UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA
CAMPUS D'ALCOI

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Politécnica Superior de Alcoy

Modelización del fraude en empresas aseguradoras

Trabajo Fin de Grado

Grado en Administración y Dirección de Empresas

AUTOR/A: Moreno Zapata, Ana

Tutor/a: Carracedo Garnateo, Patricia

Cotutor/a: Hervás Marín, David

CURSO ACADÉMICO: 2023/2024

Resumen

[ES]

La predicción del fraude en empresas aseguradoras es esencial para mitigar pérdidas y fortalecer la integridad del sistema. Por ello, la implementación de modelos estadísticos permitirá a las aseguradoras optimizar la detección de casos fraudulentos, mejorar la eficiencia en la gestión de reclamaciones y preservar la estabilidad financiera del sector. En concreto, este trabajo aplicará un modelo de regresión logística donde la variable a predecir será el fraude. Dicho modelo utilizará una muestra que proviene de una base de automóviles con variables relacionadas con reclamaciones, comportamiento de los asegurados y datos demográficos para ofrecer una evaluación proactiva del riesgo de fraude.

[VAL]

La predicció del frau en empreses asseguradores és essencial per mitigar pèrdues i enfortir la integritat del sistema. Per això, la implementació de models estadístics permetrà a les asseguradores optimitzar la detecció de casos fraudulents, millorar l'eficiència en la gestió de reclamacions i preservar l'estabilitat financera del sector. En concret, aquest treball aplicarà un model de regressió logística on la variable que s'ha de predir serà el frau. Aquest model utilitzarà una mostra que prové duna base d'automòbils amb variables relacionades amb reclamacions, comportament dels assegurats i dades demogràfiques per oferir una avaluació proactiva del risc de frau.

[EN]

Fraud prediction in insurance companies is essential to mitigate losses and strengthen the integrity of the system. Therefore, the implementation of statistical models will allow insurers to optimize the detection of fraudulent cases, improve efficiency in claims management and preserve the financial stability of the sector. Specifically, this work will apply a logistic regression model where the variable to be predicted will be fraud. This model will use a sample from an automobile database with variables related to claims, policyholder behavior and demographic data to provide a proactive assessment of fraud risk.

Palabras clave

Fraude; aseguradora; predicción; R.

Frau; asseguradora; predicció; R.

Fraud; insurance company; forecast; R

A mi familia, y en especial a Javi por estar ahí.

Agradecimientos

Este logro ha sido posible gracias al apoyo incondicional y la fe depositada en mí por aquellos que me rodean.

Quiero expresar mi más sincero agradecimiento a todas las personas que han confiado en mí. Sus palabras de aliento, respaldo y confianza han sido fundamentales en mi trayectoria. Aprecio enormemente su presencia constante, por creer en mis sueños y por impulsarme a alcanzar mis metas. Su confianza me inspira a esforzarme cada día y a perseverar con determinación.

En primer lugar, a mi compañero de viaje, Javi, por su incansable apoyo durante las interminables horas frente al ordenador, por alentarme a alcanzar lo que nunca pensé que fuera posible para mí.

En segundo lugar, quiero dar las gracias a mi familia. A mis padres y hermanos, por su ánimo y apoyo inquebrantable. A mi abuelita, cuya alegría por mis logros las siente como propia, y por enseñarme a ver la ciudad de Alcoy con nuevos ojos. A toda mi familia, por su constante respaldo y cercanía.

A Juani y Ernesto, que me han cuidado y animado como si fuera su propia hija, estoy profundamente agradecida por tenerlos en mi vida. ¡Gracias por acogerme como una más, por siempre creer en mí y por hacerme sentir amada y apoyada!

Tampoco puedo olvidar a la familia que elegí, quienes desde la distancia me han alentado y respaldado en esta aventura, demostrando que nunca es tarde para seguir creciendo. En especial a Isa, que a pesar de estar a más de miles de kilómetros de distancia te he sentido más cerca que nunca, tú siempre creíste en mí, y yo nunca lo supe ver.

Y por supuesto, un agradecimiento especial a mis compañeras de Alcoy durante estos años de desafíos, risas y alegrías: Aida, Dara, Cristian y Ximena. Gracias a ustedes, este viaje ha sido aún más gratificante gracias a vuestra compañia. Un pedacito de mí siempre quedará en Alcoy.

¡Gracias por creer en mí!

Índice

Índice de tablas

Índice de ilustraciones

1. Introducción

Una aseguradora es una entidad comercial dedicada a proporcionar cobertura financiera a individuos, empresas u otras organizaciones, mitigando los riesgos inherentes a diversas eventualidades. En esencia, actúa como intermediario entre el asegurado y el riesgo potencial al ofrecer un contrato de seguro. Este contrato, basado en la mutua confianza entre las partes involucradas, establece que, a cambio del pago de una prima periódica, la aseguradora se compromete a compensar al asegurado en caso de que ocurra el evento adverso cubierto por la póliza.

La relación entre el asegurado y la aseguradora se rige por los términos y condiciones detallados en el contrato de seguro, que define los límites de cobertura, las exclusiones y los procedimientos para presentar y gestionar reclamaciones. Además, la aseguradora evalúa los riesgos asociados con cada póliza y determina la prima adecuada en función de factores como la probabilidad de ocurrencia del evento asegurado, la magnitud del posible daño y la capacidad financiera del asegurado (Kim, 2016).

La función principal de una aseguradora es proporcionar seguridad financiera y protección contra pérdidas inesperadas, lo que contribuye a la estabilidad económica y social de los individuos y las comunidades. Al transferir el riesgo a cambio de una prima, las aseguradoras desempeñan un papel fundamental en la gestión del riesgo y la promoción de la tranquilidad y la confianza en el entorno empresarial y personal (Schuver S. S.; Schuver D. D. & Bakos T. L., 2006).

El funcionamiento de los seguros se basa en un proceso que involucra a los clientes y a las compañías aseguradoras de la siguiente manera (Mapfre, 2021):

- Prima: Los clientes pagan una prima periódica a la compañía aseguradora. Esta prima se determina considerando diversos factores, como el riesgo asociado al cliente, el tipo de cobertura requerida y otros datos relevantes, como la edad del asegurado.
- Cobertura: La compañía aseguradora se compromete a proporcionar cobertura en caso de que ocurra un siniestro o evento especificado en la póliza. Esta cobertura puede incluir la reparación de daños materiales, gastos médicos, compensación por pérdidas financieras, entre otros, dependiendo de los términos del contrato.
- Siniestro: Si ocurre un evento cubierto por la póliza, el cliente presenta una reclamación a la compañía aseguradora, proporcionando detalles sobre el incidente y los daños sufridos.
- Indemnización: La aseguradora evalúa la reclamación y, si se determina que es válida de acuerdo con los términos del contrato, procede a pagar una indemnización al cliente para cubrir los costos asociados con el siniestro.

Los seguros son instrumentos financieros diseñados para proporcionar protección y tranquilidad a individuos, familias y empresas frente a diversos riesgos y eventualidades. Estos diferentes tipos de seguros se adaptan a las necesidades y circunstancias particulares de los asegurados, brindando una variedad de opciones para la protección y la gestión de riesgos en la vida cotidiana (Asociación, 2007).Estos pueden ser clasificados en tres categorías principales, cada una dirigida a cubrir necesidades específicas (Asesorae, 2023):

- Seguros Personales:

- Enfocados en la protección del individuo y su familia:

    - **Seguro de vida:** Garantiza un respaldo financiero para los beneficiarios designados en caso de fallecimiento del asegurado.
    - **Seguro de salud:** Cubre los gastos médicos y hospitalarios del asegurado en caso de enfermedad o accidente.
    - **Seguro de accidentes:** Proporciona una compensación económica en caso de lesiones o invalidez ocasionadas por un accidente.
    - **Seguro de decesos:** Facilita la cobertura de los gastos funerarios y otros trámites relacionados con el fallecimiento del asegurado.
- Seguros de Daños o Patrimoniales:

- Orientados a proteger el patrimonio y los bienes materiales del asegurado:

    - **Seguro de coche:** Ofrece cobertura tanto para los daños propios del vehículo como para los ocasionados a terceros en caso de accidente.
    - **Seguro de hogar:** Protege la vivienda contra pérdidas por incendio, robo, daños por agua, entre otros riesgos.
    - **Seguro de viaje**: Brinda asistencia médica, repatriación y cobertura para otros imprevistos durante un viaje.
    - **Seguro de empresa:** Salvaguarda la actividad empresarial frente a riesgos como incendios, robos, responsabilidad civil, entre otros.
- Seguros de Prestación de Servicios:

- Dirigidos a proporcionar servicios adicionales en situaciones específicas:

    - **Seguro de asistencia en viaje:** Ofrece asistencia médica, legal y otros servicios en caso de contratiempos durante un viaje.
    - **Seguro de defensa jurídica:** Cubre los honorarios de abogados y gastos legales en caso de litigios.
    - **Seguro de hogar con asistencia:** Proporciona servicios de reparación de averías, limpieza del hogar, entre otros.

El sector de seguros del automóvil representa un mercado amplio y complejo, con una diversidad de actores y productos que compiten en un entorno altamente dinámico. Este ámbito se caracteriza por su elevado volumen de primas, una intensa competencia entre las compañías aseguradoras y la necesidad de gestionar el ciclo de vida del cliente de manera eficiente. Además, enfrenta diversos riesgos inherentes a la actividad, como accidentes, robos y fraudes. (Suárez Martínez, 2015)

La industria de seguros está experimentando cambios tecnológicos importantes que ofrecen tanto oportunidades como desafíos. Los consumidores desean una experiencia digital más satisfactoria y una relación más directa con sus aseguradoras. Las empresas nacidas en la era

digital tienen ventajas competitivas gracias a la automatización de procesos y modelos de distribución eficientes. La introducción de tecnologías como el internet de las cosas y el Big Data está transformando la recopilación y evaluación de datos de los clientes. Aunque el sector de seguros aún está rezagado en comparación con otras industrias en cuanto a experiencias en línea, un número significativo de consumidores utiliza canales en línea para interactuar con sus aseguradoras, incluyendo redes sociales (Convista Consulting Spain, 2015). El avance tecnológico está transformando profundamente este sector, con impacto significativo en áreas como la eficiencia operativa, el desarrollo de nuevos productos y la mejora de la experiencia del cliente. La conectividad de los vehículos también está en ascenso, permitiendo a las aseguradoras ofrecer servicios basados en datos que antes no eran posibles. Asimismo, el surgimiento de modelos de movilidad compartida está generando nuevos desafíos y oportunidades para adaptar los productos y servicios a estas nuevas formas de transporte.

El modelo de movilidad compartida ofrece un vasto campo de oportunidades valuado en 9 billones de dólares para el año 2025, según las proyecciones de Accenture Research. Esta era de transformación profunda afecta directamente al sector de seguros, el cual está concentrándose en aspectos como seguros integrados, la capacidad de prever riesgos y el impacto del comportamiento en la personalización de las pólizas (Garcia, 2023).

Sin embargo, el negocio del sector de seguros del automóvil enfrenta desafíos importantes. La rentabilidad se ve afectada por el aumento de los costos de los siniestros, la competencia agresiva y la creciente regulación. La necesidad de mantener la competitividad impulsa a las aseguradoras a ser más eficientes e innovadoras en la oferta de productos y servicios. El futuro de este sector estará marcado por la capacidad de adaptarse a las tendencias tecnológicas, la evolución de la conectividad y la integración de la movilidad compartida. Aquellas compañías aseguradoras que logren ajustarse a estos cambios y desarrollar soluciones innovadoras estarán en mejor posición para alcanzar el éxito en el futuro de este dinámico mercado.

Las industrias aseguradoras globalmente están compuestas por más de mil empresas, generando primas que suman billones de dólares anualmente. El fraude de seguros ocurre cuando individuos o entidades presentan reclamos falsos para obtener compensaciones o beneficios indebidos. Se estima que el costo total del fraude de seguros supera los cuarenta mil millones de dólares (Roy, R. & George, K. T., 2017). En el contexto español, según el informe "El fraude al seguro español. Año 2018" elaborado por Investigación Cooperativa entre Entidades Aseguradoras (ICEA), las aseguradoras en España pierden alrededor del 4% de las primas debido al fraude. Este informe también revela que se registraron 175.777 intentos de fraude en el año 2018, con un valor total de 584 millones de euros (Icea, 2018).

El fraude en automóviles se destaca como el más impactante en las aseguradoras, representando aproximadamente el 62.8% de los intentos de fraude detectados. Dentro de este ámbito, las simulaciones de accidentes y la exageración de daños en los vehículos son prácticas fraudulentas especialmente comunes. Las compañías lograron prevenir estafas por un total de 556,3 millones de euros (Arrillaga, 2023). Para combatir el fraude, las aseguradoras implementan diversas medidas, entre las que se incluyen el análisis de datos para identificar patrones sospechosos, investigaciones exhaustivas sobre reclamaciones potencialmente fraudulentas y la colaboración estrecha con las autoridades competentes para investigar y perseguir los casos de fraude.

El fraude en el sector asegurador conlleva diversas consecuencias, como el aumento de las primas para todos los clientes, las pérdidas económicas significativas para las aseguradoras y el deterioro de la confianza de los clientes en el sector. Por ello, es fundamental que las aseguradoras tomen medidas preventivas para protegerse contra este tipo de prácticas fraudulentas (Fastercapital, 2024).

El sector del automóvil enfrenta diversos tipos de fraude, que pueden ocurrir en distintas etapas del proceso asegurativo y de posesión del vehículo (Artís & Ayuso, 1999):

- Fraude en la solicitud de póliza:
    - Declaraciones falsas: El asegurado proporciona información falsa sobre datos relevantes como edad, historial de conducción o estado del vehículo.
    - Ocultación de información: El asegurado omite detalles importantes, como accidentes previos o condiciones preexistentes.
- Fraude en la reclamación:
    - Exageración de daños: El asegurado magnifica el alcance de los daños sufridos por su vehículo o por terceros.
    - Reclamaciones por siniestros inexistentes: Se presenta una reclamación por un siniestro que nunca ocurrió.
    - Robo fingido: El asegurado simula el robo de su vehículo cuando en realidad lo ha vendido o desmantelado.
- Fraude en la reparación:
    - Talleres fraudulentos: Talleres que facturan por reparaciones no realizadas o innecesarias.
    - Uso de piezas de repuesto no originales: Se utilizan piezas de repuesto de baja calidad en lugar de las originales.
    - Sobreprecio de reparaciones: Cobro excesivo por las reparaciones del vehículo.
- Fraude en la venta de vehículos:
    - Manipulación del cuentakilómetros: Se reduce el kilometraje real del vehículo para incrementar su valor de mercado.
    - Venta de vehículos con daños ocultos: Se comercializan vehículos con daños significativos no evidentes.
    - Falsificación de documentos: Documentos como el título de propiedad o la tarjeta de inspección técnica son falsificados.

