SINIESTRO, las cuales poseen 108 datos faltantes cada una. Dado que los modelos de machine learning seleccionados para el ensayo son la regresión logística y el *decision forest*, el tratamiento de estos tiene que ir en consonancia con los requisitos para implementar dichos modelos. Y si bien los bosques de decisión (que incluyen algoritmos como *decision forest*) pueden manejar y tratar los valores faltantes de manera más flexible, lo cierto es que la regresión logística no lo puede hacer y, por ende, dichos faltantes deberán ser tratados antes de entrenar el modelo. Hay diversas técnicas para tratarlos, como la imputación de la moda en variables categóricas o *k-means* (KNN), el cual utiliza los valores de las observaciones más cercanas para imputar los valores faltantes. Esta problemática y su resolución serán abordadas en el subapartado 2.3. El gráfico resultante del reporte que presenta los datos ausentes se comparte a continuación:

Figura 2 - Análisis valores nulos en dataset



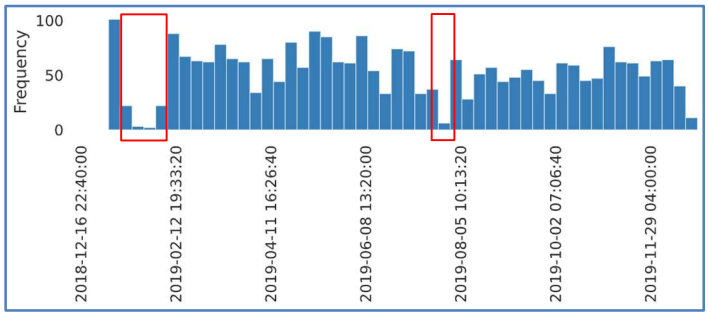
En tercer lugar, el análisis profundiza en las diferentes características de cada variable, las cuales compartimos a continuación:

1. **Número de Juicio (NRO_JUICIO)**: Naturalmente, al tratarse de un ID, no tiene datos faltantes ni duplicados, por lo que el análisis arroja que todos los valores del campo son únicos. Como la variable fue anonimizada, no se analizará otra característica provista por el reporte.

2. **Fecha de notificación del reclamo a la compañía (FEC_NOTIF_CIA)**: Dada el alcance de la base, es lógico que el mínimo sea el 01/01/2019 y el valor máximo sea el 27/12/2019 (último día hábil del año para presentar una carta documento). En total tuvo 222 valores distintos (prácticamente en línea con los días laborables de aquel año, de 246), y al ver su frecuencia en un histograma, vemos lógicamente que a principios de enero es cuando más se registraron casos (demandas que se enviaron

a fines de diciembre y que ingresaron tarde a la compañía la primera semana de enero) y el resto de enero más la segunda quincena de Julio pasó lo inverso, dada la feria judicial. Presentamos dicho histograma a continuación:

Figura 3 - Histograma campo FEC_NOTIF_CIA



3. **Abogado que representa a la compañía (ABOGADO_CIA)**: Se identificaron un total de 22 valores (o estudio de abogados) distintos, lo cual es lógico porque son abogados preseleccionados por la compañía que tienen contrato de servicio o SLA (*service level agreement*, por sus siglas en inglés) firmado. El que mayor frecuencia tiene es Hazel Davis con 393 casos asignados (que representa el 14,7% del total), y si vemos el top cinco de abogados con mayor frecuencia, ellos absorben más del 50% del total de las mediaciones asignadas (53,4%).

4. **Abogado que representa al tercero (ABOGADO_3ERO)**: Posee un total de 693 valores distintos, de los cuales 422 son únicos (sin repetición), algo lógico también por la basta cantidad de oferta de abogados en el mercado. En la base se presenta la existencia de 1.233 campos con el valor Jodie Bishop, producto de la transformación hecha por la anonimización, el cual originalmente representaba al valor "sin identificar", que lo asigna un analista cuando no se identifica al abogado que envió la demanda.

5. **Posición de la Compañía (POSICION_CIA)**: De los 2.671 registros, 2.565 (96%) corresponde con el valor G (en garantía), lo cual explica el desbalanceo mencionado previamente. No posee valores faltantes. Presenta una alta correlación de 1.0 con las variables COBERTURA y TIPO_DE_SINIESTRO.

6. **Tipo de reclamo (TIPO_RC)**: Posee 3 valores distintos: RCC, RCL y RCHO, tal lo informado previamente. De ellos, el valor RCHO únicamente apareció en nueve ocasiones (0,3%), algo lógico por lo que representa (reclamo por fallecimiento en

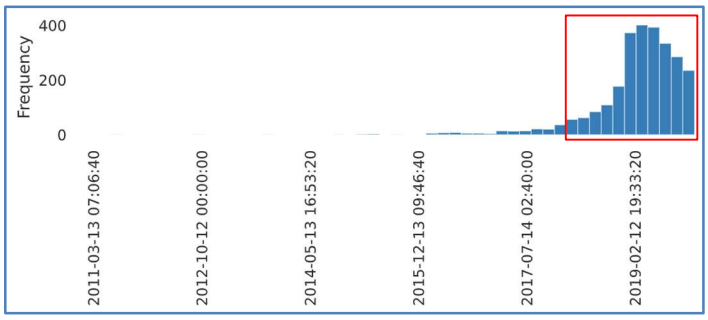
accidente). Del resto, el 53,8% corresponde con reclamos de RCL (reclamo por lesiones) y el 45,9% por RCC (reclamo por daños materiales exclusivamente).

7. **Indeterminado:** Otra variable desbalanceada, en la cual únicamente 216 reclamos (8,1%) carecen de un monto de demanda especificado en la notificación de la mediación, mientras que el resto (91,9%) si lo tiene al ser distinto de $0, y coincide con el campo DEMANDA.

8. **Cobertura**: Tercer variable desbalanceada identificada, con únicamente 29 registros (1,1%) con valores N (sin cobertura) y 108 con valores faltantes (4,0%). Se encontró una fuerte correlación de 100% con la variable POSICION_CIA.

9. **Fecha de Siniestro (FEC_STRO)**: Campo con 707 registros distintos, con un valor mínimo correspondiente con 08/06/2011 y un valor máximo de 19/12/2019, y un rango de 3.116 días u 8.54 años. Al observar el histograma, refleja una clara asimetría, ya que concentra la mayoría de los datos entre fines de 2017 y diciembre 2019. Esto es lógico dada la naturaleza de los reclamos judiciales, ya que legalmente se presenta la posibilidad de reclamar retroactivamente por eventos sucedidos hasta 10 años atrás, los cuales prescriben luego de ese plazo. Se comparte el histograma a continuación:

Figura 4 - Histograma variable FEC_STRO



10. **Código de Jurisdicción (COD_JURISDICCION)**: Se encontraron 47 jurisdicciones o valores distintos, con un valor mínimo de 1 y máximo de 125. Hay que recordar que cada código representa a una jurisdicción, y aquí se presenta otro desbalance dado que 2.201 registros (82,4%) corresponden con el código 1, que representa a la Capital Federal. Los otros códigos que tienen, de mínima, más de un 1,0% del total de los registros son: 23 (2,0% - correspondiente a Córdoba), 3 (1,7% - Lomas de Zamora), 6 (1,6% - Morón), 21 (1,4% - Rosario), 8 (1,4% - Quilmes) y 4 (1,4% - San Isidro).

11. **Código de Fuero (COD_FUERO)**: Quinta variable desbalanceada, ya que 2.200 registros (82,4%) corresponden con el código NN (no identificable). Esto es normal porque el fuero se identifica mayormente cuando traspasa a juicio. La base posee 3 valores distintos, donde el código CC tiene 463 registros, y CI ocho únicamente.

12. **Código de Juzgado (COD_JUZGADO)**: Se encontraron 38 juzgados distintos, con un valor mínimo de cero (0) y máximo de 5.004, aunque únicamente un 0,01% posee valores cero. Presenta otro desbalance ya que el 82,3% tiene el código 1111 (correspondiente con el valor no corresponde, que puede presentarse al no ser un juicio). Nuevamente, el resto de los códigos que tienen, de mínima, más de un 1,0% del total son: 2001 (5,1% - Juzg de 1° Inst en lo Civ y Com n° 1), 1001 (1,3% - Juzg. de Inst del Circuito de 1 Nom.), 2002 (1,3% - Juzg de 1° Inst en lo Civ y Com n° 2) y 2004 (1,2% - Juzg de 1° Inst en lo Civ y Com n° 4). Presentó correlaciones relativamente altas con COD_FUERO (-0,724) y COD_JURISDICCION (0,710).

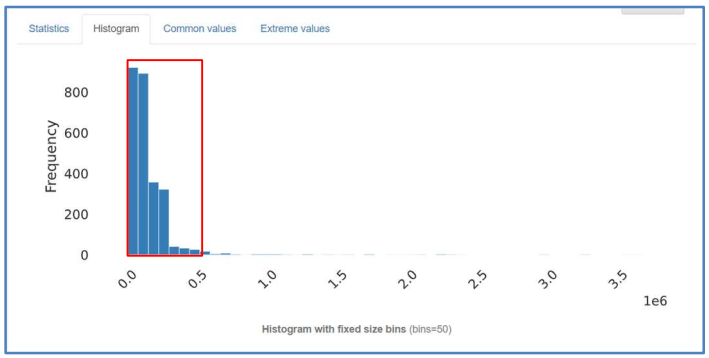
13. **Productor**: Posee 461 valores distintos, de los cuales 233 figuran únicamente en un registro o mediación.

14. **Fecha de Demanda (FECHA_DEMANDA)**: El análisis lo tomó como campo texto en lugar de fecha. Existen 231 registros distintos, algo lógico porque la fecha de la demanda (momento en que se elabora y envía) debe ser cercana a la fecha de notificación de esta (momento en que se recibe y toma conocimiento).

15. **Tipo de Siniestro (TIPO_DE_SINIESTRO)**: Variable desbalanceada, dado que, por la naturaleza de los reclamos de tercero, la mayoría son reclamos de siniestros de lesiones (1.238 registros | 46,3%) y daños materiales (1.208 registros | 45,2%), que corresponden con tipos de siniestros exclusivos del ramo automotores. En total contiene 16 valores o tipos de siniestro distintos. Presenta una alta correlación de 1,0 con la variable POSICION_CIA.

