uc3m | Universidad Carlos III de Madrid

Máster Universitario en Ciencias Actuariales y Financieras
2020-2021

Trabajo Fin de Máster

"Detección del fraude en seguros de automóvil mediante técnicas de Machine Learning"

Germán Castellanos Heras

Tutores

José Miguel Rodríguez-Pardo

Jesús Ramón Simón del Potro

Madrid, 2021

AGRADECIMIENTOS

A mis amigos, por escucharme siempre y darme alas.

A mi familia, en especial a mi hermana melliza, por todo lo que compartimos en esta vida y el apoyo incondicional que me brinda.

A la Universidad Carlos III de Madrid, por las personas que ha cruzado en mi camino y por todo lo que me ha enseñado, académica y personalmente.

A mis tutores, por transmitirme sus conocimientos y dedicarme su tiempo.

Gracias.

RESUMEN

En los últimos años se ha incrementado el fraude detectado por las compañías de seguros, sobre todo en el ramo de seguro de automóvil. Por ello, se hace necesaria la adopción de una estrategia de control y detección del fraude. Para el desarrollo de esta estrategia se implementan numerosas técnicas estadísticas de predicción, a partir de las cuales se pretende identificar diferentes factores vinculados a estas actividades ilícitas y cuantificar la probabilidad de aparición de fraude en los siniestros declarados por los asegurados.

En el presente trabajo se realiza un caso práctico, en el cual se implementan diferentes técnicas de Machine Learning, permitiendo realizar un análisis y predicción del fraude en las reclamaciones de seguros.

Palabras clave: Fraude, seguro de automóviles, técnicas estadísticas, predicción.

ABSTRACT

In recent years, fraud detected by insurance companies has increased, especially in the field of automobile insurance. Therefore, the adoption of a fraud control and detection strategy is necessary. For the development of this strategy, numerous statistical prediction techniques are implemented, from which it is intended to identify different factors linked to these illicit activities and quantify the probability of the appearance of fraud in claims declared by the insured.

In this work, a practical case is carried out, in which different Machine Learning techniques are implemented, allowing an analysis and prediction of fraud in insurance claims.

Key words: Fraud, car insurance, statistical techniques, prediction.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

El fraude es un fenómeno presente en el sector asegurador desde sus inicios, sobre todo en el ramo del seguro de automóvil. Este fenómeno ocasiona un gran impacto económico y social, "*Prácticamente puede afirmarse que todos y cada uno de nosotros resulta perjudicado por el fraude al seguro, puesto que, aunque la víctima sea la compañía, no hay que olvidarse que ante estas situaciones podría decirse que "la banca siempre gana", por lo que esas cuantías defraudadas se van a repercutir en el resto de las primas de los asegurados*" afirmaba en la revista de Actuarios de otoño 2016 Francisco Vázquez, inspector de la brigada central de delincuencia económica y fiscal.

El impacto de estas actividades está estrechamente relacionado con el contexto económico y social. Diferentes estudios realizados en 2020 y 2021 por entidades como ICEA, AXA y Línea directa, ponen de manifiesto un incremento de los casos de fraude en el escenario de crisis económica global de la última década. La información en este tema posee una elevada dificultad de interpretación, en base al análisis realizado en la revista de Actuarios se evidencia la existencia de una mayor tasa de fraude en España respecto a otros países de referencia de la Unión Europea (Actuarios, 2016).

El mayor control regulatorio derivado de Solvencia II y el aumento de los casos de fraude detectados, han hecho que las compañías tomen una mayor consciencia sobre el impacto negativo en el negocio de estos comportamientos ilícitos. En los últimos años, las entidades aseguradoras han emprendido un proceso de lucha activa contra los actos fraudulentos, destinando una mayor cantidad de recursos y aumentando las tareas de investigación. Así mismo, el desarrollo de las nuevas tecnologías ha favorecido el estudio de este fenómeno, proporcionando a los actuarios una gran cantidad de información sobre la cual construir modelos de aprendizaje automático y numerosas técnicas y algoritmos de investigación estadísticas.

Este hecho ha alentado a numerosos investigadores a realizar análisis cada vez más exhaustivos sobre este fenómeno. Uno de los centros de estudio contra el fraude más importantes a nivel internacional se encuentra en el Departamento de Economía, econometría y estadística de la Universidad de Barcelona. Este centro de investigación está dirigido por Monserrat Guillen. Ella justifica la creación de un departamento de control y el uso de herramientas de mitigación del fraude, partiendo de los resultados técnicos negativos que se han observado en el ramo de seguro de automóvil y los cuales vienen ocasionados en gran medida por las prácticas fraudulentas de los asegurados (Guillén, et al., 1999).

El análisis del fraude cada vez se está extendiendo más a todas las líneas de negocio, aunque sigue siendo el ramo de seguros de automóvil el que posee mayor implementación de técnicas cuantitativas y modelos econométricos. El análisis del fraude en el seguro de autos ha sido realizado bajo diferentes técnicas, destacando la regresión lineal múltiple, los modelos lineales generalizados, los árboles de decisión y las técnicas de Gradient Boosting.

La implementación de técnicas modernas en las entidades ha impulsado la adopción de medidas de control y la detección de este tipo de comportamientos ilícitos. Además de la adopción de medidas, los estudios revelan aspectos particulares que presentan las reclamaciones fraudulentas, como determinados patrones de comportamiento de los

asegurados y características de los siniestros. De forma general, el fraude en el seguro de autos se enmarca en la reclamación de siniestros inexistentes, la reclamación exagerada de daños y en caso de las tramas organizadas, la simulación de siniestros asociando diferentes perfiles vinculados al seguro.

A partir de la observación de este fenómeno y la implementación de técnicas estadísticas, se pretende identificar el perfil de los asegurados que comete este tipo de prácticas, así como, determinar una probabilidad sobre los factores que influyen en una mayor aparición de fraude. En este sentido, el presente trabajo busca estudiar el fraude, exponiendo un marco teórico del mismo y elaborando un caso práctico referente a las diferentes metodologías de Machine Learning que se emplean en su identificación. Se persigue identificar la metodología más efectiva en el análisis y detección del fraude, con el fin de proporcionar a la compañía una herramienta sobre la que realizar una buena política de suscripción y tratamiento de reclamaciones. Es decir, elaborar una buena estrategia de control del fraude en el seguro de automóviles.

Se seguirá el siguiente orden de presentación. En primer lugar, se elabora un marco teórico referente al fraude en el seguro de automóviles. Posteriormente, se presentarán las diferentes técnicas estadísticas que se emplean en la modelización del fraude. Por último, se implementarán las metodologías sobre una base de datos de reclamaciones automovilísticas y se analizarán los resultados obtenidos.

1. MARCO TEÓRICO: EL FRAUDE EN EL SEGURO DE AUTOMÓVIL

Como ya se ha comentado, los comportamientos fraudulentos son inherentes al negocio asegurador. Según datos de la Unión Española de Entidades Aseguradoras y Reaseguradoras (UNESPA), se estima que en España cada minuto una persona trata de defraudar a una compañía de seguros. Estas prácticas a pesar de no poseer una elevada visibilidad en la gestión diaria de una compañía constituyen uno de los principales problemas del sector. Las reclamaciones fraudulentas representan un porcentaje significativo del total de reclamaciones recibidas, sobre todo en el seguro de automóvil.

En este contexto, se debe entender con claridad el concepto de "Fraude", se podría definir a grandes rasgos como engaño o estafa. El fraude de manera estricta se define como *"la situación que se produce cuando el propio asegurado o beneficiario ha procurado intencionadamente la ocurrencia del siniestro o exagerado sus consecuencias con ánimo de conseguir un enriquecimiento injusto a través de la indemnización que espera lograr del asegurador"* (ASEPEG, 2020). Por tanto, se identifica como todas aquellas prácticas que realiza un asegurado o mediador, con el objetivo de obtener un enriquecimiento a través del siniestro declarado, el cual no le pertenece según las condiciones de la póliza. Estos comportamientos delictivos se pueden desarrollar en diferentes momentos de la vida de la póliza y a través de diferentes metodologías, siendo los más comunes aquellas en la que se proporcionan datos incorrectos, se da ocultación de información o incluso se declaran siniestros inexistentes.

De forma directa, los pagos asociados a reclamaciones fraudulentas afectan negativamente a los beneficios que obtiene una compañía de seguros. El fraude posee un coste muy elevado no sólo para las entidades aseguradoras, sino también para todos los asegurados de la compañía. El coste del fraude obliga a las compañías a cobrar una prima superior a sus clientes, perjudicando a aquellos individuos que no incurren en prácticas engañosas. El montante extra cobrado a los asegurados se destina no solo al pago de reclamaciones dolosas, sino también al desarrollo de procesos de prevención y detección del fraude.

La detección y erradicación del fraude supone una lucha constante para las entidades aseguradoras, ya que posee una gran influencia en los resultados técnicos obtenidos por la compañía. Según el estudio "Encuesta sobre Fraudes en Seguros 2020" (FRISS, 2020) compañía que proporciona software de detección de fraude a compañías de seguros, en este movimiento se identifican tres grandes tendencias, que son la automatización de procedimientos, integración de metodologías basadas en inteligencia artificial y la creación de una cultura adversa al fraude. Esta última hace referencia a la importancia de transmitir y concienciar a la sociedad de los enormes efectos negativos que poseen las actividades fraudulentas a nivel económico.

1.1.- Historia y evolución del fraude en España

Las consecuencias de las actividades fraudulentas por parte de los clientes e intervinientes se dejan ver notablemente tanto en el número de siniestros reportados como en la cuantía de estos. Estas consecuencias son la razón por la que, a lo largo de la historia las entidades

aseguradoras han destinado una gran cantidad de recursos a desarrollar mecanismos de acción frente a la detección del fraude.

Desde los inicios de la actividad aseguradora, los diferentes intervinientes en el proceso de suscripción y análisis de reclamaciones han detectado prácticas fraudulentas en los diferentes ramos de seguros. En el contexto tradicional, las entidades aseguradoras centraron sus esfuerzos en la identificación de estas prácticas a partir de anomalías en la documentación aportada por los reclamantes. De cara a aumentar la eficiencia en este ámbito, se llevan a cabo formaciones continuas a los peritos, proveedores y mediadores. De tal forma que, basándose en el juicio experto se pudieran identificar determinadas características de los siniestros que indicaran prácticas sospechosas. Una vez reconocida una posible reclamación fraudulenta sobre la base de sus características se le otorga una clasificación denominada "Red flag", por la cual se inicia un proceso de verificación de los datos, en el que se somete la reclamación a un análisis más exhaustivo.

El aumento de los casos de fraude detectados, el mayor impacto en las entidades y la entrada en vigor de determinadas normas regulatorias (Solvencia II e IFRS17) ha aumentado las exigencias de control sobre este tipo de comportamientos. En los últimos años, las compañías han modificado la orientación de sus técnicas de detección de fraude, la observación y el estudio de los daños reclamados por los asegurados pasan a un segundo plano del estudio y las tecnologías desarrolladas se centran en el análisis de patrones de comportamiento que puedan definir el perfil de un defraudador. Esta nueva orientación respecto al análisis del fraude se basa en la importancia de establecer las características generales que poseen los defraudadores, identificando los elementos que aumentan la probabilidad de aparición de fraude.

Debido a la creciente importancia que el negocio asegurador ha dado a este problema y la capacidad de estudiar el mismo desde un punto de vista analítico, ha despertado el interés de los investigadores. En los últimos años ha crecido de forma exponencial el número de estudios relacionados con la detección y prevención del fraude, los cuales cada vez son más específicos, detallados y encaminados al análisis estadístico los elementos vinculados a comportamientos fraudulentos.

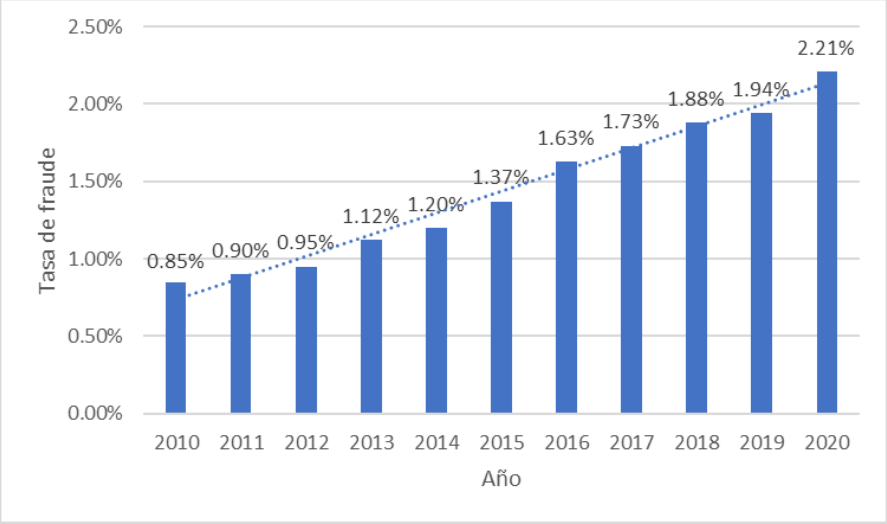
Esta trayectoria ha sido a su vez impulsada por las entidades aseguradoras, las cuales han aumentado las inversiones en tiempo, dinero e información disponible para este tipo de estudios con el fin de mejorar las aptitudes de los investigadores. A partir del informe "El Fraude al Seguro Español Año 2019" (ICEA, 2019) en el que han participado 35 entidades aseguradoras (54% del mercado asegurador), el rendimiento de la investigación obtenido por las compañías es de 46,2€ por cada euro destinado a investigar presuntos casos de fraude.

Los estudios realizados junto con los avances tecnológicos están permitiendo la utilización de manera inteligente de una cantidad extraordinaria de datos, tanto públicos como de las entidades aseguradoras. Según explicaba en 2016 para el periódico CincoDías Arturo López Linares, responsable de gestión de fraude y recobros de AXA, "Una de las claves del aumento de las cifras de fraude es que ahora se invierte mucho más en destaparlos, por lo que salen a la luz casos que en otro momento se hubieran pasado por alto". Es decir, las entidades aseguradoras son capaces de llevar a cabo una detección del fraude mucho mayor a partir de la implantación de metodologías de detección mucho más eficaces.

Según el documento (ICEA, 2019) referenciado anteriormente, el cual es el más importante en el sector sobre cifras de fraude, se estima que sólo en el año 2019 se detectaron alrededor de 181.310 intentos de fraude a las entidades aseguradoras españolas, suponiendo estos casos entorno al 70% de los intentos de estafa. Esta cantidad de siniestros fraudulentos detectados habría supuesto un gasto para las compañías de más de 454 millones de euros en reclamaciones ilegítimas.

A partir de los avances en detección de fraude derivado de la integración de nuevas metodologías por las compañías, se observa un crecimiento del número de prácticas fraudulentas con el transcurso de los años. Este hecho viene evidenciado en el informe anual de detección del fraude "VIII Mapa AXA del Fraude en España" (AXA, 2021). Este informe defiende que, aunque la mayor parte de los asegurados y demás intervinientes actúan de forma legítima se aprecia un aumento de los intentos de fraude en España, llegando a duplicar los casos de fraude en la última década.

Figura 1. Tasa de fraude en España (Casos de fraude/Siniestralidad)

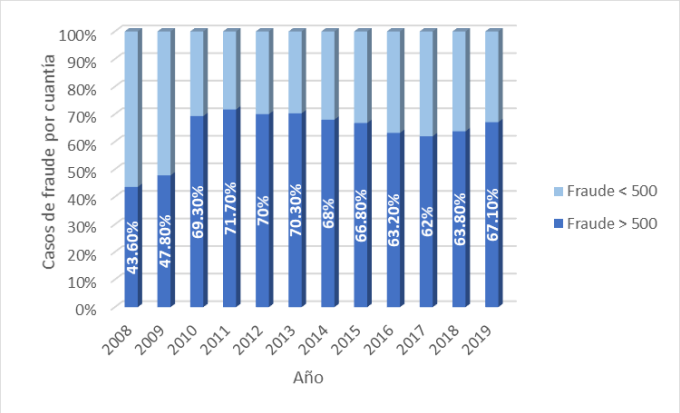


Fuente: Elaboración propia a partir de datos de "VIII Mapa AXA del Fraude en España" – AXA

Como se puede observar en la Figura 1, el incremento de la tasa de fraude en España en la última década posiciona al año 2020 como uno de los periodos donde más comportamientos deshonestos se han producido, alcanzando una tasa de fraude nacional del 2,21% respecto al total de los siniestros reclamados. Además de la mayor eficacia de las técnicas en detección de fraude llevada a cabo por las entidades que localiza un mayor número de fraudes, este fenómeno puede venir explicado por una disminución de la siniestralidad, la cual implicaría un mayor porcentaje de fraude respecto del total.

Uno de los fines de estudiar la evolución del fraude a lo largo de los años, además de observar el número de actividades delictivas que se cometen, consiste en analizar y estudiar las causas de los movimientos producidos en la cuantía que los defraudadores intentan estafar a las compañías. A partir de este análisis se puede estimar el impacto que poseen las actividades fraudulentas en el negocio asegurador, establecer patrones de cambio en los comportamientos y definir cuál es el ámbito en el que se desean centrar los esfuerzos de investigación.

Figura 2. Distribución cuantía de fraude



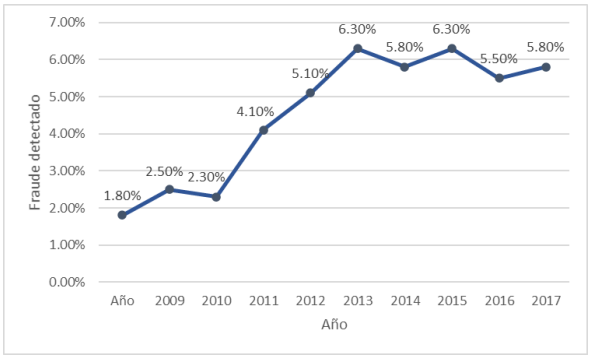
Fuente: Elaboración propia a partir de datos de "El Fraude al Seguro Español Año 2019" – ICEA

La Figura 2, correspondiente al estudio de ICEA, muestra cómo ha ido cambiando el patrón de fraude respecto a su cuantía. A principio de la década la tipología de fraude corresponde en su mayoría a reclamaciones de baja cuantía. A partir de la crisis se observa un crecimiento de los siniestros fraudulentos de importe superior a 500€, este hecho puede venir derivado por dos fenómenos. Por un lado, el crecimiento exponencial observado respecto a grupos organizados y mafias y, por otro lado, el mayor esfuerzo desarrollado por las compañías en detección de este tipo de fraudes.

En términos medios se puede afirmar que se ha experimentado un crecimiento de las reclamaciones fraudulentas con importe superior a 500€. En el año 2019, el 67,1% de los siniestros fraudulentos reclamados son de esta tipología, hecho que repercute directamente sobre el coste del fraude en la actividad aseguradora. Se estima que gracias a las técnicas de detección y prevención del fraude se ha detectado alrededor del 71,5% de esta tipología de reclamaciones, evitando el pago de alrededor de 440 millones de euros.

Cabe destacar que, el 73% de los fraudes de importe superior a 500 euros detectados por las compañías provienen del ramo de seguros de automóvil, así mismo, el 45,54% de los fraudes de importe inferior a 500 euros provienen de este ramo. Por estas razones, este es uno de los ámbitos en los que mayor lucha contra el fraude se ha realizado, ya que debido a su naturaleza obligatoria supone el ramo de seguros en el que mayor incidencia de fraude se soporta.

Figura 3. Evolución Fraude en seguros de autos



Fuente: Elaboración propia a partir de datos de "V barómetro de fraude al seguro" – Línea Directa

En la Figura 3, se puede observar cómo pese al fin de la recesión económica el fraude en el ramo de automóviles experimenta un crecimiento en la última década alcanzando máximos de 6,3% y llegando a una tasa del 5,8% de fraude en el año 2018. Es decir, 6 de cada 100 siniestros son fraudulentos, lo cuales de haberse pagado habrían costado a la compañía alrededor de 803 millones de euros en el último año.

Es de suma importancia para el estudio de prácticas fraudulentas tener siempre presente el horizonte temporal en el que se centra el estudio. Esto es debido a que el patrón de prácticas deshonestas está estrechamente relacionado con el ciclo económico. Así, cuando la economía se encuentra en un ciclo económico de crecimiento, se observa que las tasas de fraude experimentan una reducción y viceversa en periodos de recesión económica. Este hecho puede estar relacionado con la mayor necesidad económica de la sociedad, así como, con la reducción del "Moral Hazard"[1] respecto a este tipo de prácticas fraudulentas.

Acorde a esta relación del ciclo económico con el número de reclamaciones fraudulentas, en el estudio sobre fraude en España "VII Mapa AXA del Fraude en España" (AXA, 2020), se expresa cómo en el año 2008 tuvo lugar el incremento más relevante del que se tiene constancia, pasando de 30 millones de euros detectados en 2007 a los 48 millones de euros registrados en 2008. No obstante, esta afirmación se cumple en líneas generales, pero como hemos podido comprobar en el caso concreto de seguro de automóviles, este hecho no se cumple. Es más, a partir de la crisis de 2008 se ha presenciado un crecimiento, tanto en número como en coste, de los fraudes en este ramo.

El actual contexto de crisis económica y social provocada por el COVID-19, augura un escenario de fraude desfavorable para las entidades de seguros durante los próximos años. Según el estudio "Encuesta sobre Fraudes en Seguros 2020" (FRISS, 2020), la mayoría de los profesionales de la industria de seguros calculan que las reclamaciones fraudulentas se han duplicado desde el inicio de la pandemia.

Este informe sugiere que alrededor del 18% de las reclamaciones totales que se han realizado durante el COVID-19 posee algún elemento fraudulento. Debido a la adaptación que se han tenido que realizar respecto al lugar de trabajo, las compañías han incurrido en una mayor carga de trabajo y la jornada laboral en remoto ha dado lugar a un menor número de inspecciones. Estos hechos han favorecido el incremento de este tipo de actividades delictivas, en los que los defraudadores oportunistas han aprovechado estas brechas en la supervisión para realizar sus reclamaciones y obtener un beneficio.

No obstante, la rápida transformación propiciada por la pandemia ha hecho que las entidades aseguradoras adapten su modelo de negocio y también las técnicas de lucha contra el fraude. La digitalización de las diferentes operativas vinculadas a los productos posee un potencial que las compañías pueden explotar con el fin de beneficiar a sus clientes evitando prácticas delictivas, una de las operativas más potenciada durante la pandemia de Covid-19 en el ramo de seguro de automóvil ha sido la video-peritación. En definitiva, la transformación tecnológica, estrechamente conectada con la inteligencia artificial, ha acelerado el cambio de metodologías y técnicas de detección en el que las aseguradoras ya estaban trabajando.

[1] El Moral Hazard hace referencia a los superiores riesgos que pueden asumir los individuos en sus decisiones cuando las consecuencias negativas no son asumidas por ellos.

Tener en cuenta la evolución de las prácticas fraudulentas y su vinculación con el contexto económico, resulta de vital importancia de cara a implementar técnicas de detección de fraude. Estos análisis permiten identificar cuáles son los periodos en los que más probabilidad existe de recibir una reclamación engañosa y a partir de esa probabilidad orientar los procedimientos de prevención y detección.

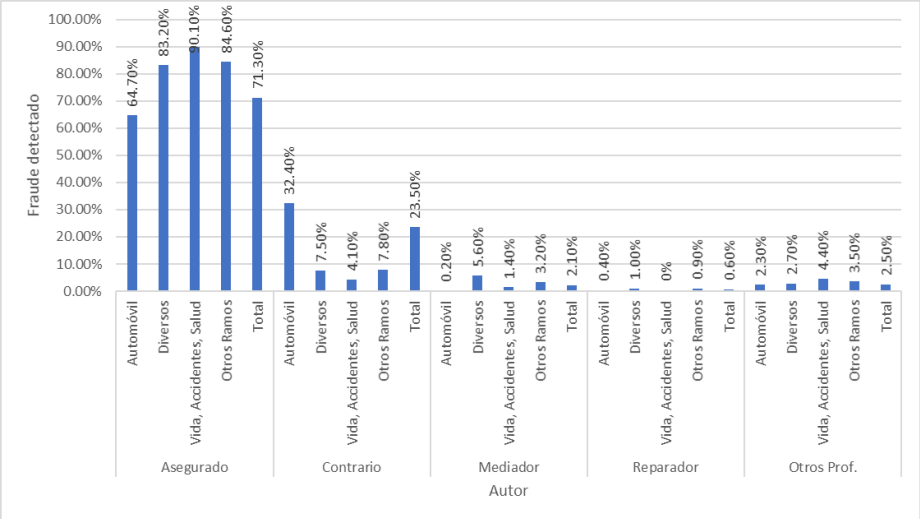
1.2.- Tipos de fraude y procesos de gestión

En aras de entender de forma adecuada las técnicas y algoritmos de detección de fraude que se analizan posteriormente en este trabajo, es de vital importancia conocer de forma detallada las diferentes prácticas deshonestas llevadas a cabo por los asegurados o demás intervinientes. Se poseen diversos criterios de clasificación, a través de los cuales se pueden identificar los comportamientos de los defraudadores y los fines de estos. Partiendo de las diferentes categorías se determina que metodología de detección aplicar en cada una de ellas. Incluso dentro de una misma categoría, es posible que las técnicas y algoritmos utilizados por la compañía sean diferentes con el fin de adaptarse a las características o modus operandi concretos que poseen los defraudadores.

Una de las premisas de los defraudadores es pasar desapercibidos. En los actos deshonestos se llevan a cabo multitud de técnicas de ocultación y colaboración entre los diferentes intervinientes en el evento, dificultando así la detección de las acciones fraudulentas. Es por ello por lo que se debe prestar mucha atención en las diferentes categorías que se presentan a continuación, ya que revelan posibles características propias de los defraudadores y actividades fraudulentas que facilitan su detección.

En primer lugar, se realiza una clasificación de las características referentes a las personas que cometen actos deshonestos. En este sentido, se debe tener en cuenta que no sólo son los asegurados quienes cometen actos fraudulentos, sino que también son personas o empresas vinculadas a la entidad quienes aprovechan su condición de empelado para obtener un beneficio a través de estas prácticas.

Figura 4. Distribución del fraude por autor y ramo



Fuente: Elaboración propia a partir de datos de "El Fraude al Seguro Español Año 2019" – ICEA

La mayor parte de los actos fraudulentos vienen desarrollados por los propios asegurados, según se puede observar en la Figura 4 (ICEA, 2020). Los fraudes realizados por los asegurados se concentran sobre todo en los ramos referentes a vida y diversos (coberturas de responsabilidad civil, hogar y empresas), siendo el ramo de seguros de automóvil el que menor incidencia posee. Por el contrario, el ramo de autos es el que mayor fraude presenta realizado por el tercero implicado. El resto de los intervinientes se puede observar que poseen un menor peso dentro del fraude analizado, siendo el mediador y otros profesionales en los ramos de vida y diversos los que mayor tasa de fraude representan.

Tener una visión de cuáles son los clientes o intervinientes que poseen mayor probabilidad de cometer fraude, supone una de las herramientas de identificación más importantes para las compañías. A partir de las siguientes categorías, las compañías pueden determinar sobre qué personas se debe realizar una investigación más exhaustiva.

- Defraudador ocasional:

  Este tipo de defraudadores normalmente llevan a cabo el engaño una vez se ha producido el siniestro de forma accidental o real. Viene desencadenado por las pérdidas económicas o patrimoniales a las cuales se enfrenta el asegurado tras la ocurrencia del evento desfavorable. Según el informe sobre fraude en España (AXA, 2020), esta tipología representa la mayor proporción de casos fraudulentos acumulando un total del 54% de los siniestros, así mismo suelen ser hechos engañosos de baja cuantía ya que la mayoría de los casos no superan los 600€.

  Por otro lado, esta categoría a su vez engloba a todos aquellos clientes que, sin tener que hacer frente a una situación económica desfavorable, realizan de forma puntual la reclamación de la indemnización vinculada a un siniestro. Este tipo de conducta es más grave y supone una tipología denominada fraude premeditado, en la cual el asegurado simula la ocurrencia del evento cubierto por la póliza de seguros, en este tipo de actos suelen estar implicadas varias personas. Este tipo de delitos suponen el 45% de los fraudes detectados en las compañías de seguros, además el coste de este es sensiblemente superior al realizado sin premeditación.

- Defraudador habitual:

  Esta categoría se caracteriza por la simulación la ocurrencia de un evento cubierto en la póliza de forma repetida, con el fin de cobrar la indemnización por dicho siniestro de forma reiterada en el tiempo. En este caso la compañía se estaría enfrentando a grupos organizados que dado su nivel de información sobre los siniestros que desean producir y su vinculación con profesionales de la industria, son muy difíciles de identificar.

  Este tipo de fraude ha sufrido un aumento, el mayor crecimiento se observó en el año 2018 en el que se identificaron un 23% más de tramas respecto al año anterior, según viene recogido en el informe de fraude (Línea Directa, 2020). Así mismo, en un análisis más prologado se ha determinado que en la última década el número de organizaciones que se dedican a estafar a las compañías aseguradoras se ha multiplicado por tres.

  Esta categoría es la que menor peso posee dentro de los defraudadores identificados. No obstante, en términos de cuantía si posee una gran relevancia ya que las cantidades

reclamadas normalmente están vinculadas a indemnizaciones por daños corporales los cuales suelen ser muy superiores a los fraudes convencionales, alcanzando el coste medio por siniestro de 10.543€. Según el informe de fraude elaborado por AXA en 2020, esta tipología de fraude supone un coste para las compañías de alrededor de 3 millones de €.

Ambos tipos de defraudadores emplean las coberturas de la póliza con el fin de obtener un beneficio a través de actos ilícitos. En términos generales los defraudadores habituales poseen técnicas mucho más maduras, es por ello por lo que este tipo de fraude es mucho más peligroso y difícil de detectar. Las compañías centran sus esfuerzos en analizar en detalle el modus operandi para poder obtener indicios de las prácticas ilícitas, incluso se puede precisar la colaboración de las fuerzas y cuerpos de seguridad del estado.

Otro de los puntos que se deben tener en cuenta a la hora de realizar estudios sobre el fraude, es la determinación de la naturaleza de estas actividades. Se pretende determinar cuáles son los tipos de daños a partir de los cuales el defraudador realiza la reclamación engañosa. En esta clasificación se encuentran la tipología referente a daños materiales y daños corporales. Bajo esta clasificación se establece que el 82% de los casos fraudulentos pertenecen a la categoría de daños materiales (AXA, 2021).