Las consecuencias del fraude en el sector del automóvil tienen como consecuencia un aumento de las primas para todos los clientes, pérdidas económicas significativas para las aseguradoras y un deterioro en la confianza de los clientes en la industria aseguradora. El fraude en el sector de seguros afecta el costo total de los seguros para todos los clientes, ya que las compañías aseguradoras deben considerar el riesgo de fraude al calcular las primas. Como resultado, los costos operativos y las primas pueden aumentar para compensar las pérdidas causadas por el fraude. Para abordar este desafío, las compañías aseguradoras implementan una variedad de estrategias, como utilizar técnicas de análisis de datos para identificar patrones y

anomalías que puedan indicar posibles casos de fraude, realizar investigaciones exhaustivas sobre reclamaciones sospechosas para determinar su validez y detectar posibles actividades fraudulentas, y trabajar en estrecha colaboración con las autoridades pertinentes, como organismos reguladores y agencias de aplicación de la ley, para investigar y perseguir los casos de fraude de manera efectiva (Mendieta, 2024). Las medidas estratégicas para enfrentar el fraude en el sector pueden ser agrupadas en tres categorías fundamentales:

- Prevención:
    - Análisis de datos: Mediante el análisis de grandes volúmenes de información, las aseguradoras buscan identificar patrones que puedan indicar posibles casos de fraude.
    - Investigación de antecedentes: Las aseguradoras llevan a cabo investigaciones exhaustivas de los antecedentes de los solicitantes de pólizas para verificar la autenticidad de la información proporcionada.
    - Medidas de seguridad: Implementan diversas medidas de seguridad destinadas a dificultar la presentación de reclamaciones fraudulentas.
- Detección:
    - Investigaciones: Se realizan investigaciones detalladas sobre reclamaciones sospechosas para determinar su legitimidad.
    - Peritajes: Se recurre a peritos especializados para evaluar los daños y determinar su autenticidad.
    - Tecnologías de detección de fraude: Se emplean herramientas tecnológicas avanzadas, como la inteligencia artificial, para identificar posibles fraudes de manera más eficiente.
- Disuasión:
    - Penalización: Las aseguradoras aplican sanciones a aquellos asegurados que son hallados culpables de fraude, como la cancelación de la póliza o el aumento de la prima.
    - Denuncias: Se denuncian los casos de fraude a las autoridades competentes para que se tomen las medidas legales correspondientes.
    - Campañas de concienciación: Se llevan a cabo campañas de sensibilización dirigidas a informar a los clientes sobre las consecuencias del fraude y la importancia de mantener una conducta honesta.

Estas acciones son cruciales para proteger a los asegurados honestos, preservar la rentabilidad del negocio y salvaguardar la confianza del público en el sector asegurador (Schulz, 1994).

La detección y prevención del fraude son esenciales para las empresas y organizaciones por múltiples razones. Primero, ayudan a salvaguardar los activos financieros al minimizar las pérdidas derivadas de prácticas fraudulentas. Además, contribuyen a mantener la integridad y la reputación de la organización, generando confianza entre clientes, inversores y socios comerciales. Por último, promueven un entorno empresarial justo y equitativo al garantizar que todas las partes involucradas sean tratadas con imparcialidad y justicia

Los modelos descriptivos son herramientas valiosas en esta lucha, ya que permiten identificar patrones y anomalías en grandes conjuntos de datos que podrían indicar posibles casos de fraude. Su capacidad para crear perfiles de riesgo y automatizar el proceso de detección los convierte en aliados clave para las aseguradoras en la lucha contra el fraude. Estudios realizados demuestran la efectividad de los modelos descriptivos en la detección del fraude, revelando que pueden identificar hasta el 90% de los casos fraudulentos. No obstante, es importante tener en cuenta que estos modelos deben complementarse con otras medidas, como investigaciones humanas y colaboración con autoridades, para ser realmente eficaces en la prevención y detección del fraude (Ayuso, M., Guillén, M., & Artís, M., 1999).

En el contexto actual de los negocios, la detección y prevención del fraude se ha vuelto una tarea esencial para garantizar la sostenibilidad y la integridad de las empresas, especialmente en sectores expuestos a riesgos significativos como el de los seguros. La creciente sofisticación de las prácticas fraudulentas, junto con su impacto económico considerable, ha generado la necesidad imperativa de implementar estrategias más efectivas para combatir este fenómeno.

Numerosos estudios han abordado el fraude en la industria de seguros, destacando la importancia de mejorar las metodologías de detección. Por ejemplo, un estudio examina el fraude en seguros de automóviles y propone mejorar la metodología considerando errores en la información previa de reclamaciones, resaltando la relevancia de estas conclusiones para modelos de detección de fraude (Artís, M., Ayuso, M., & Guillén, M., 2002).

Otro estudio destaca la detección de fraudes en reclamaciones de seguros, evidenciando que una proporción considerable de reclamaciones legítimas puede contener errores, lo que subraya la necesidad de análisis costo-beneficio para estrategias de auditoría más efectivas (Sundarkumar & Ravi, 2015).

La influencia de las características de los vehículos en los accidentes automovilísticos y las reclamaciones de seguros ha sido objeto de estudio, revelando correlaciones significativas entre dichas características y la probabilidad y severidad de las reclamaciones (Wu & Li, 2020).

Este trabajo se enfoca en la utilización de herramientas de modelización estadística para predecir y gestionar el fraude en empresas aseguradoras, con especial atención al sector de los seguros de automóviles. El objetivo principal es desarrollar un marco analítico que permita a las aseguradoras mejorar la identificación de casos fraudulentos, optimizar la gestión de reclamaciones y salvaguardar la estabilidad financiera del sector.

La aplicación de modelos estadísticos, particularmente la regresión logística, ofrece un enfoque riguroso y sistemático para evaluar y predecir el riesgo de fraude. Este método permite incorporar una amplia variedad de variables relevantes, como datos relacionados con reclamaciones, historial de comportamiento de los asegurados, características demográficas y otros factores pertinentes, posibilitando una evaluación proactiva del fraude.

En este sentido, este estudio tiene como objetivo aplicar un modelo de regresión logística a una muestra de datos de seguros de automóviles, utilizando la variable de fraude como la variable a predecir y diversas variables explicativas para construir un modelo predictivo robusto y preciso. A través de este enfoque, se busca fortalecer las capacidades de las aseguradoras para identificar

y gestionar eficazmente los riesgos asociados al fraude, promoviendo así la integridad y estabilidad del sistema asegurador en su conjunto.

En resumen, este trabajo se sitúa en la intersección de la analítica de datos y la gestión de riesgos, proponiendo un enfoque metodológico innovador para abordar el desafío del fraude en empresas aseguradoras. Mediante la aplicación de modelos estadísticos avanzados, se busca ofrecer una herramienta práctica y efectiva para mejorar la capacidad de las organizaciones para detectar, prevenir y gestionar el fraude, con el fin último de proteger sus intereses financieros y fortalecer la confianza en el sector asegurador.

1.2. Contribución en la agenda 2030

La contribución de un modelo predictivo de fraude en el sector de seguros de automóviles puede tener varios impactos en relación con los objetivos de la Agenda 2030 para el Desarrollo Sostenible de las Naciones Unidas (Gobierno, 2018):

- Objetivo 8: Trabajo decente y crecimiento económico: La detección y prevención del fraude en seguros contribuye a la estabilidad económica al reducir las pérdidas financieras de las compañías aseguradoras. Esto puede fomentar un entorno más propicio para el crecimiento económico y la creación de empleo, al tiempo que protege los recursos financieros de las empresas y los consumidores.
- Objetivo 9: Industria, innovación e infraestructura: La implementación de modelos avanzados de análisis de datos, como el modelo de regresión logística desarrollado, promueve la innovación en la industria de seguros. Esta innovación puede conducir a una mayor eficiencia en la gestión de riesgos, procesos de suscripción más precisos y una mejor experiencia para los clientes.
- Objetivo 16: Paz, justicia e instituciones sólidas: La lucha contra el fraude en el sector de seguros fortalece las instituciones y promueve la justicia al garantizar que los recursos financieros se utilicen de manera transparente y equitativa. Al mejorar la integridad del sistema de seguros, se fomenta la confianza pública en las instituciones financieras y se reduce la posibilidad de prácticas fraudulentas que puedan socavar la estabilidad social.
- Objetivo 17: Alianzas para lograr los objetivos: La colaboración entre empresas de seguros, reguladores, investigadores y otras partes interesadas es fundamental para desarrollar y aplicar soluciones efectivas contra el fraude. El intercambio de conocimientos y la cooperación en la implementación de tecnologías avanzadas pueden fortalecer las capacidades de detección y prevención del fraude a nivel global, promoviendo así el desarrollo sostenible en todo el mundo.

En resumen, el uso de tecnologías analíticas avanzadas para combatir el fraude en seguros de automóviles puede contribuir significativamente a varios objetivos de la Agenda 2030, al promover un crecimiento económico sostenible, fomentar la innovación, fortalecer las instituciones y fomentar la colaboración entre diversas partes interesadas.

2. Datos

En el presente apartado, se lleva a cabo un análisis exhaustivo de los datos recopilados en el contexto del estudio sobre detección y prevención del fraude en el sector de seguros de automóviles. Estos datos constituyen una valiosa fuente de información que permite comprender en detalle los diferentes aspectos relacionados con los reclamos de seguros y los factores asociados al fraude.

El análisis se centra en un conjunto de 22 variables que abarcan diversas dimensiones relevantes, desde características demográficas de los conductores hasta detalles específicos de los siniestros y características de los vehículos involucrados. Esta variedad de variables proporciona una visión completa y detallada del fenómeno del fraude en el ámbito asegurador.

En primer lugar, se exploran variables numéricas como el importe del siniestro, la edad del conductor y la antigüedad de la póliza, entre otras. Este análisis revela patrones y tendencias significativas que pueden ser de utilidad para identificar posibles casos de fraude y optimizar la gestión de reclamaciones. Tras ello, se examinan variables categóricas, como el sexo del conductor, el tipo de documento utilizado y la forma de pago del seguro. Estas variables ofrecen información crucial sobre las características y comportamientos de los asegurados, lo que contribuye a enriquecer la comprensión del fenómeno del fraude y a diseñar estrategias más efectivas para su detección y prevención.

El análisis detallado de estos datos constituye un paso fundamental en el proceso de investigación, ya que proporciona información valiosa que servirá de base para el desarrollo de modelos predictivos y estrategias de gestión de riesgos destinadas a combatir el fraude en el sector de seguros de automóviles.

1. Sexo (sexo): Esta variable categórica indica el género del conductor involucrado en el siniestro.
2. Fecha de intervención (fechaintervencion): Esta variable de fecha registra la fecha en que ocurrió el siniestro o la intervención en el reclamo.
3. Importe del siniestro (importesiniestro): Esta variable numérica indica el monto del siniestro reclamado por el asegurado.
4. Importe invertido (importeinvertido): Esta variable numérica registra la cantidad de dinero invertida en el reclamo.
5. Importe ahorrado (importeahorrado): Esta variable numérica indica la cantidad de dinero ahorrado en el reclamo.
6. Tipo de documento (tipodocumento): Esta variable categórica indica el tipo de documento de identificación proporcionado por el asegurado.
7. Respuesta a la reclamación (respuesta_dicot1): Esta variable categórica indica si la reclamación se considera un fraude o no. Es la variable respuesta del modelo.
8. Edad del conductor (edad_conductor1): Esta variable numérica registra la edad del conductor en el momento del siniestro.
9. Garantía agrupada (garantia_agrupada): Esta variable categórica indica la categoría de garantía asociada al reclamo.

10. **Antigüedad de la póliza (antiguedad_poliza):** Esta variable registra la antigüedad de la póliza en el momento del siniestro.
11. **Días hasta la notificación del siniestro (dias_notificacion):** Esta variable numérica indica la cantidad de días transcurridos hasta la notificación del siniestro.
12. **Forma de pago agrupada (formapago_agrupado):** Esta variable categórica indica la forma de pago utilizada por el asegurado.
13. **ID del siniestro (siniestroid):** Esta variable identifica de manera única cada siniestro.
14. **Provincia ID (provinciaid):** Esta variable categórica indica el ID de la provincia donde ocurrió el siniestro.
15. **Valor del vehículo de mercado (valorvehiculomercado):** Esta variable numérica indica el valor de mercado del vehículo asegurado.
16. **Valor del vehículo de fábrica (valorvehiculofabrica):** Esta variable numérica registra el valor de fábrica del vehículo asegurado.
17. **Potencia (potencia):** Esta variable numérica indica la potencia del vehículo.
18. **Cilindrada (cilindrada):** Esta variable numérica indica la cilindrada del vehículo asegurado en centímetros cúbicos (cc).
19. **Peso del vehículo (pesovehiculo):** Esta variable numérica indica el peso del vehículo asegurado en kilogramos (kg).
20. **Longitud (longitud):** Esta variable numérica indica la longitud del vehículo asegurado en milímetros (mm).
21. **Scoring (scoring):** Esta variable categórica indica el grupo de scoring utilizado para evaluar el riesgo del asegurado.
22. **Aceptación sin antecedentes (aceptoculpasinantecedentes):** Esta variable categórica indica si se acepta al culpable sin antecedentes.

Estos datos proporcionan una visión detallada de diversos aspectos relacionados con los reclamos de seguros de automóviles, incluidos los montos de los reclamos, la información del conductor, las características del vehículo y la respuesta al reclamo. Este análisis será fundamental para el desarrollo de modelos predictivos y estrategias de detección y prevención del fraude en el sector de seguros.

3. Metodología

La metodología a utilizar combina dos enfoques principales: el método de MissForest y la regresión logística. En primer lugar, el método de MissForest se utilizará para imputar los valores perdidos en el conjunto de datos, lo que es crucial dado que los datos incompletos son comunes en este tipo de conjuntos de datos debido a la naturaleza no estructurada de la información de reclamaciones de seguros.

Una vez que se hayan completado los datos, se procederá a aplicará la regresión logística para construir un modelo predictivo que pueda clasificar las reclamaciones de seguros como legítimas o fraudulentas.

3.2. Algoritmo missForest

La presencia de datos faltantes es una realidad omnipresente en las bases de datos. A pesar de ello, muchos estudios no detallan explícitamente cómo gestionan estos datos (Wood & White, 2004), dejando implícitos ciertos métodos a través del software estadístico. Este enfoque puede llevar a diferentes tratamientos de datos faltantes entre distintos paquetes, lo que potencialmente compromete la reproducibilidad de los resultados. A menudo, los investigadores optan por la eliminación completa de casos con datos faltantes, un método predeterminado en muchos paquetes de regresión (Demissie, LaValley, Horton, & Glynn, 2003). Sin embargo, este enfoque solo es fiable cuando la cantidad de datos faltantes es pequeña y sigue un patrón aleatorio o aleatorio condicionalmente a las otras variables.

La eliminación completa de casos conlleva la pérdida de información, un desafío particularmente relevante en conjuntos de datos con un gran número de variables. Además, puede introducir sesgos impredecibles (Masconi, Matsha, & Erasmus, 2015). Una alternativa para mitigar estas limitaciones es la imputación, donde los valores faltantes se sustituyen por valores estimados. Dado el dinamismo en este campo, existen numerosos métodos y paquetes diseñados para la imputación. Todo esto ya fue estudiado por Zhang, Z, en el estudio sobre “Imputación de datos faltantes: centrándose en la imputación única” (Zhang, 2016).