16. **Demanda**: Posee valor mínimo de $0 y máximo de $ 3.665.000, con una media de $ 148.553,92. Tuvo un total de 216 ceros (casos indeterminados – 8,1%). Se presenta histograma con distribución de frecuencias, donde se observa que la mayoría de los valores de demanda se concentran entre 0 y 500 mil (0,5 millones):

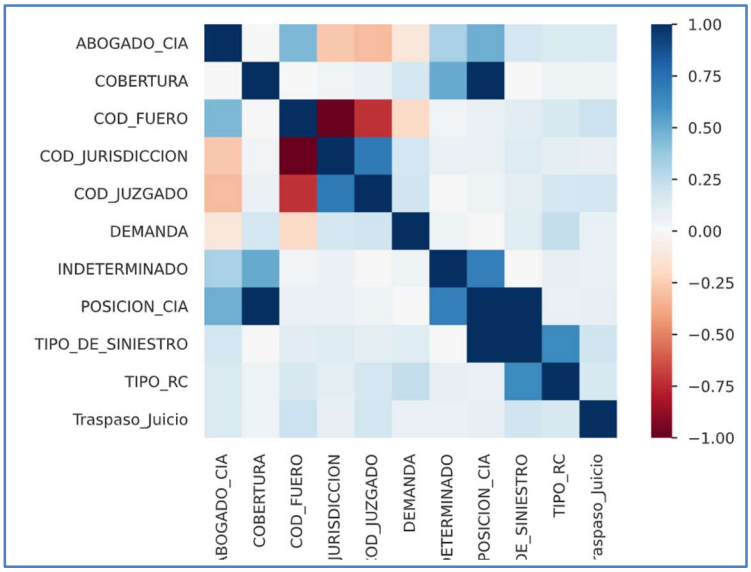
Figura 5 - Histograma variable DEMANDA



17. **Traspaso de Juicio**: Como era sabido, el campo no posee faltantes y presenta 2 valores distintos: SI o NO, que corresponden con los valores a predecir. También presenta un desbalance, ya que 2.225 valores corresponden con NO (83,3%).

Finalmente, se comparte el gráfico o mapa de calor para realizar el análisis de correlación correspondiente, y que refleja algunas cuestiones que se han mencionado previamente:

Figura 6 - Análisis de correlación de variables



Precisamente, del gráfico de correlación se ve claramente la fuerte correlación positiva en 1.0 de POSICION_CIA con las variables TIPO_DE_SINIESTRO y COBERTURA. A la vez, se ven, aunque con menor intensidad, correlaciones negativas entre COD_FUERO y COD_JURISDICCION (-0,985), y COD_FUERO y COD_JUZGADO (-0,724). Por última, y en menor medida, se encontraron correlaciones positivas entre COD_JURISDICCION y COD_JUZGADO (0,710), INDETERMINADO y POSICION_CIA (0,682),

INDETERMINADO y COBERTURA (0,501) y, finalmente, TIPO_DE_SINIESTRO y TIPO_RC (0,637).

Es sabido que, en algunos modelos, las correlaciones pueden afectar los modelos de predicción. Este es el caso de la regresión logística, debido a la multicolinealidad, que ocurre cuando dos o más variables independientes están altamente correlacionadas. No obstante, no resulta ser igual en los algoritmos de *decision forest,* los cuales son menos sensibles a dicha multicolinealidad. Esto se debe a que los árboles de decisión seleccionan subconjuntos aleatorios de características para dividir en cada nodo, lo que reduce la dependencia de variables altamente correlacionadas. Más adelante se informará que lo realizado para tratar este escenario.

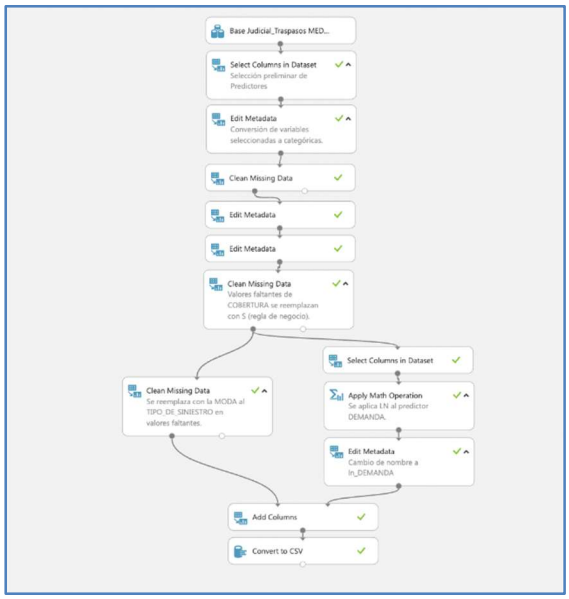
2.3. Preparación de los datos previo al entrenamiento del modelo

Para la preparación de los datos se utilizó en primer lugar código en Python (.ipynb), en donde se utilizaron las librerías *Pandas* y *anonymizedf*, con el propósito de anonimizar la base y resguardar la privacidad de los datos utilizados, como se adelantó en el subapartado 2.1. Los datos anonimizados fueron los siguientes: NRO_JUICIO, ABOGADO_CIA, ABOGADOR_3ERO, ANALISTA, PRODUCTOR, ASEGURADO, ACTOR, DEMANDADO y LIQUIDADOR. Los nuevos valores son adjuntados al final del *dataset*, y la librería *anonymizedf* los diferencia de las variables originales porque comienzan con la palabra *Fake* y luego el nombre correspondiente de la columna original, por lo que se procedió a eliminar dichas columnas originales para luego quitarle la palabra *Fake* a las creadas. La librería presenta una deficiencia y es que no detecta la fijación de una semilla o *seed*, donde se indagó en las funciones de la librería en sí, así como también con la función nativa de *python* y en la librería *numpy*, por lo que no es posible replicar la anonimización dos veces y lograr un resultado idéntico. Asimismo, a los valores nulos o códigos estándares como *no identificado* (en el campo del abogado del tercero) les asigna valores ficticios, por lo que se debió identificarlos para poder llevarlos nuevamente a nulos -en el primer caso- o valores anteriores -en el segundo- para evitar distorsionar los resultados. Se adjunta el notebook con el procedimiento descripto en el apartado Apéndices.

Posteriormente se migró el CSV resultante de la anonimización a Azure Machine Learning para realizar la 2da parte de la preparación, en donde se aplicó ingeniería de predictores, como por ejemplo se pasó a categóricas las variables correspondientes (salvo el monto de la demanda), se reemplazaron datos faltantes por la moda en el tipo de siniestro, o por el valor S en el campo cobertura (dado que se asume que si el analista no lo especifica, el reclamo debiera tener

cobertura que lo cubra) o mismo valores alterados en la exportación del CSV durante la anonimización (por ejemplo, figuraba *MediaciÃ²n* y volvió a transformarse a *Mediación*). Asimismo, se seleccionaron las variables a utilizar en ambos modelos (las 15 variables enumeradas anteriormente), y se aplicó LN (logaritmo natural) al predictor DEMANDA, para intentar mitigar el efecto inflacionario del año. Se comparte el proceso realizado en Azure Machine Learning para dar mejor visibilidad de lo realizado:

Figura 7 - Preparación y limpieza de datos en Azure Machine Learning



2.4. Diseño y evaluación de los modelos de predicción de traspaso de instancia

Como se mencionó previamente, los modelos a utilizar son *Regresión Logística* y un *Decision Forest*. En el caso de la regresión logística, se seleccionó dado que es un modelo adecuado para problemas de clasificación binaria, proporciona una salida interpretable en términos de probabilidades, y asimismo funciona bien con conjuntos de datos de tamaño moderado y cuando las variables no están correlacionadas de manera compleja. Por otra parte, el *decision forest*, como bien se sabe, es un conjunto de árboles de decisión entrenados en diferentes submuestras de los datos. Es un algoritmo de *machine learning* con muchas bondades, entre ellas: es uno de los más precisos y versátiles para problemas de clasificación; puede manejar variables numéricas y categóricas sin dificultad; es resistente al sobreajuste (*overfitting*) debido a la combinación de múltiples árboles, entre los principales. El propósito del ensayo es optimizar cada uno de los modelos para luego compararlos y elegir el más eficaz o exacto de los dos.

Para poder realizar el entrenamiento y la evaluación, se dividió la base en un setenta y treinta (70 / 30) respectivamente. Para evitar la fuga de datos y asegurar que el modelo se evalúe

correctamente, posterior al *split* de datos se implementó la función *smote* para generar datos sintéticos con el propósito de incrementar en un 200% los datos (semilla o *seed* 2886) de aquellos registros que SI se traspasaron a Juicio, dada la disparidad en la frecuencia de uno (SI) y de otro (NO) según lo ya observado durante el perfilado de datos. Es importante mencionar que se aplicó *smote* solo al conjunto de entrenamiento, el cual al crear datos sintéticos balancea las clases (relación entre valores SI y NO) y mejora la capacidad del modelo para aprender de los datos minoritarios sin introducir sesgos en el conjunto de evaluación. Estos datos se utilizarán para el entrenamiento de ambos modelos.

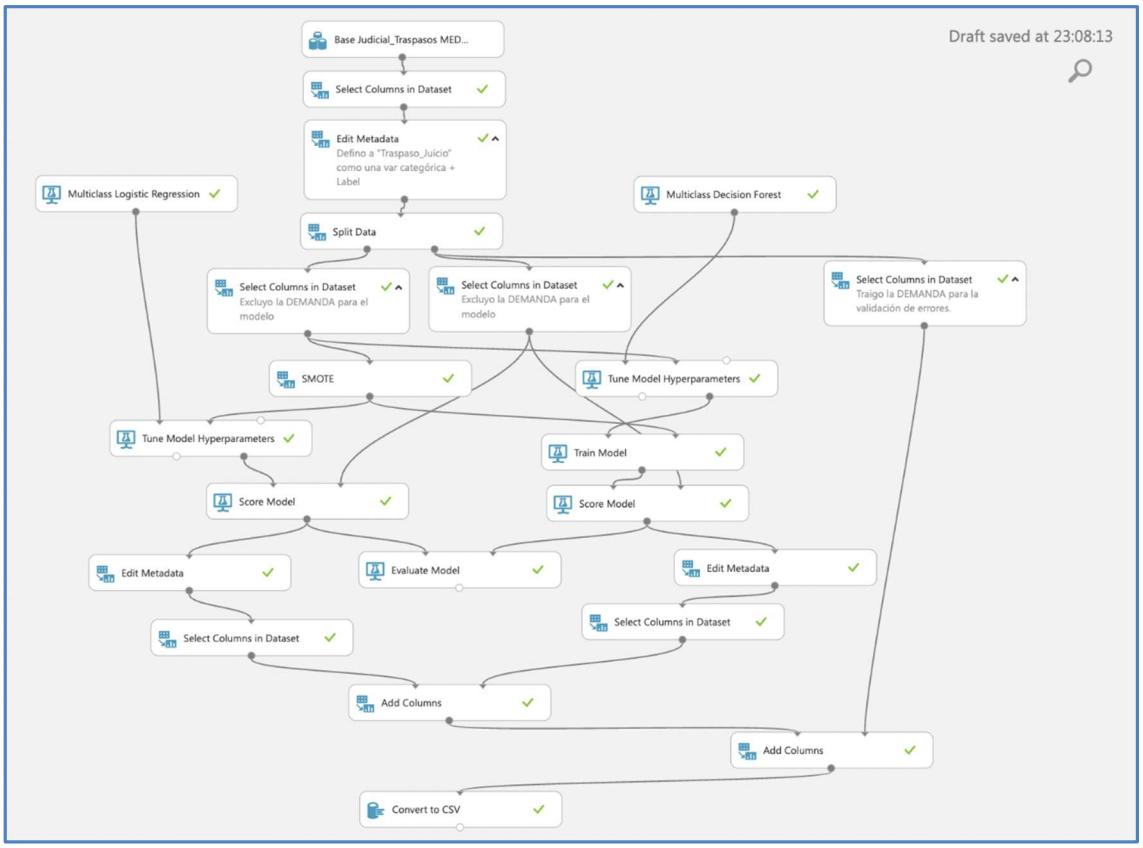
Antes de dividir los datos, se realizó una selección de columnas y edición de metadatos para asegurar que solo las variables relevantes mencionadas anteriormente se incluyeran en el modelo y que las etiquetas de las variables estuvieran correctamente definidas. Esta etapa es crucial para preparar adecuadamente el conjunto de datos para el entrenamiento del modelo. De hecho, de las 17 variables que figuraban preseleccionadas, terminaron excluyéndose en esta etapa las variables COBERTURA y POSICION_CIA, para evitar tener variables muy correlacionadas en función del análisis anterior, así como también el ID de cada JUICIO, que no debiera considerarse en el aporte del modelo.