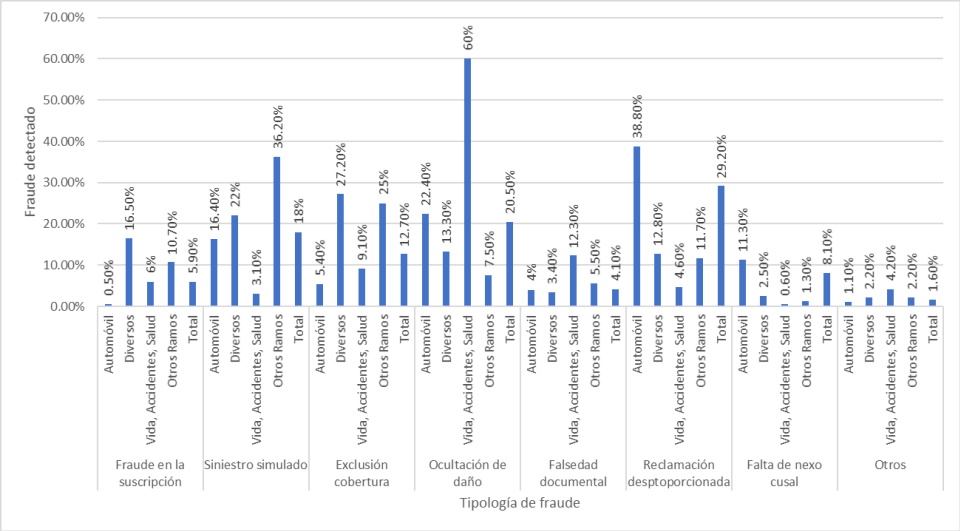
El peso que poseen los daños corporales en las prácticas fraudulentas ha ido reduciéndose en los últimos años. Este hecho es debido a determinadas normas legislativas que han entrado en vigor recientemente, como la reforma del Baremo de 2016 o la reforma del Código Penal que modificaba la ley Orgánica. Estas reformas se explicarán con más detalle en epígrafes siguientes, pero en líneas generales estas modificaciones de la ley dificultan la simulación o exageración de daños corporales.

Una vez analizada la tipología de personas y la naturaleza de los fraudes que estas realizan, se debe prestar especial atención a las características de las actividades fraudulentas que se han observado. Estas actividades están presentes en todos los ramos de seguros, adaptándose a las coberturas y debilidades propias de cada ramo. En líneas generales se puede identificar la siguiente taxonomía de actividades fraudulentas:

- Duplicidad de los seguros para la cobertura de un mismo evento. Consiste en contratar de forma deliberada varias pólizas de seguro, en diferentes compañías, que cubran un mismo riesgo. En la actualidad este tipo de prácticas está perdiendo relevancia gracias a la información cruzada que comparten las aseguradoras.
- Simulación de los eventos acaecidos y cubiertos. Esta categoría hacer referencia a la reclamación de siniestros que no han ocurrido o bien se han producido de forma intencionada. Este tipo de reclamaciones son las más difíciles de detectar, en muchos casos debido a la escrupulosa planificación y número de intervinientes que alcanza.
- Modificación de las características del siniestro. Este tipo de fraude se realiza con el objetivo de que el evento acaecido quede recogido por las condiciones generales, especificas o particulares de la póliza suscrita. El asegurado en muchas ocasiones pretende recibir compensaciones por eventos que no están recogidos dentro de su contrato de seguros. En este tipo de fraudes es relativamente fácil de detectar por las compañías ya que se suele poseer con frecuencia la falta de relación entre los daños reclamados y el siniestro ocurrido.

- Ocultar o falsear información relevante en el momento de contratación de la póliza. En este contexto se encontrarían todas aquellas reclamaciones en las que el daño o lesión sufridas por el asegurado se habían producido con anterioridad a la entrada en vigor de la póliza. Así mismo, se deben tener en cuenta todas aquellas acciones que realiza el asegurado de forma intencionada proporcionando información incorrecta, ya sea de la persona asegurada o del objeto poseedor de la cobertura.
- Agravamiento de las consecuencias del siniestro. El principal objetivo de este tipo de actividades fraudulentas es la obtención de un beneficio por parte del defraudador a través de la declaración desmedida de los efectos negativos del evento acaecido. Esta tipología de fraude es muy común encontrarla en las reclamaciones referentes a lesiones corporales, de tal manera que la entidad debe hacer frente a indemnizaciones mucho más elevadas.

Figura 5. Distribución de la tipología de fraude por ramo



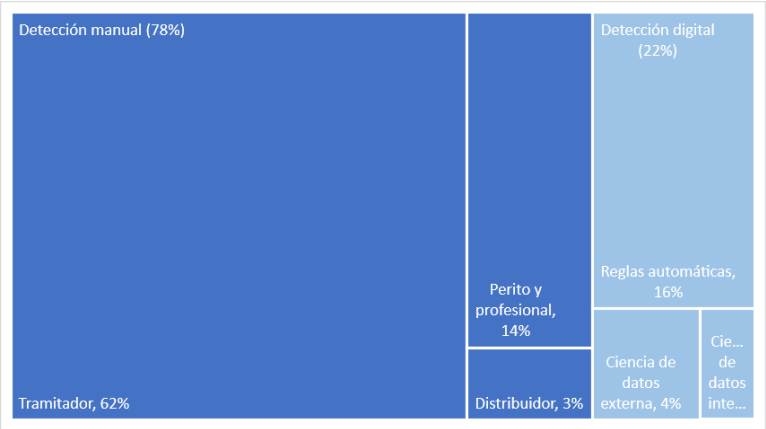
Fuente: Elaboración propia a partir de datos de "El Fraude al Seguro Español Año 2019" – ICEA

A partir de la Figura 5, se reconoce la ocultación de daño o lesión preexistente como la principal tipología de fraude en seguros de vida, es por ello por lo que siempre se precisa de la cumplimentación de un cuestionario al contratar la póliza. Respecto al ramo de automóviles la principal causa de fraude es la reclamación desproporcionada de daños, vinculada a coberturas de daños corporales. Esta tipología de fraudes en el seguro de automóvil se pudo observar en el 38,5% de los actos fraudulentos analizados. La tipología de fraudes menos comunes son la falta de nexo y la falsedad de facturas, esto es debido a que es la tipología más fácil de identificar en una primera revisión de la reclamación.

Uno de los principales hándicaps que poseen las entidades aseguradoras respecto a la detección del fraude, es el desconocimiento de la conducta de sus clientes. En un primer momento las compañías se topan con la dificultad de distinguir entre buenas y malas prácticas. Este hecho se da sobre todo en el ramo de seguro de automóviles, donde dificulta la identificación de actividades fraudulentas y el paso de la reclamación a un proceso de investigación más profundo. A partir de los tipos de fraude, se han desarrollado diferentes metodologías de actuación, adaptando los procedimientos de análisis y evaluación con el objetivo de tener en consideración determinados aspectos concretos de cada reclamación.

La detección del fraude en una compañía sigue siendo dominada por el juicio experto de los empleados, sobre todo en el ramo de seguros de auto. En la Figura 6, se puede evidenciar que el 78% de las actividades fraudulentas detectadas se han realizado de forma manual. Mientras que las actividades de honestas detectadas mediante modelos de detección o inteligencia artificial ocupa el 22% restante (AXA, 2020).

Figura 6. Clasificación de las técnicas de detección



Fuente: Elaboración propia a partir de datos de "VIII Mapa AXA del Fraude en España" – AXA

Dentro de cada una de las categorías identificadas, los tramitadores de siniestros son aquellos que mayor detección de fraude han realizado. Este hecho en parte se puede justificar por ser el primer esclavón de la cadena de detección que posee los datos del siniestro. Los peritos y profesionales son la segunda categoría que mayor detección han realizado, esta función es muy eficaz en la detección del fraude, ya que tienen la capacidad de observar de forma real cuales son las consecuencias del siniestro y si estas se ajustan a la magnitud del evento ocurrido. Por otro lado, se tienen las reglas de detección automáticas, la cuales revelan la gran implantación de nuevas tecnologías y algoritmos de detección llevada a cabo en las compañías de seguros.

A pesar de las diferencias entre ambas categorías, se puede apreciar el gran esfuerzo que han realizado las entidades aseguradoras en materia de detección de fraude. En primer lugar, en la formación del personal de la compañía y la integración de algoritmos. En este sentido las investigaciones que se realizan en este campo no son en vano y poseen un beneficio claro para las compañías. No obstante, se puede afirmar que se posee un largo camino hasta que las compañías sean capaces de procesar y utilizar el gran volumen de datos del que disponen para la detección del fraude.

En la misma línea de importancia que la detección del fraude se encuentra la prevención de este. Las compañías de seguros han desarrollado multitud de metodologías de actuación que suponen una prevención activa respecto al fraude. Estas se pueden estructurar principalmente en 4 categorías:

- Análisis del fraude. La compañía debe analizar los diferentes tipos de siniestros y las características concretas que suponen un riesgo de fraude en cada uno de los ramos. Una vez realizado este análisis se debe aplicar a través de diferentes metodologías a las reclamaciones que se reciben.

- Proceso de investigación del fraude. La formación de los empleados es considerada uno de los pilares clave para la prevención del fraude. A pesar del avance experimentado en la implantación de nuevas técnicas antifraude, el juicio experto en la investigación y prevención sigue siendo el pilar fundamental. No obstante, dada la complejidad de los procesos fraudulentos, los procesos investigación se prolongan en el tiempo, reduciendo la eficiencia en la prevención y consumiendo una gran cantidad de recursos.
- Siniestro fraudulento verificado. Una vez se identifica una posible reclamación fraudulenta, esta pasa a un proceso de análisis y verificación de datos. Tras este proceso se confirma si se trata de una relación fraudulenta y el grado de engaño que existe en dicha reclamación. A partir de este tipo de siniestros se determina si los modelos implementados funcionan de forma correcta o se deben realizar ajustes.
- Sistema de prevención en el futuro. Las compañías deben realizar estudios para determinar cuáles de los mecanismos de prevención del fraude son más eficaces. Se pretende desarrollar metodologías que no necesiten una gran cantidad de recursos y posean poco margen de error en la detención del fraude.

En líneas generales el fraude es un fenómeno versátil y dinámico, características que afectan de forma significativa en los procesos de detección y supervisión del fraude. Las metodologías adoptadas deben estar en constante evolución, permitiéndoles de adecuarse a las diferentes peculiaridades de las actividades engañosas.

1.3.- Normativa aplicable a prácticas fraudulentas en España

El principal problema vinculado a las actividades fraudulentas es que se consideran actos con un alto potencial de reportar beneficios a los defraudadores, sin necesidad de que estos corran un riesgo considerable. Por norma general, los defraudadores desconocen las consecuencias vinculadas a la realización de este tipo de actos delictivos.

La mayor parte de los fraudes detectados por las compañías no son comunicados a las autoridades policiales o judiciales, sino que dicha información se mantiene en la compañía y únicamente se comparte con otras entidades aseguradoras a través de ficheros antifraude, este hecho dificulta en muchas ocasiones la identificación de los defraudadores por parte de los entes públicos y por ende la toma de medidas sancionadoras contra el fraude.

Los organismos de regulación estatales buscan crear planes de lucha contra las actividades fraudulentas que tengan como objetivo común facilitar la prevención e investigación de las actividades desarrolladas para defraudar en los distintos ramos de seguros. En este sentido, se ha firmado un acuerdo de cooperación entre compañías aseguradoras y los organismos públicos liderado por UNESPA y la guardia civil. Este acuerdo pretende crear campañas de sensibilización contra el fraude, así como aumentar la colaboración entre las entidades y las fuerzas de seguridad del estado.

Han sido numerosas las organizaciones de la industria del seguro que han elaborado directrices y guías de actuación para detectar, prevenir y colaborar contra el fraude. Partiendo de las normas establecidas en el sector y en documentos legislativos, este tipo de actos delictivos posee tres principales consecuencias una vez han sido verificados por la compañía.

- Rescisión de la póliza. Una de las condiciones recogidas en los contratos de seguros, normalmente hace referencia a la capacidad que posee la compañía de cancelar el contrato de seguros, en caso de ser identificado un caso de fraude.
- Pérdida de la indemnización. En caso de ser detectado un caso de fraude por la entidad aseguradora y se haya verificado, esta puede negarse legalmente a abonar la indemnización al cliente referente al evento reclamado. En este sentido, los gastos en los que se hayan podido incurrir derivados del siniestro e incluso judiciales, deberán ser abonados por el defraudador.
- Imputación de delitos penales. El intento de fraude a las compañías de seguros está definido como una actuación delictiva, la cual viene detallada en documentos legislativos. Por tanto, este tipo de actividades conlleva la atribución de delitos penales de diversa índole para la persona que ha desarrollado las actividades engañosas. las consecuencias asociadas son determinadas por órganos públicos, tanto en el código penal como en la ley de contrato de seguros.

Las primeras consecuencias explicadas son recogidas en la Ley de Contratos de Seguros, donde los órganos encargados de la elaboración de esta ley han desarrollado diferentes artículos con carácter específico referente a malas prácticas en el sector de seguros. Los artículos 4, 10, y 19 de esta ley hacen referencia a la nulidad del contrato, el deber de información de ocurrencia del siniestro y la mala fe del asegurado respecto a la ocurrencia del evento asegurado. Por ello, estos son considerados los artículos que mayor vinculación con los actos fraudulentos que realiza un cliente frente a la compañía.

Las consecuencias de las prácticas fraudulentas que vienen recogidas en los artículos mencionados de la LCS principalmente están enmarcadas en la definición de los deberes y derechos que poseen las entidades, una vez se ha verificado la existencia de fraude en una reclamación. El derecho más destacado que otorgan los artículos mencionados es la exoneración del pago de la indemnización por parte del asegurador. No obstante, cabe destacar que la LCS trata el fraude desde la existencia de mala fe, ya que el dolo es considerado un concepto penal que en el derecho privado no se contempla.

En referencia a la imputación de delitos, se contemplan diferentes consecuencias legales en función de la gravedad y del número de intervinientes implicados en el fraude cometido. Según el Código Penal las sanciones a las que debe hacer frente el estafador se establecen en función de la cantidad defraudada, partiendo desde multas hasta penas de cárcel. Estas sanciones mencionadas vienen recogidas en los artículos 248 y siguientes del Código penal. En caso de haber intentado defraudar una cantidad inferior a 400€ la multa puede ser de uno a tres meses. Para aquellas personas que intentaran defraudar cantidades superiores a 400€ se determinan penas de prisión entre seis meses y cuatro años, así como una multa que oscilara entre uno y tres meses.

Para aquellos intentos de fraude que por su envergadura son considerados grandes estafas, las sanciones que se imputan se agravan es por ello por lo que vienen recogidas en el artículo 250 del Código Penal de forma más independiente. En estos casos las sanciones pueden oscilar entre uno y seis años de prisión y multas de entre seis a doce meses.

El concepto de multa es el establecido en el propio código penal, donde el juez tiene la potestad de definir la multa que debe ser abonada por el defraudador. Suele establecerse el pago diario de una cantidad durante el periodo de tiempo definido por los artículos anteriores, así mismos, esta cuantía suele estar en consonancia con la cantidad defraudada a la compañía y la gravedad de los delitos cometidos.

Otro de los artículos del código penal en este ámbito es el 457, el cual hace referencia a la simulación de siniestros. Para que el defraudador quede contemplado bajo este artículo se debe cumplir que la simulación del evento acaecido provoque la intervención policial o judicial y que se pruebe que existe una evidencia clara de dolor en los hechos acaecidos. En este tipo de delitos la sanción oscila entre seis meses y tres años de prisión y seis meses por simulación de un delito.

Cabe destacar que, uno de los mayores avances que se han realizado en materia de ley contra el fraude en el sector asegurador ha sido la reforma del Baremo de 2016, reforma del Código Penal Ley Orgánica 1/2015 de 30 de marzo, referente a las faltas en los accidentes de tráfico con lesiones leves. Como se ha venido mencionando a lo largo del estudio, el ramo de seguros de automóvil es el más afectado por este tipo de prácticas deshonestas. Esta reforma consiste en la retirada a los perjudicados de un accidente vial, de la capacidad de obtener una valoración gratuita por parte de un médico forense y la posterior celebración de los procesos judiciales a través de la vía penal. El afectado a partir de 2016 debe hacer frente a los costes del juicio y la valoración médica hasta que de forma judicial se tuviera una sentencia favorable. De esta forma se ha conseguido reducir notablemente el número de fraudes en el ramo de automóviles referentes a daños corporales.

La normativa referente a estos actos delictivos contra entidades aseguradoras está en constante evolución debido a la aparición de nuevas vías y metodologías de fraude. Las diferentes normas publicadas en el código penal español siguen directrices análogas a las normas establecidas por numerosos países de la Unión Europea, en aras de obtener un mercado homogéneo en materia de lucha contra el fraude. En este sentido, el principal objetivo de los entes públicos es adaptarse lo más rápido posible a los cambios en las actividades fraudulentas con el fin de concienciar a la sociedad de las consecuencias de este tipo de actos y de velar por los intervinientes afectados.

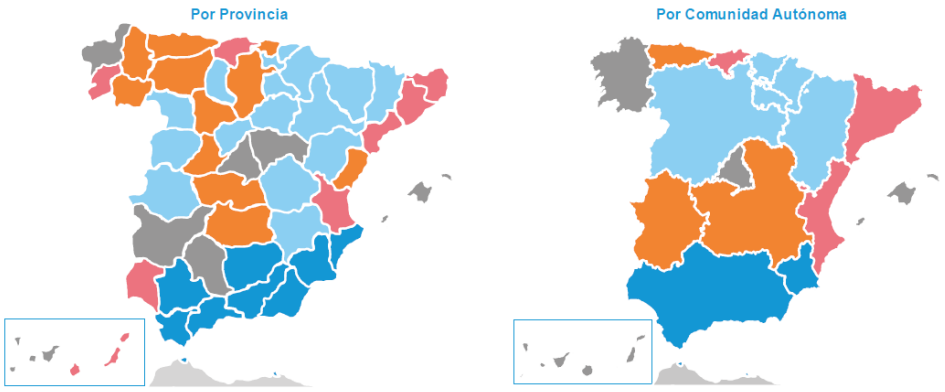
1.4.- El fraude en el contexto nacional e internacional

Una vez conocidos los aspectos importantes de los fraudes cometidos contra las entidades aseguradoras, es muy importante entender las diferentes incidencias que poseen estas actividades delictivas en función del área geográfica. La práctica de este tipo de actividades está estrechamente relacionada con la situación socioeconómica que posee cada territorio, siendo las zonas con menor desarrollo aquellas que presentan una menor aversión a la realización de este tipo de actos delictivos.

El territorio español no es ajeno a estas diferencias territoriales en materia de fraude. Según numerosos estudios, como los elaborados por AXA e ICEA en 2020, existen diferentes tasas de fraude entre las diferentes comunidades autónomas, incluso entre las provincias que componen cada comunidad. Las tasas de fraude en cada CCAA se han obtenido teniendo en

cuenta el montante asegurador de la zona geográfica, de tal manera que se adquiere un resultado transparente apto para ser comparado. La tasa de fraude media en España se encuentra entrono al 2,21% de las reclamaciones totales en 2020, siendo la diferencia entre la comunidad con mayor índice y la que menor índice de entorno a un 2,7% (AXA, 2020).

Figura 7. Número de casos de fraude por zona geográfica



Fuente: Elaboración propia a partir de datos de "El Fraude al Seguro Español Año 2019" – ICEA

(Leyenda: Azul ≥ 0.42; Rosa 0.42% - 0.35%; Gris 0.35% - 0.30%; Naranja 0.30% - 0.26; Azul claro < 0.26%)

El mapa izquierdo de la Figura 7, extraída del estudio de fraude en España, muestra un análisis territorial de las tasas de fraude entre particulares en España, en este caso clasificadas en las diferentes provincias que componen el territorio. En esta clasificación más exhaustiva se pueden identificar cuáles son las provincias que mayor tasa de fraude poseen y por ende aumentan la media de la comunidad o viceversa. En este sentido, se aprecia que las provincias con mayor tasa de fraude son las situadas al sur de la geografía española, donde Almería es la provincia española con mayor fraude alcanzando una tasa igual a 3,41%, seguida de Cádiz con un 3,39% y Málaga con un 3,20%. En el extremo opuesto, dentro del grupo de provincias que poseen menor incidencia de estas actividades delictivas se encuentran las provincias de Soria con una tasa acumulada de 0,89%, Albacete acumulando una tasa de 1,19% y Álava con un 1,22%.

El mapa derecho de la Figura 7 aporta información referente a las diferentes tasas de fraude clasificadas según la comunidad autónoma. Como se puede observar se producen determinados cambios en el ranking de incidencia de fraude, debido a que determinadas provincias con una elevada tasa son compensadas con otras que poseen una tasa inferior. En este sentido, las CCAA españolas que mayor tasa de fraude poseen son Andalucía alcanzando una tasa de fraude igual a 2,9%, seguido de Murcia con un 2,7% y Cantabria con un 2,6%. En el extremo opuesto, Aragón con una tasa de fraude acumulada entorno al 1,9%, Castilla y León con una tasa igual a 2,1%, y La Rioja con un 2,2%, son consideradas las comunidades autónomas que menor incidencia de fraude posee.

En líneas generales se puede afirmar que se ha experimentado un crecimiento del número de actividades fraudulentas detectadas en cada uno de los territorios analizados (ICEA, 2020). No obstante, cabe destacar que no todos han evolucionado de la misma manera, siendo las provincias de Almería, Ávila y Cantabria aquellas que sufrieron un mayor incremento de este tipo de actos delictivos. Por el contrario, Las provincias de Melilla, La Rioja y Guadalajara son las

provincias que mayor reducción del fraude ha efectuado. En términos de comunidades autónomas encontramos algunas diferencias, aquellas que han sufrido un aumento de la tasa de fraude en los últimos años han sido Cantabria, Andalucía y Murcia. En el extremo opuesto, encontramos las comunidades de Melilla y La Rioja siendo las que mayor reducción de fraude han experimentado.

Como se ha comentado anteriormente, el ramo de seguros de automóvil es el ámbito asegurador que mayor incidencia de prácticas delictivas acumula. En este contexto, tanto la tasa de fraude medio como la incidencia por área geográfica se ven modificadas. Como se puede observar en la Figura 8, prestando atención únicamente al ramo de seguros de automóvil, las comunidades que mayor incidencia de prácticas fraudulentas poseen son Andalucía con una tasa de fraude superior a 5,6% y Murcia. En el extremo contrario se encuentran las comunidades de las Islas Canarias con una tasa de fraude menor a 4,3%, Castilla y León y Galicia, con las menores tasas de fraude a este ramo.

Figura 8. Distribución del fraude de automóvil por provincias

Fuente: Elaboración propia a partir de datos de "V barómetro de fraude al seguro" – Línea Directa

Las prácticas delictivas son comunes en todas las entidades aseguradoras del mundo, aunque la incidencia del fraude sobre los diferentes territorios depende de forma directa del punto de vista desde el que se estudia el mismo. Por ello, se han realizado comparativas en las que se analizan los efectos del fraude en las diferentes economías y las metodologías para disuadir los tipos de fraude de cada zona geográfica, como son las elaboradas por el III (Insurance Information Institute).

Desde el punto de vista europeo el alcance de las actividades fraudulentas varía entre los diferentes países. La cifra de este tipo de actividades delictivas varía entre países dependiendo de la estructuración del mercado y de la prevalencia de un tipo concreto de seguro. No obstante, según el estudio "The impact of insurance fraud" (Insurance Europe, 2013), se estima que en líneas generales el fraude, tanto identificado como no identificado, representa el 10% de las reclamaciones recibidas en el sector.

Dependiendo del país que se desee analizar se poseen diferentes efectos sobre el sector asegurador. A continuación, se presentan los resultados observados más relevantes.

- Reino Unido. Este mercado está regido por la Asociación de Aseguradoras Británicas (ABI), la cual a partir de los estudios realizados ha estimado que pese a los esfuerzos de las diferentes entidades en detección de fraude casi 1.900 millones de libras son abonadas en reclamaciones fraudulentas.
- Alemania. La asociación de seguros alemana (GDV) es la encargada de marcar las directrices en este mercado, en el cual estimó que alrededor del 50% de la totalidad de las reclamaciones recibidas correspondían a siniestros referentes a aparatos tecnológicos, lo cual crea indicios de un elevado fraude en este ámbito. Se estima que este tipo de actos delictivos cuestan a las aseguradoras alemanas alrededor de 4.000 millones de euros al año.
- Suecia. Según estudios elaborados por el Insurance Sweden (Larmtjänst) se estima que los fraudes detectados por las compañías han evitado el pago indebido de alrededor de 40 millones de euros.
- Francia. En este mercado los estudios provienen de la asociación de seguros francesa (FFSA), la cual ha estimado que 168 millones de euros corresponden a reclamaciones fraudulentas recibidas y detectadas por las compañías.
- Finlandia. La asociación de seguros finlandesa (FFI) a partir de un estudio realizado sobre la población estimó que alrededor del 27% de los ciudadanos conocían a una persona que había realizado actos fraudulentos contra entidades de seguros.

Las respuestas del sector asegurador frente a las prácticas fraudulentas son muy variadas entre los países, cada una de las decisiones poseen una gran repercusión. En este sentido se posee un alto grado de cooperación transfronteriza, donde las diferentes asociaciones de seguros representantes de cada país se reúnen para debatir las líneas de lucha contra el fraude que se van a seguir, los desafíos a los que se enfrenta la industria y las posibles peculiaridades afines a cada uno de los mercados.

Numerosos países de la Unión Europea han creado organismos públicos específicos cuya misión es la lucha contra el fraude, suministrando a las entidades los instrumentos necesarios, elaborando planes de formación continuada para los empleados y proporcionando la información sobre las prácticas delictivas a las fuerzas y cuerpos de seguridad del estado.

A nivel mundial, según viene recogido en el informe “Global Claims Fraud Survery” (RGA, 2017), el fraude también se reparte de manera desigual entre las diferentes zonas geográficas del globo terráqueo, siendo Asia y Europa los continentes que poseen una mayor incidencia en seguros. En el extremo contrario se encuentran los continentes de Oceanía y Sur América acumulando las tasas de fraude más reducidas. En líneas generales el fraude detectado mundial se estima que supone el 10% de los beneficios de la industria.

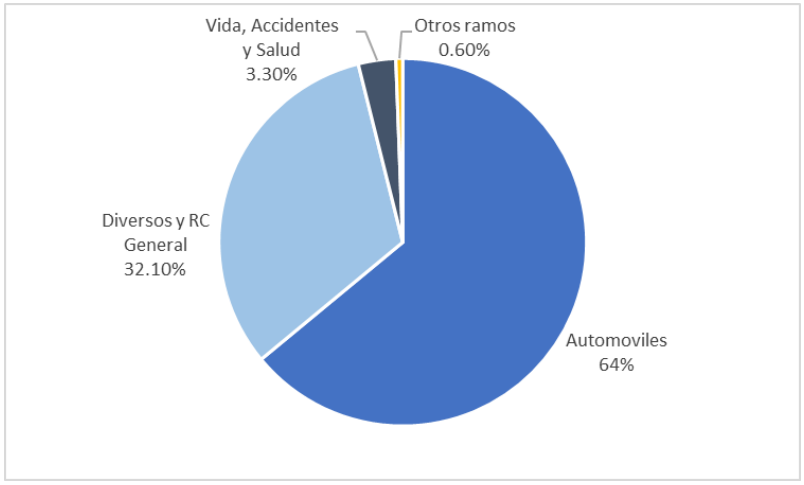
A lo largo de los años se han llevado a cabo políticas contra el fraude que pretenden lograr una lucha homogénea entre zonas geográficas, teniendo presente las peculiaridades que posee cada mercado. En líneas generales se puede identificar un aumento progresivo de la tasa de fraude a nivel mundial. Este incremento en parte viene asociado al mayor control al que están sometidas las entidades aseguradoras de forma internacional y a la adopción por parte de estas de técnicas más sofisticadas para la evaluación de este tipo de prácticas.

1.5.- Incidencia de prácticas fraudulentas según el ramo de seguro

Al igual que la incidencia de las actividades fraudulentas cambia con respecto a las diferentes zonas geográficas lo hace también con respecto a los diferentes ramos de seguros de las compañías. El fraude es un fenómeno dinámico y flexible, el cual se adapta a los ramos que mayor penetración poseen en el mercado, las características de los productos y posibles brechas legales que existen en cada una de las líneas de negocio.

A partir de la Figura 9, se puede advertir que el 64% de la totalidad de los casos de fraude detectados en 2019 en España pertenece al ramo de automóviles. "En 2012 siete de cada 10 siniestros fraudulentos se producían en este ramo" (AXA,2020). El hecho de que el ramo de seguros de automóvil sea el que mayor tasa de fraude presenta, está estrechamente vinculado a que es el que mayor presencia posee en el mercado español debido a su carácter obligatorio. Esta mayor presencia del seguro de automóviles permite a los defraudadores poseer un mayor conocimiento sobre este tipo de seguros e identificar las brechas que explotar para obtener beneficios.

Figura 9. Distribución del fraude por productos



Fuente: Elaboración propia a partir de datos de "VIII Mapa AXA del Fraude en España" – AXA

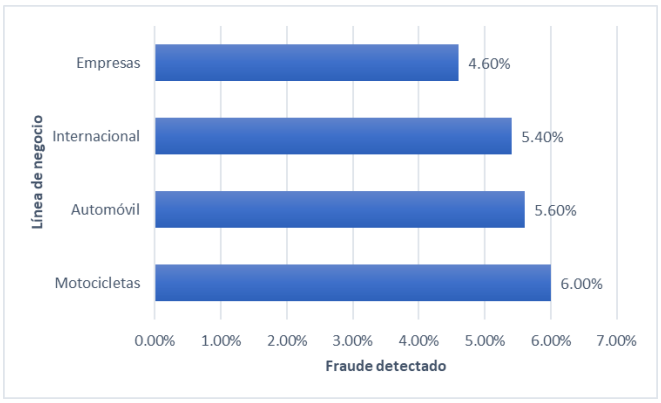
En la Figura 9, se puede identificar el ramo de seguros diversos y RC general como el segundo ramo en el que las compañías aseguradoras registraron un gran número de casos fraudulentos, acumulando el 32,1% de la totalidad de fraudes detectados. Dentro de esta categoría, cabe destacar que, más de la mitad de los casos identificados pertenecen al ramo de seguros de hogar y comunidades. Esta categoría de seguros es la más empleada por los defraudadores que actúan de forma premeditada, ya que posee una gran flexibilidad y diversidad respecto a los actos fraudulentos, pudiéndose llevar a cabo una gran variedad de prácticas con las que realizar reclamaciones fraudulentas (Daños eléctricos, daños de agua, daños climatológicos, etc.). No obstante, a pesar de suponer un número reducido de casos en el ramo de diversos, los fraudes vinculados a RC son los que mayor coste suponen para la compañía, normalmente la tipología de fraude en este tipo de línea de negocio está vinculada a la garantía de extensivos.