El uso de métodos como la media, la mediana o el método de k vecinos más cercanos son los más comunes para la imputación de datos faltantes, pero tiene limitaciones y consideraciones importantes que se deben de tener en cuenta:

- Sensibilidad a valores extremos: La media es especialmente sensible a valores atípicos (outliers) en los datos, lo que puede distorsionar las estimaciones. El método de k vecinos también puede ser afectado por valores atípicos, ya que depende de la proximidad de los vecinos en el espacio de características.
- Simplicidad: Estos métodos son relativamente simples de implementar y entender, lo que los hace atractivos en muchas situaciones. Sin embargo, su simplicidad puede conducir a una pérdida de información importante en los datos, especialmente si la estructura de los datos es compleja o no lineal.
- Supuestos subyacentes: La imputación de la media y la mediana asume que los datos faltantes son aleatorios y no están sesgados. Si esta suposición no es válida (por

ejemplo, si los datos faltantes están relacionados con ciertas características de los individuos o con el proceso de selección de la muestra), la imputación puede sesgar los resultados y reducir la precisión de las estimaciones.

- Dependencia de la distancia: El método de k vecinos más cercanos depende de la medida de distancia utilizada y del número de vecinos considerados. La elección de estos parámetros puede afectar significativamente los resultados de la imputación y puede no ser óptima en todos los casos.

Es decir, que a pesar de que la imputación de datos a través de la media, la mediana y el método de k vecinos más cercanos sean los métodos más utilizados, esto no implica que sean los correctos (Gareth, Daniela, & Trevor, 2013).

Aunque la mayoría de los estudios sobre imputación de datos se han centrado en la estimación de la media y la varianza poblacional, como en su día realizó (Bello, 1993) existen otros parámetros importantes, como la función de distribución y los cuantiles, que no han sido suficientemente explorados en este contexto. El estudio de Rosas y Verdejo (Rosas, 2009) se buscó abordar esta brecha al presentar y analizar diversas técnicas de imputación bajo un diseño general. Su objetivo fue evaluar cómo estas técnicas afectan la estimación de varios parámetros, incluyendo la media, la función de distribución y los cuantiles, proporcionando así una visión más completa de su aplicación en la práctica.

En el estudio realizado por García, Albaladejo y Fernández (García & Albaladejo, 2006) sobre métodos de inferencia estadística con datos faltantes, se encontró que la imputación mediante la media no es un procedimiento recomendable para la inferencia. Este método mostró un comportamiento altamente inestable, especialmente en lo que respecta a la cobertura de los intervalos de confianza.

Para abordar la presencia de datos faltantes representados como "NA" en las observaciones, se propone el uso de la imputación estocástica con “MissForest”, un método que preserva la variabilidad original de la variable al predecir el valor faltante basándose en los datos y otras variables. La función missForest se encuentra en el paquete de R “missForest”, creada por Daniel J. Stekhoven (Stekhoven D. J., 2022). MissForest representa una herramienta crucial en el ámbito de la imputación de datos faltantes, especialmente en conjuntos de datos mixtos que incluyen tanto variables continuas como categóricas. Este método no paramétrico se fundamenta en el uso de Random Forest, un algoritmo de aprendizaje automático que emplea un conjunto de árboles de decisión para predecir los valores faltantes. Sus características distintivas incluyen:

- Imputación robusta: MissForest demuestra una notable capacidad para imputar valores faltantes de manera robusta, incluso en presencia de datos atípicos y estructuras no lineales.
- Manejo de variables mixtas: Es capaz de abordar conjuntos de datos mixtos, que contienen tanto variables continuas como categóricas, adaptándose a la complejidad inherente de esta diversidad de datos.
- Estimación del error de imputación: Proporciona una estimación del error de imputación, lo que facilita la evaluación de la precisión de las imputaciones realizadas.

El uso de MissForest y su implementación a través del paquete 'missForest' en R ofrecen una herramienta para abordar eficazmente la imputación de datos faltantes en conjuntos de datos complejos y heterogéneos, permitiendo así un análisis más completo y preciso en diversas aplicaciones académicas y prácticas (Stekhoven, 2011).

En el estudio realizado por Stekhoven & Bühlmann, en la comparación con otros métodos de imputación, MissForest destacó especialmente en configuraciones de datos con interacciones complejas y relaciones no lineales. Además, se observó una eficiencia computacional atractiva y la capacidad de MissForest para manejar datos de alta dimensionalidad (Stekhoven, D. J. & Bühlmann, P., 2012).

3.3. Regresión logística

La regresión logística es una técnica estadística utilizada para modelar la relación entre una variable categórica binaria (por ejemplo, sí/no, éxito/fracaso) y un conjunto de variables predictoras. A diferencia de la regresión lineal, que se utiliza para variables cuantitativas, la regresión logística es adecuada cuando la variable dependiente es categórica. En la regresión logística, el objetivo es predecir la probabilidad de que ocurra un evento de interés, en nuestro caso, fraude o no fraude, en función de las variables predictoras. La relación entre las variables predictoras y la probabilidad de que ocurra el evento se modela utilizando una función logística.

La función logística transforma la variable dependiente (la probabilidad) para que esté en una escala continua entre 0 y 1. Esto se logra utilizando la transformación logarítmica de la probabilidad dividida por su complemento. Esta transformación asegura que las predicciones del modelo estén acotadas en el rango de 0 a 1, lo que es necesario para representar probabilidades (Kleinbaum, Dietz, Gail, & Klein, 2002).

$$P(Y) = \frac{e^{PL}}{1 + e^{PL}} = \frac{e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\cdots+\beta_j*x_j}}{1 + e^{\beta_0+\beta_1*x_1+\beta_2*x_2+\cdots+\beta_j*x_j}}$$

Ilustración 1. Formula de la regresión logística

Características clave de la regresión logística:

- Variable de Resultado Binario: Está diseñada para situaciones en las que la variable de resultado solo puede tomar dos valores, a menudo representados como "éxito" y "fracaso" o "presencia" y "ausencia".
- Función Logística: La regresión logística emplea la función para transformar la combinación lineal de variables predictoras en probabilidades entre 0 y 1.
- Odds Ratio: El modelo genera Odds Ratio, que indican el cambio en las probabilidades de que ocurra el evento por un aumento de una unidad en una variable predictora, manteniendo todas las demás variables constantes.
- Interpretación: Al analizar los coeficientes y su significancia en el modelo, podemos comprender cómo los cambios en las variables predictoras influyen en la probabilidad del resultado.

Para ajustar el modelo de regresión logística, se utilizan técnicas de optimización basadas en la máxima verosimilitud, que busca encontrar los valores de los coeficientes del modelo que maximizan la probabilidad de observar los datos reales. Una vez ajustado el modelo, se pueden interpretar los coeficientes para comprender cómo afectan las variables predictoras a la probabilidad de que ocurra el evento de interés.

En resumen, la regresión logística es una herramienta poderosa para modelar y predecir eventos binarios, como la presencia o ausencia de una mutación celular, utilizando variables predictoras. Su capacidad para manejar datos categóricos y producir predicciones en forma de probabilidades la hace ampliamente utilizada en campos como la biología, la medicina, la economía y más.

3.3.1. Odds Ratio

El odds ratio (OR) es una medida estadística de gran importancia en diversos campos de la investigación. Su utilidad radica en su capacidad para cuantificar la relación entre variables predictoras y eventos binarios en un marco probabilístico. En el contexto de la regresión logística, el OR emerge como una herramienta esencial para comprender cómo las diferentes variables independientes influyen en la probabilidad de que ocurra un evento específico.

El cálculo del OR implica una comparación de las probabilidades de que ocurra un evento entre dos grupos definidos por los valores de una variable predictora. Específicamente, se compara la probabilidad de ocurrencia del evento en un grupo de individuos con un valor específico de la variable predictora con la probabilidad de ocurrencia del mismo evento en otro grupo de individuos con un valor diferente de la variable predictora. Esta relación se expresa como una razón de odds entre los dos grupos.

La interpretación del OR proporciona una comprensión significativa de la asociación entre la variable predictora y el evento de interés. Un OR mayor que 1 indica que la variable predictora está asociada con mayores probabilidades de que ocurra el evento, lo que sugiere una relación positiva entre la variable y el resultado binario. Por otro lado, un OR menor que 1 señala una asociación inversa, donde la presencia de la variable predictora se asocia con menores probabilidades de ocurrencia del evento. Cuando el OR es igual a 1, no hay asociación entre la variable predictora y el evento, lo que implica que la presencia o ausencia de la variable no afecta las probabilidades de ocurrencia del evento (Menard, S. W., 2010).

En la práctica, el OR se interpreta como el factor por el cual las probabilidades de ocurrencia del evento cambian cuando la variable predictora aumenta en una unidad, manteniendo constantes las otras variables en el modelo de regresión logística. Esta interpretación es esencial para comprender el impacto relativo de la variable predictora en el evento de interés.

Además de su papel en la evaluación de la asociación entre variables, el OR también se utiliza para comparar diferentes grupos definidos por los valores de la variable predictora. Esto permite identificar diferencias significativas en las probabilidades de ocurrencia del evento entre estos grupos y evaluar la influencia relativa de la variable predictora en cada uno.

En resumen, el odds ratio es una medida estadística que proporciona una comprensión detallada de la asociación entre variables predictoras y eventos binarios. Su interpretación cuidadosa y su aplicación en la investigación son fundamentales para entender cómo las variables independientes influyen en la probabilidad de ocurrencia de eventos específicos, lo que lo convierte en un componente indispensable en el análisis estadístico (Kleinbaum, Dietz, Gail, & Klein, 2002).

La interpretación de los coeficientes en la regresión logística depende del tipo de variable predictora: categórica o numérica. A continuación, se detalla la interpretación para cada caso:

3.3.1.1. Interpretación de las variables categóricas:

Las variables categóricas, como género, estado civil o tipo de producto se codifican mediante variables dicotómicas o dummie. Para cada variable categórica con k categorías, se introducen k-1 variables dummies en el modelo. La interpretación de los coeficientes para variables categóricas se basa en la comparación de las probabilidades de la variable de resultado entre las categorías.

El coeficiente de una variable dummy representa el cambio en el logit (la transformación logarítmica de las probabilidades) de la variable de resultado para la categoría representada por la variable dummy en comparación con la categoría de referencia (que no se incluye en el modelo). El exponente del coeficiente representa la razón de probabilidades (Odds ratio) de la categoría respecto a la categoría base. En otras palabras, indica cuántas veces más probable es que ocurra el evento de interés para una categoría específica en comparación con la categoría base.

- Coeficiente positivo: Indica que la probabilidad de que ocurra el evento es mayor para la categoría representada por la variable dummy en comparación con la categoría de referencia.
- Coeficiente negativo: Indica que la probabilidad de que ocurra el evento es menor para la categoría representada por la variable dummy en comparación con la categoría de referencia.
- Coeficiente cercano a 0: Indica que no hay una diferencia significativa en la probabilidad de que ocurra el evento entre la categoría representada por la variable dummy y la categoría de referencia.

3.3.1.2. Variables Continuas:

Las variables continuas, como edad, ingresos, se incluyen directamente en el modelo. La interpretación de los coeficientes para variables continuas se basa en el cambio en el logit de la variable de resultado por una unidad de aumento en la variable continua.

El coeficiente de una variable continua representa el cambio en el logit de la variable de resultado por un aumento de una unidad en la variable continua, manteniendo todas las demás variables constantes. El exponente del coeficiente representa la Odds ratio de incrementar en una unidad el valor de la variable respecto a no hacerlo. En otras palabras, indica cuántas veces más probable es que ocurra el evento de interés por cada unidad de aumento en la variable continua.

- **Coeficiente positivo:** Indica que la probabilidad de que ocurra el evento aumenta a medida que aumenta la variable continua.
- **Coeficiente negativo:** Indica que la probabilidad de que ocurra el evento disminuye a medida que aumenta la variable continua.
- **Coeficiente cercano a 0:** Indica que no hay una relación lineal significativa entre la variable continua y la probabilidad de que ocurra el evento.

En la interpretación de coeficientes en la regresión logística, las diferencias clave radican en el tipo de variable predictora utilizada. Para variables categóricas, la interpretación se centra en comparar las categorías entre sí, evaluando cómo afecta cada categoría a la probabilidad del evento. En contraste, para variables continuas, la interpretación se enfoca en el cambio en la probabilidad del evento por cada unidad de aumento en la variable continua. Además, consideraciones adicionales, como el uso del odds ratio para comprender la fuerza de la asociación, y la evaluación de la significancia estadística del coeficiente para determinar su relevancia, son cruciales. También se destaca la importancia de contextualizar los resultados dentro del problema específico y considerar la escala de las variables para una interpretación precisa y relevante (Hosmer Jr & Lemeshow, 2013).

3.3.2. Separación perfecta:

La "Separación Perfecta" es un fenómeno que puede ocurrir en modelos de regresión logística cuando existe una relación muy fuerte y clara entre las variables predictoras y la variable de resultado. En esta situación, una combinación lineal específica de las variables predictoras puede predecir con precisión el resultado binario de la variable dependiente para todas las observaciones en los datos. Por ejemplo, si en un estudio todas las personas con una cierta característica siempre tienen el mismo resultado (por ejemplo, "fraude" o "no fraude"), entonces existe separación perfecta.

Este escenario plantea problemas significativos en el análisis estadístico. Por un lado, puede hacer que el modelo de regresión logística no converja, lo que significa que no puede estimar los parámetros del modelo de manera adecuada. Por otro lado, incluso si el modelo converge, la estimación de los parámetros puede ser extremadamente inestable, lo que hace que las inferencias sobre las relaciones entre las variables sean poco confiables (Kleinbaum, Dietz, Gail, & Klein, 2002).

Para comprender mejor el problema de la separación perfecta, se presenta un ejemplo de la misma en el caso de dos variables predictoras. En la ilustración 2 se observa que la recta que divide los dos conjuntos de datos es clara y precisa: cualquier punto por encima de esta línea se clasifica como parte de una clase, mientras que cualquier punto por debajo se clasifica como parte de la otra clase. Esto es posible porque las dos variables predictoras son independientes linealmente, lo que significa que no hay una relación lineal entre ellas. La importancia de esta separación perfecta radica en la capacidad del algoritmo de clasificación para aprender la relación entre las variables predictoras y la variable objetivo con una precisión total.

Que tenga infinitas soluciones, significa que existen infinitas rectas diferentes que pueden dividir perfectamente los dos conjuntos de datos (puntos rojos y negros) con una precisión del

100%.Es decir, que se tenga infinitas soluciones en el problema de la separación perfecta indican que hay múltiples formas válidas de separar los datos, lo que ofrece flexibilidad pero también requiere cuidado para evitar el sobreajuste y elegir la mejor solución.