Como la base entera, sin dividir entre entrenamiento y evaluación, tiene un total de 2.671 registros, se definió que cuando se haga el tuneo de hiperparámetros de la base de entrenamiento (2.492 registros, luego de aplicar *smote*) para cada modelo, se utilice la opción *Entire_Grid* en *sweeping mode*, que consume más recursos, pero realiza una búsqueda exhaustiva y no aleatoria de los valores de los hiperparámetros. Este proceso es esencial para optimizar el rendimiento de los modelos. Es relevante destacar que como se trata de modelos de clasificación, la métrica seleccionada aquí fue el *accuracy*. Asimismo, no se aplicó normalización de parámetros porque la única variable numérica es el predictor DEMANDA, al cual en su lugar se aplicó un LN para mitigar efectos inflacionarios, como se mencionó oportunamente.

Después del ajuste de hiperparámetros, se entrenaron ambos modelos y luego se evaluaron en el conjunto de datos de evaluación (el cual posee 801 registros, correspondiente con el 30% del *dataset* total), resultados los cuales se mostrarán en el subapartado siguiente. Finalmente, se realizó una edición de metadatos y selección de columnas adicionales (las columnas con los predictores de ambos modelos) para preparar los datos para la exportación. Los resultados se convirtieron a un archivo CSV, lo cual facilita el análisis posterior de los resultados obtenidos, adjunto en el apartado Apéndices.

Se comparte un esquema del diseño del experimento en Azure ML:

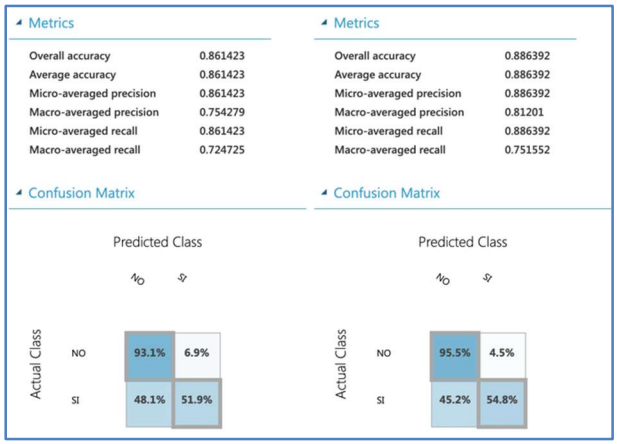
Figura 8 - Modelos de predicción de traspaso de instancia en Azure Machine Learning



2.5. Análisis de resultados y principales predictores que explican los modelos

Los resultados del entrenamiento resultaron ser claros al aplicarlos en el *dataset* de evaluación, los cuales dieron mejores resultados en el *Decision Forest*. A continuación, se comparte un resumen de los KPI's (*key performance indicators*) internos obtenidos por cada modelo: Regresión Logística (a la izquierda) y Decision Forest (a la derecha).

Figura 9 - Resultado de los modelos y matriz de confusión



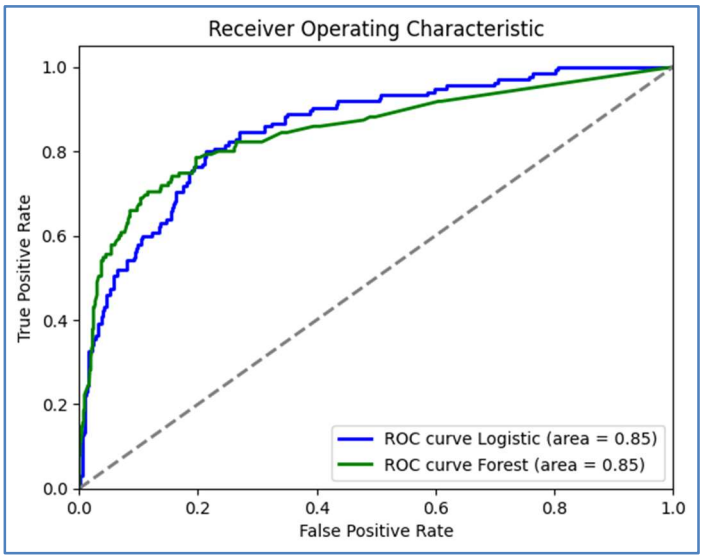
Como se puede observar, tanto el *accuracy*, como el *precision* y *recall* son mejores al aplicar el modelo *Decision Forest*, lo cual facilitó la tarea de seleccionar el mejor de ellos. Este modelo tiene una mayor cantidad de verdaderos negativos (95,5%) y verdaderos positivos (54,8%), lo que sugiere que es más efectivo en clasificar correctamente tanto los casos negativos como los positivos en comparación con la regresión logística.

Ahora bien, al observar este modelo, es claro que, a pesar de ser mejor y ser alta la cantidad de verdaderos negativos, no lo es tan así la cantidad de verdaderos positivos, lo cual al conversar oportunamente con la línea de negocio y tal como se describió al inicio del ensayo, no se percibe como algo grave a priori para implementar un primer modelo. Esto es así dado que permitiría tomar medidas de negociación diferentes y – probablemente – más agresivas, en más de un 50% de los casos que tienen alto riesgo de traspasarse a la instancia de Juicio. De esta manera, permitiría reducir en un mediano largo plazo el *transfer ratio* o tasa de traspasos. Asimismo, es tan bajo el porcentaje de falsos positivos (4,5%), que no se espera que impacte negativamente en el transfer ratio el cambio de estrategia de negociación en estos reclamantes, aunque sí puede que aumente levemente el costo medio de la instancia de mediación (indicador que deberá monitorearse).

Previo a continuar con la indagación del modelo Decision Forest, es importante hacer una pausa y focalizarnos en el análisis de la curva ROC (*receiver operating characteristic*) y el AUC (*area under the curve*) de ambos modelos. Es que, para evaluar la capacidad de discriminación de los modelos de predicción de traspasos, debemos utilizar dichas curvas. La curva ROC nos permite visualizar el rendimiento de los modelos a través de diferentes umbrales de decisión, mientras que el AUC proporciona una medida agregada del rendimiento del modelo. Compartimos el

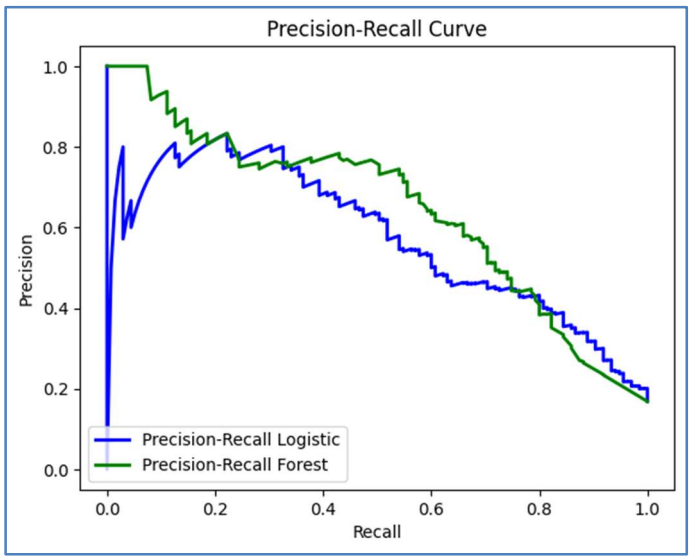
resultado a continuación, obtenido gracias a un script realizado en una notebook con Python el cual utilizó las librerías skikit-learn, matplotlib y pandas:

Figura 10 – Curva ROC de ambos modelos



En la ilustración 10, se presentan las curvas ROC para los modelos de Regresión Logística y Decision Forest. Ambas curvas están por encima de la línea diagonal (que representa un clasificador aleatorio), y muestran un AUC de 0,85, lo que indica que dichos modelos tienen una buena capacidad para discriminar entre las clases positivas y negativas y su efectividad de predicción es bueno. Sin embargo, las curvas no están consistentemente por encima de 0.9, lo que sugiere que hay margen para mejorar el rendimiento de los modelos.

Figura 11 - Curva Precision-Recall de ambos modelos



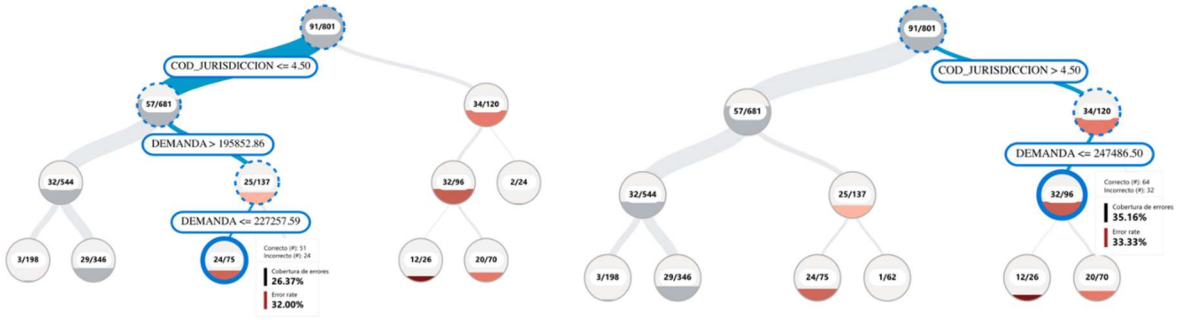
De forma complementaria, en la ilustración 11, se presenta la curva de Precisión-Recall, que muestra la relación entre la precisión y el *recall* para diferentes umbrales de decisión. Ambos

modelos muestran una alta precisión cuando el *recall* es bajo, pero la precisión disminuye significativamente a medida que el *recall* aumenta. Este comportamiento es típico en muchos modelos de clasificación y sugiere la necesidad de un balance entre precisión y *recall*.

En resumen, ambos modelos muestran un buen rendimiento, aunque aún con oportunidades de mejora. Asimismo, al tener ambos un rendimiento similar en términos de AUC y precisión-*recall*, la elección entre los dos modelos depende de otros factores, los cuales ya fueron expuestos en párrafos anteriores.