Dentro de los ramos con menor incidencia de fraude se encuentra el referente a vida, accidentes y salud. El menor número de reclamaciones fraudulentas en este ramo se debe a la mayor dificultad para la declaración ficticia de siniestros o sus consecuencias. Así mismo, al haber una menor carga de reclamaciones como, por ejemplo, en el caso de seguro de automóviles, los defraudadores identifican que existe una mayor probabilidad de ser identificados, por el tramitador.

Respecto a la incidencia del fraude experimentada en los diferentes ramos, se puede afirmar que no hay ninguna línea de negocio de seguros que no sufra los efectos negativos de las prácticas fraudulentas. Este tipo de prácticas en la última década ha experimentado cambios derivados del avance de las nuevas tecnologías y las metodologías antifraude. En función del entorno tecnológico y socioeconómico, se han inclinado las prácticas fraudulentas hacia unas líneas de negocio u otras.

A partir del estudio sobre fraude en el ramo de automóviles (Línea Directa, 2020) se extrae la Figura 10, donde se puede observar como en el ámbito de los seguros referentes a vehículos, se ha experimentado mayor crecimiento del número de reclamaciones fraudulentas es en el referente a seguros de motos. En este sentido, según se explica en el estudio se ha observado una reducción de la incidencia media en los seguros de coches, internacionales y de empresas.

Figura 10. Frecuencia de fraude por línea de negocio

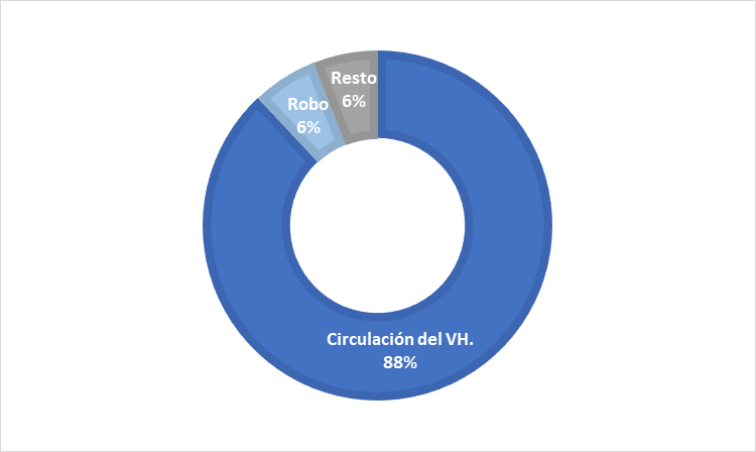


Fuente: Elaboración propia a partir de datos de "V barómetro de fraude al seguro" – Línea Directa

La reducción del número de reclamaciones fraudulentas en determinadas categorías no se traduce en una disminución del fraude en sí misma, sino que se produce una traslación del fraude hacia otros ramos (ICEA, 2020). Derivado de la pandemia del Covid-19 se ha apreciado un descenso del fraude en el ramo de autos debido a las restricciones de movilidad y se ha experimentado un aumento de los siniestros fraudulentos del ramo hogar vinculado al confinamiento domiciliario.

A partir de datos analizados (AXA, 2020), en consonancia con los anteriores estudios realizados por esta compañía, se ha establecido que el 88% de las reclamaciones fraudulentas realizadas en el ramo de automóviles se produce cuando el vehículo se encuentra en circulación. El 12% de las reclamaciones restantes se divide equitativamente entre casos de robo y reclamaciones de diversa índole. De forma visual se puede apreciar la distribución de las reclamaciones en autos en la Figura 11.

Figura 11. Distribución de las causas de fraude en Autos



Fuente: Elaboración propia a partir de datos de "El Fraude al Seguro Español Año 2019" – ICEA

Dentro de los casos expuestos en la Figura 11, el fraude oportunista es el que mayor presencia posee, estando implicado en el 77% de las reclamaciones. Respecto al resto de casos de fraude el 20% pertenecen a las tipologías de premeditado y el 3% a tramas organizadas. (AXA, 2020). Derivado de acontecimientos acaecidos durante la conducción el fraude en el seguro de automóviles es principalmente oportunista. Así mismo, estas prácticas delictivas se dan en determinadas coberturas.

Tabla 1. Distribución del fraude en autos por garantía

Ramos/Garantías	Fraude Evitado		Gasto medio investigación
	% Casos	Imp. Medio	
Automóviles			
R.C daños materiales	38.63%	799 €	12.50 €
R.C daños corporales	30.19%	4,810.40 €	139.20 €
Daños propios	12.25%	1,381.60 €	16.60 €
Robo	3.00%	1,898.80 €	41.50 €
Incendios	0.24%	4,702.40 €	106.70 €
Accidentes Personales	2.71%	1,252.80 €	17.80 €
Lunas	0.67%	573.90 €	7.60 €
Otras	12.31%	955.20 €	16.30 €
Total automóviles	**64%**	**2,154.00 €**	**52.90 €**

Fuente: Elaboración propia a partir de datos de "El Fraude al Seguro Español Año 2019" – ICEA

Como se puede extraer a partir de la Tabla 1, las principales coberturas objeto de fraude en las reclamaciones del seguro de automóvil son las referentes a la responsabilidad civil de daños, tanto material como corporal, como los daños propios. En este sentido, la cobertura de RC de daños corporales es la cobertura que mayor coste supone a las compañías, siendo el importe medio 4.810,4€, cifra que triplica el coste referente a la cobertura de RC de daños materiales.

En relación con los casos de fraude detectados, se estima que a partir de las técnicas de detección del fraude implementadas por las compañías se ha evitado el pago de en torno a 290 millones de euros provenientes de reclamaciones ilícitas. Este hecho posee un impacto positivo no sólo en el balance de las compañías sino también en la prima que deben afrontar los asegurados. Se estima que los pagos evitados derivados de fraude detectado se han convertido en un ahorro para los clientes de entre el 10% y el 15% en las garantías de RC y accidentes (INESE, 2019). Este hecho es uno de los principales beneficios explícitos de la aplicación de técnicas activas de detección y prevención del fraude en el sector asegurador.

A partir de la Tabla 2, se puede observar cual ha sido el impacto del fraude en cada una de las líneas de negocio. Los ramos de responsabilidad civil, transportes e industrias son los que mayor ahorro poseen en las primas derivada de prácticas fraudulentas.

Tabla 2. Repercusión del fraude en la prima del seguro

Auto	4%
Hogar	6%
Industrias	8%
Comercio	10%
Comunidades	8%
Responsabilidad Civil	32%
Accidentes	7%
Oficinas	6%
Transportes/Embarcaciones	19%
Incendios	7%
Tecnicos (C/Maq.)	5%

Fuente: Elaboración propia a partir de datos de "VIII Mapa AXA del Fraude en España" – AXA

La línea de negocio referente a automóviles posee un ahorro para los asegurados cercana al 4%, lo cual es reducido en comparación con otros ramos. No obstante, no se debe olvidar que es uno de los ramos en los que más ha aumentado el coste medio del fraude. El coste del fraude en el ramo de seguros de automóvil se ha incrementado un 121% en la última década pasando de 586€ en 2009-2010 a 1.296€ en 2017-2018 (Línea directa, 2020).

El aumento progresivo del coste medio ha llevado a las compañías aseguradoras a intensificar las metodologías de peritación y a incluir metodologías avanzadas, con el fin de conseguir un mayor control y precisión sobre los elementos mecánicos de los vehículos y esclarecer con mayor eficacia los siniestros reclamados por los clientes.

Aunque cada vez está ganando mayor importancia en otros ramos, la aplicación de métodos cuantitativos estadísticos y econométricos al tratamiento del fraude se ha centrado, principalmente, en el seguro de automóvil. A lo largo del trabajo se realizará un estudio sobre las diferentes metodologías y técnicas desarrolladas para el análisis de los elementos vinculados a comportamientos fraudulentos.

2. METODOLOGÍA MACHINE LEARNING EN LA DETECCIÓN DEL FRAUDE

Una vez se han comprendido las características y efectos negativos del fraude en el sector asegurador y en concreto en el ramo de seguros de automóvil. Las entidades aseguradoras han tomado cada vez mayor consciencia sobre este fenómeno, aplicando diferentes técnicas estadísticas que lo estudian y donde *"La validación estadística de los denominados indicadores de fraude es, sin duda, una pieza clave a la hora de dirigir de forma adecuada la investigación de los accidentes."* según (Guillén, et al., 1999). Las metodologías empleadas por las compañías deben ser robustas y homogéneas, con el fin de implementar medidas de detección y prevención del fraude eficaces.

La metodología Machine Learning se estructura en 3 grandes categorías: aprendizaje supervisado, aprendizaje no supervisado y, las técnicas más recientes, aprendizaje reforzado. Alpaydin (2009) fue uno de los primeros autores en recoger en un estudio las diferentes técnicas de Machine Learning, vinculando a cada metodología los problemas y soluciones que presentaba.

- En los métodos de aprendizaje supervisado la solución del problema se extrae a partir del análisis de datos del pasado, tratando de reproducir o predecir la respuesta.
- Los métodos de aprendizaje no supervisado tienen como objetivo la agrupación de la muestra en grupos homogéneos, de tal manera que se pueda identificar y aplicar la solución óptima para cada agrupación generada.
- Los modelos de aprendizaje reforzado se emplean para desarrollar acciones muy específicas. Esta metodología analiza la experiencia pasada y pretende adquirir el mayor conocimiento posible.

De forma tradicional, las técnicas aplicadas por las compañías giran en torno a modelos econométricos y estadísticos, es decir, de aprendizaje supervisado. Los cuales a partir de multitud de investigaciones y estudios a lo largo de los años se han ido perfeccionando y afinando. Este tipo de modelos son los más empleados por las entidades aseguradoras, dado que se posee una mayor literatura sobre su desarrollo pudiéndose verificar de forma más sencilla el buen funcionamiento del modelo respecto al problema que se quiere solventar.

La evolución tecnológica favorece el desarrollo y la utilización de técnicas, algoritmos y recursos estadísticos que facilitan la detección y prevención de comportamientos fraudulentos. El auge de las nuevas tecnologías de la información ha permitido el desarrollo de diversas técnicas cuantitativas, en especial, dirigidas al estudio de los elementos vinculados a la aparición de comportamientos fraudulentos. Los estudios realizados sostienen la existencia de determinadas variables que aumentan o disminuyen las posibilidades de que se materialicen actos fraudulentos.

El desarrollo tecnológico y el uso masivo de datos ha impulsado la implantación y uso generalizado de nuevos softwares de análisis, dando lugar a una mayor sofisticación y eficiencia en el tratamiento de los datos. No sólo se optimiza el proceso de cara a la detección del fraude, sino también la prevención de forma rápida y eficaz, evitando que personas que tengan buenos comportamientos se vean perjudicadas por acciones fraudulentas de otros individuos.

Múltiples investigadores, como Cummings y Tennyson (1996), Picard (2000) y Crocker y Tennyson (2002), han realizado análisis respecto al comportamiento fraudulento de los asegurados y su modelización. Gracias a ello se posee un variado repertorio de publicaciones que emplean algoritmos de aprendizaje para la cuantificación y predicción del fraude (Badal, et al., 2020). En este trabajo se estudiará la probabilidad de fraude en el seguro de automóviles desde diferentes ópticas, con el fin de verificar el grado de eficiencia que aporta cada una de las metodologías.

2.1.- Técnicas clásicas de detección de fraude

Las técnicas clásicas de predicción que a continuación se presentan, suponen un instrumento de gran importancia en las líneas de negocio de no vida. Este tipo de metodologías son integradas en multitud de actividades dentro de cada ramo como, por ejemplo, la elaboración de la tarifa, el cálculo de caída de cartera, el cálculo de la siniestralidad de la cartera y en los últimos años, el cálculo de la probabilidad de fraude.

En primer lugar, se deben conocer las metodologías clásicas que emplean las entidades aseguradoras en la detección del fraude. En la actualidad, este tipo de técnicas estadísticas suponen el eje central de las acciones llevadas a cabo para identificar, detectar y prevenir el fraude. Los estudios desarrollados por el I.F.B (Insurance Fraud Boureau) están basados en técnicas estadísticas clásicas, como el modelo de regresión lineal múltiple (Guillén, et al., 1999). Esta institución es uno de los entes más relevantes en cuanto a lucha contra el fraude en Estados Unidos. A través de técnicas cuantitativas, esta entidad pretende determinar las variables sobre las que se deben centrar las entidades aseguradoras a la hora de investigar los fenómenos fraudulentos.

Las compañías de seguros emplean este tipo de metodologías con el fin de llegar a conclusiones que les aporten información extra sobre el problema. Según viene recogido en el informe realizado por Elena Badal (Badal, et al, 2020) "*Las técnicas clásicas basadas en la estadística tradicional, particularmente la regresión logística, han demostrado su eficacia en la detección del fraude a lo largo de los años*".

La variable dependiente de los modelos recoge la probabilidad de sufrir una futura reclamación fraudulenta a partir del análisis de los siniestros fraudulentos recibidos. Es decir, este tipo de técnicas permite a las compañías realizar una predicción a partir de un conjunto de datos observado. La compañía tras la implementación de modelos estadísticos tiene la capacidad de identificar la vinculación entre las variables independientes y predecir la variable respuesta objeto de estudio sobre nuevas bases de datos.

En este sentido, se poseen dos principales técnicas estadísticas y econométricas que conforman gran parte de los estudios que se han desarrollado en este ámbito. De forma mayoritaria, las compañías de seguros emplean los llamados modelos lineales generalizados (en adelante, GLM) para el estudio de la probabilidad de fraude en las carteras de seguros de automóvil.

2.1.1.- Modelos lineales Generalizados (GLM)

Los modelos lineales generalizados fueron introducidos por John Nelder (2002), en ellos se observa la existencia de una relación lineal entre la función de las variables explicativas y la media correspondiente a variable respuesta del modelo. Estos modelos derivan de la estructura de los modelos lineales estándar, es por ello por lo que se basan en una ecuación muy similar a la presentada en dichos modelos, tomando la expresión 2.1:

$$g(E[Y]) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k; \qquad (2.1)$$

El funcionamiento de los GLM consiste en vincular el valor esperado de la variable respuesta E[Y], con las variables explicativas del modelo β_k, a través de una función de enlace g(). Por tanto, se puede expresar también como la siguiente ecuación 2.2:

$$g(\mu) = \eta = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k = \beta X; \qquad (2.2)$$

Los modelos lineales generalizados poseen las siguientes características particulares:

- La relación entre la variable dependiente del modelo y las variables explicativas no tiene por qué ser lineal.
- Este tipo de modelos se puede aplicar sobre variables respuesta que no siguen una distribución normal.
- Se elimina la condición necesaria de normalidad sobre los residuos, donde deben poseer una varianza constante.
- En este tipo de modelos, la relación lineal existe entre la transformación del valor esperado de la variable dependiente y las variables explicativas.

Además de unas determinadas propiedades, los GLM se caracterizan por poseer 3 componentes expresados en las funciones 2.3:

$$y = X\beta + \varepsilon;\ donde\ \varepsilon \sim N(0, \sigma^2 I)\ E(y) = \mu = X\beta \qquad (2.3)$$

Donde Xβ hace referencia al predictor lineal.

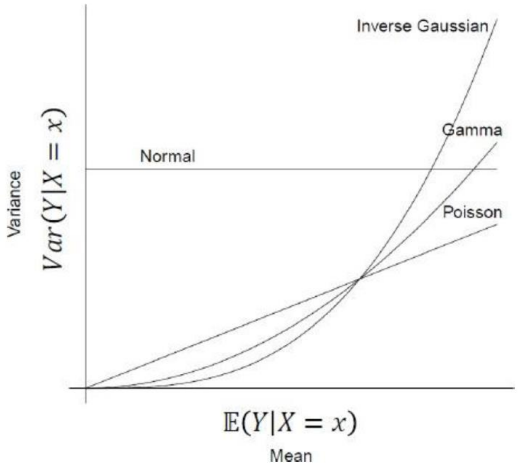
- Componente aleatorio: Hace referencia a la variable dependiente Y, comprende por tanto su distribución y probabilidad. Una de las hipótesis básicas es que esta variable respuesta pertenezca a la familia exponencial, siendo su función de densidad expresada a partir de la siguiente fórmula 2.4:

$$fi(y_i; \theta; \phi) = \exp\left\{\left[\frac{y_i\theta_i - \mathrm{b}(\theta_i)}{a_i(\phi)}\right] + c\,(y_i, \phi)\right\} \qquad (2.4)$$

Dependiendo de la finalidad de la variable respuesta (Y), se deberá utilizar una distribución para modelizar. En los casos que se posee una variable dicotómica se debe emplear una distribución binomial y para casos de conteo se puede emplear distribuciones de Poisson o binomial negativa. No obstante, todas estas metodologías siempre deben pertenecer a la familia exponencial.

La familia exponencial comprende multitud de distribuciones: Normal, Gamma, Binomial, Poisson, Binomial Negativa, etc. En la Figura 12 se pretende realizar un análisis visual de las principales distribuciones de esta familia, haciendo hincapié en la diferente variabilidad de las distribuciones.

Figura 12. Representación de las distribuciones pertenecientes a la familia exponencial



Fuente: "Generalized linear Models" – P. McCullagh

- Componente Sistemático: Comprende los factores predictivos del modelo, es decir, hace referencia a las variables explicativas definidas en la función. La combinación de las diferentes variables explicativas se denomina predictor lineal, reflejado en la expresión 2.5:

$$\alpha + \beta_1 X_1 + \ldots + \beta_k X_i \qquad (2.5)$$

- Función de enlace (Link function): Esta función lo que pretende es enlazar la media de la variable respuesta E[Y], con los factores predictivos del modelo. En un modelo de regresión lineal habitual donde μ=η, la función de enlace corresponde a la identidad. Cabe destacar, que la función de enlace transforma la media de la variable respuesta en un parámetro siguiendo la siguiente expresión 2.6:

$$g(E[y]) = \theta \rightarrow g \qquad (2.6)$$

En la Tabla 3 se definen las funciones de enlace determinadas para cada una de las distribuciones de la variable respuesta, la cuales cumplen la condición de pertenecer a la familia de distribución exponencial. Así mismo, se plantea la fórmula de cálculo determinadas para cada una de las funciones de liga.

Tabla 3. Link función según la distribución de los datos

Distribución	Link Function
Normal	Log
Gamma	Inversa
Poisson	Log
Binomial	Logit
Binomial Negativa	Log
Geométrica	Log
Inversa Gaussiana	Inversa cuadrada
Tweedie	Log

Fuente: Elaboración propia

Para el presente estudio se emplea la Distribución Binomial, ya que se pretende generar la modelización de la variable vinculada al fraude, la cual posee un indicador binario de valores "si" y "no", fraude o no fraude. La función enlace que mayor capacidad predictiva posee sobre este tipo de datos de estudio es la denominada función Link "logit". Por tanto, partiendo de la distribución binomial, se empleará un modelo conocido como modelo de regresión logística, cuya función de enlace viene definida por la siguiente ecuación 2.7:

$$g(\pi_i) = log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}; \qquad (2.7)$$

Para el correcto funcionamiento del modelo se debe realizar un análisis exhaustivo de las variables de la base de datos, con el objetivo de identificar las variables más significativas para el estudio, es decir, aquellas que más información aportan al modelo. Bajo esta selección se busca generar un GLM lo más ajustado posible a los datos. En el presente proyecto se seleccionarán de forma minuciosa las variables que se emplearán, teniendo como objetivo la reducción del ruido del modelo induciendo al mismo a ser lo más preciso posible.

Dentro de la categoría GLM, es frecuente que, a partir del conjunto de datos empleado, diferentes modelos sean eficaces y tengan una gran capacidad predictiva. En este sentido, la persona responsable debe ser capaz de desarrollar evaluaciones de juicio experto que le proporcione información sobre cuál de todos los modelos es el óptimo.

Al final del presente capítulo, se definirán de forma extensa las metodologías bajo las que se realiza la selección del modelo GLM óptimo. Principalmente se emplean los estadísticos Akaike's Information Criteria (AIC), Bayesian information criterio (BIC) y Derivance (D). Bajo estos criterios se definirá el número óptimo de variables explicativas que se deben introducir en el modelo, a partir del cual se obtiene un nivel de error reducido que permite al modelo ser preciso en la realización de predicciones. Estos criterios de verificación permiten conocer si se han obtenido buenos modelos predictivos y cuál de estos es el que mejor resultado genera.

En las técnicas y modelos desarrollados por las compañías para el análisis de las actividades fraudulentas es común que se elija el modelo según el juicio experto de los investigadores. Ya que pueden existir características particulares del fraude, que pueden hacer que determinados modelos a pesar de no ser los óptimos estadísticamente hablando proporcionen la mejor solución respecto a este fenómeno.

Es de gran importancia señalar que los modelos seleccionados por las compañías deben seguir el principio de parsimonia. Este principio defiende la simplicidad de los modelos, es decir, que el modelo posea el menor número de variables explicativas y se obtengan unos residuos lo más reducidos posible. Para obtener una precisión lo más ajustada posible se analizará la probabilidad de corte sobre los resultados del GLM, de tal manera que se identifique la probabilidad bajo la cual se realiza la mejor clasificación de estos.

En particular, los modelos GLM son modelos avanzados que presentan una forma simple y robusta, y no son modelos muy complicados de desarrollar. Este tipo de modelos predictivos son de gran importancia en las compañías aseguradoras, ya que permite identificar el riesgo que se puede asumir en el futuro, evaluar la iteración entre los diferentes factores de riesgo y analizar la importancia de las variables explicativas que mayor influencia tienen en el acaecimiento de los actos fraudulentos.

La mayor ventaja que poseen los GLM es la interpretación de los resultados generados. Este tipo de técnicas a través de una serie de cálculos devuelve un valor, el cual debe ser interpretado y analizado por los investigadores, por lo que únicamente se necesita un gran conocimiento del evento que se estudia, la posible tendencia que puede poseer y el diferente contexto en el que se implementa. El empleo de este tipo de metodología pretende dotar a la compañía de plataformas de control, a través de las cuales identificar la probabilidad de fraude de una reclamación y hacérselo saber al tramitador correspondiente para que de esta manera se preste una mayor atención al análisis de determinados siniestros.

A pesar de la robustez de esta metodología y su gran implementación en las compañías, la gran generación y explotación de datos que permiten las nuevas tecnologías está generando un crecimiento de las técnicas basadas en inteligencia artificial. Este hecho es debido a la capacidad de estas nuevas técnicas de crear diferentes algoritmos, adaptándolos en función de cambios identificados en los patrones que siguen los datos.

2.2.- Técnicas Modernas de detección de fraude

Las técnicas estadísticas presentadas hasta ahora a pesar de ser las más utilizados en el mundo asegurador poseen un problema de ajuste, ya que pierden eficacia cuando existen diversos predictores en el modelo y estos poseen algún tipo de vinculación. Es por ello por lo que se hace necesaria la implementación de nuevas metodologías que se adapten a los problemas planteados. En el presente epígrafe se pretenden presentar las metodologías a partir de las cuales se pueden desarrollar algoritmos que clasifiquen de forma recurrente actividades sospechosas que permitan una mayor eficacia en la detección del fraude.

Las técnicas modernas están basadas en algoritmos de aprendizaje que no solo buscan obtener un análisis de los datos y explicar los acontecimientos ocurridos hasta el momento, sino que poseen una gran capacidad predicción con la que se pretende obtener una amplia visión de los sucesos que puede ocurrir en el futuro con una probabilidad determinada.

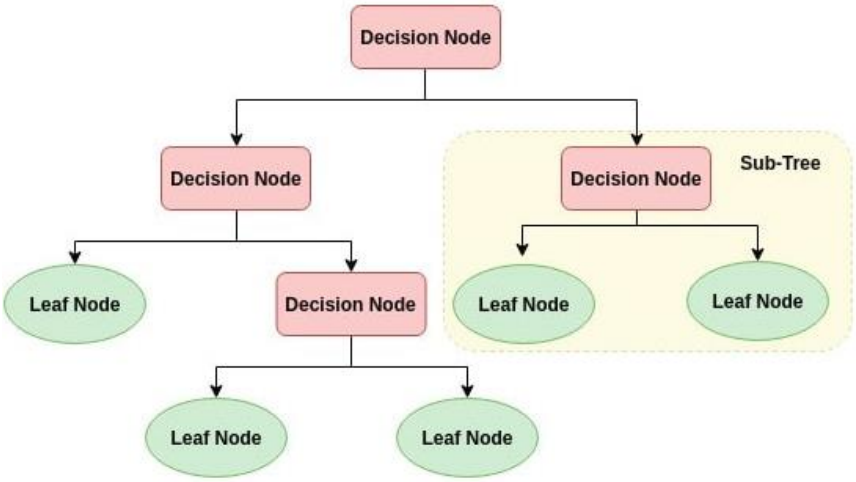
Se han desarrollado multitud de metodologías modernas de Machine Learning. No obstante, a continuación, se presentan las metodologías que mayor peso están teniendo en la investigación, detección y prevención del fraude en los seguros. Estas metodologías son ordenadas de menor a mayor complejidad interpretativa, los Árboles de Decisión, Random Forest, Gradient Boosting y Extreme Gradient Boosting.

2.2.1.- Árboles de decisión

Los árboles de decisión es una de las metodologías más implementadas de Machine Learning, la cual a partir de la segmentación de las observaciones en espacios reducidos genera una solución gráfica que considera todas las posibles soluciones al problema. El primer nodo que compone la estructura del árbol de decisión se denomina raíz y a partir de este se plantea el algoritmo.

El algoritmo se basa en condiciones sistematizadas en una estructura jerárquica que va creando las convenientes ramas, hasta llegar a la solución final del problema. El árbol de decisión se va construyendo mediante reglas de decisión, es por ello por lo que posee una interpretación muy visual. (Badal, et al., 2020). A partir de la Figura 13 se puede observar un ejemplo de la estructuración que puede poseer un árbol de decisión.

Figura 13. Estructura árbol de decisión



Fuente: Árbol de decisión en Machine Learning – Sitiobigdata

La metodología de árboles de decisión se desarrolla en dos etapas. En primer lugar, se analizan los predictores y se dividen en grupos de tal manera que no se sobrepongan entre sí, a estos grupos se les denomina nodos terminales: $R_1, R_2, R_3, \ldots, R_j$. En segundo lugar, el algoritmo predice la variable respuesta del modelo empleando los predictores ubicados en cada uno de los nodos terminales.

A partir de la aplicación del algoritmo sobre cada nodo, se identifica el conjunto de acciones óptimas para la satisfacción de una condición concreta. Los árboles de decisión tienen capacidad para resolver problemas en los que se debe estimar el valor de una variable, como aquellos en los que únicamente se necesita categorizar una variable para identificar el grupo al que pertenece, este último es el caso de este trabajo donde se pretende clasificar los casos entre fraudulentos y no fraudulentos.

Esta metodología presenta una serie de ventajas e inconvenientes que se deben tener en cuenta, la cuales se han definido en la Tabla 4.

Tabla 4. Ventajas e inconvenientes de los árboles de decisión

Ventajas	Inconvenientes
Los algoritmos basados en un único árbol de decisión son fáciles de interpretar y representar.	La interpretación de los árboles se ve dificultada en aquellos algoritmos que albergan más de un árbol de decisión.
Los algoritmos admiten predictores tanto cualitativos como cuantitativos.	Los algoritmos basados en un único árbol de decisión presentan una capacidad predictiva más baja que otros modelos.

Generalmente no necesitan una limpieza exhaustiva de la base de datos al no estar alterados por "outliers".	Al categorizar variables continuas el modelo puede incurrir en perdida de información.

Fuente: Elaboración propia

La metodología para realizar esta separación de los predictores puede variar de complejidad según la finalidad de la variable respuesta. En este contexto, los árboles de decisión pueden ser de dos tipos:

- Árboles de regresión: Poseen una variable respuesta continua.
- Árboles de clasificación: Poseen una variable respuesta discreta. Este es el caso que atañe al presente proyecto.

Los árboles de decisión de clasificación emplean la metodología CART como metodología de cálculo para elaborar árboles de clasificación binarios, la cual fue desarrollada en los años 80 por Breiman, Freidman, Olshen y Stone. Esta es una de las metodologías más empleadas en los últimos años, dada su fácil interpretación e implementación en los lenguajes de programación.

La metodología CART emplea datos históricos para generar modelos que clasifiquen o predigan nuevas observaciones. La diferencia entre el modelo CART de clasificación y de regresión reside en el criterio o factor de explicativo de división de los nodos que posee el árbol. La creación del árbol se realiza a través de la división del espacio muestral y de sucesivas divisiones o evaluaciones, cada nodo representa una determinada variable de entrada y un punto de división. Los diferentes puntos de división marcan las relaciones entre las variables. Cada uno de los datos va siendo comprobado por las subdivisiones hasta situarse en un nodo terminal específico, donde supone una variable o probabilidad de salida que se emplea para realizar la predicción.

Este algoritmo manipula con facilidad variables numéricas o categóricas, por ellos es especialmente eficaz en la predicción de variables dicotómicas cuyos valores se encuentran en el intervalo [0,1]. Así mismo, posee una gran robustez en el tratamiento de outliers. Respecto al caso estudiado en el presente trabajo, este es el algoritmo identifica situaciones discriminantes alejadas de la linealidad, como es que el asegurado haya defraudado o no.