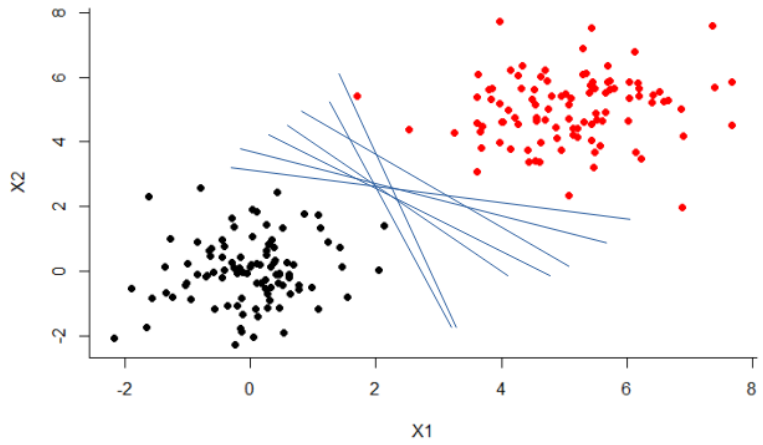

Ilustración 2. Explicación gráfica de la separación perfecta. Fuente: Elaboración propia

Este sesgo puede ocurrir debido a varias razones, incluida la asimetría en la distribución de los datos, la presencia de valores atípicos o la violación de supuestos de distribución.Firth (Firth, 1993) propuso un enfoque para reducir este sesgo en las estimaciones de máxima verosimilitud. Su método se basa en agregar una corrección al logaritmo de la función de verosimilitud, que penaliza los valores de los parámetros que generan una función de verosimilitud demasiado pequeña. Esta penalización ayuda a compensar el sesgo en las estimaciones y produce estimaciones más cercanas a los verdaderos valores de los parámetros. El enfoque de Firth se ha utilizado en una variedad de contextos estadísticos, incluida la regresión logística, la regresión de Cox y otros modelos de regresión. Se ha demostrado que es efectivo para reducir el sesgo en las estimaciones de máxima verosimilitud, especialmente en situaciones donde los datos son escasos o hay problemas de separación.

El paquete "logistf" en R proporciona una solución a este problema. Utiliza algoritmos y métodos numéricos avanzados que permiten estimar los parámetros del modelo incluso en presencia de separación perfecta. Esto se logra mediante la aplicación de técnicas como la estimación de máxima verosimilitud penalizada o la utilización de métodos de regularización para estabilizar las estimaciones de los parámetros. Además, el paquete "logistf" ofrece herramientas para diagnosticar la presencia de separación perfecta en los datos y opciones para manejarla de manera efectiva, lo que lo convierte en una herramienta esencial para investigadores y analistas de datos que enfrentan este desafío en sus estudios (Heinze, Ploner, Dunkler, & Southworth, 2023).

3.3.3. Validación del modelo

La validación del modelo se realizará bajo el estudio de su curva ROC (Receiver Operating Characteristic), esta es una herramienta gráfica utilizada en el análisis de clasificadores binarios para evaluar su rendimiento. Esta curva representa la tasa de verdaderos positivos (sensibilidad) en el eje y la tasa de falsos positivos (1 - especificidad) en el eje x (Hoo & Candlish, 2017).

En la curva ROC, cada punto en el gráfico representa un umbral de decisión diferente para clasificar las observaciones en positivas o negativas. A medida que el umbral de decisión cambia, la sensibilidad y la especificidad del clasificador también cambian, lo que da como resultado diferentes puntos en la curva ROC.

Un clasificador perfecto tendría una curva ROC que se elevaría rápidamente hacia el extremo superior izquierdo del gráfico, lo que indicaría una alta sensibilidad y una baja tasa de falsos positivos. Por otro lado, un clasificador que clasifica al azar tendría una curva ROC que se aproxima a la línea diagonal (conocida como línea de referencia), lo que indica que no hay diferencia entre la tasa de verdaderos positivos y la tasa de falsos positivos.

La curva ROC también se utiliza para calcular el área bajo la curva (AUC-ROC), que es una medida de la capacidad predictiva global del clasificador. Un AUC-ROC de 1 indica un clasificador perfecto, mientras que un AUC-ROC de 0.5 indica un clasificador que clasifica al azar. En general, cuanto mayor sea el AUC-ROC, mejor será el rendimiento del clasificador.

Para la validación del modelo se utilizará la función "roc" de la librería "pROC" en R, que es una herramienta útil para evaluar la validez de un modelo de clasificación, como un modelo de regresión logística. "roc" se utiliza para crear y visualizar curvas ROC (Receiver Operating Characteristic), que son gráficos que muestran la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1 - especificidad) para diferentes umbrales de clasificación (Robin, X., Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J. C. & Doering, M., 2018).

Características y usos importantes de la función "roc":

- Visualización de Curvas ROC: La curva ROC es una representación gráfica fundamental en la evaluación de modelos de clasificación. La función "roc" genera estas curvas ROC a partir de las probabilidades de predicción obtenidas de un modelo, lo que permite visualizar cómo varía la tasa de verdaderos positivos (sensibilidad) en comparación con la tasa de falsos positivos (1 - especificidad) al ajustar el umbral de clasificación. Esta visualización proporciona una imagen completa del rendimiento del modelo en diferentes puntos de corte y ayuda a comprender su capacidad discriminativa.
- Área bajo la Curva (AUC): Además de trazar la curva ROC, la función "roc" calcula automáticamente el área bajo la curva ROC (AUC). El AUC es una métrica clave que resume la capacidad discriminativa global del modelo. Un valor de AUC cercano a 1 indica un modelo altamente preciso, mientras que un valor cercano a 0.5 sugiere un rendimiento similar al azar. Esta medida es fundamental para comparar modelos y determinar cuál es el más efectivo para una tarea de clasificación específica.
- Evaluación de Sensibilidad y Especificidad: La curva ROC y el AUC también proporcionan información sobre la sensibilidad y especificidad del modelo en diferentes puntos de corte de probabilidad. La sensibilidad se refiere a la capacidad del modelo para identificar correctamente los casos positivos, mientras que la especificidad indica su habilidad para evitar falsos positivos. La curva ROC visualiza este equilibrio y ayuda a encontrar el punto de corte óptimo que maximiza la precisión del modelo.

- Comparación de Modelos: Una ventaja significativa de la función "roc" es su capacidad para comparar múltiples modelos de clasificación de manera eficiente. Al trazar varias curvas ROC en la misma figura y calcular sus respectivas áreas bajo la curva, la función "roc" facilita la identificación del modelo más efectivo para una tarea de clasificación particular. Esto es crucial en situaciones donde se están considerando varios enfoques de modelado o variables predictoras.

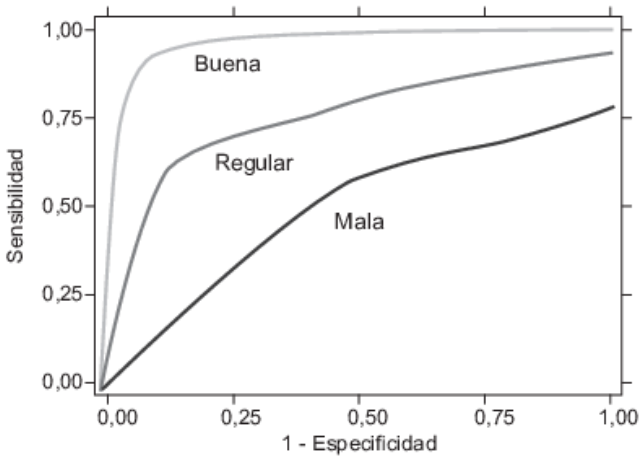

Ilustración 3. Esquema explicativo de distintas posibilidades de curvas ROC. Fuente: ResearchGate

En el análisis de las curvas ROC mostradas en la ilustración superior, podemos identificar tres posibles escenarios que reflejan el rendimiento de los modelos de clasificación binaria (Pérez, J.M. & Martín, P.P., 2023).

En primer lugar, una "Curva ROC Buena" se caracteriza por una alta sensibilidad y especificidad en todo el rango de umbrales de clasificación. Esto sugiere que el modelo tiene una capacidad excelente para distinguir entre casos positivos y negativos, lo cual se refleja en un área bajo la curva (AUC) cercana a 1, indicando un rendimiento óptimo.

Por otro lado, nos encontramos con una "Curva ROC Regular", donde la sensibilidad y especificidad del modelo son moderadas en todo el rango de umbrales. Si bien el modelo tiene una capacidad aceptable para diferenciar entre casos positivos y negativos, el AUC se sitúa entre 0,5 y 1, lo que sugiere un rendimiento moderado y áreas potenciales de mejora.

Finalmente, una "Curva ROC Mala" exhibe una baja sensibilidad y especificidad en todo el rango de umbrales. Esto indica que el modelo tiene dificultades para distinguir entre casos positivos y negativos, y su AUC es cercano a 0,5, lo que sugiere un rendimiento similar al azar.

En resumen, el análisis de las curvas ROC proporciona información valiosa sobre la capacidad de un modelo para realizar clasificaciones precisas en problemas de clasificación binaria. La forma de la curva ROC permite inferir el rendimiento del modelo y proporciona indicadores sobre posibles áreas de mejora, lo que resulta fundamental en la evaluación y selección de modelos en contextos académicos y prácticos. Es por ello por lo que la función "roc" de la librería "pROC" es una herramienta valiosa para evaluar la validez y el rendimiento predictivo de modelos de clasificación, como los modelos de regresión logística. Proporciona una forma intuitiva de

visualizar y comparar el rendimiento de diferentes modelos y ayuda a tomar decisiones informadas sobre su uso en aplicaciones prácticas.

3.4. Otros conceptos de interés:

Cuando se realiza un análisis de regresión logística, además de interpretar los coeficientes del modelo, hay varios otros aspectos importantes que deben considerarse para evaluar la calidad y el ajuste del modelo. Aquí están algunos de los elementos clave que pueden ser examinados:

3.4.1. P- Valor:

Este valor indica la significancia estadística de cada coeficiente en el modelo. Usualmente, se considera que un p-valor menor a 0.05 el coeficiente proporciona suficiente evidencia para rechazar la hipótesis nula, siendo la hipótesis nula que el coeficiente sea igual a cero. Por otro lado, si es mayor a 0.05, no hay suficiente evidencia para rechazar la hipótesis nula.

Podemos respaldar esta idea mediante una selección de libros de texto publicados en la última década. Esta variedad de fuentes académicas y prácticas puede brindar una perspectiva amplia y sólida sobre la relevancia y la aplicabilidad de estos conceptos en el campo de la estadística y el análisis de datos.

"El valor p se puede ver simplemente como la posibilidad de obtener este conjunto de datos dado que las muestras provienen de la misma distribución…., la aproximación del valor p como ayuda en la toma de decisiones es bastante natural, ya que casi todos los paquetes computacionales que ofrecen el cálculo de prueba de hipótesis dan valores p junto con valores del estadístico de prueba adecuado. Un valor p es el nivel (de significancia) más bajo donde es significativo el valor observado del estadístico de prueba" (Walpole, Myers, & Myers, 2007).

"El informe de valores p como parte de los resultados de una investigación proporciona más información al lector que afirmaciones como la hipótesis nula se rechaza con un nivel de significación de 0,05 o los resultados no fueron significativos en el nivel 0,05. Al informar el valor p asociado con una prueba de hipótesis se permite al lector saber con exactitud qué tan extraño o qué tan común es el valor calculado de la estadística de prueba dado que H_0 es verdadera" (Wayne, 2009).

"El valor p es la probabilidad de que el estadístico de prueba asuma un valor que sea al menos tan extremo como el valor observado del estadístico cuando la hipótesis nula H0 es verdadera. Por lo tanto, un valor p transmite mucha información acerca del peso de la evidencia en contra de H_0 y, por consiguiente, el responsable de la toma de decisiones puede llegar a una conclusión con cualquier nivel de significación especificado" (Montgomery, 2010).

3.4.2. AIC:

El AIC (Criterio de Información de Akaike) es un criterio utilizado para comparar la calidad relativa de diferentes modelos. El AIC (Criterio de Información de Akaike) es una medida estadística utilizada para comparar diferentes modelos y evaluar su calidad relativa en términos

de ajuste y capacidad predictiva. Fue desarrollado por Hirotugu Akaike y se utiliza ampliamente en el contexto de la selección de modelos en estadística y ciencia de datos.

La idea central detrás del AIC es encontrar un equilibrio entre la capacidad predictiva de un modelo y su complejidad. El AIC toma en cuenta tanto la bondad del ajuste del modelo como la cantidad de información utilizada por el modelo para ajustarse a los datos. En esencia, busca encontrar el modelo que ofrezca un buen ajuste a los datos con la menor cantidad de parámetros (Martınez, Albın, Cabaleiro, Pena, & Rivera, 2009). Cuanto menor sea el valor del AIC, mejor se considera el ajuste del modelo. El AIC toma en cuenta la capacidad predictiva del modelo y penaliza la complejidad de este, lo que lo hace útil para la selección de modelos.

3.4.3. VIF:

El VIF (Factor de Inflación de la Varianza) es una medida utilizada para evaluar la multicolinealidad entre las variables predictoras en un modelo de regresión. La multicolinealidad se refiere a la alta correlación entre dos o más variables predictoras en el modelo, lo que puede distorsionar las estimaciones de los coeficientes y afectar la interpretación del modelo (Mandeville, 2008).

El VIF se calcula para cada variable predictora en el modelo y proporciona una medida de cuánto aumenta la varianza de un coeficiente de regresión debido a la alta correlación con otras variables predictoras. Un VIF alto indica una alta multicolinealidad y sugiere que la variable predictora correspondiente puede estar redundante o altamente correlacionada con otras variables en el modelo.

Interpretar el VIF implica considerar su magnitud. Por lo general, se considera que un VIF superior a 10 indica una multicolinealidad significativa (Belsley, 1991), aunque este umbral puede variar dependiendo del contexto y la naturaleza de los datos. Un VIF muy alto puede sugerir que la variable predictora correspondiente está altamente correlacionada con otras variables en el modelo, lo que puede hacer que la interpretación de los coeficientes sea difícil y poco confiable.

Para abordar la multicolinealidad, es importante identificar las variables con VIF alto y considerar opciones como la eliminación de variables redundantes o la combinación de variables relacionadas antes de ajustar el modelo. Reducir la multicolinealidad puede mejorar la estabilidad y la precisión de las estimaciones de los coeficientes, así como facilitar la interpretación del modelo.

4. Resultado

En este apartado, nos adentramos en el proceso del análisis de datos: el estudio exhaustivo de las variables y su influencia en un modelo de regresión logística. Comenzaremos realizando un Análisis Descriptivo de las Variables antes de llevar a cabo cualquier manipulación de los datos. Este análisis nos proporcionará una visión inicial de la distribución y características de nuestras variables, brindándonos perspectivas fundamentales para comprender la naturaleza de nuestros datos. Posteriormente, procederemos a una fase de limpieza de la base de datos, abordando posibles problemas como datos faltantes o inconsistencias. Acto seguido, llevaremos a cabo la imputación de los datos faltantes, asegurándonos de conservar la integridad y la validez de nuestra información.