Finalmente, y en función de la selección final del modelo *Decision Forest* por su mayor poder de predicción, se realizó en un notebook de Google Colab el análisis de errores con la herramienta ***ErrorAnalysisDashboard*** de la librería ***raiwidgets***, a modo de entender cuáles son las oportunidades de mejora y los criterios del modelo para la segmentación:

Figura 12 - Análisis de errores modelo Decision Forest



De los gráficos se desprende que de los 801 registros del *dataset* de evaluación, el 11,36% son errores, por lo que va en línea con el *accuracy* del modelo, de un 88,64%.

Si bien los errores están bastante bien distribuidos, con los predictores COD_JURISDICCION y DEMANDA como los más utilizados para el criterio de apertura del árbol de decisión, es claro que el más crítico es el de la derecha (33,33% del total de esta rama, con un 35,16% del total de los errores del modelo), dado que el *accuracy* del modelo predice muy por debajo para este conjunto de datos que representa una minoría.

La división presentada por el análisis de errores, con una frecuencia mayor en la rama de la izquierda, es totalmente lógica al ver que las jurisdicciones menores o iguales a 4,5 son las que van del uno al cuatro, que representan a Capital Federal, Avellaneda, Lomas de Zamora y San Isidro respectivamente, jurisdicciones con mayor volumen de siniestralidad y reclamos, y al disponer de más datos, es comprensible que el modelo pueda predecir un poco mejor los comportamientos de estas jurisdicciones. Es evidente que las jurisdicciones con mayor

demanda y menor frecuencia de casos tienden a tener una mayor tasa de error, lo que refuerza la necesidad de desarrollar modelos específicos para estas áreas.

De aquí se deduce que lo más conveniente quizá sea separar el modelo y realizar uno puntual para las jurisdicciones de menos frecuencia, correspondientes en su mayoría al interior del país, así se puede ver de deducir los comportamientos no explicados por los predictores seleccionados y disponibles en la escaza información que se recibe al momento de recibir la notificación de la mediación. Pero, para ello, y dada la escaza frecuencia, quizá sea necesario realizar otro ensayo y capturar una muestra aún más grande para tener resultados más robustos en estas jurisdicciones.

En conclusión, el análisis de los principales predictores y la segmentación de errores nos permite identificar áreas de mejora y ajustar estrategias para optimizar el rendimiento del modelo. El resultado del ensayo obtenido alcanzó un buen *accuracy* si se tiene presente que es una primera iteración del modelo, el cual es efectivo para las jurisdicciones de mayor frecuencia, ergo, para jurisdicciones donde la cartera de la compañía tiene mayor frecuencia. Asimismo, la implementación de modelos específicos para jurisdicciones con menor frecuencia podría mejorar significativamente la precisión y reducir los errores, lo podría contribuir a una mejor gestión de los reclamos y a una reducción del Transfer Ratio.

Net Promoter Score: Análisis de la voz del cliente

> La ocurrencia de un siniestro es el momento de la verdad para una aseguradora, y la gestión del siniestro se convierte en el punto clave de la relación con el asegurado.
>
> El acontecimiento de un siniestro es un momento crítico para el asegurado, el cual requiere una solución inmediata a su problema, y por lo tanto una mala gestión o una tardanza en la resolución puede provocar una insatisfacción por parte del asegurado y una pérdida de dicho cliente.
>
> Además […] la experiencia cliente ha pasado a ser un objetivo prioritario para mejorar la relación del cliente, y la ocurrencia de un siniestro para el asegurado es el momento en el que se pueden generar más experiencias tanto positivas como negativas para el cliente. (Muñoz Pardo, 2018, p. 51)

En el presente capítulo se abordará la realización del segundo ensayo del TFI, el cual consiste en intentar identificar grupos de clientes del segmento automotor para evaluar la viabilidad de

brindar un servicio segmentado en función de una o más variables fijadas previo al servicio, sujeto a lo que resulte de dicho análisis. El objetivo final es el de mejorar la tasa de retención de clientes y la satisfacción media de los asegurados a través de la personalización del servicio, sin perder de vista el control del costo medio por siniestro.

Para poder realizar dicho ensayo, se utilizará el lenguaje R en RStudio con el propósito de aplicar un análisis de clústers a través de diversos métodos tanto jerárquicos como no jerárquicos mediante el uso de k-medias, en un caso de negocio que en esta oportunidad está enfocado en el servicio de cara al asegurado. A tal fin, en el primer subapartado se presentará la base de datos de muestra utilizada para dicho análisis; en un segundo subapartado se realizará el análisis exploratorio de las variables seleccionadas; y se culminará con un tercer subapartado donde se presentarán todos los métodos utilizados para agrupar la información seleccionada, a modo de entender si hace sentido las agrupaciones brindadas por los algoritmos.

3.1. Origen y descripción de la base de encuestas a asegurados

Para la aplicación de este ensayo se optó por analizar un *dataset* confidencial perteneciente a una reconocida compañía aseguradora del mercado argentino. Dicho *dataset* proviene del sistema de encuestas de calidad de servicio de la compañía, por lo que el propósito del análisis será el poder entender la factibilidad de agrupar los asegurados encuestados en grupos con comportamientos homogéneos, aunque con la existencia de heterogeneidad entre ellos.

El *dataset* que analizaremos y que provee un sistema especializado de encuestas, consta originalmente de 126 columnas y responde a la metodología de análisis denominado NPS (*Net Promoter Score*), el cual clasifica las calificaciones finales en 3 categorías: *promotores* (calificaciones 9 y 10), *neutrales* (calificaciones 7 y 8) y *detractores* (calificaciones de 6 hacia abajo). Para obtener la calificación NPS se descartan los valores neutrales, por lo que se focalizó únicamente en los *promotores* y *detractores*, donde el score que se obtiene oscila entre valores de - 100 y 100.

Ahora bien, en el presente análisis cambiaremos el foco y, para poder aplicar el método de componentes principales, los *neutrales* se incluirán. Esto se considera así por 2 cuestiones fundamentales: en primer lugar, cada persona tiene un criterio subjetivo general de una calificación, donde quizá alguien que califica con califica con siete u ocho puede que sea una persona más exigente con la calificación y en su mentalidad esa calificación encuadra como

alguien promotor, a modo de ejemplo; en segundo lugar, se sesgaría parte de la muestra, y no se desea en esta instancia hacerlo.

Con relación al *dataset*, de las 126 columnas originales se redujo la base a 24, y se quitaron varias columnas vacías, discontinuadas, con información confidencial (como ser el nombre del Asegurado, email, teléfono, número de siniestro, entre otros) o que no aporten valor al análisis. A su vez, se volvió a reducir su tamaño y de estas 24 columnas resultó finalmente una matriz de 7 variables métricas, las cuáles focalizaremos en el presente análisis. El propósito de haber acotado el número de variables se basó principalmente en entender cuáles son esenciales para la agrupación de los asegurados (y que aporten valor a dicha segmentación), por lo que se eliminaron aquellas que puedan ser redundantes y que puedan ser explicadas por las demás variables seleccionadas del modelo.

Como veremos más adelante, las variables que se presentarán son heterogéneas entre sí, por lo que se optará trabajar con valores normalizados o estandarizados, dada la alta heterogeneidad en los rangos de cada una de las variables de análisis, y por ende se trabajará con la matriz de correlación R y no con la de varianzas y covarianzas.

Por otra parte, el período comprende las respuestas de encuestas recibidas durante el 3er trimestre de 2021 del negocio de autos, más específicamente de siniestros de daños parciales y reposiciones de cristales y ruedas. La elección del período se basó en que el negocio, dada su alta frecuencia, no suele tener estacionariedad relevante, y menos aún si se trata de un único trimestre y con un foco el cual se centra en el servicio y no en otras variables de mayor volatilidad en el tiempo.

En cuanto a los observaciones o registros de la base, estos representan a individuos asegurados en la compañía. Por otra parte, y como se mencionó anteriormente,

La base contiene un total de 143 registros o individuos observados, los cuales no han sufrido un recorte o quita dado que todas las variables seleccionadas carecieron de la particularidad de poseer valores faltantes, característica que sí poseen algunas de las otras variables finalmente no seleccionadas como, por ejemplo, sub-preguntas que debían calificarse y tenían valores NA, es decir, no contestada.

Por otra parte, al aplicar la metodología de análisis de clústers se optó por estandarizar o tomar valores tipificados para el análisis, dado que cada variable o dimensión está expresada en medidas muy diferentes (meses, días, años, pesos, calificaciones, etc.).

Por último, se adjunta en los Apéndices el script de lo ejecutado y analizado en R. Allí se podrán observar: las librerías utilizadas; la estandarización de dicha matriz; la metodología para identificar y eliminar *outliers* de la base; las funciones para aplicar en primer lugar métodos jerárquicos, para luego tomar los centroides de dicho modelo y aplicar métodos no jerárquicos (*k-means*); como así también las funciones utilizadas y aprendidas para hacer análisis de componentes principales (PCA, por sus siglas en inglés) sobre los grupos conformados para graficar su distribución.

3.2. Análisis exploratorio de las variables seleccionadas

Al igual que en el ensayo anterior, en el presente se procederá realizar nuevamente un EDA de la base de datos a analizar, el cual resulta ser un paso crucial en la preparación de datos para técnicas de clustering, debido a varias razones:

- **Comprensión de la Distribución de las Variables:** ayuda a entender cómo se distribuyen las variables en el conjunto de datos. Esto incluye identificar si las variables siguen distribuciones normales o si presentan sesgos significativos, lo cual puede influir en la elección del algoritmo y en la necesidad -o no- de transformaciones de datos.
- **Detección de Relaciones entre Variables:** Analizar las relaciones entre variables puede revelar correlaciones que deben ser consideradas antes de aplicar clustering. Por ejemplo, variables altamente correlacionadas pueden ser redundantes y podrían ser combinadas o eliminadas para simplificar el modelo.
- **Evaluación de la Calidad de los Datos:** permite evaluar la calidad de los datos al identificar problemas como datos faltantes, errores de entrada o inconsistencias. Estos problemas deben ser abordados para asegurar que el clustering se base en datos precisos y fiables.
- **Selección de Variables Relevantes:** ayuda a identificar las variables más significativas que contribuyen a la diferenciación de los clústers, lo que permite una selección más informada de características.
- **Identificación de Outliers:** Los outliers pueden distorsionar los resultados del clustering, ya que pueden formar clústers propios o influir en la formación de otros clústers. El EDA permite detectar y manejar estos valores atípicos de manera adecuada.