El objetivo principal de la división es determinar el número "*j*" de nodos terminales que minimizan el error del modelo. Para el cálculo del error se emplea la siguiente expresión 2.8:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2 \qquad (2.8)$$

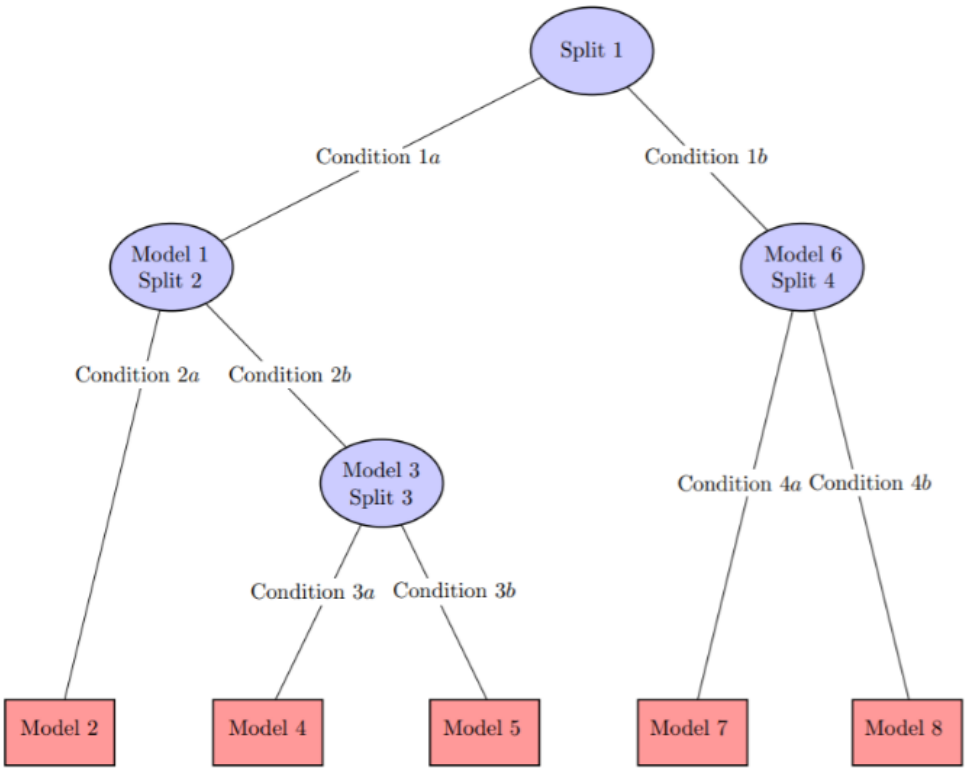
Siendo $\hat{y}_{R_j}$ la media de la variable respuesta R_j.

Una vez construido el árbol, se emplean diferentes técnicas de regresión sobre los nodos finales. Las regresiones que se llevan a cabo se fundamentarán en los predictores de cada uno de los nodos terminales que se encuentran en las ramas intermedias del árbol.

La construcción del árbol de decisión se realiza mediante la técnica denominada recursive binary splitting (División binaria recursiva). Estableciéndose la construcción en las siguientes etapas:

- El algoritmo comienza a desarrollarse a partir de la raíz o nodo inicial, donde se encuentran todas las observaciones en una misma agrupación.
- Se toman los valores de corte para cada uno de los predictores, tanto cualitativos como cuantitativos.
- Se calcula el nivel de error de cada una de las subdivisiones a partir de la fórmula 2.8. Una vez calculado el RSS, se agregan todos los resultados.
- Se selecciona el predictor que genera un menor RSS dando lugar a un nodo terminal.
- Se repiten las etapas anteriores hasta que únicamente quede un predictor, el cual será el que reporte el modelo final.

Figura 14. Esquema nodos árbol de decisión.



Fuente: "Rules Rules Rules! Cubist Regression Models" – HUHN M.

La estructura de los modelos finales se determina en función de las condiciones que se han ido satisfaciendo en los nodos intermedios, como se puede observar en la Figura 14. Cada split divide un nodo en dos subnodos, este tipo de división es denominada “binary split”. La metodología CART evalúa todos los posibles splits para las variables de entrada y determina el número óptimo de divisiones.

En este sentido, cabe destacar el proceso de “podado” de los árboles de decisión. Bajo esta estrategia se busca eliminar aquellos nodos del árbol de decisión que presentan una menor robustez. Cuando un árbol de decisión presenta una gran cantidad de nodos, este será un buen predictor de las observaciones estudiadas. Sin embargo, a la hora de realizar predicciones sobre una nueva base de datos, al poseer tanta especificidad no realizará un ajuste efectivo sobre estas nuevas observaciones, esto se conoce como “overfitting”. Bajo la técnica de podado se busca minimizar la siguiente ecuación 2.9:

$$\sum_{j=1}^{/T/} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2 + \alpha /T/ \qquad (2.9)$$

Siendo T el número de nodos terminales del árbol.

En definitiva, se trata de añadir al concepto RSS una restricción que penaliza el mayor número de nodos finales en el árbol de decisión. Este tipo de metodologías también está regido por el principio de parsimonia.

2.2.2.- Random Forest

La metodología de bosques aleatorios o Random Forest supone el progreso de la técnica de árboles de decisión, fue creada por Tim Kam Ho (1995). Es de gran importancia el reto al que está sometida la metodología anterior y que da pie a la generación de nuevas técnicas. Este desafío consiste en equilibrar dos tipos de errores del modelo:

- Error de "Bias": Hace referencia al error que posee el modelo respecto a los datos empleados en la modelización. Se calcula como la diferencia media entre las predicciones realizadas por el modelo y las observaciones que se poseen.
- Error de "Varianza": Hace referencia al error de predicción en el que incurre el modelo al emplear nuevos datos en la modelización.

En este sentido, los árboles que poseen pocas ramificaciones o nodos suelen presentar un alto error de "Bias" y un reducido error de "varianza". Y viceversa, aquellos árboles que poseen un mayor número de ramificaciones poseen un reducido error de "Bias" y un elevado error de "varianza". Para conseguir un punto intermedio entre los diferentes tipos de error, surgen las denominadas estrategias de ensemble, de las cuales las más relevantes son:

- Bagging: Se ajustan diferentes árboles o modelos formando un bosque. Esta estrategia consiste en la creación de un gran número de árboles con muchas ramificaciones por lo que tiene un reducido "bias" y una elevada varianza. A partir de la generación de más árboles, se pretende reducir la "varianza" y mantener el error "bias".
- Boosting: Se ajustan arboles de forma secuencial, de tal manera que el algoritmo va a aprendiendo de forma sucesiva. Esta estrategia se basa en árboles con poca ramificación, por lo que poseen un alto grado de error "bias" y una reducida "varianza". Este algoritmo lo que busca es la reducción del nivel de "bias" a partir de la generación de árboles o modelos muy semejantes.

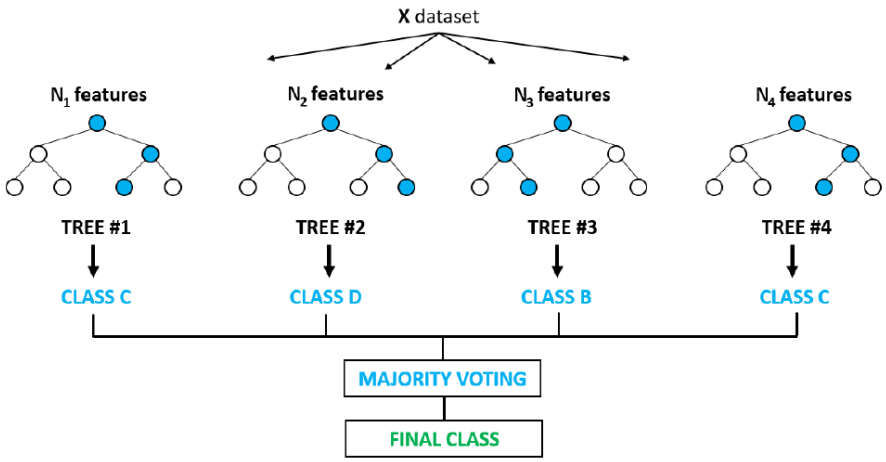
Referente a la metodología Random Forest, emplea la estrategia definida como bagging. De manera más ampliada, este algoritmo consiste en el desarrollo de forma repetida de árboles de decisión, con el objetivo de valorar los diferentes modelos de forma agregada y obtener modelos con reducida "varianza". De forma matemática se puede expresar a partir de la siguiente ecuación 2.10:

$$V(\bar{Y}) = \frac{\sigma^2}{n} \qquad (2.10)$$

Las diferentes variables respuesta poseen la misma varianza σ^2, la cual a partir de la generación de diferentes árboles se consigue reducir aumentando la precisión del modelo.

Los bosques aleatorios suponen a grandes rasgos metaclasifcadores. Como se ha mencionado anteriormente, la estrategia bagging consiste en generar diferentes árboles o modelos a partir de una división de las observaciones. Se desarrollan de forma reiterada un gran número de árboles de decisión, los cuales poseen una clasificación o resultado determinado a partir de partes diferentes de la muestra.

Figura 15. Estructura Random Forest



Fuente: Técnica Random Forest – RPubs

Partiendo de la Figura 15, cada modelo reporta un predictor diferente el cual está vinculado a la clase muestral más predominante bajo la que se construye el árbol. La predicción del modelo final se calculará como la media ponderada de los nodos finales de cada uno de los árboles. La elección aleatoria de variables hace que este tipo de modelos sea el idóneo cuando existe una base de datos con un gran número de variables.

La mayor dificultad con la que se encuentra este modelo es la creación reiterada de diferentes agrupaciones de observaciones respecto a la muestra original. Estas pseudomuestras se generan a través de la técnica de bootstrapping, la cual emplea entorno al 70% de la muestra original para generar los nuevos grupos de observaciones. A la parte restante de la muestra se le conoce como “out-of-lag”.

Este tipo de metodología posee una serie de ventajas en inconvenientes, los cuales se expresan en la Tabla 5:

Tabla 5. Ventajas e inconvenientes de la metodología Bagging

Ventajas	Inconvenientes
Mejora significativamente la capacidad predictiva de los árboles.	La interpretación y representación de los árboles se ve dificultada.
Al existir numerosos árboles se puede calcular la influencia de los predictores en cada segmento de la muestra.	La identificación de los predictores no es tan intuitiva. Se necesita un conocimiento estadístico básico.

A partir del out-of-bag de cada pseudomuestra se puede calcular la influencia de los predictores en el error del modelo.	No se trabaja de forma eficaz sobre conjuntos de datos que poseen una gran heterogeneidad.

Fuente: Elaboración propia

La técnica Random Forest posee una peculiaridad respecto a la estrategia bagging. Esta particularidad consiste en la independencia de los árboles que se han ido creando de forma sucesiva, al no estar correlacionados se puede lograr una mayor reducción del error de "varianza". En este sentido, en caso de que existiera un predictor muy relevante en las observaciones iniciales, si no se realiza una selección correcta de los predictores, los diferentes arboles creados de forma sucesiva poseerían resultados muy similares y además presentarían una fuerte correlación entre sí.

El método Random Forest, antes de generar las muestras pseudoaleatorias, realiza una selección aleatoria de "m" predictores para cada uno de los modelos. A partir de esta acción se consigue que otros predictores sean seleccionados para los modelos y que por ende los diferentes árboles no estén correlacionados, consiguiendo una mayor disminución de la varianza y una mayor eficiencia predictiva del modelo.

A partir de la expresión 2.11, se establece cual es el número óptimo de predictores que se debe seleccionar para cada uno de los modelos.

$$m = \sqrt{p} \qquad (2.11)$$

Siendo p el número de predictores total.

Una de las metodologías para saber si se ha seleccionado bien el número óptimo de predictores es comparar el valor de la variable "m" con el valor del "out-of-bag" error para el valor concreto de "*m*". Normalmente los valores óptimos son los situados por debajo del valor calculado "*m*".

2.2.3.- Gradient Boosting Machine (GBM)

Tanto la metodología Gradient Boosting como Extreme Gradient Boosting, desarrolladas por por Jerome Friedman y Tianqi Chen, son las técnicas más relevantes que emplean la estrategia de ensemble boosting, presentada anteriormente y cuyo objetivo es conseguir un equilibrio "bias – varianza". La estrategia boosting consiste el aprendizaje del algoritmo a través de la creación de forma reiterada de modelos

El algoritmo Gradient Boosting (en adelante, GBM) genera un primer árbol sencillo, a partir del cual extrae información para la elaboración del siguiente árbol y así sucesivamente. Esta metodología permite al algoritmo adaptarse y mejorar su capacidad predictiva, ya que los modelos creados parten de un análisis de los aciertos y los errores de modelos anteriores. La velocidad de aprendizaje del algoritmo se denomina "Learning rate", de forma estándar se propone un ritmo de aprendizaje que se encuentre entre los valores 0.01 y 0.001.

En este tipo de metodología, al revés que la estrategia de bagging vista hasta ahora, busca la creación de árboles sencillos con muy pocas ramificaciones, de manera que tenga poco error de varianza. Así mismo, uno de los objetivos es generar árboles que posean una gran correlación entre sí.

Dentro de la estrategia boosting, una de las principales metodologías de cálculo es Adaboost, la cual es considerada como un algoritmo clasificador. Es considerado así ya que el objetivo de este algoritmo busca la clasificación de las observaciones en dos únicos grupos. Esta metodología a través de un conjunto de árboles creados mediante la estrategia Boosting, genera un nuevo clasificador, obteniendo como resultado un modelo clasificador más eficiente.

Para implementar el algoritmo AdaBoost se deben establecer las siguientes hipótesis:

- Se debe definir un modelo sencillo a partir del cual se realizarán el resto de los cálculos, este modelo se denomina base learner. Se pretende realizar una predicción de la variable respuesta que sea un poco más eficiente que si se realizará de forma aleatoria.
- La codificación de la variable respuesta se realiza de tal manera que solo se admiten los resultados +1 y -1.
- Sobre el total del conjunto de observaciones se debe fija un peso único. La asignación de este peso seguirá la siguiente expresión 2.12:

$$w_i = \frac{1}{N}, \qquad donde\ i = 1, 2, \ldots, N \tag{2.12}$$

Siendo w_i el peso de la observación iésima y N el número total de observaciones.

Una vez se han fijado las anteriores hipótesis, se sigue el siguiente procedimiento repetitivo de M veces (m=1 hasta M):

1. A partir de los pesos definidos en las hipótesis se ajusta el modelo sobre el conjunto de observaciones.
2. Mediante las predicciones realizadas sobre las observaciones, se analizan los fallos y éxitos del algoritmo. Teniendo en cuenta los fallos se puede calcular el error del modelo, el cual sigue la siguiente ecuación 2.13:

$$err_m = \frac{\sum_{i=1}^{N} w_i\, I\,(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i} \tag{2.13}$$

Siendo $G_m(x_i)$ la predicción de cada modelo.

3. En función de la eficiencia del modelo, es decir, del número de aciertos que ha obtenido, se le adjudica una ponderación sobre el total de modelos desarrollados. El peso de cada modelo esta positivamente relacionado con el número de acierto que ha obtenido. El cálculo de los pesos de cada modelo sigue la expresión 2.14:

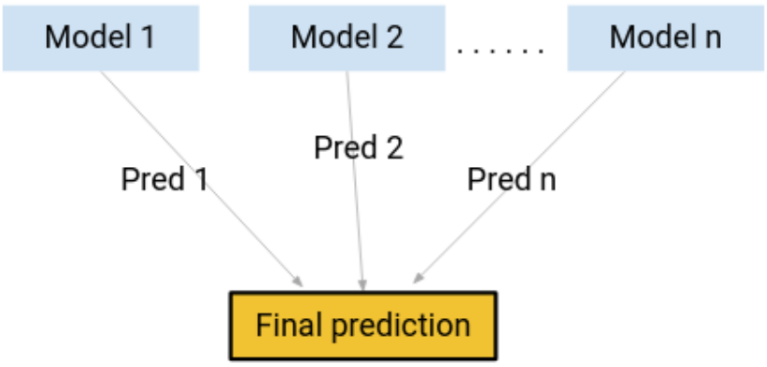
$$\alpha_m = log\,(\frac{1-err_m}{err_m}) \tag{2.14}$$

Siendo α_m el total del peso asignado al modelo.

4. Para obtener una mayor eficiencia en los sucesivos modelos desarrollados, se le otorga un mayor peso a aquellas observaciones que han resultado erróneas y se reduce el de aquellas observaciones que han sido acertadas. El principal objetivo de este cambio de ponderaciones es que los nuevos modelos se centren en aquellas observaciones que modelos anteriores no ha sabido predecir. La ecuación de cambio de ponderación sigue la siguiente expresión 2.15:

$$w_i = w_i \, exp\,[\alpha_m \, I\,(y_i \neq G_m\,(x_i)] \qquad (2.15)$$

Figura 16. Obtención predicción final GBM



Fuente: "4 Boosting Algorithms you should know–GBM, CGBoost, LightGBM & CatBoost" – Analytics Vidhya

El presente proceso compuesto por 4 etapas se repite M veces, generando un modelo en cada iteración. Como se puede apreciar en la Figura 16, para obtener la clasificación definitiva, el algoritmo tomará la predicción realizada por cada uno de los modelos generados y los ponderará por su peso respecto del total.

La técnica Gradient Boosting Machine (GBM) supone una generalización del algoritmo explicado AdaBoost con respecto a la función de medida de error del modelo. En este sentido, se puede afirmar que el funcionamiento de este modelo es muy semejante al explicado, incluyendo las siguientes correcciones:

- Se mantiene el primer paso, donde se ajusta el modelo a las observaciones y se definen los pesos correspondientes a cada una de las observaciones.
- Al realizar la predicción sobre las observaciones se calcula el error correspondiente al modelo. En este caso puede seguir la siguiente función de residuos cuadrados 2.16:

$$obj = \sum_{k=1}^{n} I\,(y_i, \hat{y}_i) \qquad (2.16)$$

- A partir de los residuos identificados del modelo predecesor, se desarrolla el nuevo modelo, con el fin de minimizar la función de error obtenida.
- Como en el caso de AdaBoost este proceso se repite M veces, de forma que la función de error se vaya reduciendo sucesivamente. En este sentido, se establece en los modelos el concepto "Learning rate (λ)" con el fin de limitar la influencia de los modelos sobre el conjunto final y evitar por ende un posible problema de overfitting.

2.2.4.- Extreme Gradient Boosting (XGB)

La técnica Extreme Gradient Boosting (en adelante, XGB) supone una de las metodologías más implementadas por las compañías en la actualidad, debido a su fácil interpretación y eficiencia. En este sentido, este algoritmo sigue una estructura muy parecida a la técnica Gradient Boosting. No obstante, posee una serie de peculiaridades con respecto a la técnica anterior:

- Este tipo de técnica se basa en un procesamiento paralelo de los algoritmos y observaciones, por lo que obtiene una mayor rapidez y eficiencia computacional.
- Este algoritmo define una variable de regularización más evolucionada y eficiente para evitar el overfitting en sus modelos.

El modelo XGB supone un algoritmo con una mayor adaptación y flexibilidad que los desarrollados anteriormente, debido a que cada nuevo árbol clasificador se basa en los errores cometidos por el anterior otorgando pesos a cada una de las observaciones que componen la muestra de entrenamiento. Esta estructura proporciona una mayor capacidad predictiva, ya que es una metodología cuyo pilar base es la reducción de forma sostenible el error de cualquier técnica de aprendizaje. (Hidalgo,2014).

Las innovaciones de la metodología Extreme Gradient Boosting la han convertido en una de las más tractivas e idóneas de cara a la implantación por las compañías. Así mismo, a priori supone la técnica que mayor concreción y exactitud posee en sus resultados.

2.3.- Medidas de verificación de resultados

Las diferentes metodologías explicadas en los epígrafes anteriores necesitan de la existencia de técnicas de evaluación que permitan identificar cuál de los diferentes modelos es el más eficiente y cuál de ellos es el que mejor se ajusta a las observaciones. La verificación de los modelos se realiza desde dos ópticas: Elección del modelo óptimo de cada metodología y posteriormente elección de la metodología que mejor predicción realiza.

El análisis de resultados realizado por las nuevas técnicas de Machine Learning se basa en la división de la base de datos en dos categorías: Una primera parte destinada al entrenamiento del modelo y una segunda parte destinada a la verificación de la eficacia del modelo. De forma general, el 70% de los datos de la muestra original se tienen en cuenta para la construcción del modelo y el 30% restante de la muestra se emplea para comprobar la capacidad predictiva.

Se pretende que los modelos se ajusten bien a los datos de construcción y a partir de ellos sea capaz de realizar una predicción lo más similar posible a los datos de verificación, es decir, el objetivo es dotar al modelo de una base de datos a partir de la cual se pueda obtener una predicción. En el presente trabajo la muestra posee datos sobre las prácticas fraudulentas de los asegurados de una cartera de autos. La determinación de la capacidad predictiva del modelo se determina mediante la comparación de los sucesos predichos por el modelo y los realmente ocurridos en la parte de la muestra de verificación, de esta manera se obtiene cuál de los modelos desarrollados posee menos errores.

En definitiva, se concluye que el mejor algoritmo de cálculo es aquel que, de la forma más optima posible, obtiene un menor número de fallos. En este sentido, se debe conocer el peso que poseen cada uno de los errores y las técnicas de evaluación que se emplean para verificar el comportamiento de los modelos.

2.3.1.- Criterios de selección de modelos

De manera general no se obtienen modelos que se ajustan de forma perfecta a las observaciones de la base de datos, es por ello por lo que se deben poseer metodologías que permitan identificar cuáles de los modelos desarrollados por cada una de las técnicas es el que mayor eficacia y mejor ajuste presenta, es decir, cual es el modelo óptimo de cada metodología.

Referente a las metodologías clásicas de Machine Learning es importante saber que existen diferentes modelos que tienen una gran capacidad predictiva y ajuste sobre las observaciones, por lo que hay varios modelos válidos. En este sentido se debe saber cuál de los modelos desarrollados es el más adecuado y contempla una mayor proporción de variabilidad de los datos. Para realizar la evaluación y escoger el modelo óptimo existen diferentes criterios de selección, de los cuales se destacan los siguientes: Akaike's Information Criteria (AIC), Bayesian information criterio (BIC) y Derivance (D).

- AKAIKE (AIC): Esta técnica fue desarrollada por Akaike (1974). Esta técnica pretende se basa en el ajuste obtenido sobre la muestra para cotejar la eficacia de un modelo prediciendo valores futuros. Es decir, para una muestra concreta, este estadístico presenta una medida relativa de la calidad o ajuste del modelo. Este estimador define una relación entre la complejidad del modelo, variables que posee, y el ajuste del modelo a las observaciones. Por otro lado, se posee el estadístico BIC, el cual se fundamenta sobre parte de la función de probabilidad del modelo y está estrechamente relacionado con el criterio AIC. Debido a esta estrecha relación, para la comparación de modelos de regresión se empleará únicamente el AIC, ya que aporta información suficiente para realizar el estudio oportuno. Se busca minimizar este indicador y se expresa a partir de la siguiente ecuación 2.17:

$$AIC = 2k - 2\ln(L) \qquad (2.17)$$

  Donde k es el número de variables recogidas en el modelo y L hace referencia al máximo valor de la función de verosimilitud.

- DEVIANCE (D): Spiegelhalter et al. (2002) fueron de los primeros autores que desarrollaron el denominado "Deviance information criterion". Este estimador realiza una evaluación entre los datos observados y los valores ajustados que se obtienen a partir del modelo, midiendo la distancia entre los diferentes puntos. Se define a partir de la expresión 2.18, donde λ hace referencia al cociente entre el modelo saturado y el modelo calculado.

$$D = -2\ln\lambda = -2\left[\ln f(y;\beta;\varphi) - \ln f(y;\beta_{sat};\varphi)\right] \qquad (2.18)$$

Además de los estadísticos mencionados, se puede realizar una representación gráfica de las variables más influyentes en cada uno de los modelos, con el fin de identificar la composición del algoritmo y comprobar la existencia de diferentes variables significativas.

Existen algunas situaciones en la que la elección del modelo óptimo se basa en el juicio experto del investigador. Esto es debido a que pueden existir parámetros o factores de riesgo que sean particulares de las reclamaciones fraudulentas y que no se puedan integrar en los modelos desarrollados.

2.3.2.- Medidas de desempeño

A partir de las observaciones de la parte muestral de verificación, se generan una serie de estimadores que permiten conocer cuáles de los modelos realizan una mejor predicción de los datos. Se debe saber identificar cuál de los modelos desarrollados realiza una mejor detección de las reclamaciones fraudulentas.

Una de las metodologías empleadas para evaluar los resultados obtenidos mediante las técnicas de Machine Learning se denomina Matriz de confusión. Esta técnica no solo muestra aquellas observaciones que han sido correctamente clasificadas, sino también aquellas que lo han sido de forma errónea indicando el tipo de error cometido. Esta técnica define cuatro tipologías de errores:

- **TP**: Hace referencia a las observaciones que han sido identificadas por el modelo de forma correcta, es decir, verdaderos positivos.
- **FP**: Se conocen como errores de tipo I, hacen referencia a aquellas observaciones que identifica el modelo como fraudulentos cuando realmente no lo son. Este tipo de error también es conocido como falsos positivos.
- **TN**: Esta categoría engloba aquellas observaciones que el modelo ha identificado como negativas y se realmente son negativas, es decir, verdaderos negativos.
- **FN**: Este tipo de categoría se conoce como error de tipo II y engloba aquellas observaciones que han sido identificadas por el modelo como no fraudulentas y realmente si suponen un caso de fraude. También se denomina falsos negativos, es el error más grave en el que puede incurrir el modelo.

Esta metodología de verificación fue desarrollada por los autores Kohavi y Provost (1998), los cuales desarrollan la matriz de confusión de forma similar a la definida en la Tabla 6.

Tabla 6. Matriz de confusión

		Predicción	
		Positivo	Negativo
Real	Positivo	TP	FN
	Negativo	FP	TN

Fuente: Elaboración propia

En este contexto, se define el error como la diferencia entre la predicción realizada por el modelo y el suceso observado en la realidad. Se les otorgan diferentes codificaciones a los errores, en el caso de que el modelo no prediga un fraude que se ha realizado, el error sería codificado como 1. En el caso de que el modelo prediga que se produce un fraude cuando en realidad no se produce, el error sería codificado como -1 y en caso de que se produzca un acierto el error es codificado como 0.

A partir de la matriz de confusión se pueden establecer diferentes estimadores para establecer la precisión de los modelos. La metodología más empleada es la Accuracy, la cual expresa de forma global la exactitud del modelo en cada una de las clases, obteniendo este resultado a partir de los errores cometidos en cada clase dividido por el total de predicciones realizadas. Estas medidas miden la sensibilidad del modelo, obteniéndose a partir de la identificación de las observaciones que poseen fraude que han sido categorizadas de forma correcta.

Otro estadístico muy empleado en el cotejo del resultado de este tipo de modelos es el Kappa, el cual define la precisión del modelo como la división entre la exactitud obtenida y la exactitud esperada dividida por 1 menos la exactitud esperada. Cuanto más próximo a 1 sea el resultado mayor precisión posee el modelo.

Además de las métricas expuestas es importante realizar un análisis de forma gráfica, ya que permite identificar de forma muy intuitiva la capacidad predictiva de los modelos. Cuando la predicción del fraude realizada por el modelo se ajuste de forma correcta a los datos los valores se reflejará una elevada efectividad, mientras que por el contario cuanto peor sea la predicción del modelo menor será la efectividad representada.

3. ANÁLISIS EXPLORATORIO DE DATOS

En el presente apartado se realizará un análisis de la base de datos empleada para la modelización de la probabilidad de fraude. Antes de la introducción de la información en los modelos es conveniente estudiar la naturaleza de las diferentes variables, eliminando aquellas que aporten poca información al modelo o sean redundantes para el estudio.

La información referente a las pólizas y a los asegurados son datos de gran delicadeza por lo que no es habitual que las compañías los compartan. Debido a la incapacidad de obtener una fuente de información apropiada sobre el fraude en una cartera de autos en España, se ha aplicado la metodología de detección de fraude empleando información referente a los Estados Unidos. La BBDD utilizada ha sido obtenida de la página web Kaggle, en la cual se encuentra información referente a reclamaciones fraudulentas en una cartera de automóviles en EE. UU durante el primer cuatrimestre del año 2015. La BBDD es un fichero de texto en formato ".csv" que contiene un total de 39 variables referentes a las diferentes características de los siniestros reclamados, siendo una de las variables indicadora de fraude en las reclamaciones realizadas.

La BBDD está compuesta por un total de 1.000 registros, donde cada una de las observaciones corresponde a un siniestro. Este espacio muestral se ha considerado suficiente para poder estimar el comportamiento de la población respecto a reclamaciones fraudulentas. En primer lugar, sobre estas observaciones se ha llevado a cabo una limpieza de los datos, con el fin de obtener un análisis de las variables lo más ajustado posible y evitar "ruido" en los modelos que dé lugar resultados distorsionados. La limpieza de los datos ha consistido en la eliminación de errores y de aquellos valores que son considerados outliers.

En este estudio el análisis de las distribuciones de las variables y el posterior desarrollo de las distintas técnicas de predicción se ha desarrollado en el programa de R-Studio, empleando scripts ad-hoc en R. Las representaciones gráficas se han realizado con el paquete *ggplot2* mientras que los modelos de Maching Learning están implementados en el paquete *caret*. El código se encuentra en el anexo del presente trabajo, permitiendo verificar los resultados obtenidos a los efectos oportunos.

A continuación, se estudian las variables que posee la BBDD, con el objetivo de realizar una descripción sobre estas e identificar aquellas que es conveniente emplear para realizar una modelización ajustada de la variable de detección de fraude.

3.1.- Variables de estudio

Las principales variables que se seleccionan para el análisis preliminar son aquellas que a priori se identifican como relevantes respecto a la capacidad predictiva de los modelos. Estas variables son divididas en diferentes categorías: referentes a las características de la póliza, referentes a las características del asegurado, vinculadas a los siniestros reclamados y por último variables referentes a las características del vehículo.

Dentro de la BBDD también se encuentra la variable respuesta de los modelos que se van a desarrollar, la cual determinará la distribución que seguirán las técnicas para realizar la modelización y se empleará para realizar el entrenamiento y la validación de estas.

Tabla 7. Variables referentes a la póliza

Variables de la póliza			
	Nomenclatura BBDD	Tipología	Definición
1	Months_as_customer (Meses en cartera)	Numérica	Tiempo que lleva el asegurado en la compañía.
2	Policy_annual_premium (Prima anual)	Numérica	Prima anual abonada por el asegurado.
3	Umbrella_limit (Cobertura paraguas)	Numérica	Cobertura que garantiza un importe extra sobre aquellas reclamaciones que exceden una cantidad determinada en la póliza.