Una vez completada la imputación, realizaremos un segundo Análisis Descriptivo de las Variables para evaluar cómo han cambiado las características de nuestras variables. Este paso nos permitirá identificar cualquier impacto que la imputación haya tenido en la distribución y en la relación entre las variables. Finalmente, nos sumergiremos en el análisis de regresión logística, explorando varios modelos y ajustes para encontrar aquel que optimice el criterio de información de Akaike (AIC). Este proceso nos permitirá identificar el modelo que mejor se ajuste a nuestros datos, proporcionándonos una base sólida para la interpretación y la toma de decisiones basada en evidencia.

4.1. Análisis descriptivo

En este apartado, nos sumergimos en un proceso fundamental para comprender la estructura y la naturaleza de nuestros datos: el Análisis Descriptivo. Inicialmente, exploraremos las características de las variables antes de cualquier manipulación, obteniendo una visión integral de su distribución y comportamiento. Esta fase incluirá un análisis comparativo entre las variables y la variable respuesta "Fraude", proporcionándonos información crucial sobre posibles relaciones y patrones iniciales.

Posteriormente, nos adentraremos en una etapa de limpieza y preparación de datos, donde abordaremos cualquier inconsistencia o falta de integridad en nuestra base de datos. Tras identificar y corregir problemas potenciales, nos centraremos en la imputación de los datos faltantes, asegurando así la completitud y la calidad de nuestra información.

Una vez completada la imputación, realizaremos un segundo Análisis Descriptivo de las Variables, esta vez evaluando cómo han cambiado las características después de la imputación. Este análisis nos permitirá comprender cualquier impacto que la imputación haya tenido en la distribución y en las relaciones entre las variables, proporcionándonos una visión actualizada de nuestros datos. Además, continuaremos comparando estas variables con la variable respuesta "Fraude" para identificar posibles cambios o nuevas tendencias que hayan surgido después del proceso de imputación. Este enfoque holístico nos brindará una comprensión completa de nuestros datos, sentando las bases para análisis posteriores y la toma de decisiones informada

4.1.1. Análisis Descriptivo de las Variables antes de la imputación

Este análisis descriptivo se basa en un conjunto de datos que contiene información sobre siniestros de seguros de automóviles. El objetivo es presentar una descripción completa y precisa de las variables que componen el conjunto de datos, utilizando diferentes técnicas estadísticas y visualizaciones.

4.1.1.1. Variables Numéricas:

1. Fecha intervención: La variable representa fechas de intervención. Las fechas varían desde el 18 de febrero de 2015 hasta el 19 de septiembre de 2023. Esta variable contiene 881 valores perdidos (NA).
2. Importe siniestro: La variable muestra una amplia dispersión, con una media de 2.887,097€ y una desviación estándar de 11.636,540€. Esto sugiere una variabilidad significativa en los montos de los siniestros. Sin embargo, la presencia de valores atípicos o extremos es evidente, con un valor máximo observado de 174.842,16€, considerablemente mayor que la media. Además, la existencia de valores negativos, como el mínimo de -1.117€, plantea la posibilidad de errores en el registro de datos.
3. Importe invertido: La variable muestra una distribución sesgada hacia la derecha, con una media de 193,363€ y una desviación estándar de 702,553€. Esto sugiere una concentración de valores más altos en el extremo derecho de la distribución, con una considerable dispersión alrededor de la media. El valor máximo observado es de 7.401,03€, considerablemente mayor que la media, indicando la presencia de valores extremadamente altos. Por otro lado, el valor mínimo es de 0€, lo que sugiere casos errores en el registro.
4. Importe ahorrado: La variable exhibe una distribución sesgada hacia la derecha, con una media de 1.084,748€ y una desviación estándar de 2.934,860€, lo que indica una concentración de valores más altos en el extremo derecho de la distribución y una considerable dispersión alrededor de la media. El valor máximo observado es de 28.282,210€, significativamente mayor que la media, señalando la presencia de valores extremadamente altos. Por otro lado, el valor mínimo es de 0€, lo que indica la existencia de casos errores en el registro.
5. Edad conductor: La variable muestra una distribución relativamente simétrica, con una media de 45,499 años y una desviación estándar de 14,015, lo que indica una dispersión moderada alrededor de la media. El rango de edades va desde 14,5 años hasta 87,858 años, lo que refleja una amplia variabilidad en las edades de los conductores en el conjunto de datos. La presencia de conductores tan jóvenes como de 14,5 años sugiere la inclusión de seguros no solo para coches, sino también para ciclomotores
6. Antigüedad de la póliza: La variable representa los años transcurridos desde la emisión de la póliza hasta la fecha del siniestro. La antigüedad promedio de las pólizas es de aproximadamente 2,613 años, con una dispersión moderada de 3,247 años alrededor de esta media. La mediana, que es de 1,407 años, indica que la mitad de las pólizas tienen una antigüedad inferior a este valor. Sin embargo, la media es mayor que la mediana, sugiriendo una posible asimetría positiva en la distribución.

7. Días notificación: La variable muestra una distribución altamente sesgada hacia la derecha, con la mayoría de los valores concentrados en el extremo izquierdo y pocos valores más altos extendiéndose hacia la derecha. La media de esta variable es de 8,486 días, indicando que, en promedio, los siniestros son notificados después de aproximadamente 8,5 días. Sin embargo, la desviación estándar es alta, con un valor de 24,626, lo que sugiere una gran dispersión alrededor de la media. El valor máximo observado es de 322 días, sugiriendo casos de notificación significativamente más tardía.
8. Valor vehículo mercado: La variable exhibe una amplia variabilidad en los precios de los vehículos en el conjunto de datos. Con una media de aproximadamente 20.224,84€, indica el valor promedio de los vehículos en el mercado en el conjunto de datos analizado. Sin embargo, la alta desviación estándar, aproximadamente de 12.832,602€, señala una dispersión considerable alrededor de la media, posiblemente debido a la presencia de valores atípicos o extremos. El rango de valores abarca desde 0 hasta 161.000€, lo que resalta la amplitud de precios representados. La inclusión de vehículos con valores tan bajos como 0€ sugiere posibles errores en el registro del valor del vehículo, mientras que valores tan altos como 161.000€ indican una diversidad considerable en los tipos y condiciones de los vehículos incluidos en el análisis.
9. Valor vehículo fabrica: La variable muestra una amplia dispersión de datos, indicando una considerable variabilidad en los valores de los vehículos. Con una media de aproximadamente 22.094,515€, representa el promedio de estos valores. Sin embargo, la desviación estándar extremadamente alta, aproximadamente de 135,082.829€, sugiere una dispersión significativa alrededor de la media, posiblemente debido a la presencia de valores atípicos. El rango de valores va desde 0€ hasta 3.633.645€, resaltando la amplia gama de valores indica errores en el registro de datos.
10. Potencia: La variable representa la potencia del vehículo y muestra una distribución relativamente simétrica, indicando una distribución uniforme alrededor de la media. Con una media de aproximadamente 100,042 unidades, refleja el promedio de la potencia de los vehículos en el conjunto de datos. La desviación estándar, de aproximadamente 48,609, señala una dispersión moderada alrededor de la media. El rango de valores va desde 2.0 hasta 500, resaltando una amplia gama de potencias representadas. La inclusión de vehículos con potencias tan bajas como 2.0 y tan altas como 500 indica una diversidad considerable en los tipos y capacidades de los vehículos analizados.
11. Cilindrada: La variable representa la capacidad de los cilindros del motor de un vehículo y muestra una distribución sesgada hacia la derecha, con una concentración de valores más bajos en el extremo izquierdo y pocos valores más altos que se extienden hacia la derecha. Con una media de aproximadamente 1.795,331cc, indica que, en promedio, la cilindrada de los vehículos en el conjunto de datos es alrededor de 1.795cc. La desviación estándar, de aproximadamente 1.081,833cc, señala una dispersión considerable de los datos alrededor de la media. El rango de valores va desde 49cc hasta 12.882cc, destacando una amplia variabilidad en las cilindradas representadas en el conjunto de datos. La inclusión de vehículos con cilindradas de 49cc indica presencia de motocicletas.
12. Peso vehículo: La variable muestra una dispersión considerable en los datos, lo que indica una amplia variabilidad en los pesos de los vehículos en el conjunto de datos. Con una media de aproximadamente 1.329,610Kg, sugiere que, en promedio, el peso de los vehículos es alrededor de 1.329,61Kg. Sin embargo, la alta desviación estándar, de aproximadamente 968,228Kg, señala una dispersión considerable alrededor de la media,

posiblemente debido a la presencia de valores atípicos o extremos. El rango de valores va desde 0 hasta 20.500Kg, resaltando la amplia gama de pesos representados. La inclusión de vehículos con un peso 0 sugiere problemas en el registro de datos.

13. Longitud: La variable representa la longitud del vehículo y muestra una distribución relativamente simétrica, indicando una distribución uniforme alrededor de la media. Con una media de aproximadamente 4.326,461mm, refleja el promedio de la longitud de los vehículos en el conjunto de datos. La desviación estándar, de aproximadamente 484,133mm, señala una dispersión moderada de los datos alrededor de la media. El rango de valores va desde 2.582mm hasta 7.345mm, resaltando la variabilidad en las longitudes de los vehículos representadas. La inclusión de vehículos con una longitud mínima de 2.582mm y una longitud máxima de 7.345mm indica una diversidad considerable en los tamaños y tipos de vehículos analizados.

4.1.1.2. Variables Categóricas:

14. Sexo: La variable "Sexo" indica el género de los asegurados y se divide en tres clases: masculino (m), femenino (f) y otros (a). La mayoría de los asegurados son hombres, representando aproximadamente el 63,7% de los casos, seguidos por los asegurados femeninos 22,1% y otros géneros 14,4%.
15. Tipo documento: La variable "Tipo de Documento" indica el tipo de documento utilizado por los asegurados para identificarse en las pólizas de seguro. Según el análisis, la gran mayoría de los asegurados, aproximadamente el 88,8% de los casos, utilizan el Documento Nacional de Identidad (DNI) o el Número de Identificación de Extranjero (NIE) para este fin.
16. Respuesta dicot1: La variable "respuesta_dicot1" es categórica y binaria, dividiendo los casos en dos clases: "NO FRAUDE" y "FRAUDE". En el conjunto de datos proporcionado, ambas clases tienen una proporción igual del 50%. Esto indica que la distribución de casos entre las dos categorías es equilibrada.
17. Garantía agrupada: La variable representa diferentes tipos de garantías de seguros. En el conjunto de datos, hay un total de 10 clases diferentes, siendo "obligatorio RC" la más común, abarcando aproximadamente el 60,6% de las observaciones. La clase menos frecuente es "atropello animales", con una proporción del 0% de las observaciones. Además, se registra un valor perdido (NA) en la variable "garantía agrupada".
18. Formato pago agrupado: La variable en el conjunto de datos presenta dos clases: "domiciliado" y "efectivo". "Domiciliado" es la clase más frecuente, abarcando aproximadamente el 52,3% de las observaciones, mientras que "efectivo" representa el 47,7%. No hay valores perdidos en esta variable, lo que indica datos completos.
19. Siniestro id: La variable contiene números de registro únicos para los siniestros registrados, representando así una categoría categórica. Con un total de 1.000 clases diferentes, cada una identifica un siniestro específico.
20. Provincia id: Esta variable aunque se presenta como numérica, los valores no representan cantidades sino identificadores únicos para cada provincia. El rango de valores va desde 2 hasta 51, con un total de 26 provincias en el conjunto de datos.
21. Acepto culpa sin antecedentes: La variable es binaria e indica si el asegurado acepta la culpabilidad sin antecedentes. Según los datos, aproximadamente el 95% de los casos no

aceptan la culpabilidad sin antecedentes. Esto muestra una distribución desigual en la variable, donde la clase "No" (indicando que el asegurado no acepta la culpabilidad sin antecedentes) es la más frecuente, representando aproximadamente el 95% de los casos, mientras que la clase "Sí" (indicando que el asegurado acepta la culpabilidad sin antecedentes) es menos frecuente, representando aproximadamente el 5% de los casos.

22. **Scoring:** La variable "scoring" del conjunto de datos es una característica categórica que refleja diferentes tipos de puntuaciones asignadas a los registros. Con un total de seis clases distintas, se observa una distribución heterogénea en las puntuaciones asignadas. La clase más común, "Grupo PROYECTA", representa aproximadamente el 53,2% de los registros, mientras que la clase minoritaria, "Grupo INICIA", constituye cerca del 1,3%.

4.1.1.3. Comparación de las variables con la variable respuesta Fraude.

En este caso procederemos a comparar las variables con la variable respuesta fraude, observando por medio de las gráficas sus distribuciones y relaciones potenciales, lo que nos permitirá identificar patrones y posibles correlaciones con el comportamiento fraudulento.

1. Sexo:

sexo	No fraude	Fraude
a	82	60
f	111	110
m	307	330

Tabla 1. Tabla de contingencia de la variable independiente y la variable sexo. Fuente: Elaboración propia

La tabla de contingencia muestra la distribución de las observaciones según las variables "sexo" y "fraude". Para la categoría "No fraude", hay 82 observaciones para el sexo "a", 111 para el sexo "f" y 307 para el sexo "m". En cuanto a la categoría "Fraude", se registran 60 observaciones para el sexo "a", 110 para el sexo "f" y 330 para el sexo "m".

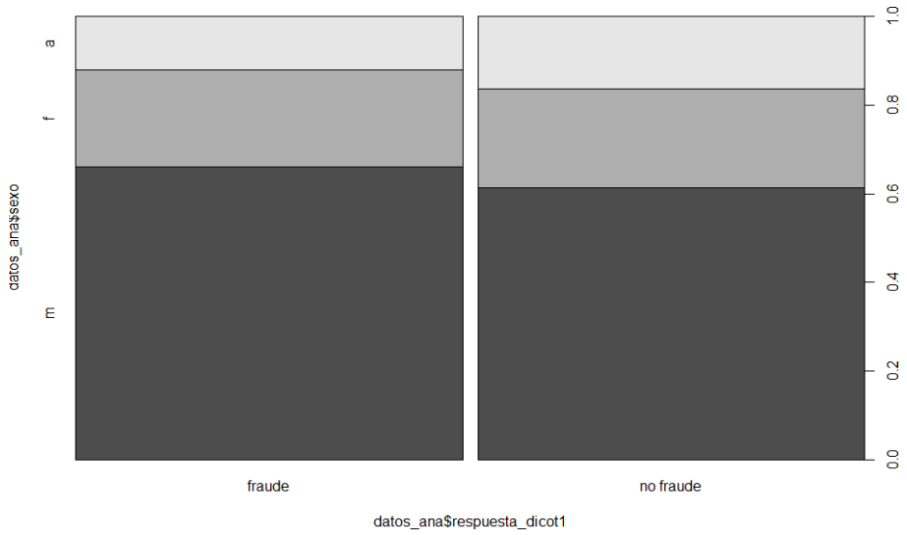

Ilustración 4. Gráfico de la variable independiente y la variable sexo. Fuente: Elaboración propia en programa R

2. Fecha de intervención:

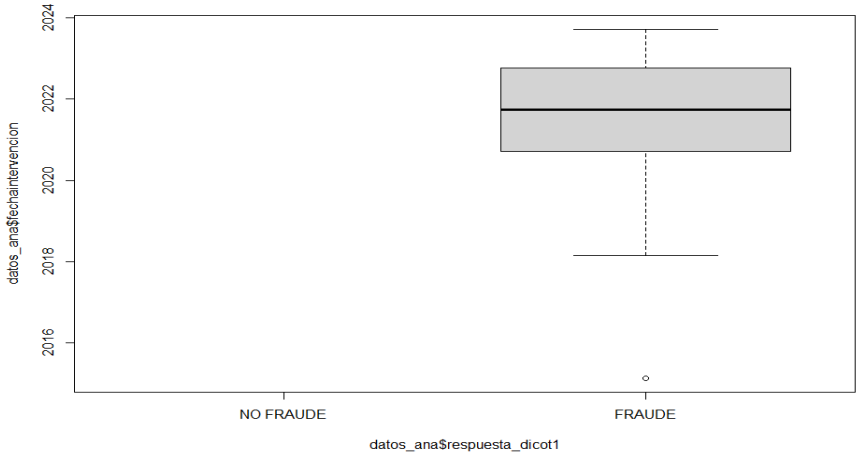

Ilustración 5. Diagrama de cajas para la variable independiente y la variable fecha de intervención. Fuente: Elaboración propia en programa R

Después de analizar la tabla, se evidencia la ausencia de datos sobre no fraude en las fechas de intervención. Este vacío se atribuye a que el 80% de los datos están marcados como faltantes (NA), lo que impide un análisis adecuado de esta variable.