Con relación a las siete variables seleccionadas que adelantábamos en el subapartado anterior, estas representan las siguientes dimensiones evaluadas para entender su correlación, y se agrega una octava la cual representa al código de ID de la encuesta (que representa a un asegurado), la cual enunciaremos pero que no será posteriormente parte del análisis:

1. **Número de Encuesta:** ID que representa al código de encuesta enviado y respondido por un asegurado.
2. **Edad:** Representa la edad del Asegurado, el cual – a priori – puede ayudar a entender si el comportamiento de la calificación se debe a una cuestión generacional, si se toma en cuenta a la vez que el proceso del servicio es igual en todos los casos.
3. **Antigüedad Relación (meses):** Antigüedad que tiene el asegurado como tal dentro de la compañía. Esta variable puede ayudar a entender – a priori – si la fidelización o grado de pertenencia sesgan la calificación del asegurado, donde puede que resulte -o no- que aquellos de mayor antigüedad tengan mayor nivel de tolerancia.
4. **NPS:** Calificación general de la encuesta, con la cual relacionaremos el resto de las variables en un primer análisis.
5. ***Vehicle Age*:** Edad del vehículo asegurado. La intención de incluir esta variable es entender si la calificación en vehículos más antiguos puede variar en la calificación, ya que pueden presentar más complicaciones a la hora de encontrar repuestos e indemnizar dentro de los plazos previstos.
6. ***Premium Amount*:** Prima que paga el asegurado mensualmente de la póliza. Su fin es similar a la variable anterior, por lo que también se quiere ratificar si hay correlación entre ambas, donde se asumiría que la política de pricing de la compañía está altamente basada – o no – en la suma asegurada del vehículo.
7. **Monto pagado:** Esta variable representa el monto indemnizado al asegurado encuestado respecto del siniestro sufrido, y ayudaría a entender la real correlación existente en la presunción de que, a mayor monto indemnizado, mayor es la satisfacción del asegurado.
8. ***Lag* Aviso a Pago:** Vida del siniestro medido en cantidad de días, en la cual se evalúa el tiempo entre el aviso del siniestro a la compañía y el pago de la indemnización. Como se puede deducir en base a la descripción de cada una y sus rangos, todas las variables son del tipo numéricas, las cuales poseen rangos heterogéneos dado que cada una se basa en diferentes unidades de medida (días, meses, unidades monetarias, calificaciones y edades).

Para profundizar en el EDA, se procedió a realizar un perfilado de datos (o en inglés, *profiling report*) en R mediante la utilización de la librería *DataExplorer*, que provee mayor detalle del set de datos a analizar, en el cual se incluyen estadísticas descriptivas, existencia de valores nulos, la distribución de las variables y la relación entre ellas. De este se puede reflejar parte de lo que ya hemos mencionado como la inexistencia de valores nulos o el hecho de que todas las variables son numéricas y continuas.

Con relación a las estadísticas descriptivas y de distribución de valores arrojados por el *profiling report*, se pasa a presentar el histograma resultante, al cual se le adiciona un cuadro confeccionado mediante código con las principales estadísticas a evaluar para cada variable, los cuales se comparten a continuación:

Figura 13 - Histogramas de variables seleccionadas

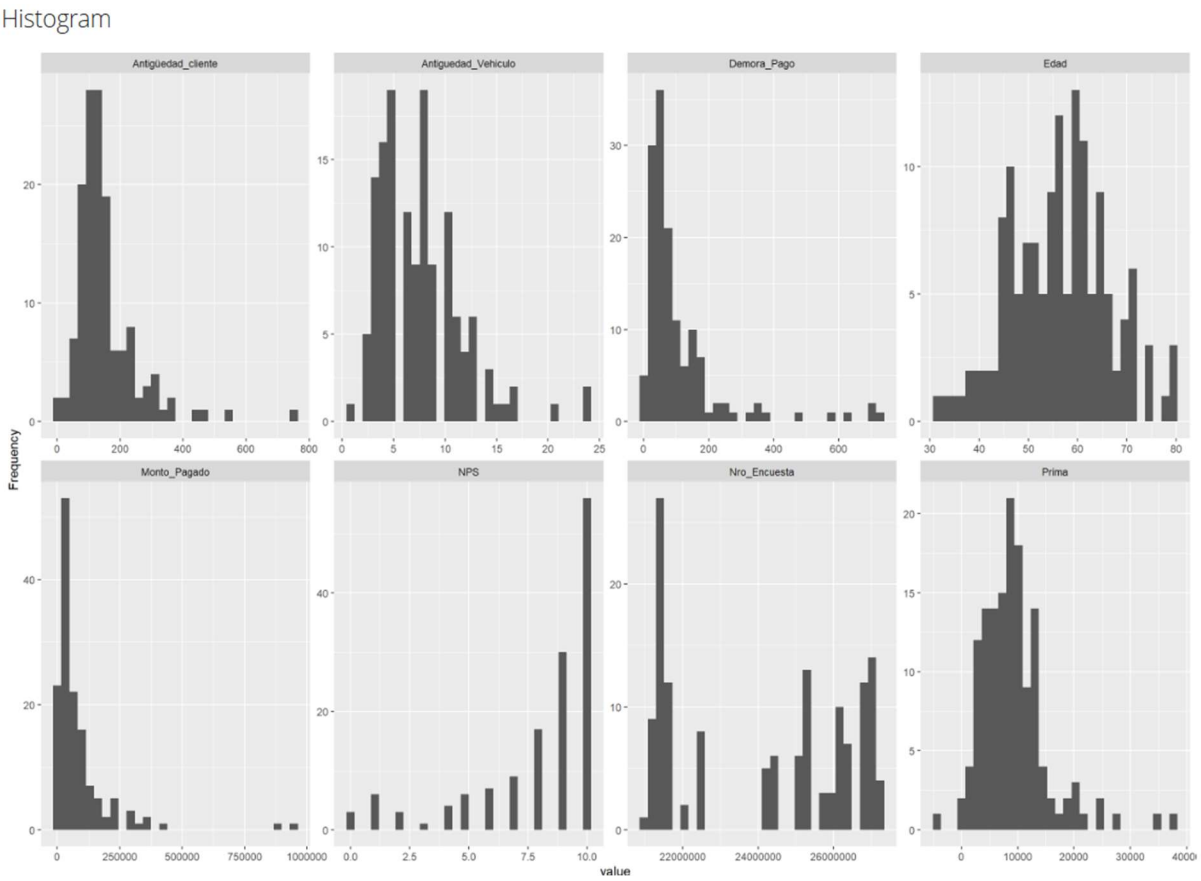


Figura 14 - Estadísticas descriptivas de variables seleccionadas

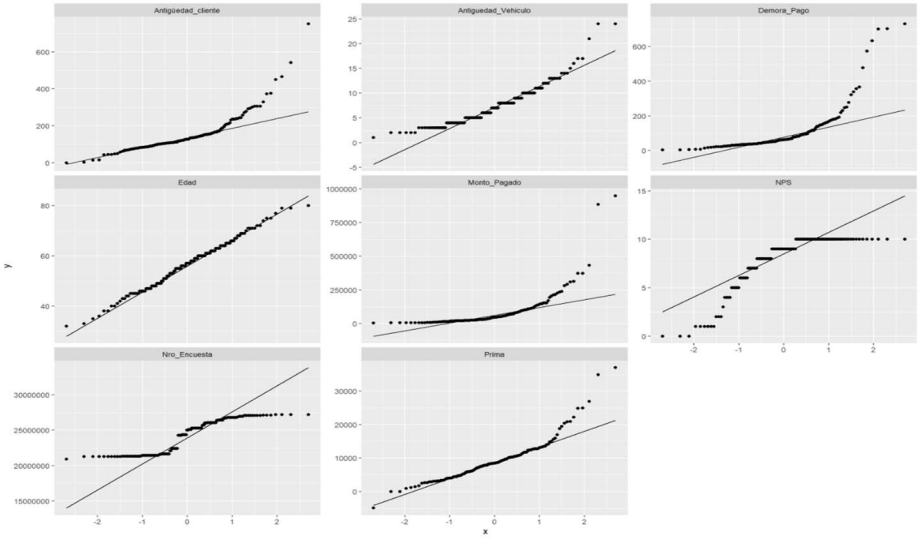
variable	mean_value	sd_value	min_value	max_value	range	median_value	q25	q75
Edad	56.436620	9.997492	32.00	80.00	48.00	57.00	49.000	63.00
Antigüedad_cliente	151.943662	100.618285	0.00	752.00	752.00	127.00	98.500	169.75
NPS	7.971831	2.699996	0.00	10.00	10.00	9.00	7.000	10.00
Antiguedad_Vehiculo	7.521127	4.142357	1.00	24.00	23.00	7.00	4.250	10.00
Prima	9261.782324	6052.287339	-4970.28	37019.54	41989.82	8449.37	5355.477	11709.61
Monto_Pagado	85755.118592	128845.285630	3022.50	950000.00	946977.50	45983.68	22649.795	100530.21
Demora_Pago	108.091549	133.467020	4.00	732.00	728.00	62.50	38.250	116.75

De estos gráficos se concluyen algunas cuestiones con relación a las variables de análisis:

1. **Edad:** Su recorrido total es de entre los 32 y los 80 años, aunque la mayoría de los clientes tienen entre 40 y 70 años, con un pico notable alrededor de los 60 años.
2. **Antigüedad del cliente:** Su recorrido total va de entre los 0 y los 752 meses. La mayoría de los clientes tienen una antigüedad menor a 200 meses, con una disminución significativa a partir de ese punto.
3. **NPS:** Su recorrido, como era de esperarse, es de calificaciones entre 0 y 10. La mayoría de las puntuaciones NPS están concentradas entre 5 y 10, por lo que indica una tendencia hacia puntuaciones altas.
4. **Antigüedad del vehículo (o Vehicle Age):** Su recorrido total va de entre 1 y 24 años de antigüedad. La mayoría de los vehículos tienen una antigüedad de entre 5 y 10 años, con una distribución decreciente a medida que aumenta la antigüedad.
5. **Demora del Pago:** Su recorrido parte de los 4 hasta los 732 días. La mayoría de los pagos se retrasan menos de 100 días, con algunos casos extremos de hasta 600 días.
6. **Monto pagado:** Su recorrido comprende valores entre $3.022,5 y $950.000. La mayoría de los pagos son menores a 250.000.
7. **Premium Amount (o prima):** Su recorrido comprende valores entre -$ 4.970 y $37.020. La mayoría de las primas están entre 5.000 y 15.000, con algunos valores extremos más altos.