Fuente: Elaboración propia

Tabla 8. Variables referentes al asegurado

Variables del asegurado			
	Nomenclatura BBDD	Tipología	Definición
4	Age (Edad)	Numérica	Edad del asegurado.
5	Insured_education_level (Nivel de educación)	Literal	Nivel de educación del asegurado.
6	Insured_occupation (Profesión)	Literal	Actividad laboral despeñada por el asegurado.
7	Insured_hobbies (Aficciones)	Literal	Aficiones del asegurado.
8	Insured_relationship (Estado civil)	Literal	Estado civil del asegurado.

Fuente: Elaboración propia

Tabla 9. Variables referentes al siniestro

Variables del siniestro			
	Nomenclatura BBDD	Tipología	Definición
9	Incident_hour_of_the_day (Hora del siniestro)	Numérica	Hora en la que se produjo el siniestro reclamado.
10	Incidet_type (Tipo de siniestro)	Literal	Tipología de siniestro
11	Collision_type (Tipo de colisión)	Literal	Zona del vehículo en la que se ha producido el siniestro
12	Incident_severity (Severidad del siniestro)	Literal	Alcance del siniestro

13	Authorities_contacted (Autoridades contactadas)	Literal	Cuerpo de seguridad del estado que ha sido contactado en el siniestro
14	Number_of_vehicles_involved (Nº de vehículos involucrados)	Numérica	Número de vehículos involucrados en el siniestro
15	Witnesses (Testigos)	Numérica	Testigos en el momento del siniestro
16	Injury_claim (Reclamación por lesiones)	Numérica	Cantidad reclamada por lesiones de los ocupantes
17	Property_claim (Reclamación de propiedad)	Numérica	Cantidad reclamada por daños sobre la propiedad
18	Vehicle_claim (Reclamación de vehículo)	Numérica	Cantidad reclamada por daños en el vehículo
19	Total_claim_amount (Cuantía total reclamada)	Numérica	Cantidad total reclamada a la compañía

Fuente: Elaboración propia

Tabla 10. Variables referentes al vehículo

Variables del vehículo			
	Nomenclatura BBDD	Tipología	Definición
20	Auto_make (Marca del vehículo)	Literal	Marca del vehículo siniestrado
21	Auto_model (Modelo del vehículo)	Literal	Modelo del vehículo siniestrado
22	Auto_year (Año del vehículo)	Numérica	Año de compra del vehículo siniestrado

Fuente: Elaboración propia

Tabla 11. Variable de estudio

Variable de estudio			
	Nomenclatura BBDD	Tipología	Definición
23	Fraud_reported (Fraude reportado)	Dicotómica	Variable indicadora de presencia de fraude en la reclamación

Fuente: Elaboración propia

Dentro de la BBDD las 23 variables anteriormente definidas vienen informadas de forma correcta y en consonancia al siniestro ocurrido en todas las observaciones, de tal forma que la base de datos no impacta negativamente sobre los modelos. Aquellas observaciones que

incluían datos erróneos o inexactos han sido ajustadas, sustituyendo los valores erróneos por valores obtenidos a partir de la distribución empírica de la variable en la que se poseía el error.

Así mismo, sobre las variables cualitativas se realiza una unificación de las categorías propias de cada variable con el fin de simplificar el estudio de estas y evitar la sobre parametrización de los modelos desarrollados. Una vez corregidas y definidas las 23 variables de estudio se realiza un análisis más profundo de la información almacenada en cada uno de los factores.

La depuración de las variables de la Tabla 9 para la implementación de los modelos se ha realizado siguiendo diversos criterios, entre ellos está la falta de capacidad predictiva como ocurre con el número de póliza o el DNI, materia legislativa que prohíbe el uso de la variable sexo para la tarificación (Sentencia Test-Achats, 2011), incapacidad de definir el efecto de estacionalidad como ocurre con la fecha, así como, la falta de comprensión de ciertas variables debido a la codificación específica de la compañía aseguradora propietaria de la BBDD.

La variable objeto de estudio es aquella que expresa si se ha detectado un comportamiento fraudulento en el siniestro reclamado o no. Esta variable se considera una variable discreta, por lo que se lleva a cabo un determinado análisis de la distribución de probabilidad para el estudio de esta variable.

3.2.- Distribución y tratamiento univariable

Una vez se han definido las diferentes variables independientes que a priori pueden intervenir en los modelos, el primer paso es estudiar las características que poseen los datos de cada una de ellas. El principal objetivo de este análisis es obtener una mayor información sobre los datos contenidos en las variables y el comportamiento de estos.

En primer lugar, se procede a examinar el comportamiento de cada una de las variables numéricas basándose en estadística descriptiva básica, empleando tanto métricas como representaciones gráficas.

En la Tabla 12, se recogen las principales métricas estadísticas referentes a determinadas variables numéricas del modelo.

Tabla 12. Medidas estadísticas básicas

	Min.	1st Qu.	Mediana	Media	3rd Qu.	Max.
Months_as_customer	0,0	115,8	199,5	204,0	276,2	479,0
Policy_annual_premium	538,2	1.090,6	1.257,2	1.258,0	1.415,7	2.047,6
Age	19	32	38	38,95	44	64
Injury_claim	0	4.330	6.780	7.441	11.305	21.450
Property_claim	0	4.480	6.755	7.407	10.885	23.670
Vehicle_claim	1.440	30.438	42.100	37.967	50.823	79.560
Total_claim_amount	1.920	41.963	58.055	52.815	70.593	114.920

Fuente: Elaboración propia

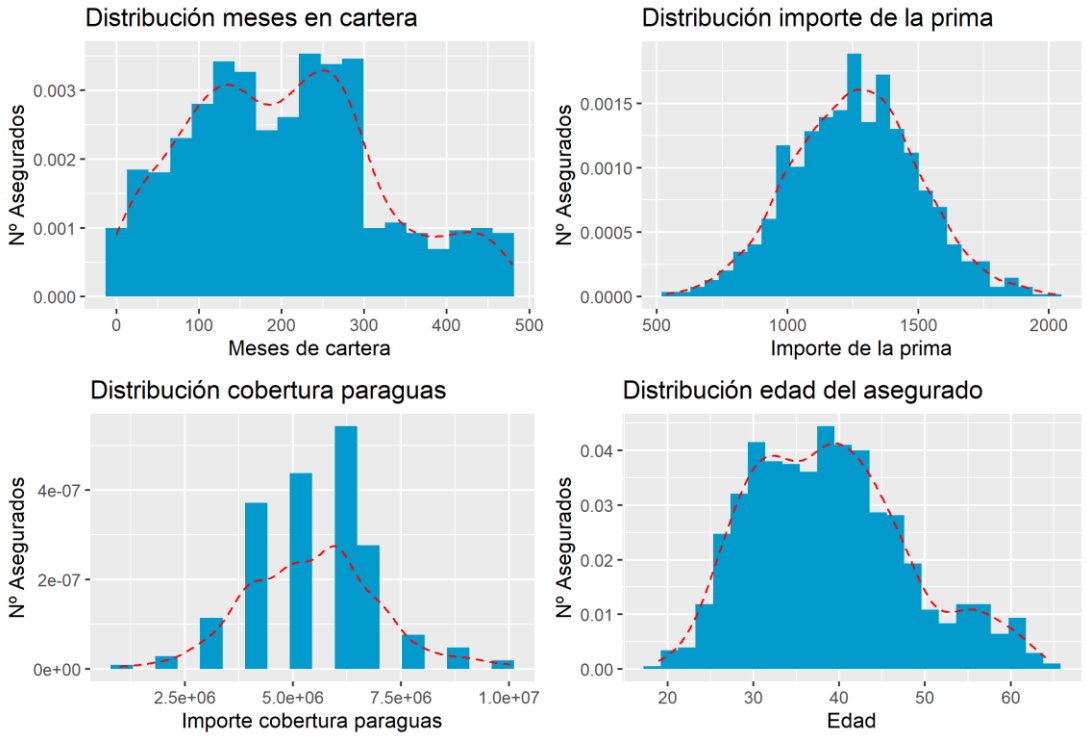
A partir de los datos contenidos en la Tabla 12 se puede determinar que en las variables numéricas presentadas se han corregido de forma correcta los valores atípicos. La media y mediana, que suponen medidas de tendencia central, poseen valores próximos. En este sentido, existen pequeñas diferencias debido a la existencia de ciertos valores elevados que al considerarse correctos se han mantenido en el espacio muestral. La mayor diferencia entre valores máximos y mínimos corresponde a la variable "Total_claim_amount" cuya diferencia es de 133.000 unidades monetarias.

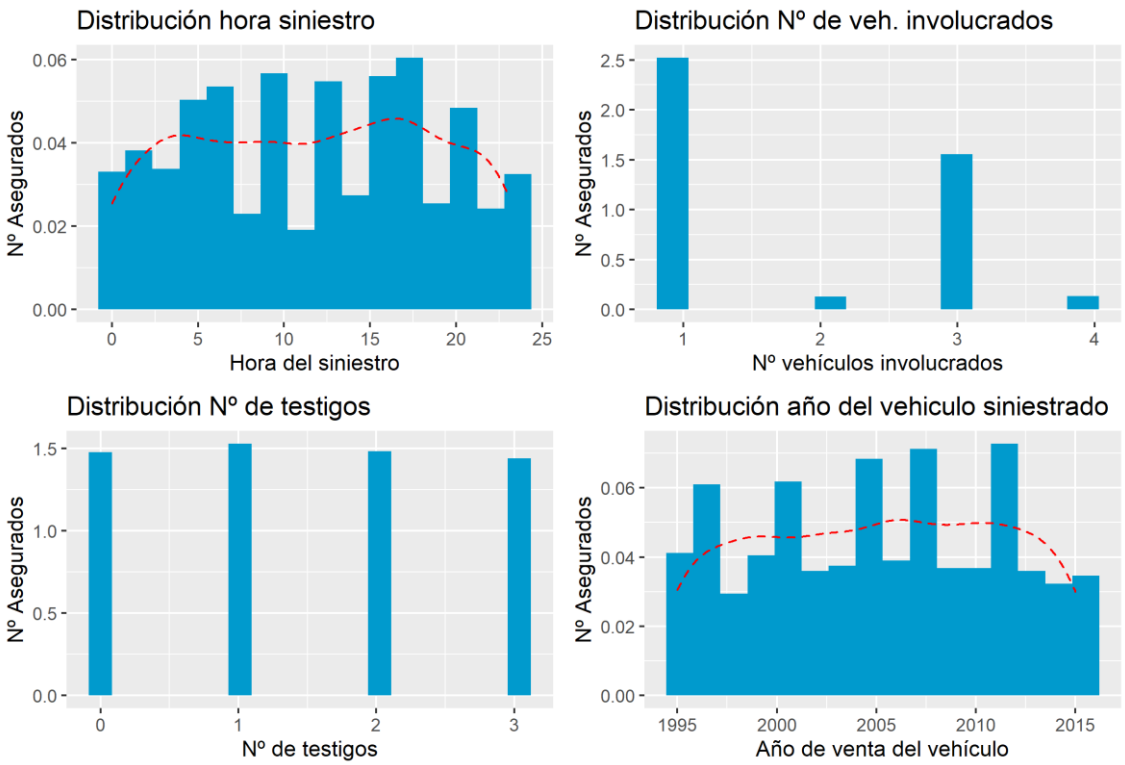
Respecto a las variables referentes a los asegurados, se puede determinar que se trata de una cartera en la que existe una gran fidelidad de estos, ya que de media posee una permanencia de 204 meses o 17 años. Así mismo, los asegurados tienen una edad de 39 años aproximadamente, en términos generales se puede considerar como una muestra intermedia, no siendo joven ni extremadamente madura.

Respecto a variables referentes a importes, la prima media de la cartera asciende a 1.258$. Respecto a los importes de las reclamaciones realizadas se identifica la cobertura del vehículo como aquella que mayor coste supone para la compañía, alcanzando un montante reclamado de 37.967$ en promedio. Cabe señalar que en las reclamaciones por lesiones y propiedad existen siniestros que suponen coste 0 para la compañía, mientras que la cobertura de vehículo siempre posee una determinada cantidad reclamada. Este hecho también explica la gran diferencia entre el máximo y el mínimo respecto a los importes totales reclamados, variable que surge de la suma de las tres anteriores.

En la Figura 17, se representa el comportamiento de las diferentes variables discretas e importe de la prima (única variable continua de esta representación), seleccionadas a priori para el desarrollo del modelo.

Figura 17. Distribuciones variables discretas e importe de la prima





Fuente: Elaboración propia

A partir de la Figura 17, y en línea con la Tabla 12, se observa que una masa importante de los asegurados de esta cartera se mantiene en la compañía durante un periodo de 200 y 300 meses, es decir, en torno a 16 años. Lo cual indica que la caída de asegurados de la cartera es muy reducida una vez se superan los primeros meses. Respecto al importe de la prima, y en consonancia con la Tabla 12, se puede observar que sigue una posible distribución normal, donde la mayor parte de las primas se encuentran en la parte central de la gráfica.

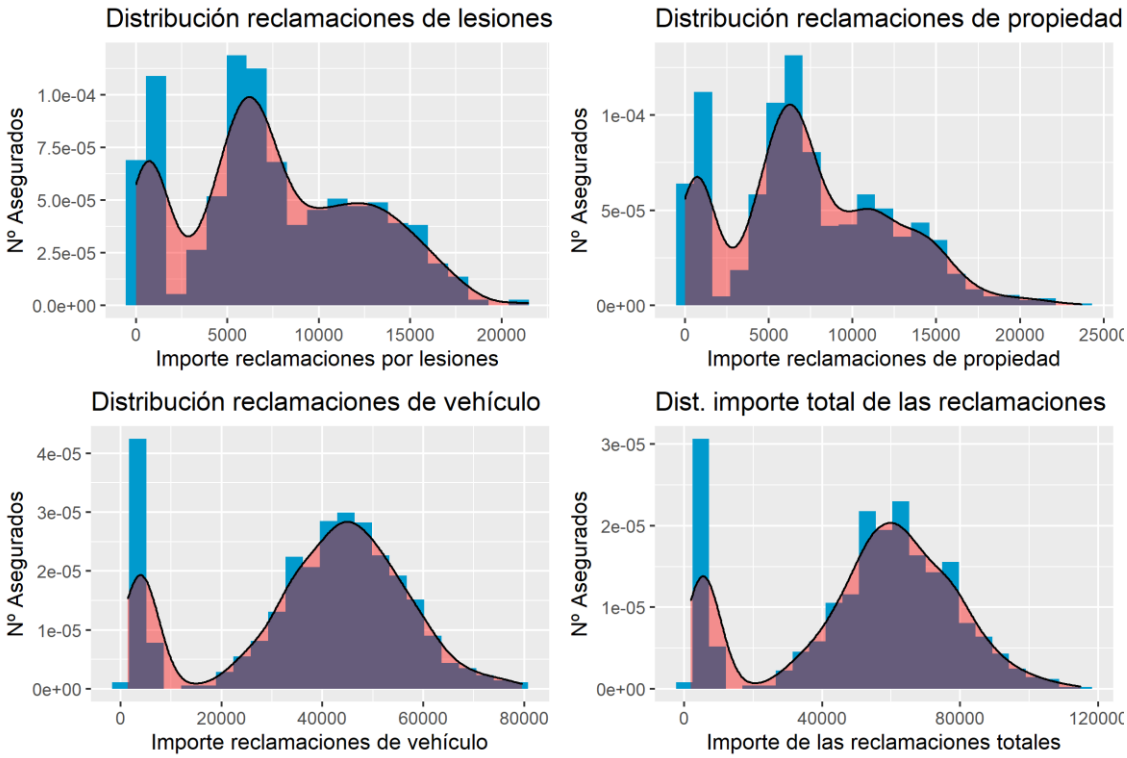
Respecto a la cobertura paraguas, en la Figura 17 se han eliminado de la distribución aquellas personas que no tienen contratada una cobertura paraguas. Se obtiene una distribución en la que la mayoría opta por una cuantía media, la cual se encuentra entre los 5.000.000$ y los 6.000.000$.

A partir de la Figura 17, se puede determinar que la mayor parte de los asegurados perteneciente a la cartera estudiada posee una edad comprendida entre los 30 y 45 años. En este sentido se poseen pocos asegurados noveles o menores de 25 lo que reduce la capacidad de estudio sobre este colectivo.

Respecto a las 4 variables inferiores representadas en la Figura 17 no se puede visualizar una distribución tan clara como en los factores superiores. No obstante, estas gráficas si permiten identificar aquellos valores más frecuentes en cada una de las variables. Respecto a los siniestros, se estima que aquellas horas con mayor número de accidentes son las primeras horas de la mañana y las horas después de comer donde se encuentra el máximo. La mayor parte de los siniestros reclamados presentan un único vehículo implicado en el accidente y entre 0 y 3 testigos. Por último, los años de venta del vehículo están muy equidistribuidos habiendo años punta en los que se adquirieron más unidades.

El análisis de las variables continuas referente a los importes de las reclamaciones queda representado en la Figura 18. Estas variables siguen una estructura de análisis similar a la planteada en las representaciones anteriores.

Figura 18. Distribuciones variables continuas



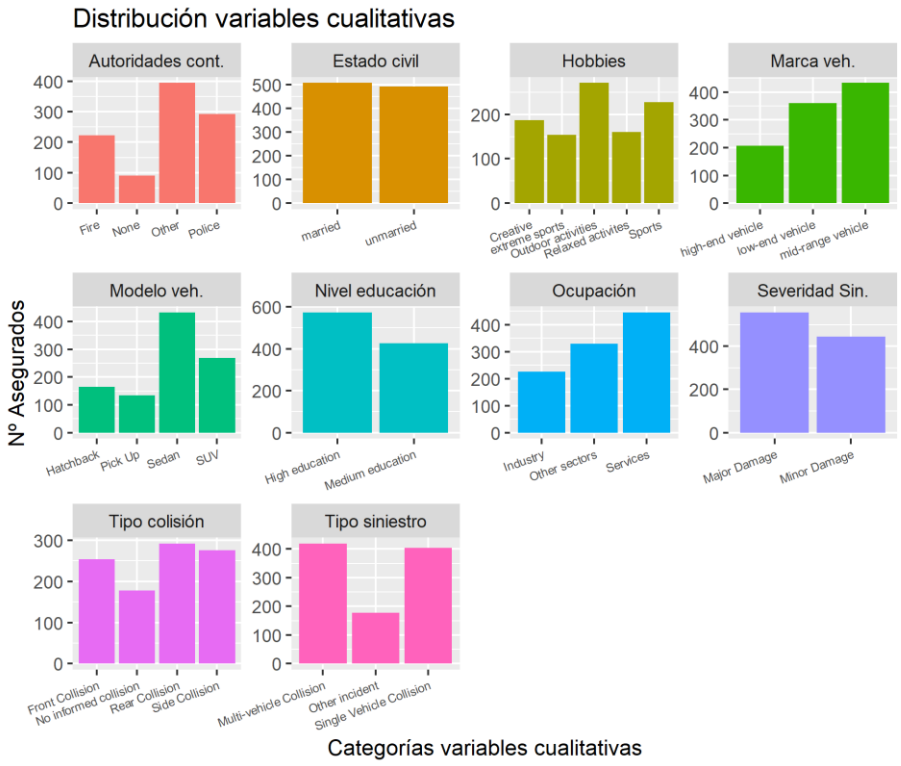
Fuente: Elaboración propia

En la figura 18 se puede observar que gran parte de los casos de reclamaciones de lesiones y propiedad presentan reclamaciones muy reducidas incluso 0. El resto de las reclamaciones presentadas en estas categorías se distribuye mayormente entre las horquillas de 5.000$ y 15.000$. La variable referente a los importes reclamados por vehículo presenta una mayor dispersión en las cantidades reclamadas. Donde a pesar de que existe un gran número de reclamaciones de cuantía reducida, la mayor parte de las cuantías oscilan entre los 20.000$ y 65.000$.

Sobre la variable referente al total de las reclamaciones, se puede identificar una distribución de los importes muy similar a la observada en la variable referente a las reclamaciones de vehículo. Es por ello por lo que se determina que el factor de estas últimas reclamaciones es el que mayor influencia posee sobre la variable de importe total. Esto puede venir explicado porque es una categoría que siempre está presente en las reclamaciones realizadas, mientras que las categorías referentes a lesiones y propiedad pueden no haberse materializado.

En segundo lugar, se lleva a cabo un análisis de las variables cualitativas. Para ello de genera la Figura 19, a partir de la cual se permite obtener una visión más profunda de la información y subcategorías albergadas en las variables categóricas.

Figura 19. Distribuciones variables cuantitativas



Fuente: Elaboración propia

Para este análisis las variables categóricas han sido organizadas por orden alfabético. Como se puede extraer de la Figura 19, al tratarse de variables categóricas se posee una escala diferente para cada una de las variables, lo cual no permite realizar una comparación tan exhaustiva como la realizada para las variables numéricas. No obstante, las variables que mayor dispersión poseen en sus categorías son la referente al modelo del vehículo y autoridades contactadas. En el lado opuesto, las variables cualitativas más equilibradas son la referente al estado civil del asegurado y los hobbies de este.

Una vez se han analizado las distribuciones que siguen las diferentes variables incluidas BBDD, se realiza un estudio de la variable dependiente del modelo "Fraude reportado". En referencia a la variable respuesta, es de suma importancia conocer la distribución que sigue el fraude en la cartera de estudio.

La variable dependiente recoge cuales de los siniestros reclamados en la cartera de autos presentan fraude. Esta variable supone una variable dicotómica, la cual devuelve un valor "Si" en caso de que haya sido detectado un comportamiento fraudulento en la reclamación estudiada y un valor "No" en los casos en los que las reclamaciones sean lícitas. Al tratarse de una variable aleatoria con dos únicas respuestas posibles, se puede identificar que sigue una distribución Binomial. Esta será la distribución que se empleará para desarrollar los modelos con los que se estimará comportamiento fraudulento de la cartera.

La distribución Binomial es una distribución probabilística discreta, la cual realiza un conteo del número de éxitos observados en un espacio muestral de *n* observaciones independientes entre sí. Los resultados son dicotómicos siguiendo una Bernoulli, donde el valor

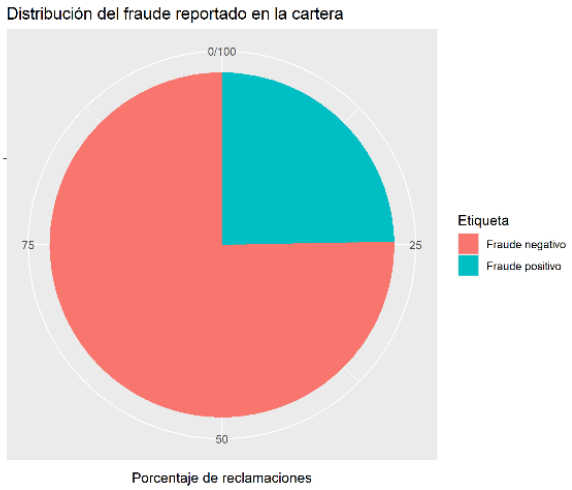
"Éxito" tiene una probabilidad de ocurrencia ρ y la probabilidad de fracaso su opuesto. La distribución Binomial se basa en la siguiente expresión matemática 2.19:

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x} \qquad (2.19)$$

Para x=1, 2, 3, …, n

Se presenta la Figura 20, a partir de la cual se recoge la frecuencia relativa que poseen las dos categorías contenidas en la variable respuesta del modelo.

Figura 20. Distribución casos de fraude



Fuente: Elaboración propia

Como se observa en la Figura 20, sobre la base de la información presente en la BBDD la mayor parte de las reclamaciones recibidas por la compañía, el 75,3% (753 observaciones), no poseen elementos fraudulentos. No obstante, casi un cuarto de la muestra, el 24,7% de las reclamaciones, si incurre en este tipo de actos delictivos. A pesar de suponer un reducido porcentaje respecto al total de la muestra, las reclamaciones que incurren en fraude suponen un gran impacto para las compañías aseguradoras.

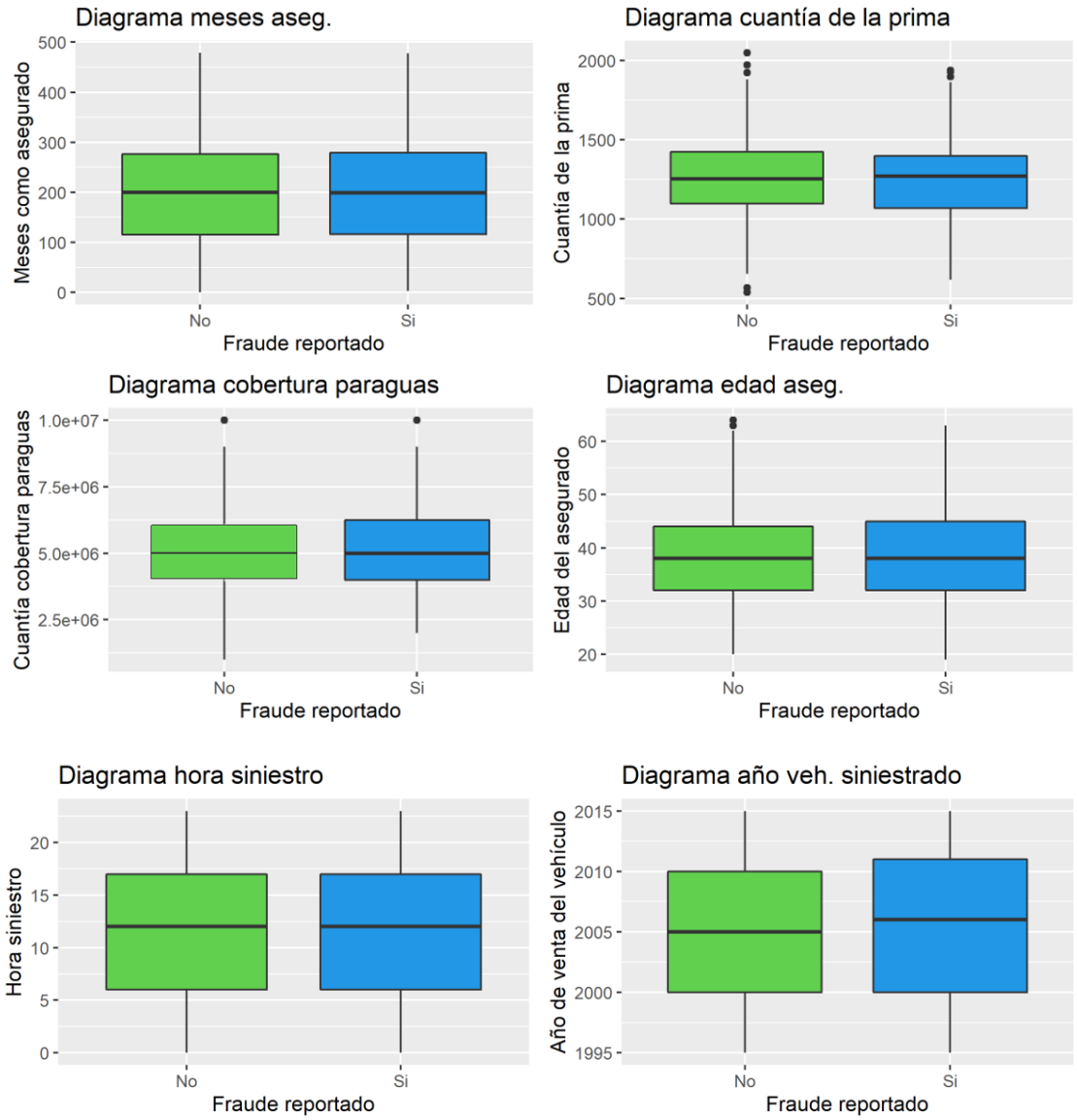
El resultado obtenido en el estudio del fraude en esta BBDD está en consonancia con los datos expuestos en el primer epígrafe sobre el fraude en el mercado asegurador, en base a los cuales se establece que los siniestros fraudulentos representan un reducido peso en el total de siniestros recibidos por las aseguradoras. Este hecho dota de veracidad la muestra con la que se desarrollaran los modelos, de tal manera que se puede extrapolar el paradigma de este trabajo sobre una cartera del mercado español.

3.3.- Distribución y tratamiento bivariable

Además del estudio del comportamiento de las variables de forma independiente, uno de los fines del presente análisis es conocer la asociación que poseen las variables estudiadas respecto a la variable de detección de fraude. Se pretende identificar la influencia de dichas variables a la hora de realizar predicciones sobre dichos comportamientos. El presente estudio se realiza empleando gráficas y diagramas.

En la Figura 21 se representan diferentes diagramas de caja de las variables numéricas siguiendo el mismo orden que las representaciones del epígrafe anterior. Este tipo de diagramas se emplean para analizar la influencia de una variable sobre los diferentes conjuntos de datos definidos en la BBDD. Cuanto más grandes son los "bigotes" de los diagramas más valores atípicos se posee en esa variable, como se puede observar aquella que mayor dispersión posee en sus datos es la variable referente a la cuantía de la prima.