3. Importe del siniestro:

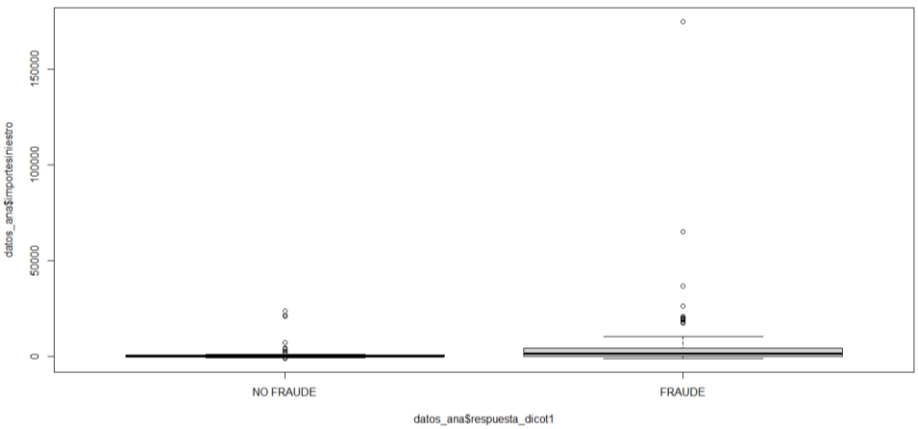

Ilustración 6. Diagrama de cajas para la variable independiente y la variable Importe del siniestro. Fuente: Elaboración propia en programa R

La mediana y el rango intercuartílico (IQR) son más altos en los casos de fraude en comparación con los no fraudulentos, indicando que los casos de fraude tienden a tener montos de reclamación más altos y una mayor dispersión de datos. Hay valores atípicos en ambos grupos, especialmente en fraudes, que podrían representar casos graves o errores de medición. Aunque hay una superposición considerable entre las distribuciones de reclamaciones fraudulentas y no fraudulentas, lo que dificulta predecir el fraude basándose únicamente en el monto de la reclamación.

4. Importe invertido:

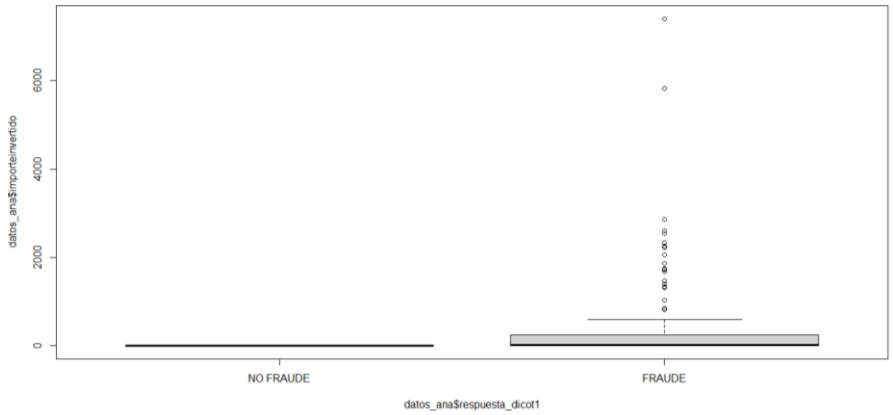

Ilustración 7. Diagrama de cajas para la variable independiente y la variable Importe invertido. Fuente: Elaboración propia en programa R

La mediana del importe invertido es similar para los casos de fraude y para los casos sin fraude. Esto indica que, en general, no hay una diferencia significativa en el importe invertido entre los dos grupos. El IQR es mayor para los casos de fraude, lo que indica que la dispersión de los datos es mayor en esos casos. Hay algunos valores atípicos en ambos grupos, especialmente en el grupo de fraude. Estos valores pueden ser casos de fraude con inversiones muy altas o errores en la medición.

La superposición entre las dos distribuciones es considerable. Esto significa que hay algunos casos de fraude con un importe invertido bajo y algunos casos sin fraude con un importe invertido alto. Esto hace que sea difícil usar el importe invertido para predecir con precisión si un caso es fraude o no.

5. Importe ahorrado:

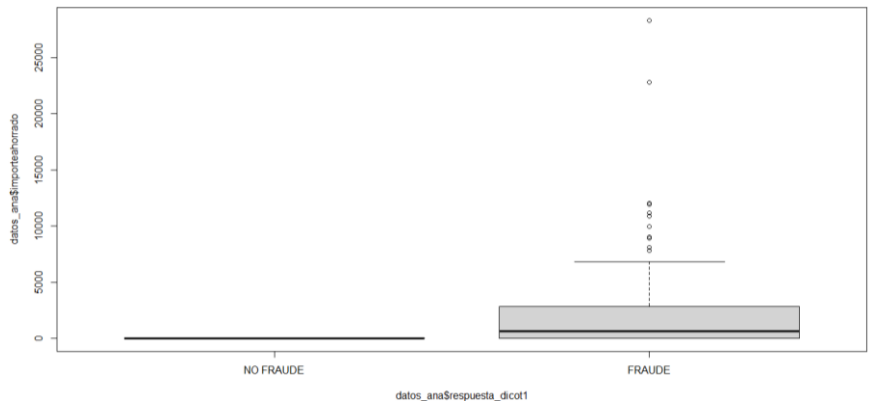

Ilustración 8. Diagrama de cajas para la variable independiente y la variable Importe ahorrado. Fuente: Elaboración propia en programa R

En los casos de no fraude , el importe ahorrado es de 0 debido, a la ausencia de ahorros. Sin embargo, en el grupo de fraude, se observan algunos valores atípicos. La superposición entre las distribuciones es notable, lo que sugiere la presencia de casos de fraude con montos de ahorro significativos.

6. Tipo de documento:

Documento	No fraude	Fraude
CIF	71	41
DNI/NIE	429	459

Tabla 2. Tabla de contingencia para la variable independiente y la variable tipo de documento. Fuente: Elaboración propia

La tabla de contingencia revela la distribución de observaciones según el "tipo de documento" y "fraude". En la categoría "No fraude", hay 71 observaciones con "CIF" y 429 con "DNI/NIE". En la categoría "fraude", hay 41 observaciones con "CIF" y 459 con "DNI/NIE".

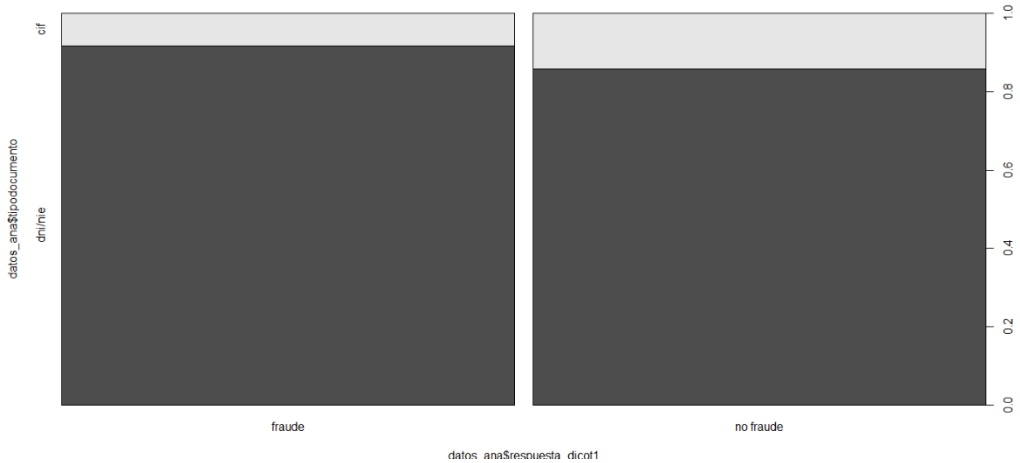

Ilustración 9. Gráfico de mosaico para la variable independiente y la variable Tipo de documento. Fuente: Elaboración propia en programa R.

7. Edad del conductor:

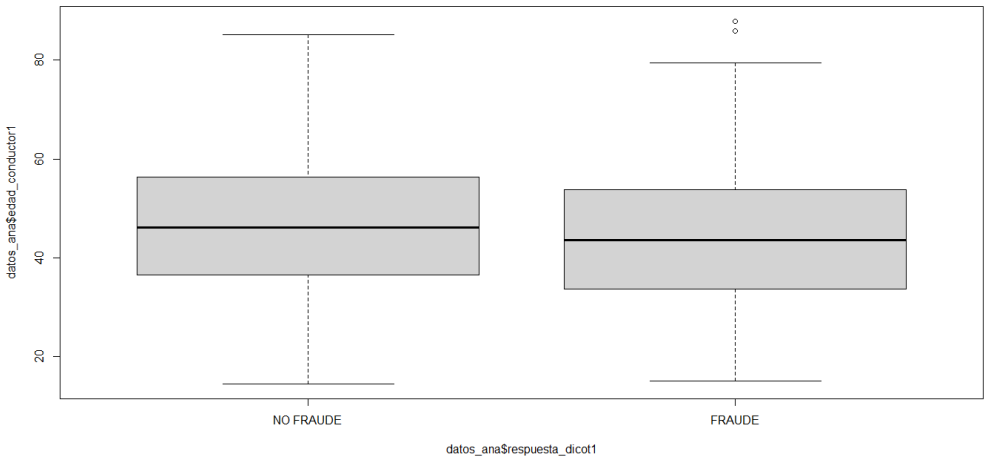

Ilustración 10. Diagrama de cajas para la variable independiente y la variable Edad del conductor. Fuente: Elaboración propia en programa R

La mediana de la edad del conductor es similar para los casos de fraude y para los casos sin fraude. Esto indica que, en general, no hay una diferencia significativa en la edad del conductor entre los dos grupos. El IQR es similar para ambos grupos, lo que indica que la dispersión de los datos es similar en ambos grupos. Hay algunos valores atípicos en ambos grupos, especialmente en el grupo sin fraude. Estos valores pueden ser casos de conductores muy jóvenes o mayores que cometen fraude, o errores en la medición de la edad.

La superposición entre las dos distribuciones es considerable. Esto significa que hay algunos casos de fraude en todas las edades y algunos casos sin fraude en todas las edades. Esto hace que sea difícil usar la edad del conductor para predecir con precisión si un caso es fraude o no.

8. Garantía agrupada:

Garantía agrupada	No fraude	Fraude
Asistencia	263	8
Atropello animales	0	0
Daños propios	24	15
F. Meteorológicos	0	0
Incendio	0	1
Lunas	52	13
Multa	0	0
Obligatorio RC	153	452
Robo	5	10
Viajeros	2	1

Tabla 3. Tabla de contingencia para la variable independiente y la variable garantía agrupada. Fuente: Elaboración propia

La tabla de contingencia revela que las categorías más comunes, como "asistencia", "daños propios", "lunas" y "obligatorio rc", tienen una concentración significativa tanto en la categoría "no fraude" como en "fraude". Sin embargo, se observa que algunas categorías, como "atropello animales", "f. meteorológicos" y "multas", no tienen ninguna observación en ninguna de las dos categorías. Además, se destaca que la categoría "incendio" solo tiene una observación en la categoría "fraude", mientras que "viajeros" presenta una observación tanto en "no fraude" como en "fraude".

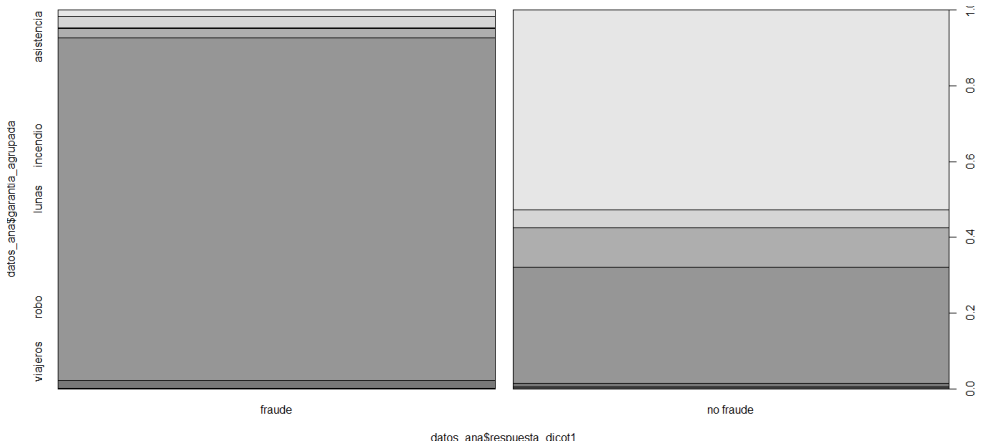

Ilustración 11. Gráfico de mosaico para la variable independiente y la variable garantía agrupada. Fuente: Elaboración propia en programa R

9. Antigüedad de la póliza:

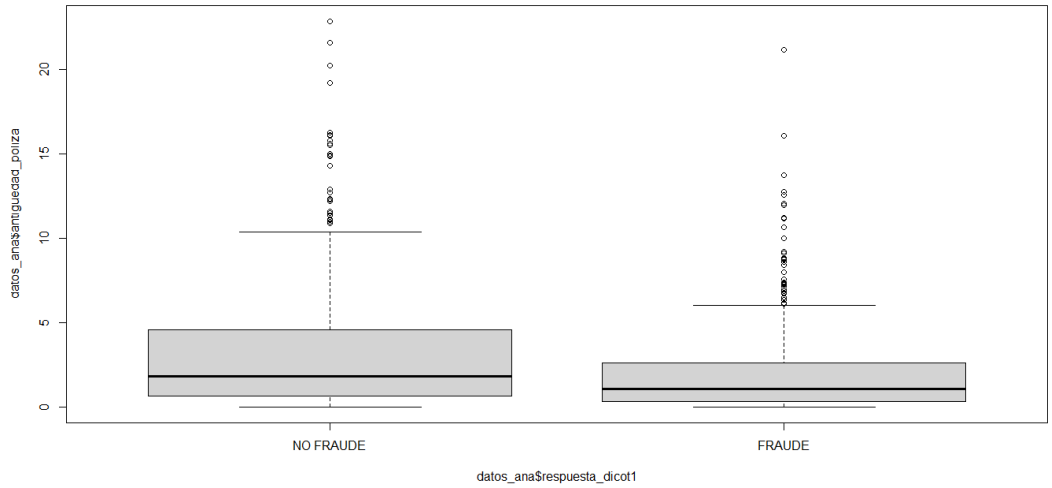

Ilustración 12. Diagrama de cajas para la variable independiente y la variable Antigüedad de la póliza. Fuente: Elaboración propia en programa R

La mediana de la antigüedad de la póliza es mayor en los casos sin fraude que en los casos de fraude, lo que sugiere que, en general, los clientes sin fraude tienen pólizas más antiguas. Aunque el rango intercuartílico (IQR) es similar en ambos grupos, indicando una dispersión de datos comparable, se observan valores atípicos en ambas categorías, especialmente en los casos sin fraude. La superposición entre las distribuciones de antigüedad de la póliza para casos de fraude y no fraude es considerable, lo que dificulta utilizar este factor para predecir con certeza si un caso es fraude o no.