El perfilado también nos arroja un QQ *plot* para entender si la distribución de los valores de cada variable se acerca o no a una distribución normal, lo cual nos va a ayudar a entender cómo proceder con dichas variables y con los modelos a utilizar a posteriori:

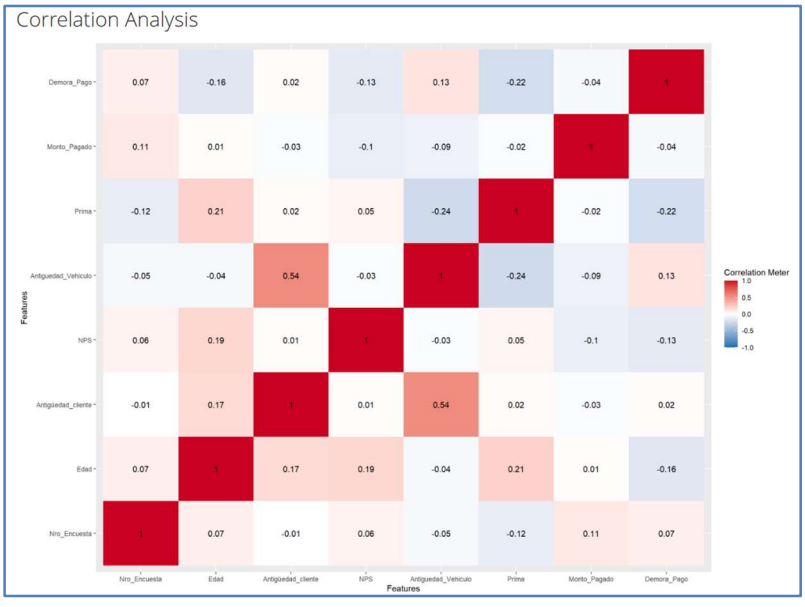
Figura 14 - QQ plot de variables seleccionadas



Como se puede observar en el conjunto de gráficos, la mayoría de las variables en tu conjunto de datos no siguen una distribución normal, ya que los puntos en los QQ Plots se desvían significativamente de la línea de referencia. En la única variable que si se observa una mayor alineación respecto de la línea de referencia es en la edad de los asegurados, a pesar de identificar algunas desviaciones en los extremos, lo que sugiere que esta sigue una distribución normal de forma aproximada. Esto respalda la necesidad de normalizar las variables previo a comenzar a aplicar métodos jerárquicos y no jerárquicos para el análisis de clústers.

Otro gráfico de importancia existente en el *profiling report* es el de análisis de correlaciones, que se presenta a continuación:

Figura 15 - Análisis de correlación de variables seleccionadas



Las correlaciones observadas son en su mayoría débiles, con algunas excepciones moderadas como la relación entre la antigüedad del vehículo y la edad del cliente, la cual es una correlación positiva y deja a entrever que los clientes mayores tienden a tener vehículos más antiguos. Este panorama general sugiere que, aunque hay algunas tendencias, no hay relaciones extremadamente fuertes entre la mayoría de las variables analizadas, por lo que no se considera redundante la existencia de ninguna de ellas y por ende no se modificó la cantidad de variables existentes en el *dataset*.

Finalmente, y si bien no se eliminaron observaciones/encuestas por contener al menos un valor NA, si se analizó la posible existencia de Outliers en la base al calcular la distancia de Mahalanobis, previo a la obtención o cálculo de las distancias euclídeas. Para su análisis y, en primer lugar, se calcularon las varianzas y las medias de las variables para el conjunto de las 142 observaciones, y se calcularon las distancias de Mahalanovis apoyadas en estas.

A tal fin, lo que se realizó fue un test de hipótesis para entender si algunos valores son atípicos o no: en primer lugar, se establece en el test que la hipótesis nula consiste en afirmar que ningún valor es atípico, y por ende la hipótesis alternativa confirma lo contrario; en segundo lugar, se establece que para rechazar la hipótesis nula asumimos que el p-valor obtenido en cada observación de la base (con 7 grados de libertad == cantidad de variables métricas de la base) debe ser inferior a 0,01; por último, se creó un estadístico Xi2 con 7 grados de libertad (de valor 18,4753), el cual utilizó la función qchisq que será el umbral para determinar si la distancia de Mahalanobis para cada observación supera o no al estadístico, y si lo supera, debe excluirse el valor. Si ambas condiciones se cumplen, se rechaza la hipótesis nula, y si solo se cumple una, se conserva la observación y se afirma que no es robusto o sólido el criterio de rechazo.

Dicho esto, de la prueba resultante se observa que 11 de los 142 registros son Outliers, por lo que se quitan en la base y se deja finalmente 131 observaciones o encuestas de Asegurados en esta. La tabla con el test lo pueden observar en la tabla “Maha.xlsx” que figura en la sección Apéndices. Para obtener esa subtabla, se procedió a eliminar los registros en el script principal.

3.3. Aplicación de métodos jerárquicos para identificar clústers de asegurados

Luego de definir la base final a analizar las posibles agrupaciones entre observaciones, se decide que, como no se trata de observaciones binarias sino de observaciones métricas heterogéneas, la medida a utilizar para medir distancias es la de distancia euclídea, pero lo hago sobre los datos estandarizados tal como se adelantó en la sección anterior. Para ello, realizo su cálculo en

R y armo la matriz de distancias euclídeas. La salida en R es poco amigable de ver ya que se trata de una salida de 131x131 dadas las cantidades de observaciones.

Una vez ejecutada la matriz de distancias Matriz.Dist.Euclid.norm.NPS.DF, se utilizó para empezar a aplicarla en el método de agrupamiento seleccionado: en esta oportunidad, se determinó primero avanzar con métodos jerárquicos para luego utilizar sus centroides como semillas para aplicar a posteriori el método no jerárquico que utilizará k-medias.

Ahora bien, se utilizaron las distancias euclídeas para proceder con la aplicación de los 5 métodos no jerárquicos existentes (centroide; vecino más cercano; vecino más lejano; de la vinculación promedio y de Ward), y así entender tanto la cantidad de grupos óptimos recomendados por cada uno como las observaciones que recomiendan conformen cada grupo sugerido.

A modo ilustrativo y con fines que se explayarán posteriormente al mostrar los resultados, se procederá a reflejar lo hecho mediante el uso del método de Ward, y luego se mostrarán únicamente los dendogramas y las recomendaciones del resto de los métodos para abordar la primera conclusión en base a estas metodologías jerárquicas.

Para avanzar acorde a lo establecido en el párrafo anterior, se carga la librería nbClust y de ella aplicamos la función hclust a la matriz de distancias euclídeas, indicándole precisamente en uno de sus argumentos que le aplicaremos el método de Ward, el cual crea el objeto clust.nps.ward. Este objeto se utiliza para luego reflejar el dendograma respectivo.

Posteriormente, gracias a la aplicación de la función NbClust, se crea otro objeto denominado res.Ward, la cual nos arrojará el número óptimo de grupos recomendados en función de la aplicación de 30 índices que miden las distancias de las observaciones con diferentes ópticas, como el caso de del índice DB, el pseudo t2 o el de Hubert, entre otros. R aplica la regla de la mayoría para recomendar el grupo óptimo, e indica la cantidad de agrupaciones que recibió el visto bueno de la mayoría de estos 30 índices, bajo la condición de que como máximo puede armar 10 grupos, ya que en para un análisis y posterior aplicación práctica, segmentar el servicio en más grupos no sería viable.

En primera instancia, se presenta el análisis de la salida de R una vez que se aplicó NbClust:

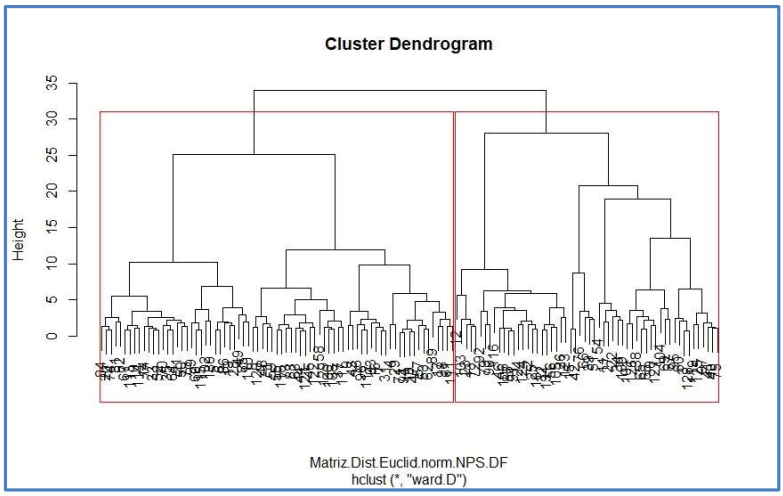
Figura 16 - Resumen resultados de índices con NBClust

```
*****************************************************************
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 3 proposed 3 as the best number of clusters
* 4 proposed 4 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 5 proposed 6 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 4 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2


*****************************************************************
```

En este resumen se indica que, luego de haber calculado los 30 índices con la aplicación del método de Ward, siete de ellos (23,33%) recomienda dos clústers como los óptimos.

Al saber que el número recomendado por "regla de la mayoría" es de dos clústers, se procedió a graficar el dendograma, el cual separa gráficamente los dos conglomerados con dos rectángulos en color rojo:

Figura 17 - Dendograma con dos clusters



Del gráfico se pueden visualizar dos agrupaciones con datos distribuidos de una forma un poco asimétrica, grupos con datos robustos en ambos casos, con una leve inclinación por el grupo de la izquierda ya que el grupo 1 concentra 71 registros homogéneos, y el 2 concentra 60 de ellos, acorde a lo calculado en R.

Adicionalmente, se comparte de forma complementaria el resultado gráfico que arrojan dos de los 30 índices que brindan el plus de analizar los resultados visualmente. Ellos son los índices de "Hubert" y el "D Index", donde en el primero podemos observar un recodo en el valor o

grupo 4, donde en apariencia la mejor agrupación es esa. Ahora bien, con el D Index ya no es tan claro dónde figura el recodo para establecer el grupo recomendado por este:

Figura 18 - Análisis gráfico de índice Hubert

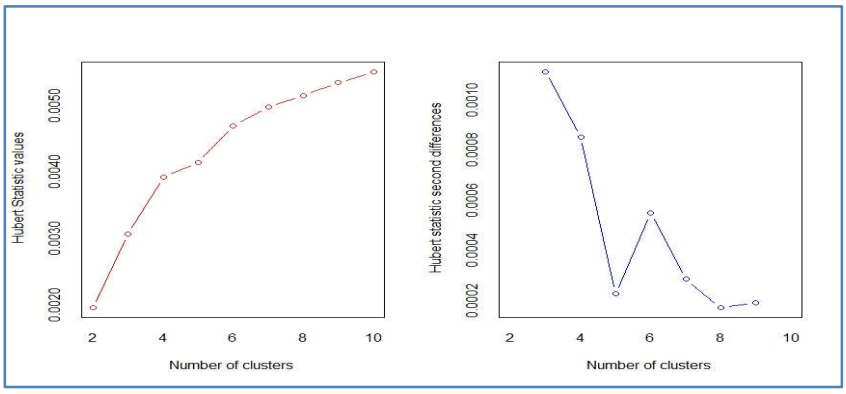
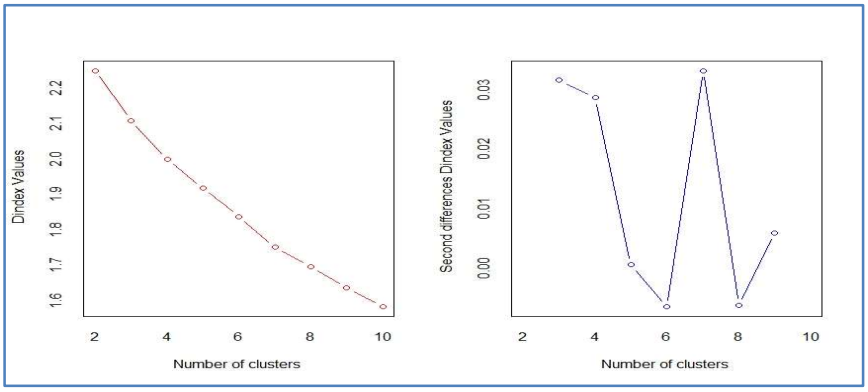


Figura 19 - Análisis gráfico de índice D Index



En los cuatro métodos restantes también arrojó la recomendación de agrupar por dos conglomerados principales, pero con las siguientes características distintivas:

- El **método del centroide** indica que 12 índices recomiendan dos grupos (40%), pero con la particularidad que las agrupaciones son: 130 observaciones para el grupo 1 y 1 observación para el grupo 2.
- El **método de la vinculación promedio** arroja un análisis idéntico al anterior: 12 índices recomiendan dos grupos, con 130 observaciones para el grupo 1 y 1 observación para el grupo 2.
- El **método del vecino más cercano** brinda nuevamente un análisis coincidente con sus 2 predecesores.
- El **método del vecino más lejano difiere** vagamente de los 3 anteriores, donde 11 índices recomiendan ahora a dos cómo el número óptimo de agrupaciones, y los grupos los divide en 126 y 5 observaciones.