Figura 21. Diagramas de caja variables discretas e importe prima vs. fraude reportado



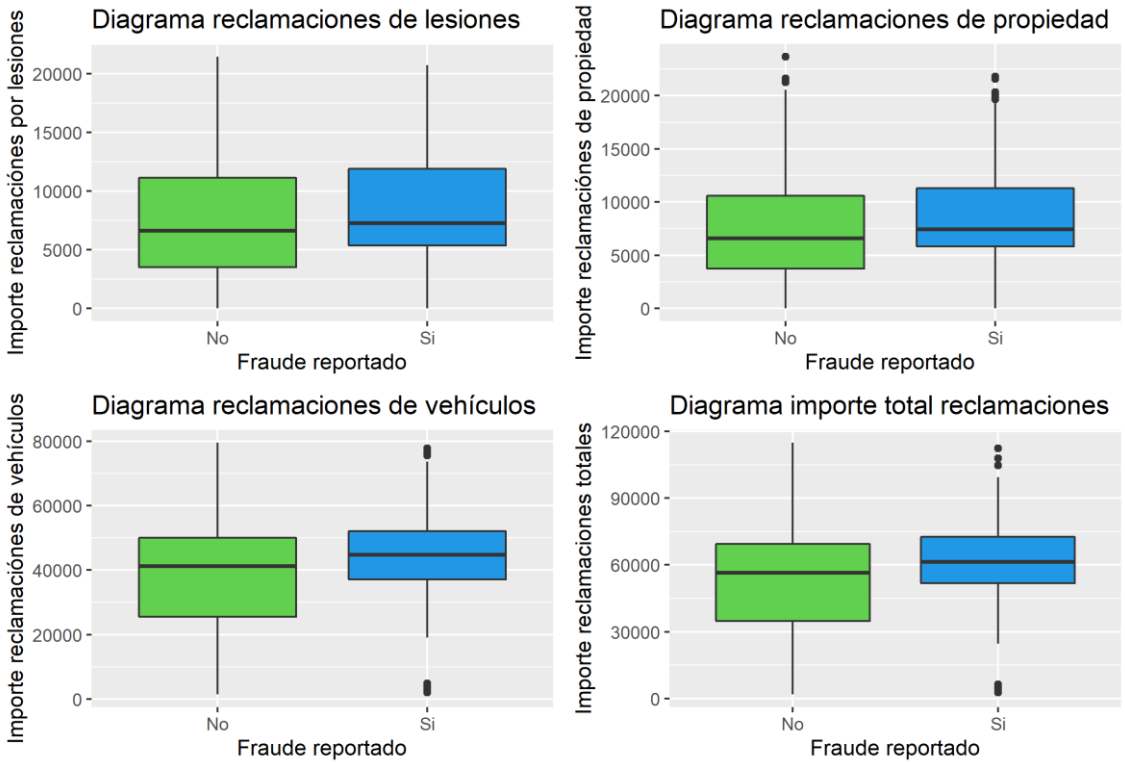
Fuente: Elaboración propia

La mayor parte de las relaciones representadas en la Figura 21 poseen un elevado equilibrio, es decir, presentan medianas y cuartiles muy similares. Es por ello por lo que en un primer momento se podría afirmar que no poseen una gran relevancia en la predictibilidad de la variable de fraude. Así mismo, se han omitido del presente análisis las variables referentes a los vehículos involucrados y el número de testigos del siniestro, debido a la detección de anomalías en las representaciones graficas derivado de la distribución de sus datos.

No obstante, respecto a la variable referente a la cobertura adicional, si presenta una asimetría en la mediana lo cual podría indicar una relación o tendencia entre la cobertura paraguas contratada y el comportamiento fraudulento de los asegurados.

En cuanto a la clasificación de los importes de las reclamaciones frente a la variable dependiente se extrae la Figura 22. A continuación, se representan las reclamaciones en función del importe reclamado en cada una de las coberturas con el fin de obtener una visión desagregada de la influencia de los importes sobre las actividades fraudulentas. A pesar de la corrección realizada sobre la BBDD, la variable reclamaciones de propiedad parece ser que es la que mayor número de valores atípicos presenta respecto al resto de variables estudiadas.

Figura 22. Diagramas de caja variables continuas vs. Fraude reportado



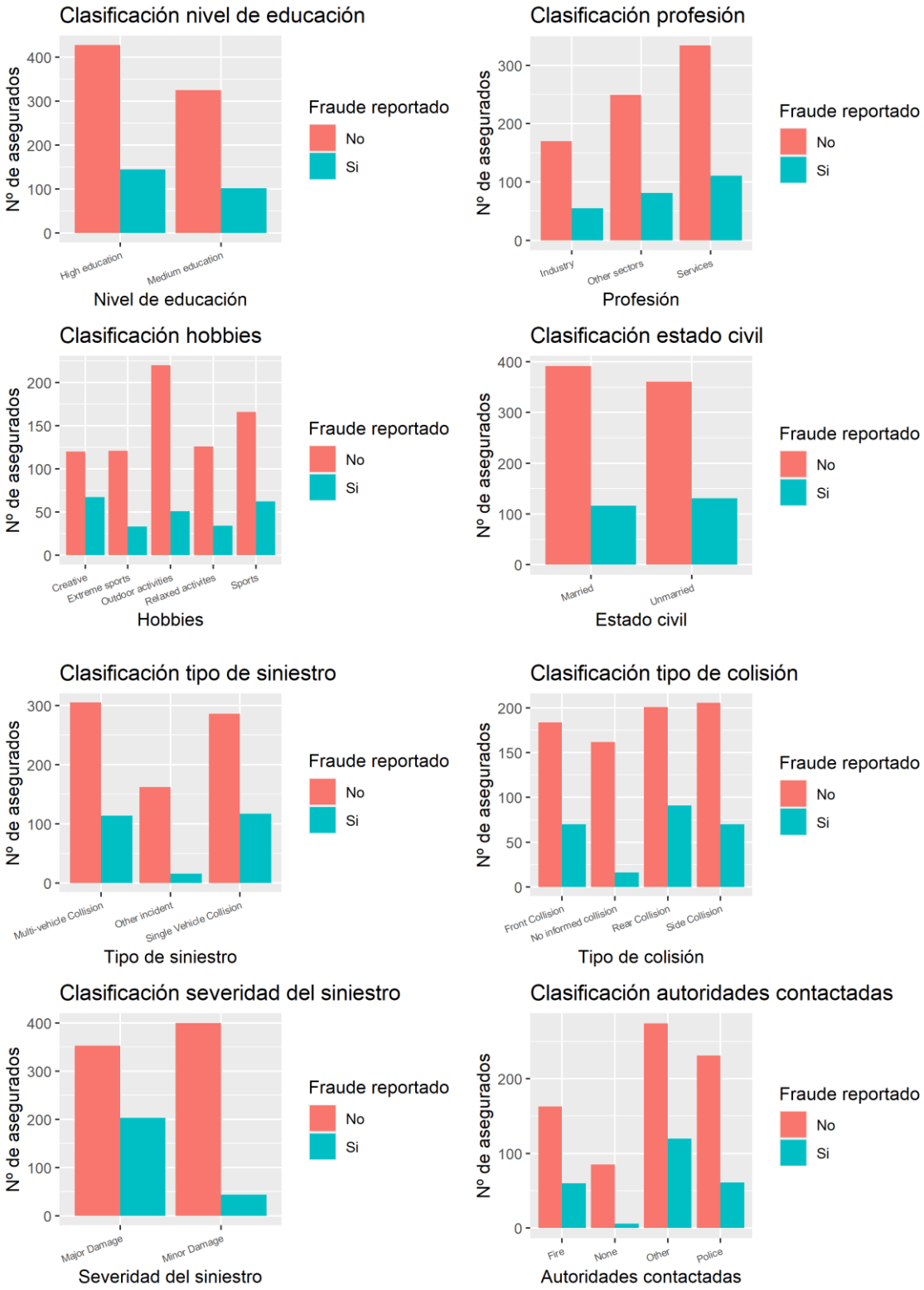
Fuente: Elaboración propia

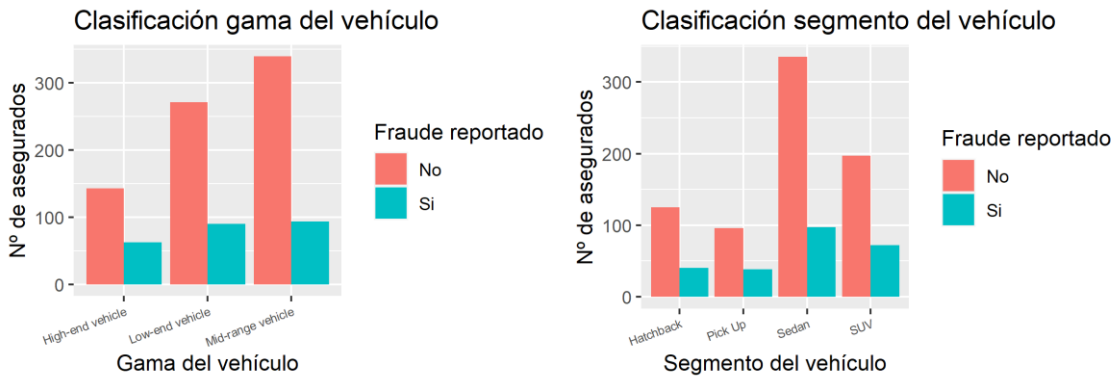
Al contrario que ocurre en el caso anterior, en las variables estudiadas en la Figura 22 se puede observar una clara relevancia respecto al fraude detectado. Se puede afirmar que existe una clara relación positiva entre el importe de la reclamación realizada y el fraude reportado. En este sentido, a medida que aumenta la cantidad reclamada por el asegurado aumenta la probabilidad de que ese caso concreto presente fraude. Las variables referentes a los importes de las reclamaciones de lesiones y del vehículo son aquellas en las que se aprecia una mayor relación positiva respecto al fraude detectado.

Debido a la gran relación que poseen las diferentes variables referentes a las coberturas de las reclamaciones, la cuantía total reclamada representada en la Figura 22, hereda una relación positiva sobre la variable respuesta de los modelos. Además, sobre esta representación realizada se puede observar la existencia de outliers en el importe, sobre todo en aquellos casos en los que si se ha detectado una práctica delictiva.

Referente a las variables categóricas y la relación con la variable respuesta se genera la Figura 23. A partir de la cual se puede extraer la vinculación de cada una de las subcategorías de las variables con el fraude detectado.

Figura 23. Diagramas variables cualitativas vs. fraude reportado





Fuente: Elaboración propia

Las diferentes variables cualitativas presentan en general una estructura muy equilibrada en cuanto a división de la muestra por subcategorías y el fraude detectado en cada una de ellas. No obstante, existen algunas variables en las que la incidencia de fraude supera la medida determinada para el conjunto de las variables de la BBDD.

Las variables que presentan un mayor porcentaje de fraude en alguna de sus categorías son aquellas referentes a los hobbies de los asegurados, al tipo de colisión en la que se ha visto envuelta el vehículo, la severidad del siniestro, la gama del vehículo y el segmento de este. Basándonos en el juicio experto, se puede afirmar que las variables mencionadas anteriormente pueden presentar una relación positiva con las actividades fraudulentas detectadas en la careta.

De forma general, se puede asumir que se cumplen los supuestos presentados en el marco teórico del presente trabajo, ya que la relación de los datos con la variable respuesta sigue los mismos patrones que los definidos en dicho epígrafe.

Además de la relación de las diferentes variables con la variable respuesta, es de gran importancia estudiar la relación entre los factores. Con el análisis de la vinculación de las variables entre sí, se pretende identificar aquellos factores que poseen una elevada correlación procediendo a su eliminación y evitar el efecto de la multicolinealidad. A partir de este análisis se identifican las variables que añaden información al modelo y por tanto poseen un gran potencial de ser finalmente añadidas para la estimación del fraude.

Para este estudio se ha empleado el coeficiente de correlación de Pearson, el cual se emplea sobre aquellas variables que sean numéricas. Para el análisis de asociación de las variables cualitativas se necesita una metodología diferente y es por ello por lo que se empela el coeficiente de V de Cramer, se denomina así por su creador el estadístico sueco Harald Crámer (1946).

En la Tabla 13 basada en el coeficiente de Pearson, se encuentra información referente a la correlación entre diferentes variables numéricas. La correlación mide la intensidad de los cambios que se producen en las variables con respecto a otras. La elevada correlación que presentan determinadas variables indica la posible existencia de una tendencia o patrón entre los datos contenidos por dichas variables. Por el contrario, la correlación negativa presentada por algunas de las variables indica la existencia de movimientos contrarios entre los factores.

Tabla 13. Correlación de variables continuas

	Months_as_customer	Policy_annual_premium	Umbrella_limit	Age	Injury_claim	Property_claim	Vehicle_claim	Total_claim_amount
Months_as_customer	1	-0.002	0.016	0.922	0.064	0.034	0.059	0.06
Policy_annual_premium	-0.002	1	-0.009	0.007	-0.018	-0.012	0.016	0.007
Umbrella_limit	0.016	-0.009	1	0.018	-0.046	-0.023	-0.038	-0.04
Age	0.922	0.007	0.018	1	0.074	0.059	0.061	0.068
Injury_claim	0.064	-0.018	-0.046	0.074	1	0.562	0.722	0.804
Property_claim	0.034	-0.012	-0.023	0.059	0.562	1	0.731	0.81
Vehicle_claim	0.059	0.016	-0.038	0.061	0.722	0.731	1	0.982
Total_claim_amount	0.06	0.007	-0.04	0.068	0.804	0.81	0.982	1

Fuente: Elaboración propia

Sobre la Tabla 13 cabe destacar las variables referentes al tiempo "months_as_customer" y "age", las cuales si poseen una gran vinculación entre ellas no siendo así con el resto de los factores estudiados. Así mismo, sobre la variable prima, se puede asumir que no posee correlación con ninguna de las variables introducidas en el modelo. En el lado contario, encontramos las variables referentes a las cuantías reclamadas en las diferentes garantías, la cuales poseen una elevada correlación entre sí. Dentro de esta agrupación, la variable referente a las reclamaciones del vehículo posee una correlación de hasta un 0.982 sobre la variable de reclamaciones totales, por lo que se podría asumir que ambas variables aportan la misma información al modelo.

Respecto a las variables categóricas su relación es estudiada desde dos puntos de vista: existencia de relación e intensidad de la relación. En primer lugar, se lleva a cabo un contraste de hipotesis con la Chi-cuadrado a partir del cual se identifica si las diferentes variables poseen relación entre sí. Esta metodología está compuesta por las siguientes hipótesis:

$$H_0 = \text{Independencia de variables}$$

$$H_1 = \text{Dependencia de variables}$$

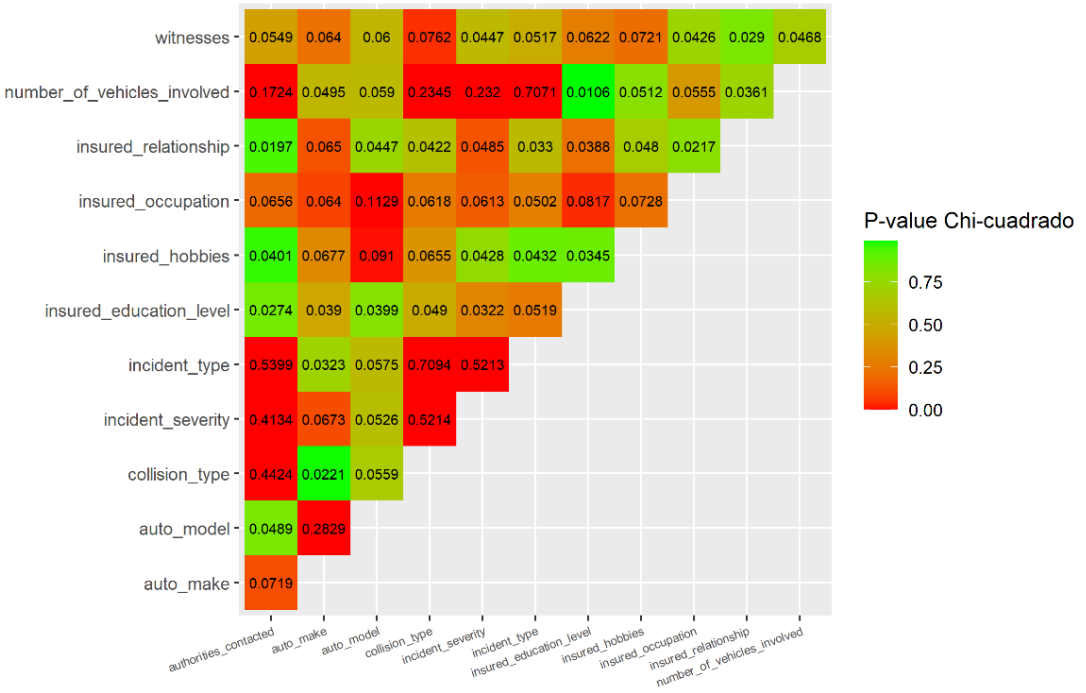
El coeficiente de V de Cramer indica la intensidad de la vinculación entre dos o más variables cualitativas. El resultado de este coeficiente oscila entre 0 y 1, de tal manera que cuanto más elevado sea el resultado mayor intensidad en la relación existe entre las variables. Cuando dicho resultado es 1 o muy próximo a este valor, indica una vinculación total, por lo que se podría decir que ambas variables contienen la misma información para el modelo. Se considera una fuerte relación entre dos o más variables cuando su coeficiente se encuentra entrono a un 0.6.

Para conseguir una mayor comprensión, se ha elaborado la Figura 24, en la cual se presenta el resultado obtenido por cada uno de los factores en este coeficiente. Se han señalado por colores las relaciones entre las diferentes variables siendo el color rojo aquel que menor relación presenta y el color verde aquel que refleja la existencia de vinculación entre las variables. Respecto a la intensidad de la relación, en el mapa de calor quedan representados de forma expresa en cada una de las variables.

Como se puede observar en la Figura 24, se posee una gran intensidad de correlación entre variables como el número de testigos, el número de vehículos involucrados, el tipo de autoridades contactadas, el tipo de colisión, tipo de incidente y la severidad de este. Esta relación puede venir explicada porque todas estas variables hacen referencia a las características del siniestro ocurrido.

Por otro lado, según la Figura 24 observamos que la mayoría de las variables que son características del asegurado tienen una alta relación entre sí. Entre estas variables encontramos el nivel de educación adquirida por los asegurados, los hobbies y la ocupación de estos, incluso el modelo del vehículo. Esto puede venir explicado por la aversión a las prácticas fraudulentas que pueden tener los asegurados según la personalidad que poseen.

Figura 24. Correlación V de Cramer



Fuente: Elaboración propia

El análisis del comportamiento de las diferentes variables y el estudio de la relación entre las mismas influirá en la construcción de los modelos de predicción. Eliminado aquellas que no aporten gran información al modelo y queden representadas por otras variables que recojan una mayor información, como es el caso de las variables referentes al tipo de siniestro, tipo de colisión y severidad del siniestro, las cuales poseen la misma capacidad explicativa sobre la variable respuesta. No obstante, previo al planteamiento de los modelos finales, se procede a confeccionar un modelo de selección.

3.4.- Selección de variables

No todas las variables analizadas son necesarias para obtener un buen modelo predictivo. De forma general, la utilización de determinadas variables contenidas en un subconjunto de la muestra genera resultados más ajustados que la utilización de la totalidad de las variables.

El objetivo del presente apartado es identificar aquellos factores que mayor precisión y capacidad predictiva otorgan al modelo. Una adecuada selección de variables permitirá reducir en gran medida la complejidad de los modelos desarrollados, facilitando su comprensión y aumentando la precisión de las estimaciones realizadas por estos. No obstante, bajo ningún algoritmo de selección se garantiza la definición de un modelo óptimo.

Como se ha comentado, la variable de estudio del presente trabajo es una variable dicotómica, es por ello por lo que esta variable se ajusta de forma muy favorable a la distribución Binomial, la cual es empleada de forma habitual por las compañías para modelizar este tipo de variables. Por tanto, el modelo GLM desarrollado para la selección de variables sigue dicha distribución.

A partir del modelo que se presenta en la Figura 25 se realiza una selección de aquellas variables que poseen una mayor relación con la variable respuesta de detección del fraude. Este proceso de selección de variables se realiza de forma independiente al empleo de las metodologías de aprendizaje automático implementadas.

Figura 25. Modelo GLM selección de variables

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6296  -0.8196  -0.4444  -0.2768   2.5242

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -8.688e-03  3.731e-01  -0.023 0.981419
umbrella_limit                  7.868e-08  3.345e-08   2.352 0.018661 *
insured_hobbiesExtreme sports  -8.540e-01  2.678e-01  -3.189 0.001428 **
insured_hobbiesOutdoor activities -9.919e-01  2.344e-01  -4.231 2.33e-05 ***
insured_hobbiesRelaxed activites  -9.031e-01  2.657e-01  -3.398 0.000678 ***
insured_hobbiesSports          -5.856e-01  2.314e-01  -2.531 0.011362 *
insured_relationshipUnmarried   2.998e-01  1.591e-01   1.885 0.059425 .
incident_severityMinor Damage  -1.653e+00  2.039e-01  -8.104 5.30e-16 ***
incident_hour_of_the_day       -1.859e-02  1.144e-02  -1.625 0.104138
vehicle_claim                   8.506e-06  5.448e-06   1.561 0.118493
auto_makeLow-end vehicle       -2.397e-01  2.097e-01  -1.143 0.253093
auto_makeMid-range vehicle     -4.254e-01  2.055e-01  -2.070 0.038420 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1118.03  on 999  degrees of freedom
Residual deviance:  976.76  on 988  degrees of freedom
AIC: 1000.8