10. Días hasta la notificación del siniestro:

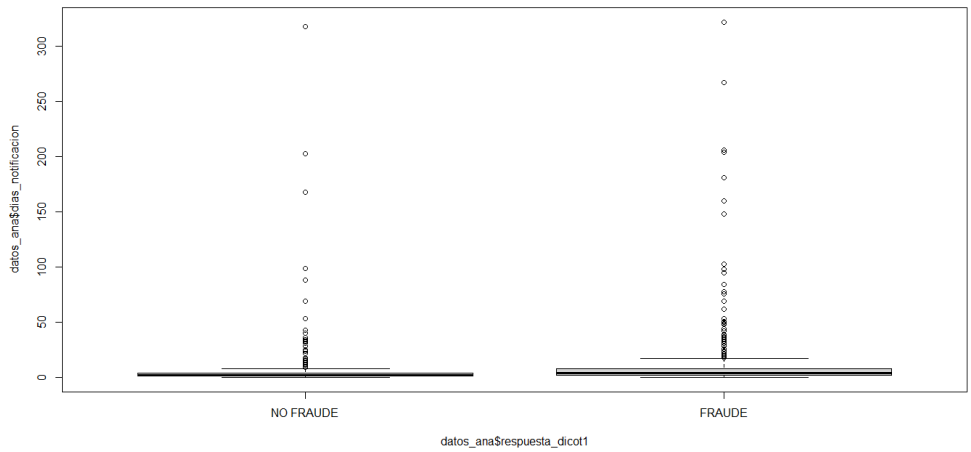

Ilustración 13. Diagrama de cajas para la variable independiente y la variable días hasta la notificación del siniestro. Fuente: Elaboración propia en programa R

La mediana del número de días de notificación del siniestro es menor en los casos de fraude que en los casos sin fraude, lo que sugiere que, en general, los casos de fraude se notifican más rápidamente. Aunque el rango intercuartílico (IQR) es similar en ambos grupos, indicando una dispersión de datos comparable, se observan valores atípicos en ambas categorías, especialmente en los casos sin fraude. La superposición entre las distribuciones de tiempo de notificación del siniestro para casos de fraude y no fraude es considerable, lo que dificulta utilizar este factor para predecir con certeza si un caso es fraude o no.

11. Forma de pago agrupada:

Forma de pago agrupada	No fraude	Fraude
Domiciliado	270	253
No domiciliado	230	247

Tabla 4. Tabla de contingencia para la variable independiente y la variable forma de pago agrupada. Fuente: Elaboración propia

La tabla de contingencia presenta la distribución de observaciones según las variables "forma de pago agrupado" y "fraude". En la categoría "NO FRAUDE", se registran 270 observaciones para la forma de pago "DOMICILIADO" y 230 para "EFECTIVO". Por otro lado, en la categoría "FRAUDE", se encuentran 253 observaciones para "DOMICILIADO" y 247 para "EFECTIVO". Esto indica que, en general, hay una distribución relativamente equilibrada entre las formas de pago para ambos grupos, tanto para casos de fraude como para aquellos sin fraude.

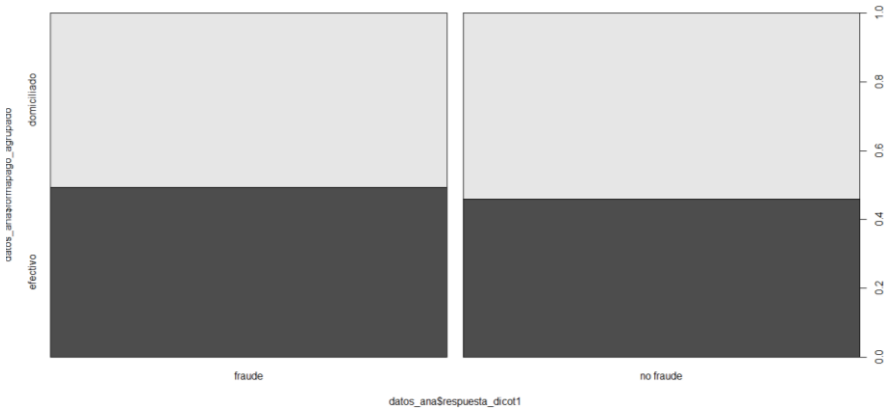

Ilustración 14. Gráfico de mosaico para la variable independiente y la variable forma de pago agrupada. Fuente: Elaboración propia en programa R

12. ID del siniestro:

Dado que esta variable es el número de identificación de cada siniestro, no se realizará la comparación con la variable fraude, ya que aporta relevancia para el estudio.

13. Provincia ID:

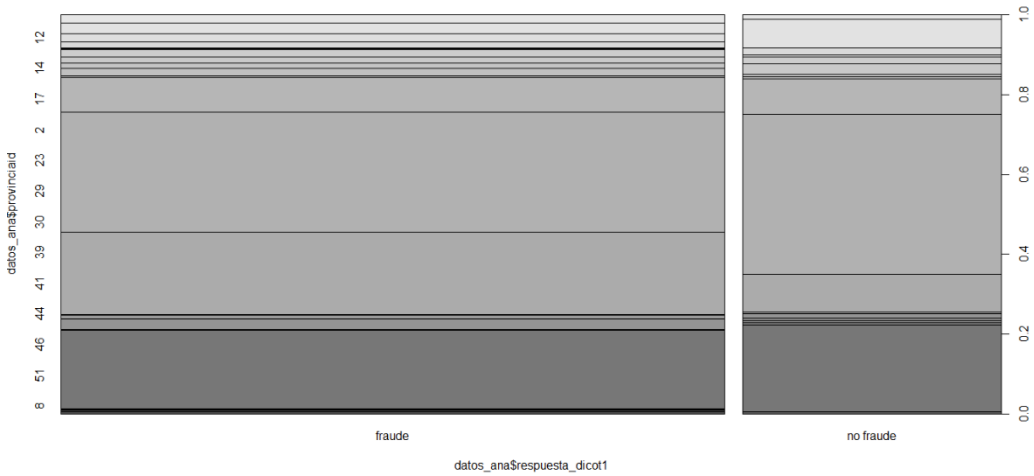

Ilustración 15. Gráfico para la variable independiente y la variable Provincia. Fuente: Elaboración en programa R

ID Provincia	No fraude	Fraude
11	9	2
12	13	13
13	9	0
14	7	3
16	1	1
17	1	0
18	9	3
2	7	5
21	6	1
23	9	1
28	1	0
29	41	16
3	139	72
30	95	17
33	0	1
39	1	0
4	4	0
41	12	2
43	0	1
44	1	1
45	0	1
46	91	39
5	1	0
21	1	0
7	2	1
8	2	0

Tabla 5. Tabla de contingencia para la variable independiente y la variable provincia. Fuente: Elaboración propia

La tabla de contingencia muestra la frecuencia conjunta de las categorías de las variables "provinciaid" y "respuesta_dicot1", representando la cantidad de casos clasificados como fraude o no fraude en diferentes provincias de España. Los números en la tabla revelan variaciones significativas en la ocurrencia de fraudes a nivel nacional, destacando la disparidad entre provincias. Por ejemplo, en la provincia 29 se registran 41 casos de fraude y 16 de no fraude. Este análisis proporciona una comprensión detallada de la distribución de casos fraudulentos y no fraudulentos, siendo fundamental para identificar patrones regionales en la incidencia de fraudes y para la toma de decisiones en la gestión y prevención de fraudes en el sector asegurador.

14. Valor del vehículo de mercado:

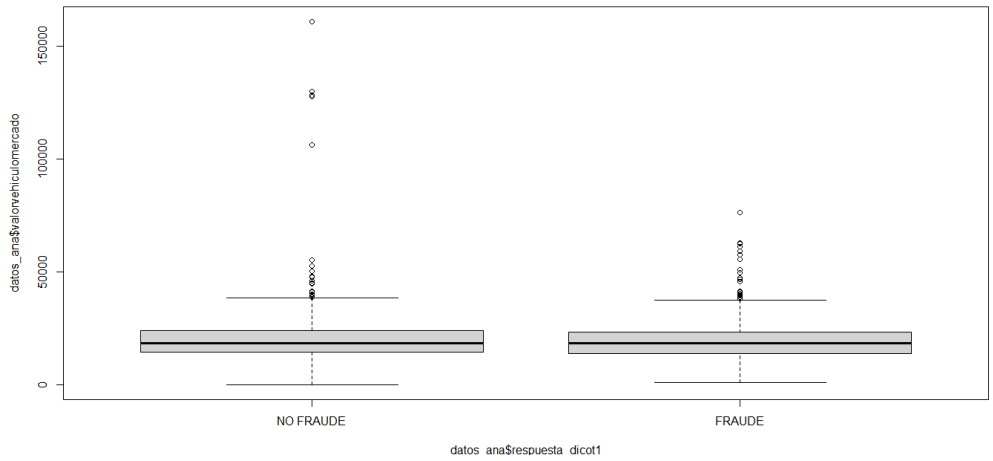

Ilustración 16. Diagrama de cajas para la variable independiente y la variable Valor del vehículo de mercado. Fuente: Elaboración propia en programa R

La mediana del valor del vehículo en el mercado es similar para casos de fraude y no fraude, indicando una falta de diferencia significativa en el valor del vehículo entre ambos grupos. Sin embargo, el rango intercuartílico (IQR) es mayor para casos de fraude, lo que sugiere una mayor dispersión de datos en estos casos. Se observan valores atípicos en ambos grupos, especialmente en el grupo de fraude, lo que podría representar vehículos de alto valor o errores en la medición de este. La considerable superposición entre las distribuciones de valores del vehículo para casos de fraude y no fraude dificulta su uso como predictor preciso de fraude.

15. Valor del vehículo de fábrica:

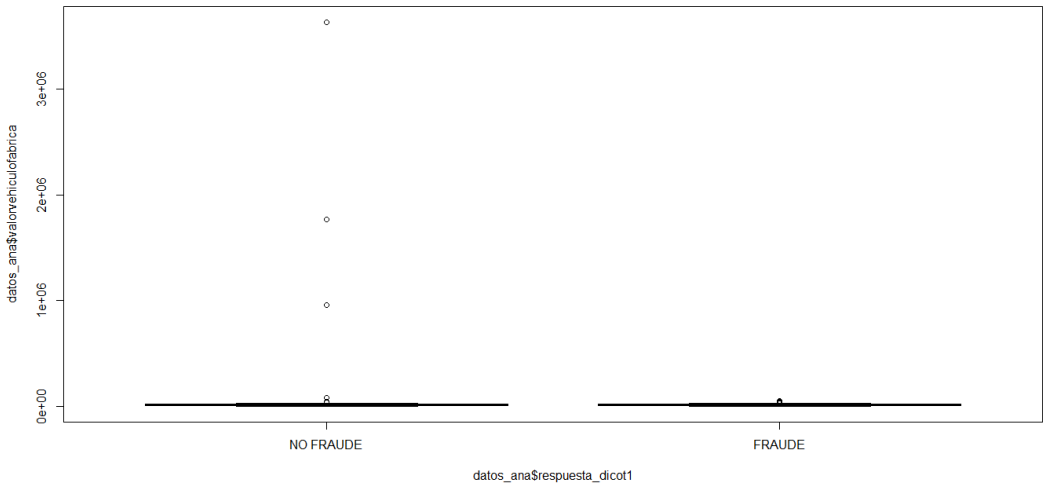

Ilustración 17. Diagrama de cajas para la variable independiente y la variable Valor del vehículo de fábrica. Fuente: Elaboración propia en programa R

La mediana del valor del vehículo de fábrica es mayor para los casos sin fraude que para los casos de fraude, sugiriendo que aquellos que no cometen fraude tienden a poseer vehículos con un valor de fábrica más alto. Aunque el rango intercuartílico (IQR) es similar en ambos grupos, indicando una dispersión de datos comparable, se observan valores atípicos en ambas categorías, particularmente en el grupo sin fraude. Esta discrepancia puede deberse a vehículos de fábrica muy costosos pertenecientes a personas que cometen fraude, o a errores en la medición del valor del vehículo. La considerable superposición entre las distribuciones de valores del vehículo de fábrica para casos de fraude y no fraude dificulta su uso como predictor preciso de fraude.

16. Potencia:

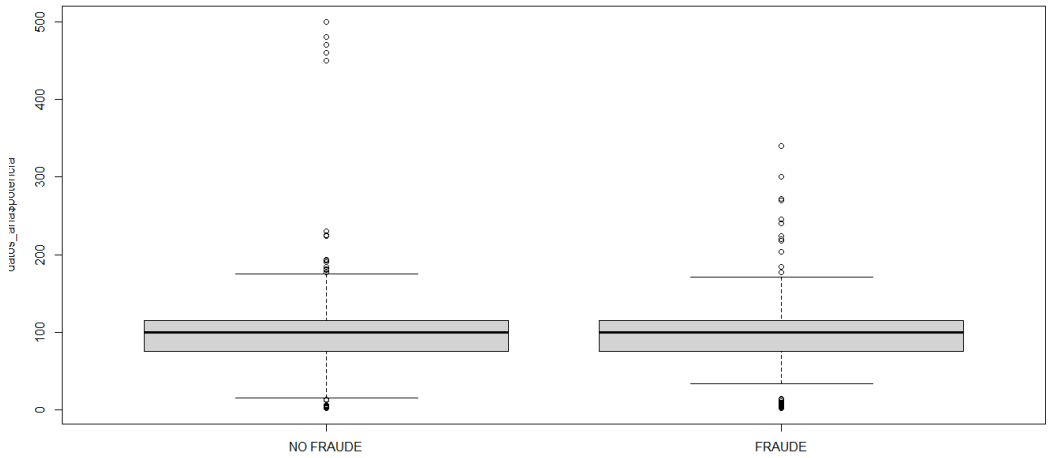

Ilustración 18. Diagrama de cajas para la variable independiente y la variable Potencia. Fuente: Elaboración propia en programa R

La mediana de la potencia es similar tanto para los casos de fraude como para los casos sin fraude, lo que sugiere que, en general, no existe una diferencia significativa en la potencia entre ambos grupos. El rango intercuartílico (IQR) también es similar en ambos grupos, indicando una dispersión de datos comparable. Se observan valores atípicos en ambos conjuntos de datos, que podrían ser casos de fraude o no fraude con potencia muy alta o baja, o simplemente errores en la medición de la potencia. La superposición entre las distribuciones de potencia para casos de fraude y no fraude es considerable, lo que dificulta utilizar este factor como predictor preciso de fraude.