Los dendogramas de estos cuatro métodos representan a la perfección lo antedicho, los cuales se adjuntan en el correspondiente apéndice.

Con lo hecho hasta el momento, se concluye en una primera instancia que no caben dudas que el grupo recomendado con los cinco métodos para ejecutar clústers es de dos grupos, pero con la salvedad que cuatro de ellos recomiendan grupos muy dispares (1 / 130 o 5 / 126), lo cual no permite realizar un análisis concluyente. Por tal motivo nos quedamos con Ward que no sólo presenta una distribución más equitativa, sino que también consiste en la minimización de la varianza dentro de cada grupo en lugar de medir distancias, lo cual lo hace desde el punto de vista del autor, un método confiable.

Lo llamativo de estos cuatro métodos es que en el grupo "pequeño" siempre aparecía la observación 12 (u observación 13 si incluyéramos los *outliers*) correspondiente a la encuesta ID "26875806", que en la prueba estuvo al límite de ser excluida pero que en ninguno de los 2 test superaba los umbrales (distancia de Mahalanobis de 17,845 y p-valor de 0,012). Si bien se pudo haber quitado, se quiso respetar dicho umbral y optó por Ward dado que ya brindaba información relevante del modelo, como se comentará en los párrafos subsiguientes.

3.4. Aplicación de método no jerárquico mediante uso de *k-means*

Finalmente, se procede a utilizar los centroides de los grupos conformados por Ward bajo el método jerárquico, para utilizarlo como semilla y aplicar ahora los métodos no jerárquicos al momento de utilizar *k-means*.

Una vez aplicado, R nos indica que ahora el grupo 1 concentra 84 observaciones, y el grupo 7 concentra 47. Recordemos que en Ward la distribución era un poco más equitativa: 71 y 60, respectivamente. Por otra parte, se creó y exportó una base que contiene una comparación entre cómo agrupó las observaciones Ward vs. como los capturó el método de *k-means*. El resultado está adjunta en la base *Comparación J Ward con NJ Kmeans.xlsx* adjunta en los Apéndices, y en este se aprecia que 41 encuestas fueron reasignadas de grupo.

Sobre este nuevo grupo reasignado volvemos a calcular sus medias como hicimos con el cálculo del centroide anterior:

Figura 20 - Análisis de medias de los dos grupos resultantes

	Grupo	Encuesta	Edad	Antiguedad_cliente	NPS	Antiguedad_Vehiculo	Prima	Monto_Pagado	Demora_Pago	kmeans
1	1	24202048	58.26	117.07	8.02	5.62	10858.77	84419.73	68.95	1.30
2	2	24023378	52.23	187.66	7.74	10.51	5609.10	50328.66	136.94	1.66

Este nuevo recuadro cambió las relaciones de las medias de las variables en cada grupo:

1. En algunos casos acentúa la diferencia de las medias, haciéndolos más heterogéneos entre sí. Por ejemplo, la brecha de la Edad pasó de ser de solo 0,7 años a casi seis años.
2. En otros casos, la brecha se redujo, como en el caso de la antigüedad del cliente que pasó de ser 88,3 meses a 70,6 meses, o la del NPS pasó de ser de 2,55 a 0,28.
3. En cambio, en el caso de monto pagado se revirtió la relación, por lo que perdió un poco fuerza la conclusión preliminar de que es más fuerte a la hora de dar un *feedback* la demora media en el servicio, que el monto promedio de pago.

Conclusión

El primer ensayo llevado adelante, con foco en la predicción del riesgo judicial en traspasos de mediación a juicio, fue un caso de éxito si se tiene presente que se trata de la primera iteración y que se aplicaron únicamente dos algoritmos de clasificación de los varios existentes en la actualidad. Ambos algoritmos trabajados, es decir, *decisión forest* y regresión logística, tuvieron tasa de éxito aceptables y similares, con una leve ventaja del primero al tener un mejor *accuracy* (88,3%), *precision* y *recall*, modelo el cual fue el que luego se indagó con mayor profundidad. Esto no implica que sea perfecto, ya que aún es baja la cantidad de verdaderos positivos (54,8%), pero dado que hasta el momento no existe un modelo de predicción con foco en el riesgo judicial, el hecho de implementarlo con estos hiperparámetros es un avance significativo para reducir dicho riesgo.

Se recuerda que el propósito final, una vez aplicado el algoritmo e identificados las mediaciones ingresadas con riesgo de traspaso, es tomar medidas de negociación más agresivas con el tercero y su abogado representante para evitar esto. En otras palabras, si al menos se logra identificar poco más del 50% de casos con un riesgo real de traspaso, la compañía podrá reducir sus costos financieros y operativos de manera significativa. Esto es así dado que, si bien puede que inicialmente se deba acudir a abogados propios con mayor *expertise* y honorarios para defender a la compañía en mediación, se evitarán todos los costos que conlleva el juicio. Y en cuanto a los falsos positivos que se identifiquen, no se ve en una primera instancia un riesgo aún mayor en aplicar otra clase de estrategia más agresiva en casos en donde quizá no sea necesario.

Por otra parte, el análisis de errores nos permitió ver de manera explícita que los predictores COD_JURISDICCION y DEMANDA fueron los más utilizados para el criterio de apertura del árbol de decisión y, por tal motivo, nos mostraron que las jurisdicciones con mayor volumen de siniestralidad y reclamos (Capital Federal, Avellaneda, Lomas de Zamora y San Isidro), fueron las que tuvieron menor tasa de error y una mejor precisión en la predicción de sus comportamientos. Esto dejó en evidencia que el modelo no es demasiado efectivo para el resto de las jurisdicciones, y en algunos casos, también para algunas demandas con valores de entre 196 mil y 247 mil pesos (a valores de 2019, los cuales habría que indexar o actualizar por tasa activa judicial). Es por tal motivo que se mencionó y alentó la posibilidad de realizar un segundo modelo, más específico, para las jurisdicciones de menos frecuencia, correspondientes en su mayoría al interior del país. A tal fin, se alienta a continuar la investigación y realizar otro ensayo que continúe en la búsqueda de un modelo más robusto para este escenario.

En resumen, este primer ensayo pretende dar un puntapié inicial para que las aseguradoras comiencen a indagar en esta problemática que, aunque preocupa al mercado (en especial al argentino), no se ha visualizado ningún accionar explícito al respecto, y en donde en estos últimos años ha aumentado la frecuencia de demandas. La clave consistirá no solo en mejorar y desplegar los modelos en producción, así como pensar en cuáles serán las mejores estrategias de negociación para estos casos, sino principalmente en mejorar los procesos de captura de información iniciales, ya que los modelos se basarán en estos para realizar sus predicciones.

Con relación al segundo ensayo realizado, el haber podido aplicar *clusters* con diferentes metodologías sobre el grupo de asegurados encuestados permitió hacer un análisis interesante sobre cómo es el servicio actual y cómo se puede encarar el servicio a futuro.

Si nos basamos en Ward, obtenemos grupos distribuidos de manera más equitativa, y las medias de las variables en cada grupo indican que los asegurados ponderan mucho más los tiempos del servicio que el monto pagado a la hora de calificar. Asimismo, el hecho de tener vehículos más nuevos mejora la calificación dada la menor dificultad de valuar cada uno de ellos y de conseguir los repuestos necesarios para repararlos.

Ahora bien, al observar las medias obtenidas al haber aplicado *k-means*, nos muestra que algunas brechas se acentuaron, otras se atenuaron, y algunas relaciones se revirtieron. Entre las características más destacadas, se presenta una relación más clara por ejemplo en el grupo 1, donde a mayor prima pagada, mayor es el monto pagado, mucho menor es la demora en el pago, más nuevo es su vehículo y mayor es su edad. Esto puede entenderse a que a mayor edad promedio, más aumenta el poder adquisitivo y la posibilidad de comprar autos 0KM (que tienen un costo mayor), y la compañía presta un mejor servicio de atención, acompañado por el posible hecho de que la probabilidad de obtener repuestos sea mayor en este grupo.

En conclusión, si se quiere mejorar la experiencia y percepción del asegurado, y en función de lo analizado en los modelos, la compañía deberá focalizarse más en los tiempos del servicio, entender estrategias de obtención de repuestos, ampliar la oferta de proveedores y mejorar el sistema de licitación de estos. Las compañías deben aliarse con Insurtech y proveedores de tecnologías para atender estas oportunidades de mejora, donde pueden obtener soluciones como la que presenta ClaimsServices, mencionada en la página 25.

Lo que se intentó realizar en este ensayo fue abordar el interés de las aseguradoras en segmentar a los clientes con foco en el servicio brindado y el *feedback* recibido por este, y no desde el

comportamiento de consumo y selección de productos que suele abordar el equipo de *marketing*.