Number of Fisher Scoring iterations: 5
```

Fuente: Elaboración propia

La técnica empleada consiste en la comparación de medidas de bondad del ajuste de los diferentes modelos. Es decir, se compara el ajuste obtenido por cada uno de los modelos desarrollados con las diferentes variables partiendo de criterios de información como el AIC y Deviance. Observando la evolución de los valores en cada una de las métricas mencionadas, se selecciona como mejor modelo aquel que minimiza los valores de los criterios de información mencionados.

Como se puede observar en la Figura 25, se obtiene un grado AIC que asciende a 1.000,8, el cual se ha reducido con respecto al modelo con todas las variables el cual obtiene un AIC igual a 1.029,9. Por otro lado, se posee la Deviance, la medida Null Deviance hace referencia al modelo con la constante y el Residual deviance al modelo ajustado. En la Figura 25 se puede observar una reducción igual a 141,27 (1.118,03-976,76) con una pérdida de 11 grados de libertad, sobre la cual se puede interpretar como una disminución significativa respecto al modelo inicial que

arrojaba una reducción igual a 156,14 (1.118,03-961,89) con una pérdida de 33 grados de libertad. El valor obtenido en ambas métricas corresponde al mejor valor obtenido por el conjunto de las comprobaciones realizadas, es decir, corresponde al mejor modelo.

A partir de las diferentes comprobaciones sobre los factores realizadas con el modelo GLM de selección de variables, se identifican aquellas variables que explican en mayor grado el fraude reportado en la cartera. A través de la Figura 25, se puede realizar una clasificación de la capacidad explicativa de las variables:

4 variables relevantes: aficiones (deportes extremos, actividades al aire libre, actividades relajadas) y severidad del siniestro.

4 variables con cierto poder predictivo: cobertura paraguas, aficiones (deportes), marca del vehículo (vehículo de gama media) e estado civil (soltero).

2 variables menor nivel predictivo: hora del siniestro y reclamación de vehículo.

Las variables anteriores son aquellas que reducen los criterios de información del modelo y por ende poseen una mayor capacidad predictiva sobre los valores recogidos por la variable de estudio. En definitiva, las variables con las que se realizarán las predicciones son aquellas contenidas en estas categorías.

Con el fin de obtener una mejora de la capacidad predictiva del modelo se ha llevado a cabo una normalización de las variables numéricas que mayor dispersión presentan, como es el caso de las variables referentes a la “cobertura paraguas” y a la “cuantía de reclamación de vehículo”. Así mismo, la variable “hora del siniestro” se ha transformado en una variable categórica con el fin de reducir el espacio muestral, las clases contempladas por esta variable son “Noche”, “Mañana” y “Tarde”.

Este proceso se ha realizado ya que, al poseer una gran dispersión en sus valores, las variables mencionadas transmiten a los modelos una elevada volatilidad, afectado negativamente a las predicciones realizadas. Una vez se ha garantizado la correcta escala de las variables se procede a implementarlas en los modelos de predicción.

4. RESULTADOS Y COMPARATIVA DE TÉCNICAS

Tras el análisis del fraude en el sector asegurador, las técnicas predictivas y el estudio de las diferentes variables vinculadas a este evento, se procede a desarrollar los diferentes modelos de detección del fraude, identificando cuál de ellos es el más idóneo de cara a estimar este tipo de conductas. El objetivo de la modelización de los siniestros fraudulentos reside en la capacidad de predecir que puede suceder en el futuro bajo un umbral de probabilidad.

La idea de desarrollo mantenida por los modelos estadísticos de predicción consiste en la construcción de los modelos y entrenarlos con información conocida aplicando lo aprendido posteriormente para nuevos datos desconocidos. Es por ello por lo que, el análisis de información realizado por las nuevas técnicas de Machine Learning se basa en la división de la base de datos en dos categorías: Una primera parte destinada al entrenamiento del modelo y una segunda parte destinada a la verificación de la eficacia del modelo. De forma aleatoria, se han seleccionado el 70% de los datos de la muestra original con los cuales se realizará la construcción y entrenamiento de los modelos y el 30% restante de la muestra se emplea para comprobar la capacidad predictiva.

El objetivo de esta sección es que el modelo se ajuste bien a los datos de construcción y a partir de ellos sea capaz de realizar una predicción lo más similar posible a los datos de verificación. Es decir, se pretende desarrollar modelos que no solo realicen una buena interpretación de los datos de entrenamiento, sino que también gocen de una elevada capacidad predictiva.

La determinación de la capacidad predictiva del modelo se realiza mediante la comparación de los sucesos predichos por el modelo y los casos de fraude realmente ocurridos en la parte de la muestra de verificación.

De cara a verificar que la muestra no se ha distorsionado en gran medida tras la separación en los dos datasets, se genera la Tabla 15. Determinados estudios, como el realizado por Monserrat Guillén, et. Al (1999), hacen hincapié en la gran relación entre la separación equilibrada de la muestra y la capacidad de los algoritmos de aprendizaje para realizar estimaciones precisas.

Tabla 14. Comprobación no distorsión de la muestra

	Indicador fraude positivo	Indicador fraude negativo
Muestra original	75.3%	24.7%
Set entrenamiento	75.14%	24.86%
Set validación	75.67%	24.33%

Fuente: Elaboración propia

A partir de la Tabla 15, se puede observar que no se ha distorsionado la muestra, manteniendo balanceados los porcentajes de fraude negativo y positivo que conforman la variable dependiente en todos los conjuntos muestrales.

Una vez se ha verificado que las muestras generadas están idénticamente distribuidas que la muestra original se comienza a desarrollar los diferentes modelos de predicción de la variable dependiente.

4.1.- Resultados de entrenamiento y validación

A continuación, se procede a implementar las distintas técnicas de Machine Learning explicadas y referenciadas a lo largo del trabajo. Se comienza realizando el estudio por las metodologías más sencillas como el GLM y se finaliza con técnicas más complejas.

En cada uno de los epígrafes siguientes se hará referencia a la metodología empleada, la ejecución del algoritmo y los resultados obtenidos en la fase de entrenamiento y de validación de los modelos.

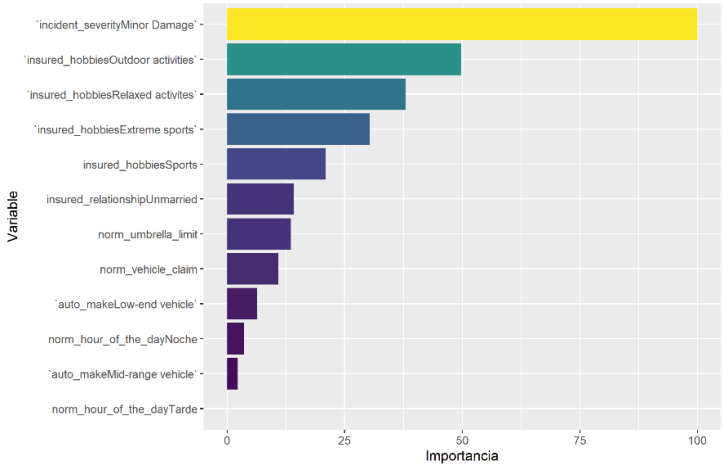
Los algoritmos de ML empleados en el presente trabajo poseen su propia forma de evaluar la importancia de los factores, posibilitando la comparación entre los diferentes modelos. El análisis en la etapa de entrenamiento el análisis se basa en la importancia que han obtenido las diferentes variables en la ejecución del modelo. Mediante una representación gráfica se permite identificar aquellas variables con mayor capacidad predictiva sobre los resultados.

Así mismo, en cada uno de los modelos se ha elaborado una matriz de confusión, a partir de la cual se puede identificar el grado de precisión que ha generado cada modelo en la predicción del fraude, tanto en los casos positivos como negativos.

4.1.1. Modelo GLM

Referente al modelo generalizado de regresión se emplea un modelo similar al de selección de variables. Como se ha mencionado la distribución seleccionada para la modelización es la Binomial, ya que la variable respuesta es dicotómica. En la etapa de implementación del modelo se ha empleado en R el algoritmo “glm” aplicando una función de liga binomial sobre el set de entrenamiento.

Figura 26. Importancia variables - Modelo GLM



Fuente: Elaboración propia

En la Figura 26, se representa la importancia relativa de cada variable sobre la capacidad predictiva del modelo de entrenamiento. Como se puede observar, las variables que mayor

importancia han adquirido son la severidad de siniestro (Daño menor) y la referente a los hobbies de asegurado (Actividades al aire libre y relajadas o sin esfuerzo físico).

Una vez se han obtenido los resultados sobre la base de datos de entrenamiento se aplica el modelo sobre la base de datos de validación, obteniendo así una predicción de la variable respuesta. La predicción obtenida es analizada para establecer la probabilidad de corte del modelo y realizar una mejor clasificación de los casos. Esta clasificación es cotejada con los resultados existentes en el set de validación, obteniendo la Tabla 15. Donde se realiza una matriz de confusión referente a los resultados obtenidos por el modelo.

Tabla 15. Matriz de confusión - Modelo GLM

		Predicción	
		Negativo	Positivo
Real	Negativo	219	8
	Positivo	60	13

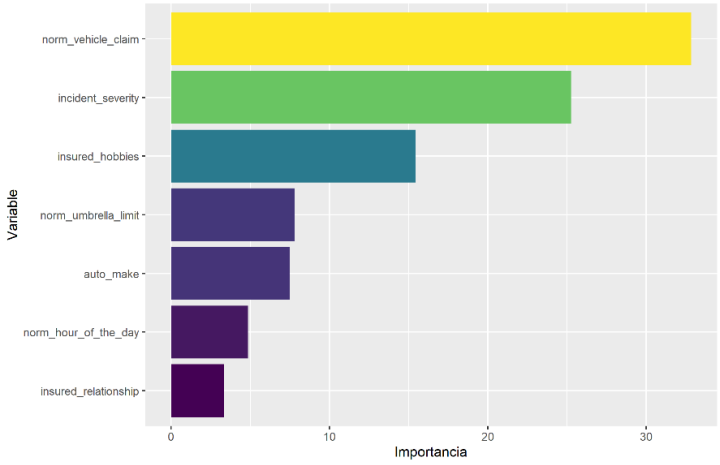
Fuente: Elaboración propia

Cuanto mayor número de casos se encuentren bien catalogados mejor será la predicción realizada por el modelo. Como se puede extraer de la Tabla 15, la mayor parte de los casos negativos han sido catalogados de forma correcta. No siendo así para los casos positivos de fraude, donde el modelo realiza una catalogación errónea de gran parte de los casos analizados.

4.1.2. Modelo Árbol de decisión

La técnica referente al Árbol de decisión se desarrolla en R mediante el algoritmo “rpart” con el método “class”. Esta función realiza un encasillamiento de las diferentes variables, en la fase se realiza sobre el set entrenamiento. A partir de su aplicación se obtienen los siguientes resultados.

Figura 27. Importancia variables - Árbol de decisión



Fuente: Elaboración propia

En la Figura 27, se representa la importancia relativa de cada variable sobre la capacidad predictiva del modelo de entrenamiento. Como se puede observar, las variables que mayor importancia han adquirido son la cuantía reclamada de vehículo, la severidad de siniestro y la referente a los hobbies de asegurado, todas ellas si la especificación de subcategorías. En este sentido cabe destacar la relevancia de la variable cuantía reclamada, la cual en el modelo anterior se había estimado con una importancia reducida.

Una vez se han obtenido los resultados sobre la BBDD de entrenamiento se aplica el modelo sobre la BBDD de validación, obteniendo así una predicción de la variable respuesta. La predicción obtenida es cotejada con los resultados existentes en el set de validación, obteniendo la Tabla 16. Donde se realiza una matriz de confusión referente a los resultados obtenidos por el modelo.

Tabla 16. Matriz de confusión - Árbol de decisión

		Predicción	
		Negativo	Positivo
Real	Negativo	206	21
	Positivo	59	14

Fuente: Elaboración propia

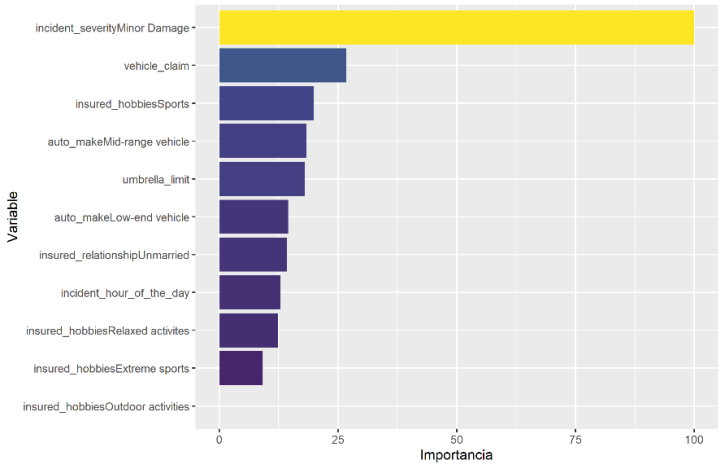
Como se puede extraer de la Tabla 16, la mayor parte de los casos negativos han sido catalogados de forma correcta. No siendo así para los casos positivos de fraude, donde el modelo realiza una catalogación errónea en su mayoría. Con respecto al modelo anterior incurre en un mayor error en los casos de fraude negativo y un aumento en un punto de los aciertos en los casos positivos de fraude, aunque este hecho no supone una gran mejora respecto al objetivo del estudio.

4.1.3. Random Forest

La técnica de Random Forest es muy similar a la anterior, consiste en la reiteración del algoritmo de clasificación de variables de tal forma que se obtienen una gran cantidad de árboles. Para el desarrollo de esta metodología en R se ha empleado el algoritmo "train" aplicando la metodología "ranger" sobre la BBDD de entrenamiento. Mediante este algoritmo se selecciona de manera aleatoria las variables a partir de las cuales se van a construir cada uno de los árboles, para ello se ha divido el set de datos de entrenamiento en 10 espacios muestrales.

En la Figura 28, se representa la importancia relativa de cada variable sobre la capacidad predictiva del modelo de entrenamiento. Como se puede observar, las variables que mayor importancia han adquirido son la severidad de siniestro (Daño menor), la cuantía del siniestro reclamada y la referente a los hobbies de asegurado (Sports). Siendo la variable de severidad aquella que posee una importancia casi total en el modelo.

Figura 28. Importancia variables - Modelo Random Forest



Fuente: Elaboración propia

Una vez se han obtenido los resultados sobre la BBDD de entrenamiento se aplica el modelo sobre la BBDD de validación, obteniendo así una predicción de la variable respuesta. La predicción obtenida es cotejada con los resultados existentes en el set de validación, obteniendo la Tabla 17. Donde se realiza una matriz de confusión referente a los resultados obtenidos por el modelo.

Tabla 17. Matriz de confusión - Random Forest

		Predicción	
		Negativo	Positivo
Real	Negativo	211	16
	Positivo	63	10

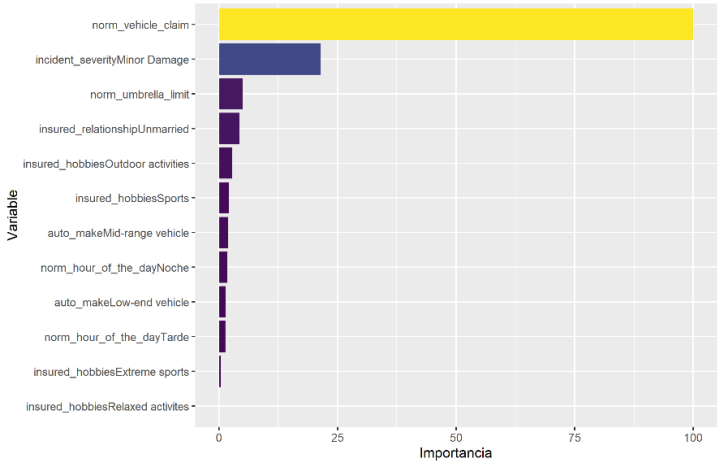
Fuente: Elaboración propia

Como se puede extraer de la Tabla 17, la mayor parte de los casos negativos han sido catalogados de forma correcta. No siendo así para los casos positivos de fraude, donde el modelo realiza una catalogación errónea de la mayor parte de los casos analizados. Con respecto al modelo anterior se puede identificar una mejora en la estimación de los casos negativos, mientras que para los casos positivos se obtiene un menor grado de ajuste.

4.1.3. Modelos Gradient Boosting

El modelo Gradient Boosting es desarrollado en R a partir del método “gbm” sobre el algoritmo train. El cual permite el desarrollo de árboles sencillos y la creación de los nuevos árboles aprendiendo de los errores cometidos sobre los árboles anteriores. Tras su aplicación sobre el set de entrenamiento se obtienen los siguientes resultados.

Figura 29. Importancia variables - Modelo GBM



Fuente: Elaboración propia

En la Figura 29, se representa la importancia relativa de cada variable sobre la capacidad predictiva del modelo. Como se puede observar, las variables que mayor importancia han adquirido son la cantidad reclamada de vehículo con una importancia casi total y la severidad de siniestro (Daño menor).

Una vez se han obtenido los resultados sobre la BBDD de entrenamiento se aplica el modelo sobre la BBDD de validación, obteniendo así una predicción de la variable respuesta. La predicción obtenida es cotejada con los resultados existentes en el set de validación, obteniendo la Tabla 18. Donde se realiza una matriz de confusión referente a los resultados obtenidos por el modelo.

Tabla 18. Matriz de confusión - GBM

		Predicción	
		Negativo	Positivo
Real	Negativo	210	17
	Positivo	60	13

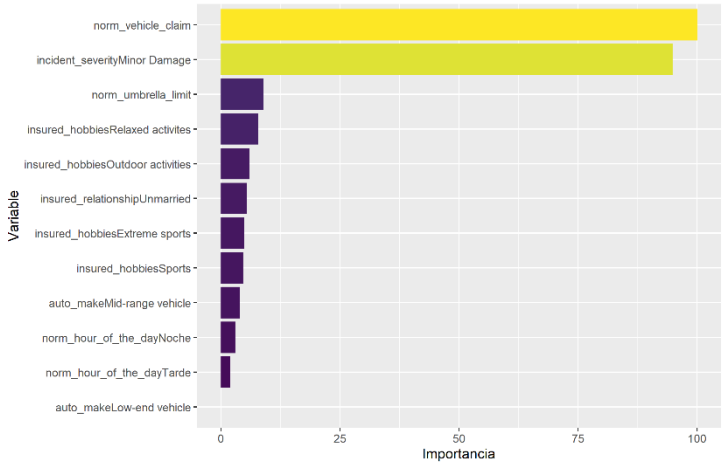
Fuente: Elaboración propia

Como se puede extraer de la Tabla 18, la mayor parte de los casos negativos han sido catalogados de forma correcta. Por el contrario, para los casos positivos de fraude, el modelo realiza una catalogación errónea de la mayor parte de los casos analizados. Con respecto al modelo anterior esta técnica aumenta la precisión en 3 puntos de la estimación de aquellos casos de fraude positivos sin apenas reducir la precisión en los casos negativos de fraude.

4.1.4. Modelo Extreme Gradient Boosting

La metodología Extreme Gradient Boosting supone una mejora en cuanto al ajuste sobre la muestra respecto a la metodología anterior. Esta técnica ha sido implementada en R mediante el algoritmo “train” con la metodología “xgb”. Tras su aplicación sobre el set de entrenamiento se extraen los siguientes resultados.

Figura 30. Importancia variables - Modelo XGB



Fuente: Elaboración propia

En la Figura 30, se representa la importancia relativa de cada variable sobre la capacidad predictiva del modelo. Como se puede observar, las variables que mayor importancia han adquirido son la cantidad reclamada de vehículo con una importancia casi total y la severidad de siniestro (Daño menor) en este caso adquiriendo una gran importancia también.

Una vez se han obtenido los resultados sobre la BBDD de entrenamiento se aplica el modelo sobre la BBDD de validación, obteniendo así una predicción de la variable respuesta. La predicción obtenida es cotejada con los resultados existentes en el set de validación, obteniendo la Tabla 19. Donde se realiza una matriz de confusión referente a los resultados obtenidos por el modelo.

Tabla 19. Matriz de confusión - XGB

		Predicción	
		Negativo	Positivo
Real	Negativo	215	12
	Positivo	62	11

Fuente: Elaboración propia

Como se puede extraer de la Tabla 19, para los casos positivos de fraude, el modelo realiza una catalogación errónea de la mayor parte de los casos analizados. Por el lado contrario encontramos los casos negativos de fraude, los cuales el modelo realiza una identificación muy ajustada. Este modelo con respecto a la técnica anterior mejora la precisión en la estimación de

los casos negativos de fraude, aunque reduce en un punto los aciertos obtenidos para aquellos casos de fraude positivo.

4.2.- Comparativa de efectividad de los modelos

Tras la implementación de los modelos sobre los diferentes sets de datos, se recopilan las medidas expuestas en el epígrafe 2.3.2 de este trabajo las cuales suponen una evaluación de los resultados generados. En primer lugar, se elabora una comparativa entre las consideradas técnicas modernas de Machine Learning y posteriormente, se comprará el mejor modelo con el GLM obtenido.

4.2.1. Selección mejor técnica de Machine Learning moderna

En la Tabla 20 se exponen de forma agregada las magnitudes seleccionadas para la realización del estudio, la cuales son: Accuracy, Kappa, TN y TP. A partir de la cual se pretende comparar la calidad de las predicciones generadas por cada una de las técnicas de Machine Learning modernas. Partiendo de esta comparación se obtienen conclusiones sobre el modelo que mejor predicción de los casos de fraude ha realizado.

Tabla 20. Medidas de comparación técnicas modernas ML

	AR	RF	GBM	XGB
Accuracy	0,7333	0,7348	0,7357	0,747
Kappa	0,1205	0,0086	0,1455	0,1196
TN	0,9074	0,9295	0,9251	0,9471
TP	0,1917	0,1369	0,178	0,1506

Fuente: Elaboración propia

Como se puede extraer de la Tabla 20 los diferentes modelos implementados sobre la BBDD no presentan grandes diferencias, generando una capacidad de predicción muy similar. En términos generales el modelo Extreme Gradient Boosting es el que mejores predicciones realiza, obteniendo una precisión igual 74,7% y la mayor tasa de aciertos en aquellos casos de fraude negativo 94,71%.

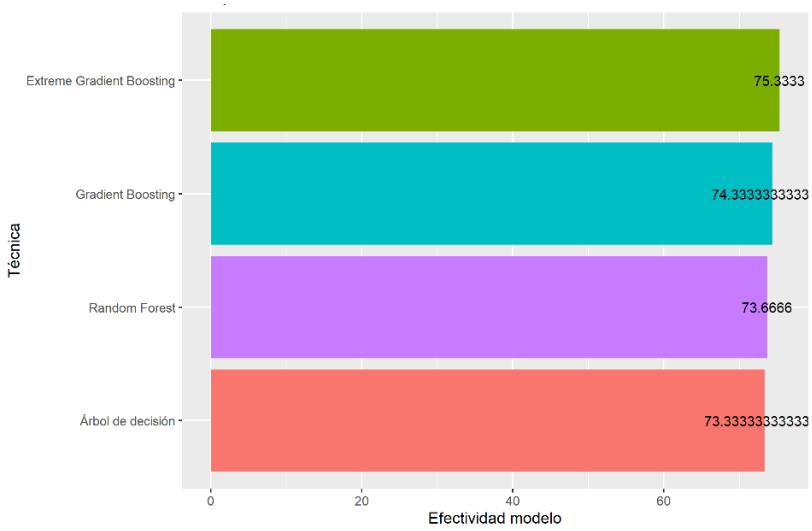
Respecto al resto de modelos, se puede identificar la técnica de árboles de decisión como aquella que obtiene una mayor tasa de aciertos sobre los casos de fraude positivos. La técnica de Gradient Boosting es la que mejor kappa posee, por lo que es el modelo que más se aproxima a los resultados esperados. No obstante, todas las kappas obtenidas por los modelos son lejanos a 1 por lo que se puede establecer que no generan resultados ajustados a los esperados. Por último, el modelo que peores predicciones realiza es el random forest ya que obtiene los peores coeficientes en la kappa y la clasificación de casos positivos de fraude.

Tomando como base las métricas expuestas se puede realizar una selección del modelo que será finalmente comparado con el modelo GLM. No obstante, para obtener una mayor

confianza en la elección realizada se expone la Figura 31, en la cual se recoge la efectividad total generada por los modelos en sus predicciones.

La efectividad total hace referencia a la calidad de las predicciones realizadas por el modelo de forma absoluta, es decir, cuantas predicciones del modelo coinciden con los resultados que se poseen en la base de datos. En la Figura 31 se representa la efectividad de los modelos mediante un diagrama de barras, en el cual se identifica la técnica Extreme Gradient Boosting como aquella que mayor efectividad ha generado sobre los casos de fraude. Esta técnica mejora en un 0,33 la probabilidad basal de la que se parte en un inicio, por la cual la probabilidad de acierto en caso de identificar todos los casos de fraude como negativos queda definida en un 75%.

Figura 31. Comparativa efectividad técnicas modernas ML



Fuente: Elaboración propia

En definitiva, basándose en los coeficientes generados y la Figura 31 se puede seleccionar el modelo Extreme Gradient Boosting como aquel que mejores predicciones genera. Es por ello por lo que este modelo será la técnica de machine learning moderna que se comprará con el modelo GLM desarrollado.

4.2.2. Comparación modelo seleccionado con GLM

Las métricas se exponen de forma agregada en la Tabla 21, a partir de la cual se pretende comparar la calidad de las predicciones generadas por la técnica de machine learning seleccionada y el modelo GLM. Partiendo de esta comparación se obtienen conclusiones sobre el modelo que mejor predicción de los casos de fraude ha obtenido.

Como se puede comprobar en la Tabla 32, no existe una gran diferencia entre los coeficientes mostrados entre los dos modelos estudiados. Se podría afirmar que el modelo que mejor evaluación obtiene es el GLM, ya que en los coeficientes kappa, TN y TP obtiene mejores resultados. Por el lado contrario se observa el coeficiente accuracy donde obtiene un resultado peor con muy poca diferencia respecto al XGB.

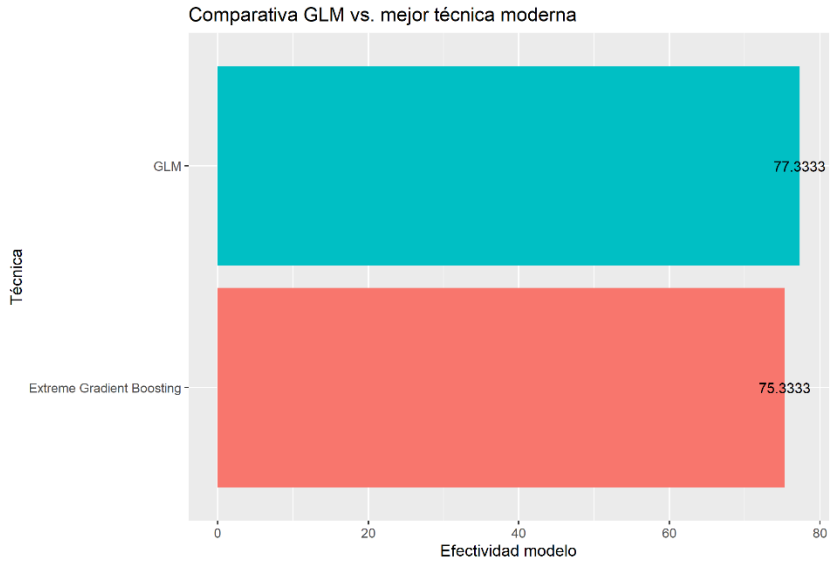
Tabla 21. Medidas de comparación modelo final

	GLM	XGB
Accuracy	0,74428	0,747
Kappa	0,13817	0,1196
TN	0,9648	0,9471
TP	0,178	0,1506

Fuente: Elaboración propia

Con el fin de obtener una base más sólida para la elección del mejor modelo de predicción del fraude, se procede a representar la efectividad absoluta obtenida por ambos modelos.

Figura 32. Comparativa efectividad modelo final



Fuente: Elaboración propia

Como se puede extraer de la Figura 32, la metodología que mayor precisión obtiene respecto a la predicción de casos de fraude es el GLM. Esta metodología obtiene una probabilidad de acierto 2.33 superior a la probabilidad basal la cual se situaba en el 75%. Esta mejora de la probabilidad de detección mejora la identificación de aquellos casos positivos de fraude y por tanto aumenta la capacidad de la compañía de tomar decisiones de mitigación respecto a este tipo de actividades.

5. CONCLUSIONES

La implementación de técnicas de detección del fraude no pretende sustituir la labor del tramitador. El desarrollo de modelos de predicción sobre los casos de fraude tiene como fin la obtención de una plataforma que favorezca la tramitación de siniestros sospechosos, facilitando el análisis de las reclamaciones y basando la toma de decisiones en los datos recopilados.

Para llevar a cabo una lucha efectiva contra el fraude, las compañías deben adoptar una estrategia firme de lucha frente a estos comportamientos. Deben ser conscientes de la trascendencia que poseen estas actividades ilícitas en el negocio asegurador y en la solvencia de la compañía, con el fin de poder tomar las medidas oportunas para erradicar estos comportamientos o minimizar el impacto de estos.

La finalidad del presente trabajo es el estudio de la detección del fraude en las compañías de seguros, partiendo del análisis teórico de este tipo de actividades y realizando una comparación de las diferentes técnicas de Machine Learning que se emplean en su identificación.

Tras analizar y comparar los diferentes modelos mediante un cuadro de métricas de evaluación se puede afirmar que, a pesar de los esfuerzos computacionales realizados, el modelo GLM es el que mejor estimación genera de las prácticas fraudulentas. No obstante, se observa una mejora en la predicción realizada por aquellas técnicas con algoritmos más sofisticados, como Gradient Boosting y Extreme Gradient Boosting.

La estimación generada por las diferentes técnicas puede venir explicada por el porcentaje de casos positivos y negativos que poseen los modelos para el entrenamiento. Al dividir la BBDD en entrenamiento y validación el modelo únicamente posee 175 observaciones para entender la estructuración de los casos positivos, mientras que para los casos de fraude negativo posee un total de 525 observaciones. El gran número de observaciones de casos negativos posibilita un mejor aprendizaje. Es por ello por lo que en las matrices de confusión de todos los modelos estudiados se obtienen predicciones más ajustadas sobre los casos negativos de fraude y peores estimaciones sobre los casos positivos.

Como se ha mencionado, la falta de observaciones puede ser una de las de las principales causas de la predicción errónea de los casos positivos de fraude. Es por ello por lo que se hace de vital importancia para las compañías el disponer de fuentes de datos completas y extensas, alimentando así la capacidad de aprendizaje de los modelos.

Realizando una lectura más profunda del modelo GLM, sobre la base de la información empleada se ha estimado que los factores más influyentes sobre la predicción del fraude son la severidad de siniestro (Daño menor) y el referente a los hobbies de asegurado (Actividades al aire libre, relajadas o sin esfuerzo físico y vinculadas al deporte). Siendo la severidad del siniestro aquella que mayor influencia posee sobre la aparición de fraude en las reclamaciones.

Se debe tener en cuenta que los modelos de detección desarrollados en este ámbito tienen que ir actualizándose de forma continua. Ya que el fraude es un fenómeno dinámico y con la incorporación de nuevas observaciones, pueden producirse modificaciones en referencia a las variables que participan en las actividades fraudulentas, tanto por parte de las características del siniestro como por las características de los asegurados.

Paralelamente a los aspectos técnicos, uno de los posibles caminos que puede seguir el estudio en este entorno, es el análisis en profundidad del fraude desde una perspectiva más próxima a los clientes, sus características y, sobre todo, las características de los siniestros ocurridos. Es decir, cuáles son los elementos siniestrales que mayor presencia tienen en las reclamaciones fraudulentas y cuál es su vinculación con los asegurados.

En definitiva, el estudio realizado indica que los modelos tradicionales como el GLM siguen siendo efectivos a la hora de realizar predicciones sobre ciertos aspectos, como es el caso de la detección del fraude. Sobre este tipo de modelos las compañías aseguradoras deberán realizar una valoración de la solución obtenida, verificando que el resultado sigue las pautas establecidas por la estrategia de control de fraude definida.

6. BIBLIOGRAFÍA

AXA España. (2020). *"VII Mapa AXA del Fraude en España".*

AXA España. (2021). *"VIII Mapa AXA del Fraude en España".*

Informe Nº 1598. ICEA. (2020). "*El Fraude al Seguro Español Año 2019*".

Línea Directa. (2020). "*V Barómetro Línea Directa*".

Ayuso, M., Guillén, M., y Artís, M. (1999). *"Técnicas cuantitativas para la detección del fraude en el seguro del automóvil".* Departamento de Econometría, Estadísticas y Economía Española Universitat de Barcelona.

Badal Valero, E., Sanjuán Díaz, A., y Segura Gisbert, J. (2020). *"Algoritmos de machine learning para la detección del fraude en el seguro de automóviles".* Anales del Instituto de Actuarios Españoles 4ª época, 26, 2020/23-46.

ASEPEG (2020). "*Glosario de Términos"*. Disponible En:

https://www.apeseg.org.pe/glosario-de-terminos/

RGA. (2017). *"Global Claims Fraud Survey".*

Insurance Europe. (2013). *"The impact of insurance fraud"*.

FRISS. (2020). *"Encuesta de fraude en seguros".*

Casares San José-Marti, I. (2013). *"El fraude en los seguros"*. Fundación Inade, Instituto Atlántico del Seguro.

Art. 248, 249 y 250. Código Penal. (1996) Ley Orgánica 10/1995, de 23 de noviembre. Jefatura del estado (BOE-A-1995-25444).

(2020). *"El fraude al seguro se ha duplicado en la última década".* La Vanguardia. Disponible en:

https://www.lavanguardia.com/seguros/coches/20200304/473963888339/fraude-engano-seguro-simulacion-perito-axa.html

Martínez de la Puente, F. *"El fraude en los seguros".* SegurosCEA. Disponible en:

https://www.seguroscea.es/blog/280-el-fraude-en-los-seguros

Instituto de Actuarios Españoles (2016). *"Fraude en el seguro". Revista Actuarios. Número 39. Otoño de 2016.*

FRISS (2020). "*Encuesta de fraude en seguros".*

Metz, J. (2021). *"Pandemic fires up insurance fraud, here's what to watch for".* Forbes. Disponible en:

https://www.forbes.com/advisor/car-insurance/pandemic-insurance-fraud/

Inese. (2019). *"Los intentos de fraude aumentan en los seguros particulares".* Disponible en:

https://www.inese.es/los-intentos-de-fraude-aumentan-en-los-seguros-particulares/

Meraviglia, A. (2016). *"¿Qué comunidad autónoma defrauda más al seguro?"*. Cinco Días. Disponible en:

https://cincodias.elpais.com/cincodias/2016/02/29/mercados/1456759516_656493.html

Inese. (2019). *"Las 5 tendencias en la lucha contra el fraude".* Disponible en:

https://www.inese.es/las-5-tendencias-en-la-lucha-contra-el-fraude-en-seguros-para-2019/

P. McCullagh. y J. A. Nelder. (1989). *"Generalized linear Models".* Second edition. Chapman and Hall.

Tse, Y.-K. (2009). *"NonLife Actuarial Models. Theory, Methods and Evaluation".* International Series on Actuarial Science. Cambrigde.

Dunn, P. K., y Smyth, G. K. (2018). "Generalized Linear Models with examples in R". Springer texts in statistics. Springer.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2017). *"An Introduction to Statistical Learning with applications in R"*. Springer texts in statistics. Springer.

Fernandes de Mello, R., y Antocelli Ponti, M. (2018). *"Machine Learning. A Practical Approach on the Statistical Learning Theory".* Springer.

Hidalgo Ruiz-Capillas, S (2014). "*Random Forests para detección de fraude en medios de pago"*. Trabajo Final de Máster. Universidad Autónoma de Madrid.

Nam Tran, Q., y Arabnia, H. (2015). *"Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology. Algorithms and Software Tools".* Morgan Kaufmann.

Hastie, T., Tibshirani, R., y Friedman, J. (2017). *"The Elements of Statistical Learning. Data Mining, Inference, and Prediction".* Springer series in statistics. Springer.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., y Van der Linde, A. (2013). *"The deviance information criterion: 12 years on".* Journal of the Royal Statistical Society. Disponible en:

http://pointer.esalq.usp.br/departamentos/lce/arquivos/aulas/2014/WORKSHOP_ON_BAYESIAN_BIOSTATISTICS/SpiegelhalterEtAl_TheDevianceInformationCriterion12YearsOn_JRSSB2014.pdf

SitioBigData. (2019). *"Árbol de decisión en Machine Learning".* Disponible en:

https://sitiobigdata.com/2019/12/14/arbol-de-decision-en-machine-learning-parte-1/

Avalos, F. y Pilco, V. (2020). *"Proyecto 3er parcial".* Disponible en:

https://rpubs.com/Avalos42/randomforest

Singh, A. (2020). *"4 Boosting Algorithms you should know – GBM, CGBoost, LightGBM & CatBoost".* Analytics Vidhya. Disponible en:

https://www.analyticsvidhya.com/blog/2020/02/4-boosting-algorithms-machine-learning/

Amat Rodrigo, J. (2017). *"Árboles de decisión, random forest, gradient boosting y C5.0".* Ciencia de datos. Disponible en:

https://www.cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_c50#Boosting

7. ANEXO

```
################################################################################
### CÓDIGO TFM: DETECCIÓN DEL FRAUDE MEDIANTE TÉCNICAS DE MACHINE LEARNING ###
################################################################################


###############
###Paquetes###
##############

  library(ggplot2)
  library(dplyr)
  library(RColorBrewer)
  library(plotly)
  library(gridExtra)
  library(tidyverse)
  library(rcompanion)
  library(lsr)
  library(InformationValue)
  library(lattice)
  library(caret)
  library(ISLR)
  library(rpart)
  library(rpart.plot)
  library(randomForest)
  library(MASS)
  library(gbm)
  library(Metrics)
  library(xgboost)
  library(party)


##################
###Lectura BBDD###
##################

  setwd("C:/Users/germa/OneDrive/Escritorio/Germán/Uc3m/TFM/BBDD")

  data_fraud <- read.csv (file="insurance_claims_V.csv", sep=";", dec=".")

  head(data_fraud)

  setwd("C:/Users/germa/OneDrive/Escritorio/Germán/Uc3m/TFM/BBDD")

  data_fraud_dist <- read.csv (file="insurance_claims_V_Distribution.csv", sep=";",

       dec=".")

  head(data_fraud_dist)


#########################
###Análisis univariante###
#########################

  #Medidas estadisticas de las variables

  str(data_fraud)
```

```
table_statistics <- summary(data_fraud)

table_statistics

summary(data_fraud_dist)


#Distribución de las variables continuas y discretas:

Asegurados=seq(1:1000)

a <- ggplot(data=data_fraud, aes(x=months_as_customer))+geom_histogram(
      aes(y=..density..),
      binwidth=density(data_fraud$months_as_customer)$bw,fill="deepskyblue3")+
      geom_density(linetype="dashed",color="red",alpha=0.4)+ylab("Nº Asegurados")+
      xlab("Meses de cartera")+ggtitle("Distribución meses en cartera")

b <- ggplot(data=data_fraud, aes(x=policy_annual_premium))+geom_histogram(
      aes(y=..density..),
      binwidth=density(data_fraud$policy_annual_premium)$bw,fill="deepskyblue3")+
      geom_density(linetype="dashed",color="red",alpha=0.4)+ylab("Nº Asegurados")+
      xlab("Importe de la prima")+ggtitle("Distribución importe de la prima")

Asegurados_1=seq(1:202)
umbrella <- data_fraud[data_fraud$umbrella_limit > 0,]

c <- ggplot(data=umbrella,aes(x=umbrella_limit))+geom_histogram(aes(y=..density..),
      binwidth=density(data_fraud$umbrella_limit)$bw,fill="deepskyblue3")+
      geom_density(linetype="dashed",color="red",alpha=0.4)+ylab("Nº Asegurados")+
      xlab("Importe cobertura paraguas")+ggtitle("Distribución cobertura paraguas")

d <- ggplot(data=data_fraud, aes(x=age))+geom_histogram(aes(y=..density..),
      binwidth=density(data_fraud$age)$bw,fill="deepskyblue3")+
      geom_density(linetype="dashed",color="red",alpha=0.4)+ylab("Nº Asegurados")+
      xlab("Edad")+ggtitle("Distribución edad del asegurado")

e <- ggplot(data=data_fraud, aes(x=incident_hour_of_the_day))+geom_histogram(
      aes(y=..density..), binwidth=density(data_fraud$incident_hour_of_the_day)$bw,
      fill="deepskyblue3")+
      geom_density(linetype="dashed",color="red",alpha=0.4)+ylab("Nº Asegurados")+
      xlab("Hora del siniestro")+ggtitle("Distribución hora siniestro")

f <- ggplot(data=data_fraud, aes(x=number_of_vehicles_involved))+geom_histogram(
      aes(y=..density..),
      binwidth=density(data_fraud$number_of_vehicles_involved)$bw,
      fill="deepskyblue3")+
      ylab("Nº Asegurados")+xlab("Nº vehículos involucrados")+
      ggtitle("Distribución Nº de veh. involucrados")

g <- ggplot(data=data_fraud, aes(x=witnesses))+geom_histogram(aes(y=..density..),
      binwidth=density(data_fraud$witnesses)$bw,fill="deepskyblue3")+
      ylab("Nº Asegurados")+xlab("Nº de testigos")+
      ggtitle("Distribución Nº de testigos")

h <- ggplot(data=data_fraud, aes(x=auto_year))+geom_histogram(aes(y=..density..),
      binwidth=density(data_fraud$auto_year)$bw,fill="deepskyblue3")+
      geom_density(linetype="dashed",color="red",alpha=0.4)+ylab("Nº Asegurados")+
      xlab("Año de venta del vehículo")+
      ggtitle("Distribución año del vehiculo siniestrado")
```

```
tiff("Dist_v_continuas_1.tiff", height = 5.44, width = 8, units = 'in',
      compression = "lzw", res =300)

  grid.arrange(a, b, c, d, e, f, g, h, ncol=2)

dev.off()

i <- ggplot(data=data_fraud, aes(x=injury_claim))+geom_histogram(
      aes(y=..density..),
      binwidth=density(data_fraud$injury_claim)$bw,fill="deepskyblue3")+
      geom_density (fill="red",
      alpha=0.4)+ylab("Nº Asegurados")+xlab("Importe reclamaciones por lesiones")+
      ggtitle("Distribución reclamaciones de lesiones")

j <- ggplot(data=data_fraud, aes(x=property_claim))+geom_histogram(
      aes(y=..density..),
      binwidth=density(data_fraud$property_claim)$bw,fill="deepskyblue3")+
      geom_density(fill="red",
      alpha=0.4)+ylab("Nº Asegurados")+xlab("Importe reclamaciones de propiedad")+
      ggtitle("Distribución reclamaciones de propiedad")

k <- ggplot(data=data_fraud, aes(x=vehicle_claim))+geom_histogram(
      aes(y=..density..),
      binwidth=density(data_fraud$vehicle_claim)$bw,fill="deepskyblue3")+
      geom_density(fill="red",
      alpha=0.4)+ylab("Nº Asegurados")+xlab("Importe reclamaciones de vehículo")+
      ggtitle("Distribución reclamaciones de vehículo")

l <- ggplot(data=data_fraud, aes(x=total_claim_amount))+geom_histogram(
      aes(y=..density..),binwidth=density
      (data_fraud$total_claim_amount)$bw,fill="deepskyblue3")+
      geom_density(fill="red",alpha=0.4)+
      ylab("Nº Asegurados")+  xlab("Importe de las reclamaciones totales")+
      ggtitle("Dist. importe total de las reclamaciones")

tiff("Dist_v_continuas_2.tiff", height = 5.44, width = 8, units = 'in',
      compression = "lzw", res =300)

  grid.arrange(i, j, k, l, ncol=2)

dev.off()


#Distribución de las variables cualitativas:

tiff("Dist_Variable_cualitativa.tiff", height = 5.44, width = 8, units = 'in',
      compression = "lzw", res =300)

  ggplot(data=data_fraud_dist, aes(x=category, y=Number, fill= Variables, Pos))+
      geom_bar(stat="identity")+ facet_wrap(~Variables, scales="free") +
      ylab("Nº Asegurados")+
      xlab("Categorías variables cualitativas") +
      ggtitle("Distribución variables cualitativas") +
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))

dev.off()
```

```
#Distribución variable de estudio:

Porcentaje <- as.numeric(prop.table(table(data_fraud$fraud_reported))*100)
Etiqueta <- c("Fraude negativo", "Fraude positivo")
Ind_fraud <- paste(Etiqueta, Porcentaje, "%", sep=" ")
Ind_fraud <- as.data.frame(Ind_fraud)

tiff("Dist_Variable_respuesta.tiff", height = 5.44, width = 8, units = 'in',
     compression = "lzw", res =300)

  ggplot(data= Ind_fraud, aes(x = "", y = Porcentaje, fill = Etiqueta)) +
      geom_bar(width = 1, stat = "identity") + coord_polar(theta = "y",
      start = 0) +
      ylab("Porcentaje de reclamaciones")+ xlab("")+
      ggtitle("Distribución del fraude reportado en la cartera")

dev.off()


#########################
###Análisis bivariante###
#########################

#Análisis bivariante de las variables Continuas y discretas:

m <- ggplot(data=data_fraud, aes(fraud_reported, months_as_customer))+
      geom_boxplot(fill=3:4)+
      ylab("Meses como asegurado")+xlab("Fraude reportado")+
      ggtitle("Diagrama meses aseg.")

n <- ggplot(data=data_fraud, aes(fraud_reported, policy_annual_premium))+
      geom_boxplot(fill=3:4)+
      ylab("Cuantía de la prima")+xlab("Fraude reportado")+
      ggtitle("Diagrama cuantía de la prima")


o <- ggplot(data=umbrella, aes(fraud_reported, umbrella_limit))+
      geom_boxplot(fill=3:4)+
      ylab("Cuantía cobertura paraguas")+xlab("Fraude reportado")+
      ggtitle("Diagrama cobertura paraguas")

p <- ggplot(data=data_fraud, aes(fraud_reported, age))+geom_boxplot(fill=3:4)+
      ylab("Edad del asegurado")+xlab("Fraude reportado")+
      ggtitle("Diagrama edad aseg.")

tiff("Caja_v_continuas_1.tiff", height = 5.44, width = 8, units = 'in',
      compression = "lzw", res =300)

grid.arrange(m, n, o, p, ncol=2)

dev.off()

q <- ggplot(data=data_fraud, aes(fraud_reported, incident_hour_of_the_day))+
      geom_boxplot(fill=3:4)+
      ylab("Hora siniestro")+xlab("Fraude reportado")+
      ggtitle("Diagrama hora siniestro")

r <- ggplot(data=data_fraud, aes(fraud_reported, number_of_vehicles_involved))+
      geom_boxplot(fill=3:4)+
      ylab("Nº vehículos involucrados")+xlab("Fraude reportado")+
      ggtitle("Diagrama Nº de veh. involucrados")
```

```
s <- ggplot(data=data_fraud, aes(fraud_reported, witnesses))+
      geom_boxplot(fill=3:4)+
      ylab("Nº de testigos")+xlab("Fraude reportado")+
      ggtitle("Diagrama Nº de testigos")

t <- ggplot(data=data_fraud, aes(fraud_reported, auto_year))+
      geom_boxplot(fill=3:4)+
      ylab("Año de venta del vehículo")+xlab("Fraude reportado")+
      ggtitle("Diagrama año veh. siniestrado")

tiff("Caja_v_continuas_2.tiff", height = 5.44, width = 8, units = 'in',
      compression = "lzw", res =300)

  grid.arrange(q, r, s, t, ncol=2)

dev.off()

u <- ggplot(data=data_fraud, aes(fraud_reported, injury_claim))+
      geom_boxplot(fill=3:4)+
      ylab("Importe reclamaciónes por lesiones")+xlab("Fraude reportado")+
      ggtitle("Diagrama reclamaciones de lesiones")

v <- ggplot(data=data_fraud, aes(fraud_reported, property_claim))+
      geom_boxplot(fill=3:4)+
      ylab("Importe reclamaciónes de propiedad")+xlab("Fraude reportado")+
      ggtitle("Diagrama reclamaciones de propiedad")

w <- ggplot(data=data_fraud, aes(fraud_reported, vehicle_claim))+
      geom_boxplot(fill=3:4)+
      ylab("Importe reclamaciónes de vehículos")+xlab("Fraude reportado")+
      ggtitle("Diagrama reclamaciones de vehículos")

x <- ggplot(data=data_fraud, aes(fraud_reported, total_claim_amount))+
      geom_boxplot(fill=3:4)+
      ylab("Importe reclamaciones totales")+xlab("Fraude reportado")+
      ggtitle("Diagrama importe total reclamaciones")

tiff("Caja_v_continuas_3.tiff", height = 5.44, width = 8, units = 'in',
      compression = "lzw", res =300)

  grid.arrange(u, v, w, x, ncol=2)

dev.off()


#Análisis bivariante de las variables cualitativas:

aa <- ggplot(data=data_fraud, aes(x=insured_education_level,
      fill=fraud_reported)) +
      geom_bar(position = "dodge")+ ggtitle("Clasificación nivel de educación") +
      labs( x = "Nivel de educación",y = "Nº de asegurados",
      fill = "Fraude reportado")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))

ab <- ggplot(data=data_fraud, aes(x=insured_occupation, fill=fraud_reported)) +
      geom_bar(position = "dodge")+ggtitle("Clasificación profesión") +
      labs( x = "Profesión",y = "Nº de asegurados", fill = "Fraude reportado")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))
```

```
ac <- ggplot(data=data_fraud, aes(x=insured_hobbies, fill=fraud_reported)) +
      geom_bar(position = "dodge")+ggtitle("Clasificación hobbies") +
      labs( x = "Hobbies",y = "Nº de asegurados", fill = "Fraude reportado")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))

ad <- ggplot(data=data_fraud, aes(x=insured_relationship, fill=fraud_reported)) +
      geom_bar(position = "dodge")+ggtitle("Clasificación estado civil")+
      labs( x = "Estado civil",y = "Nº de asegurados", fill = "Fraude reportado")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))

tiff("Caja_v_cualit_1.tiff", height = 5.44, width = 8, units = 'in',
      compression = "lzw", res =300)

  grid.arrange(aa, ab, ac, ad, ncol=2)

dev.off()

ae <- ggplot(data=data_fraud, aes(x=incident_type, fill=fraud_reported)) +
      geom_bar(position = "dodge")+ ggtitle("Clasificación tipo de siniestro") +
      labs( x = "Tipo de siniestro",y = "Nº de asegurados",
      fill = "Fraude reportado")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))

af <- ggplot(data=data_fraud, aes(x=collision_type, fill=fraud_reported)) +
      geom_bar(position = "dodge")+ ggtitle("Clasificación tipo de colisión") +
      labs( x = "Tipo de colisión",y = "Nº de asegurados",
      fill = "Fraude reportado")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))

ag <- ggplot(data=data_fraud, aes(x=incident_severity, fill=fraud_reported)) +
      geom_bar(position = "dodge")+
      ggtitle("Clasificación severidad del siniestro") +
      labs( x = "Severidad del siniestro",y = "Nº de asegurados",
      fill = "Fraude reportado")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))

ah <- ggplot(data=data_fraud, aes(x=authorities_contacted, fill=fraud_reported)) +
      geom_bar(position = "dodge")+
      ggtitle("Clasificación autoridades contactadas") +
      labs( x = "Autoridades contactadas",y = "Nº de asegurados",
      fill = "Fraude reportado")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))

tiff("Caja_v_cualit_2.tiff", height = 5.44, width = 8, units = 'in',
      compression = "lzw", res =300)

  grid.arrange(ae, af, ag, ah, ncol=2)

dev.off()

ai <- ggplot(data=data_fraud, aes(x=auto_make, fill=fraud_reported)) +
      geom_bar(position = "dodge")+ ggtitle("Clasificación gama del vehículo") +
      labs( x = "Gama del vehículo",y = "Nº de asegurados",
      fill = "Fraude reportado")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))

aj <- ggplot(data=data_fraud, aes(x=auto_model, fill=fraud_reported)) +
      geom_bar(position = "dodge")+ggtitle("Clasificación segmento del vehículo") +
      labs( x = "Segmento del vehículo",y = "Nº de asegurados",
      fill = "Fraude reportado")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1, vjust = 1))
```

```
tiff("Caja_v_cualit_3.tiff", height = 5.44, width = 8, units = 'in',
      compression = "lzw", res =300)

  grid.arrange(ai, aj, ncol=2, nrow=2)

dev.off()


########################################
###Estudio de correlación y covarianza###
########################################

#Variables numéricas

Var_num <- c(1, 2, 3, 4, 16, 17, 18, 19)

data_fraud_num <- data_fraud[Var_num]

table_num <- cor(data_fraud_num)

head(table_num)


#Variables cualitativas (V de Cramer)

#Función para obtener chi-cuadrado p-value y V de Cramer

vari <- c(5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 20, 21)

data_fraud_cual <- as.data.frame(data_fraud[,vari])

func = function(x,y) {
      tbl = data_fraud_cual %>% select(x,y) %>% table()
      chisq_pval = round(chisq.test(tbl)$p.value, 4)
      cramV = round(cramersV(tbl), 4)
      data.frame(x, y, chisq_pval, cramV) }

data_fraud_comb = data.frame(t(combn(sort(names(data_fraud_cual)), 2)),
                  stringsAsFactors = F)

data_fraud_res = map2_df(data_fraud_comb$X1, data_fraud_comb$X2, func)

#Tabla V de Cramer:

head(data_fraud_res)

#Gráfico V de Cramer:

tiff("V_Cramer.tiff", height = 5.44, width = 8, units = 'in',
     compression = "lzw", res =300)

data_fraud_res %>%
      ggplot(aes(x,y,fill=chisq_pval))+ geom_tile()+ geom_text(
      aes(x,y,label=cramV),
      check_overlap = TRUE,size=2.5) + scale_fill_gradient(low="red",
      high="green")+
      theme(axis.text.x = element_text(angle = 20, size = 6,hjust = 1,
      vjust = 1)) +
      ggtitle("Correlación variables cualitativas V de Cramer") +
```

```
       labs( x = "",y = "",
       fill = "P-value Chi-cuadrado")

  dev.off()


############################
###Selección de variables###
############################

  #Presentación modelos GLM de selección de variables:

  modelo_general <- glm(formula = Fraud_report ~
                        months_as_customer+policy_annual_premium+umbrella_limit+age+
                        insured_education_level+insured_occupation+insured_hobbies+
                        insured_relationship+
                        incident_hour_of_the_day+collision_type+incident_type+
                        incident_severity+
                        authorities_contacted+number_of_vehicles_involved+witnesses+
                        injury_claim+
                        property_claim+vehicle_claim+total_claim_amount+auto_make+
                        auto_model+auto_year,
                        data=data_fraud, family = binomial("logit"))

  summary(modelo_general)

  #Modelo 1: sin variables que en su conjunto están por encima del 75%
              #(insured_education_level,insured_occupation, property_claim,auto_year)

    #Se quitan las varibles collision_type y total_claim_amount por contener NA.

  modelo_1 <- glm(formula = Fraud_report ~
                  months_as_customer+policy_annual_premium+umbrella_limit+age+
                  insured_hobbies+insured_relationship+
                  incident_hour_of_the_day+incident_type+incident_severity+
                  authorities_contacted+number_of_vehicles_involved+witnesses+
                  injury_claim+
                  vehicle_claim+auto_make+auto_model,
                  data=data_fraud, family = binomial("logit"))

  summary(modelo_1)

  #Modelo 2: sin variables que en su conjunto están por encima del 70%
              #(policy_annual_premium,authorities_contacted,auto_model)

  modelo_2 <- glm(formula = Fraud_report ~
                  months_as_customer+umbrella_limit+age+
                  insured_hobbies+insured_relationship+
                  incident_hour_of_the_day+incident_type+incident_severity+
                  number_of_vehicles_involved+witnesses+injury_claim+
                  vehicle_claim+auto_make,
                  data=data_fraud, family = binomial("logit"))

  summary(modelo_2)

  #Modelo 3: sin variables que en su conjunto están por encima del 20%
              #(months_as_customer, age,incident_type, number_of_vehicles_involved, i
njury_claim)

  modelo_3 <- glm(formula = Fraud_report ~
                  umbrella_limit+
                  insured_hobbies+insured_relationship+
```

```
                incident_hour_of_the_day+incident_severity+
                witnesses+
                vehicle_claim+auto_make,
                data=data_fraud, family = binomial("logit"))

summary(modelo_3)

#Modelo 4: sin variables que en su conjunto están por encima del 20%
           #(incident_hour_of_the_day,witnesses,vehicle_claim)


modelo_4 <- glm(formula = Fraud_report ~
                umbrella_limit+
                insured_hobbies+insured_relationship+
                incident_severity+incident_hour_of_the_day+
                vehicle_claim+auto_make,
                data=data_fraud, family = binomial("logit"))

summary(modelo_4)


#Tratamiento variables seleccionadas(Normalización):

data_fraud$norm_umbrella_limit <-(data_fraud$umbrella_limit-mean(
                                  data_fraud$umbrella_limit))/
                                  sd(data_fraud$umbrella_limit)

data_fraud$norm_vehicle_claim <- (data_fraud$vehicle_claim-mean(
                                  data_fraud$vehicle_claim))/
                                  sd(data_fraud$vehicle_claim)

data_fraud$norm_hour_of_the_day <-if_else(data_fraud$incident_hour_of_the_day>=0&
                                  data_fraud$incident_hour_of_the_day<=7,"Noche",
                                  if_else(data_fraud$incident_hour_of_the_day>7&
                                  data_fraud$incident_hour_of_the_day<=13,"Mañana",
                                  if_else(data_fraud$incident_hour_of_the_day>13&
                                  data_fraud$incident_hour_of_the_day<=21, "Tarde",
                                  if_else(data_fraud$incident_hour_of_the_day>21&
                                  data_fraud$incident_hour_of_the_day<25, "Noche",
                                  "ERROR"))))


#########################
###División de la BBDD###
#########################

#División de la BBDD en sección de entrenamiento y de validación

set.seed(330)

train <- sample(nrow(data_fraud), 0.7*nrow(data_fraud), replace = FALSE)

M_TrainSet <- data_fraud[train,]
M_ValidSet <- data_fraud[-train,]

summary(M_TrainSet)
summary(M_ValidSet)

Porcentaje_train <- as.numeric(prop.table(table(M_TrainSet$fraud_reported))*100)
Etiqueta_train <- c("Fraude negativo", "Fraude positivo")
```

```
Ind_fraud_train <- paste(Etiqueta_train, Porcentaje_train, "%", sep=" ")
Ind_fraud_train <- as.data.frame(Ind_fraud_train)

Porcentaje_valid <- as.numeric(prop.table(table(M_ValidSet$fraud_reported))*100)
Etiqueta_valid <- c("Fraude negativo", "Fraude positivo")
Ind_fraud_valid <- paste(Etiqueta_valid, Porcentaje_valid, "%", sep=" ")
Ind_fraud_valid <- as.data.frame(Ind_fraud_valid)

#Validación que no se ha trastocado la BBDD

head(Ind_fraud)
head(Ind_fraud_train)
head(Ind_fraud_valid)



################################################
###Entrenamiento y validación de los modelos ###
################################################

#GLM

  #FASE ENTRENAMIENTO:

  set.seed(15095)

  modelo_glm <- glm(Fraud_report ~ norm_umbrella_limit+
                      insured_hobbies+insured_relationship+incident_severity+
                      norm_hour_of_the_day+norm_vehicle_claim+auto_make,
                      data = M_TrainSet,
                      family = binomial("logit"))

  summary(modelo_glm)
  print(modelo_glm)

  #Importancia de las variables

  Var_imp_glm <- varImp(modelo_glm)$importance
  Var_imp_glm <- Var_imp_glm %>%
                 rownames_to_column(var = "predictor")

  tiff("Var_imp_glm.tiff", height = 5.44, width = 8, units = 'in',
       compression = "lzw", res =300)

  ggplot(Var_imp_glm,aes(x=reorder(predictor,Overall), y= Overall, fill=Overall))+
         geom_col()+
         coord_flip()+scale_fill_viridis_c()+theme(legend.position = "none")+
         labs(x="Variable", y= "Importancia",
         title="Importancia variables - Modelo GLM")

  dev.off()


  #FASE VALIDACIÓN:

  #Verificación modelo

  prediction_glm <- predict(modelo_glm,M_ValidSet, type="response")
```

```
    #MEDIDAS DE DESEMPEÑO:

    #Matriz de confusión

    opcutoff <- optimalCutoff(M_ValidSet$Fraud_report, prediction_glm)

    head(opcutoff)

    error_clas <- misClassError(M_ValidSet$Fraud_report, prediction_glm,
                                threshold = opcutoff)

    head(error_clas)

    M_conf_glm <- confusionMatrix(M_ValidSet$Fraud_report, prediction_glm,
                                  threshold = opcutoff)

    head(M_conf_glm)

    #Evaluación del modelo

    val_glm <- sum(prediction_glm == M_ValidSet$fraud_reported)/
                   length(M_ValidSet$fraud_reported)*100

    head(val_glm) #Efectividad de predicción que arroja el modelo


#Árbol de decisión

    #FASE ENTRENAMIENTO:

    set.seed(15395)

    modelo_ar <- rpart(formula = fraud_reported ~ norm_umbrella_limit+auto_make+
                         insured_hobbies+insured_relationship+incident_severity+
                         norm_hour_of_the_day+
                         norm_vehicle_claim+auto_make,method="class",data =M_TrainSet)

    summary(modelo_ar)
    print(modelo_ar)

    tiff("Arbol_decision.tiff", height = 5.44, width = 8, units = 'in',
          compression = "lzw", res =300)

    rpart.plot(modelo_ar, type=2, extra=104, nn= TRUE, fallen.leaves = TRUE,
               faclen=4,varlen=8, shadow.col="gray")

    dev.off()

    #Importancia de las variables

    Var_imp_ar <- varImp(modelo_ar)
    head(Var_imp_ar)
    Var_imp_ar <- Var_imp_ar %>%
                  rownames_to_column(var = "predictor")

    tiff("Var_imp_ar.tiff", height = 5.44, width = 8, units = 'in',
          compression = "lzw", res =300)

    ggplot(Var_imp_ar, aes(x=reorder(predictor,Overall), y= Overall, fill=Overall))+
           geom_col()+
           coord_flip()+scale_fill_viridis_c()+theme(legend.position = "none")+
           labs(x="Variable", y= "Importancia",
```

```
          title="Importancia variables - Árbol de decisión")

    dev.off()



    #FASE VALIDACIÓN:

    #Verificación modelo

    prediction_ar <- predict(modelo_ar, M_ValidSet, type = "class")


    #MEDIDAS DE DESEMPEÑO:

    #Matriz de confusión

    conf_mat_ar <- confusionMatrix(table(prediction_ar, M_ValidSet$fraud_reported))

    M_conf_ar <- table(M_ValidSet$fraud_reported, prediction_ar)

    head(conf_mat_ar)

    head(M_conf_ar)

    #Evaluación del modelo

    val_ar <- sum(prediction_ar == M_ValidSet$fraud_reported)/
                  length(M_ValidSet$fraud_reported)*100

    head(val_ar)


#Random Forest

    #FASE ENTRENAMIENTO:

    set.seed(15895)

    modelo_rf<- train(fraud_reported~norm_umbrella_limit+
                        insured_hobbies+insured_relationship+incident_severity+
                        norm_hour_of_the_day+norm_vehicle_claim+auto_make,
                        data =M_TrainSet,
                        method = "ranger", trControl = trainControl("cv", number=10))

    modelo_rf

    getTrainPerf(modelo_rf)

    #Importancia de las variables

    Var_imp_rf <- varImp(modelo_rf)$importance
    head(Var_imp_rf)
    Var_imp_rf <- Var_imp_rf %>%
                  rownames_to_column(var = "predictor")

    tiff("Var_imp_rf.tiff", height = 5.44, width = 8, units = 'in',
```

```
        compression = "lzw", res =300)

    ggplot(Var_imp_rf, aes(x=reorder(predictor,Overall), y= Overall, fill=Overall)) +
           geom_col()+
           coord_flip()+scale_fill_viridis_c()+theme(legend.position = "none")+
           labs(x="Variable", y= "Importancia",
           title="Importancia variables - Modelo Random Forest")

    dev.off()


    #FASE VALIDACIÓN:

    #Verificación modelo

    prediction_rf <- predict(modelo_rf, M_ValidSet)


    #MEDIDAS DE DESEMPEÑO

    #Matriz de confusión

    M_conf_rf <- table(M_ValidSet$fraud_reported, prediction_rf)

    head(M_conf_rf)

    #Evaluación del modelo

    val_rf <- sum(prediction_rf == M_ValidSet$fraud_reported)/
                  length(M_ValidSet$fraud_reported)*100

    head(val_rf)


#Gradient Boosting

    #FASE ENTRENAMIENTO:

    set.seed(18395)

    modelo_GBM<- train(fraud_reported ~ norm_umbrella_limit+
                         insured_hobbies+insured_relationship+incident_severity+
                         norm_hour_of_the_day+norm_vehicle_claim+auto_make,
                         data =M_TrainSet,
                         method = "gbm", trControl = trainControl("cv", number = 10),
                         tuneGrid =expand.grid(interaction.depth = c(4,6,8),
                         n.trees=c(1,3,5,7,9)*100,
                         shrinkage = c(0.1) , n.minobsinnode = 20))

    getTrainPerf(modelo_GBM)

    #Importancia de las variables

    Var_imp_GBM <- varImp(modelo_GBM)$importance
    head(Var_imp_GBM)
    Var_imp_GBM <- Var_imp_GBM %>%
                   rownames_to_column(var = "predictor")

    tiff("Var_imp_GBM.tiff", height = 5.44, width = 8, units = 'in',
```

```
          compression = "lzw", res =300)

    ggplot(Var_imp_GBM, aes(x=reorder(predictor,Overall), y= Overall, fill=Overall))+
           geom_col()+
           coord_flip()+scale_fill_viridis_c()+theme(legend.position = "none")+
           labs(x="Variable", y= "Importancia",
           title="Importancia variables - Modelo GBM")

    dev.off()


    #FASE VALIDACIÓN:

    #Verificacion modelo

    prediction_GBM <- predict(modelo_GBM, M_ValidSet)


    #MEDIDAS DE DESEMPEÑO

    #Matriz de confusión

    M_conf_GBM <- table(M_ValidSet$fraud_reported, prediction_GBM)

    head(M_conf_GBM)

    #Evaluación del modelo

    val_GBM <- sum(prediction_GBM == M_ValidSet$fraud_reported)/
                   length(M_ValidSet$fraud_reported)*100

    head(val_GBM)


#Extreme Gradient Boosting

    #FASE ENTRENAMIENTO:

    set.seed(15895)

    modelo_XGB<- train(fraud_reported ~ norm_umbrella_limit+insured_hobbies+
                         insured_relationship+
                         incident_severity+norm_hour_of_the_day+norm_vehicle_claim+
                         auto_make,
                         data =M_TrainSet, method = "xgbTree",
                         trControl = trainControl("cv",number=10),
                         tuneGrid =expand.grid(max_depth = c(4,6,8), nrounds =
                         c(30,40,60,80,100,200,300),
                         eta = c(0.1, 0.12, 0.14, 0.16, 0.18, 0.2) ,
                         gamma=0,colsample_bytree= 1,min_child_weight= 1,
                         subsample= 1))

    modelo_XGB

    getTrainPerf(modelo_XGB)

    #Importancia de las variables

    Var_imp_XGB <- varImp(modelo_XGB)$importance
    head(Var_imp_XGB)
```

```
    Var_imp_XGB <- Var_imp_XGB %>%
                   rownames_to_column(var = "predictor")

    tiff("Var_imp_XGB.tiff", height = 5.44, width = 8, units = 'in',
         compression = "lzw", res =300)

    ggplot(Var_imp_XGB, aes(x=reorder(predictor,Overall), y= Overall, fill=Overall))+
           geom_col()+
           coord_flip()+scale_fill_viridis_c()+theme(legend.position = "none")+
           labs(x="Variable", y= "Importancia",
           title="Importancia variables - Modelo XGB")

    dev.off()


    #FASE VALIDACIÓN:

    #Verificacion modelo

    prediction_XGB <- predict(modelo_XGB, M_ValidSet)


    #MEDIDAS DE DESEMPEÑO

    #Matriz de confusión

    M_conf_XGB <- table(M_ValidSet$fraud_reported, prediction_XGB)

    head(M_conf_XGB)

    #Evaluación del modelo

    val_XGB <- sum(prediction_XGB == M_ValidSet$fraud_reported)/
                   length(M_ValidSet$fraud_reported)*100

    head(val_XGB)


##################################
###Comparación de los modelos ###
##################################

  #Comparación técnicas modernas

    tecnicas <- c("Árbol de decisión", "Random Forest", "Gradient Boosting",
                  "Extreme Gradient Boosting")
    test <- c(val_ar, val_rf, val_GBM, val_XGB)
    comparacion <- data.frame(tecnicas, test)
    head(comparacion)

    tiff("Comp_tec_1.tiff", height = 5.44, width = 8, units = 'in',
         compression = "lzw", res =300)

    ggplot(comparacion, aes(x=reorder(tecnicas, test), y = test))+geom_bar(
           stat = "identity",
           aes(fill=tecnicas))+geom_text(aes(label= test),check_overlap = TRUE,
           size=3.5)+
           labs(x="Técnica", y="Efectividad modelo", title =
           "Comparativa técnincas modernas ML")+coord_flip()+
```

```
          theme(legend.position = "none")

    dev.off()

#COmparación con GLM

    tecnicas_1 <- c("GLM", "Extreme Gradient Boosting")
    test_1 <- c(val_glm, val_XGB)
    comparacion_1 <- data.frame(tecnicas_1, test_1)
    head(comparacion_1)



    tiff("Comp_tec_2.tiff", height = 5.44, width = 8, units = 'in',
         compression = "lzw", res =300)

    ggplot(comparacion_1, aes(x=reorder(tecnicas_1, test_1), y = test_1))+
           geom_bar(stat = "identity",
           aes(fill=tecnicas_1))+geom_text(aes(label= test_1),check_overlap = TRUE,
           size=3.5)+
           labs(x="Técnica", y="Efectividad modelo", title =
           "Comparativa GLM vs. mejor técnica moderna")+coord_flip()+
           theme(legend.position = "none")

    dev.off()
```