Por supuesto, se sabe que ambos ensayos realizados tienen oportunidades de mejora y han explorado solo una porción de las opciones que hoy por hoy brindan las técnicas de *machine learning*. El poder realizar pruebas sobre bases de datos más robustas, aplicar validaciones cruzadas en las bases cuando se dividen los datos en *testing* y evaluación, utilizar otros algoritmos como ser redes neuronales, *vector support machines* (SMV), naive bayes o *gradient boosting machines* (GBM), entre otros, son prácticas que pueden hacerse para seguir profundizando en los ensayos realizados. A la vez, al que desee indagar más en la presente investigación, se le sugiere aumentar el volumen o período de la muestra para obtener resultados más sólidos, ya que los períodos de un trimestre (NPS) y un año (ingreso de mediaciones) pueden resultar algo acotado en sus respectivos campos dada la naturaleza de lo investigado, así como también el riesgo analizado, ya que únicamente se indagó en los seguros de automotores y no en multirriesgo o vida, por dar algunos ejemplos. Y se reafirma lo dicho anteriormente, al sugerir la posibilidad de desarrollar los modelos en grupos por regiones, ya que la región de AMBA (área metropolitana de Buenos Aires) presenta una frecuencia y comportamientos diferentes al resto del interior del país. Finalmente, y hasta que las compañías no cuenten con un gobierno del dato que vele por la implementación de mejores procesos de solicitud, captura y calidad de datos en producción, se recomienda no solo evaluar la posibilidad de eliminar registros con valores faltantes o inexactos, sino también evaluar las variables seleccionadas y probar con nuevas o eliminar algunas de ellas ya que, por sus características, puedan empeorar la efectividad del modelo. Algunas opciones son las de incluir la antigüedad del asegurado (modelo de predicción de traspasos), cantidad de riesgos contratados por el asegurado en la compañía (ambos modelos), cantidad de mediaciones y juicios en los que el abogado del tercero ha participado (modelo de predicción de traspasos), mostrando cantidad de casos abiertos, así como litigios ganados y perdidos, entre algunos ejemplos.

El presente autor desafía a las aseguradoras, y en especial al departamento de siniestros, a continuar con esta clase de investigaciones y comenzar a idear soluciones y servicios personalizados en función de segmentos de clientes, tal como hacen otros mercados. De avanzar de tal forma, no solo se cortaría con la marcada tendencia de brindar un servicio homogéneo por tipo de producto, cobertura, o zona geográfica, sino que lograría ampliar de esa forma la manera actual de segmentar al cliente.

Para finalizar, este trabajo final integrador persigue el propósito de brindar visibilidad del estado de situación del mercado asegurador, tal como se detalló en el primer apartado, para dar un diagnóstico de este. Pero, principalmente, el propósito es mostrar que la información está y que, si abrazan la transformación, se invierte de manera inteligente, se alían a proveedores de tecnología o *Insurtech y* fomentan la curiosidad y resolución de problemáticas desde otra perspectiva, el progreso en el mercado va a acelerarse de manera exponencial.

Este autor entiende que tanto el contexto económico y financiero actual, como el hecho de que las aseguradoras no tienen tanto margen de utilidad como sucede en otras industrias, muchas veces frenan algunas posibilidades de inversión. Pero esto no quita que las aseguradoras puedan, en el durante, impulsar una cultura de datos para incentivar el cambio de mentalidad, el estudio de nuevos casos de uso y la curiosidad en innovar, en lugar de defender la forma tradicional de gestionar el negocio por el mero hecho de que por décadas dio resultados y porque hasta ahora no apareció un jugador en el mercado que cambiara las reglas de juego. Si abrazan estas ideas, con el tiempo seguramente se conviertan en ese jugador que el mercado aún no encontró.

Referencias bibliográficas

Álvarez Plá, B. (2018). Insurtech, La revolución tecnológica del seguro. *Estrategas*, *170*, 76-84. https://www.revistaestrategas.com.ar/_revista/3769_estrategas_170.pdf

Blanco, F., Castro, J., Gayoso, R. y Santana, W. (2019). *Las claves de la Cuarta Revolución Industrial: Cómo afectará a los negocios y a las personas* (1era. ed.). Libros de Cabecera.

Carelli, Eliana. (10 de Noviembre de 2020) 'Los estudios aconsejan a sus clientes reducir el stock judicial'. *Estrategas*. https://www.revistaestrategas.com.ar/contenidos/6237/los_estudios_aconsejan_a_sus_clientes_reducir_el_stock_judicial

CIO España. (2017, 22 de febrero). El sector seguros invierte más de 160 mil millones de euros en transformación digital en los últimos 5 años. *CIO*. https://www.cio.com/article/2083882/el-sector-seguros-invierte-mas-de-160-mil-millones-de-euros-en-transformacion-digital-en-los-ultimos-5-anos.html

Díaz Zettelmann, V. (2023, 7 de junio). *Aumento de los casos de fraude.* Cobertura. https://cobertura.com.ar/2023/06/07/aumento-de-los-casos-de-fraude/

Infobae. (2023, 27 de enero). Inteligencia artificial en el aula: cómo es la tecnología que va a revolucionar la educación. https://www.infobae.com/educacion/2023/01/27/inteligencia-artificial-en-el-aula-como-es-la-tecnologia-que-va-a-revolucionar-la-educacion/

Jiménez, J. P. (2023). La inteligencia artificial y la nube: los dos pilares de la banca del futuro. *Estrategas*, *212*, 27. https://www.revistaestrategas.com.ar/digital/revista.php?revista=212

Lotko, A. (2015). *Classifying customers according to NPS index: cluster analysis for contact center services.* Central European review of economics & finance. 2(8), 5-30.

Lovagnini, A. (2023). ¿Pueden convivir inteligencia artificial, marketing y seguros? *Estrategas*, *212*, 16-18. https://www.revistaestrategas.com.ar/digital/revista.php?revista=212

Morselli, C.; Masias, V. H.; Crespo, F.; Laengle, S. (2013). *Predicting sentencing outcomes with centrality measures.* Security Informatics, a SpringerOpen Journal, 2(4), 1-9.

Muñoz Pardo, Alex (2018). *Aplicación de las nuevas tecnologías a la gestión de siniestros multirriesgos.* [Tesis de Maestría. Universitat de Barcelona]. Dipòsit Digital de la Universitat de Barcelona. http://diposit.ub.edu/dspace/handle/2445/144687

Perazo, C. (2023). Actualización del core, IA y ciberseguridad, grandes targets de las inversiones tecnológicas. *Estrategas*, *213*, 76-98. https://www.revistaestrategas.com.ar/digital/revista.php?revista=213

Redacción Clarín. (2017, 24 de febrero). Dos de cada tres casos que van a mediación se resuelven sin juicio. *Clarín*. (Publicado originalmente el 7 de julio de 2006). https://www.clarin.com/sociedad/casos-van-mediacion-resuelven-juicio_0_S16zUXNkRYg.html

Superintendencia de Seguros de la Nación. (2022). *Siniestros de casco del ramo automotores 2022*. https://www.argentina.gob.ar/sites/default/files/ssn_2022_siniestro_casco_0.pdf

Superintendencia de Seguros de la Nación. (2023a). *Situación del mercado asegurador a diciembre 2023*. https://www.argentina.gob.ar/sites/default/files/ssn_202312_sit_mercado_asegurador_anexo.pdf

Superintendencia de Seguros de la Nación. (2023b). *Indicadores del mercado asegurador a diciembre 2023*. https://www.argentina.gob.ar/sites/default/files/ssn_202312_indicadores_mercado_anexo.pdf

Todo Riesgo. (2024, 6 de marzo). Las 50 aseguradoras que encabezan el ranking de ventas del ramo automotores a diciembre de 2023. https://www.todoriesgo.com.ar/50-aseguradoras-ranking-automotores-diciembre-2023/

Todo Riesgo. (2024, 4 de marzo). Patrimoniales y mixtas: las 50 compañías líderes a diciembre de 2023. https://www.todoriesgo.com.ar/patrimoniales-mixtas-50-companias-lideres-diciembre-2023/

Torres Ayala, Meritxell (2020). *IT y Machine Learning en Seguros. Aplicación práctica en Fraudes.* [Tesis de Maestría. Universitat de Barcelona]. Dipòsit Digital de la Universitat de Barcelona. http://hdl.handle.net/2445/172022

Vitagliano, C. (2012). Determinar a priori el escenario de sentencia. *Revista Estrategas*. https://www.revistaestrategas.com.ar/contenidos/3204/determinar-a-priori-el-escenario-de-sentencia

Zapiola Guerrico, Martín. Las nuevas tecnologías en la actividad aseguradora, 53 Rev. Ibero-Latinoam.Seguros, 137-160 (2020). https://doi.org/10.11144/Javeriana.ris53.ntaa

Apéndices

Link al repositorio:

https://drive.google.com/drive/folders/1SlaJ26g9NwZnFM2T3PBeJXanu0Lbo5Tm?usp=sharing

Reporte del tutor

El presente trabajo aborda un tema relevante y actual en el ámbito de los seguros: la aplicación de técnicas de machine learning para mejorar la toma de decisiones en la gestión de siniestros. La investigación plantea un enfoque innovador, integrando métodos cuantitativos para analizar bases de datos complejas y predecir comportamientos en la gestión de siniestros, lo cual tiene un alto potencial para optimizar procesos en el sector asegurador y reducir costos.

Evaluación de la Metodología

La metodología empleada en el estudio es sólida y adecuada para el tema investigado. El autor utiliza técnicas de análisis de datos y modelos predictivos como el análisis de clústers y el algoritmo *k-means*, aplicando herramientas como R y Python para el procesamiento de datos. La metodología refleja un enfoque riguroso en el tratamiento de los datos, asegurando la confidencialidad mediante técnicas de anonimización. No obstante, la investigación podría beneficiarse de una mayor exploración en la validación cruzada de los modelos, lo cual fortalecería la precisión y confiabilidad de los resultados obtenidos.

Contenido y Relevancia de los Resultados

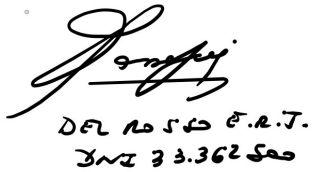
El contenido es extenso y cubre diversos aspectos clave, desde la evolución de los seguros hasta las aplicaciones específicas de inteligencia artificial en la predicción de siniestros. Además, el análisis del Net Promoter Score (NPS) y la segmentación de clientes son aportes significativos para mejorar la experiencia del cliente en el sector. Los resultados obtenidos son valiosos para la industria, particularmente en el contexto argentino, donde la transformación digital en seguros aún enfrenta barreras. Sin embargo, una discusión más detallada sobre las limitaciones del *dataset* y los posibles sesgos en el modelo contribuiría a una mejor interpretación de los hallazgos.

Consideraciones para Futuras Investigaciones

El trabajo abre diversas vías para futuras investigaciones, principalmente en la mejora de la calidad de los datos en la notificación de siniestros y en la incorporación de sensores IoT para una mejor prevención. La implementación de modelos de predicción de traspaso de instancias judiciales es otro aspecto de gran potencial, que podría extenderse en estudios futuros para reducir la judicialización de reclamos de terceros en el mercado. Es recomendable también explorar otras técnicas de machine learning que incluyan redes neuronales o modelos de

aprendizaje profundo, que podrían ofrecer aún más precisión en la predicción de siniestros complejos.

En conclusión, el trabajo presenta un análisis exhaustivo y metodológicamente robusto sobre un tema innovador y de gran relevancia en el mercado asegurador. Es un valioso aporte para futuras investigaciones en la gestión de siniestros y la toma de decisiones, evidenciando el potencial transformador de la inteligencia artificial en la industria.



Aclaración: EZEQUIEL RODRIGO JAVIER DEL ROSSO
Lugar y fecha: CABA, 09 de Noviembre de 2024