17. Cilindrada:

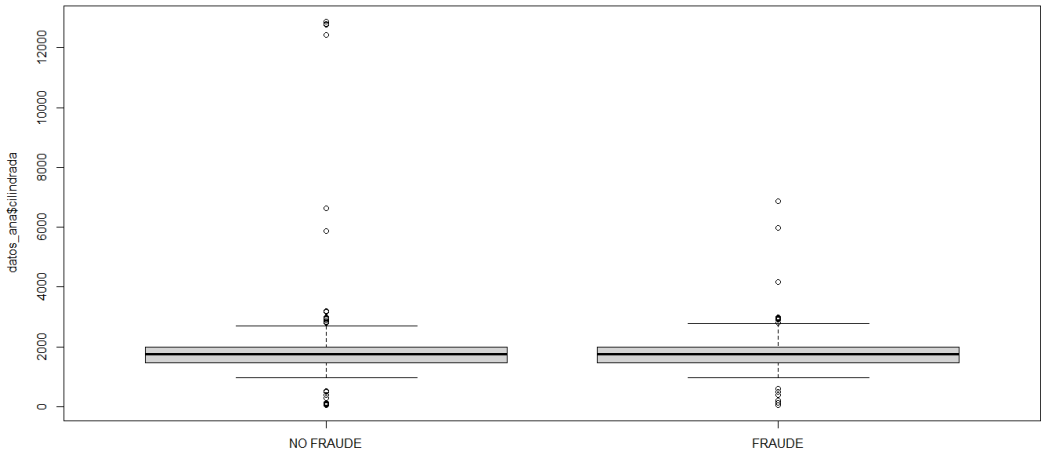

Ilustración 19. Diagrama de cajas para la variable independiente y la variable Cilindrada. Fuente: Elaboración propia en programa R

La mediana de la cilindrada es similar tanto para los casos de fraude como para los casos sin fraude, lo que sugiere que, en general, no hay una diferencia significativa en la cilindrada entre ambos grupos. El rango intercuartílico (IQR) también es similar en ambos grupos, lo que indica una dispersión de datos comparable. Se observan valores atípicos en ambos conjuntos de datos, que podrían ser casos de fraude o no fraude con cilindrada muy alta o baja, o simplemente errores en la medición de la cilindrada. La considerable superposición entre las distribuciones de cilindrada para casos de fraude y no fraude dificulta su uso como predictor preciso de fraude.

18. Peso del vehículo:

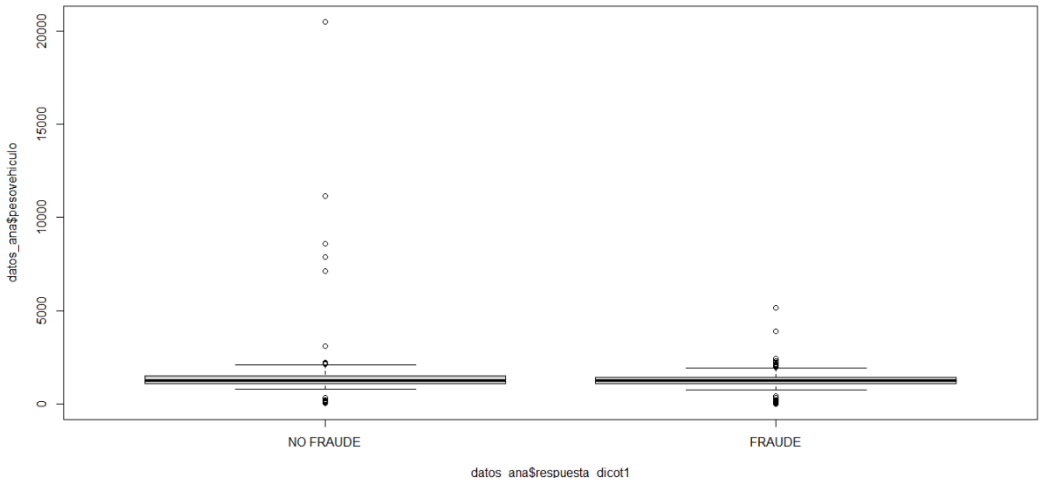

Ilustración 20. Diagrama de cajas para la variable independiente y la variable peso del vehículo. Fuente: Elaboración propia en programa R

La mediana del peso del vehículo es similar tanto para los casos de fraude como para los casos sin fraude, lo que sugiere que, en general, no hay una diferencia significativa en el peso del vehículo entre ambos grupos. El rango intercuartílico (IQR) también es similar en ambos grupos, lo que indica una dispersión de datos comparable. Se observan valores atípicos en ambos conjuntos de datos, que podrían ser casos de fraude o no fraude con peso del vehículo muy alto o bajo, o simplemente errores en la medición del peso. La considerable superposición entre las distribuciones de peso del vehículo para casos de fraude y no fraude dificulta su uso como predictor preciso de fraude.

19. Longitud:

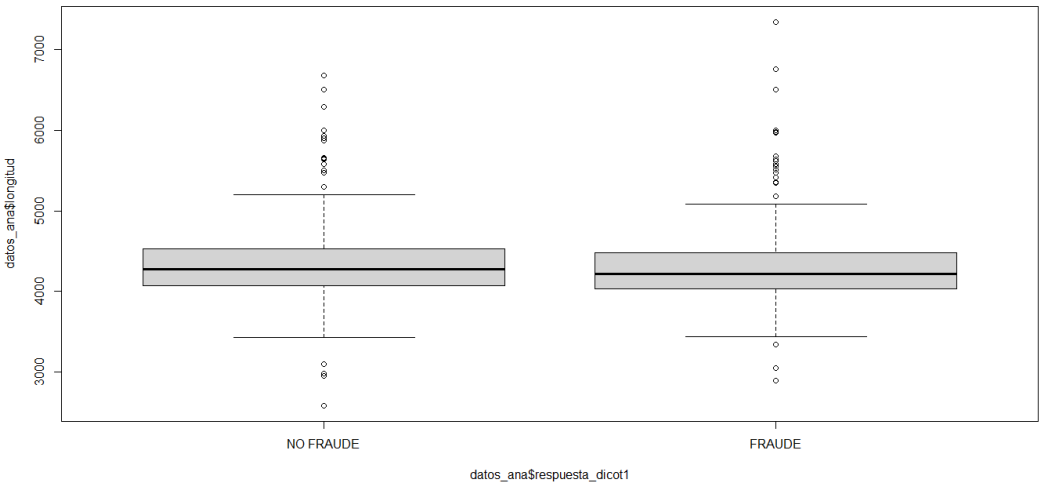

Ilustración 21. Diagrama de cajas para la variable independiente y la variable longitud. Fuente: Elaboración propia en programa R

La mediana de la longitud del vehículo es similar tanto para los casos de fraude como para los casos sin fraude, lo que sugiere que, en general, no hay una diferencia significativa en la longitud del vehículo entre ambos grupos. El rango intercuartílico (IQR) también es similar en ambos grupos, lo que indica una dispersión de datos comparable. Se observan valores atípicos en ambos conjuntos de datos, que podrían ser casos de fraude o no fraude con longitud del vehículo muy alta o baja, o simplemente errores en la medición. La considerable superposición entre las distribuciones de longitud del vehículo para casos de fraude y no fraude dificulta su uso como predictor preciso de fraude.

20. Scoring:

Scoring	No fraude	Fraude
Grupo avanza	18	38
Grupo expande	29	38
Grupo inicia	6	6
Grupo óptima	141	102
Grupo óptima plus	40	23
Grupo proyecta	235	266

Tabla 6. Tabla de contingencia para la variable independiente y la variable Scoring. Fuente: Elaboración propia

La tabla de contingencia revela la distribución de observaciones en relación con las variables "scoring" y "fraude" en distintos grupos. En el grupo "avanza", se registran 18 observaciones sin fraude y 38 con fraude, mientras que en "expande" hay 29 y 38 respectivamente. En el grupo "inicia", se encuentran 6 observaciones tanto sin fraude como con fraude. Por otro lado, en "optima" se observan 141 casos sin fraude y 102 con fraude, mientras que en "optima plus" se cuentan 40 y 23 respectivamente. Finalmente, en "proyecta", se registran 235 observaciones sin fraude y 266 con fraude.

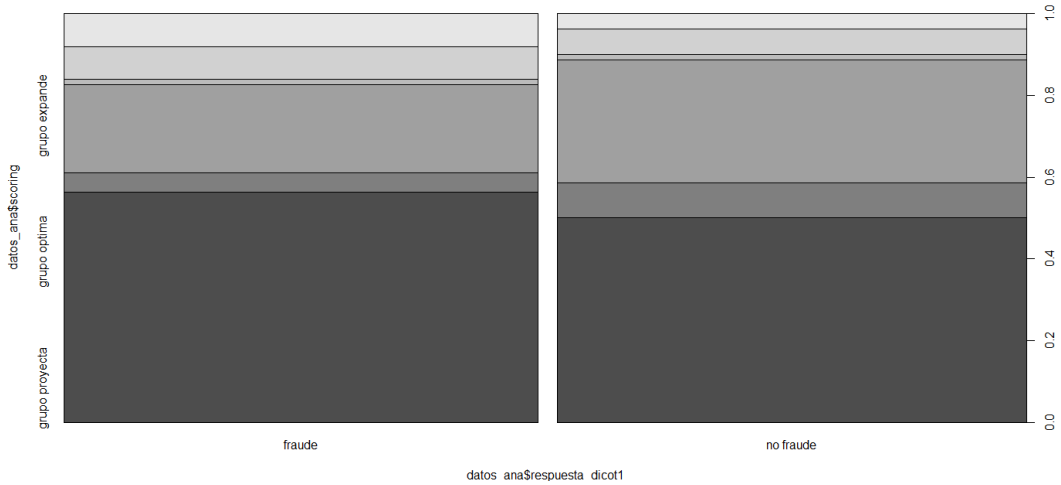

Ilustración 22. Gráfico de mosaicos para la variable independiente y la variable Scoring. Fuente: Elaboración propia en programa R

21. Aceptación sin antecedentes:

Aceptación sin antecedentes	No fraude	Fraude
No (0)	457	491
Si (1)	43	9

Tabla 7. Tabla de contingencia para la variable independiente y la variable Aceptación sin antecedentes. Fuente: Elaboración propia

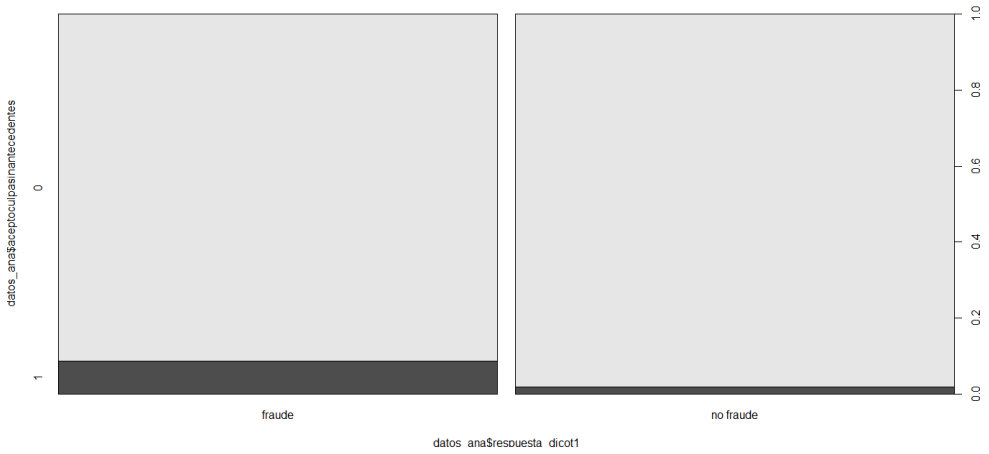

Ilustración 23. Gráfico de mosaico para la variable independiente y la variable Aceptación sin antecedentes. Fuente: Elaboración propia en programa R

La tabla de contingencia muestra la frecuencia conjunta de las categorías de las variables "aceptoculpasinantecedentes" y "respuesta_dicot1", revelando las proporciones de casos clasificados como fraude y no fraude en cada categoría. Para aquellos que no aceptaron la culpabilidad sin antecedentes (0), se observaron 457 casos clasificados como fraude y 491 casos como no fraude. Por otro lado, entre aquellos que sí aceptaron la culpabilidad sin antecedentes (1), se registraron 43 casos clasificados como fraude y 9 casos como no fraude.

4.1.2. Limpieza de base de datos

A continuación, se llevará a cabo un proceso de limpieza de la base de datos, debido a la detección de inconsistencias y errores durante el análisis descriptivo. Este procedimiento es esencial para garantizar la integridad y la fiabilidad de los datos, lo cual es fundamental para análisis posteriores y la toma de decisiones informadas (Godínez, 2011).

Durante el análisis descriptivo, se identificaron datos que parecen estar mal recogidos o que presentan inconsistencias. Estos pueden incluir valores atípicos, valores nulos, datos incorrectos o faltantes, entre otros. Por lo tanto, la limpieza de la base de datos implica corregir estos problemas para asegurar que los datos sean precisos, coherentes y útiles para análisis futuros.

Este proceso de limpieza puede implicar varias acciones, como:

1. **Identificación y eliminación de valores atípicos:** Se verificarán los valores extremos que puedan ser errores de entrada o errores de medición, y se corregirán o eliminarán según corresponda.
2. **Manejo de valores nulos:** Se evaluarán y manejarán los valores nulos de manera adecuada, ya sea eliminándolos, imputándolos o utilizando técnicas de interpolación según el contexto de los datos y el impacto en el análisis.
3. **Verificación de la consistencia de los datos:** Se revisarán las relaciones entre diferentes variables para identificar posibles inconsistencias o errores lógicos, y se corregirán según sea necesario.
4. **Estandarización de formatos:** Se asegurará de que los datos estén en el formato correcto y consistente, como fechas, números y texto, para facilitar su análisis y comprensión.

5. Validación de datos: Se realizarán comprobaciones adicionales para garantizar que los datos cumplan con los criterios establecidos y representen con precisión el fenómeno o proceso que se está estudiando.

En resumen, la limpieza de la base de datos es un paso crucial en el proceso de análisis de datos, ya que asegura la calidad y la fiabilidad de los datos utilizados para la toma de decisiones. Al abordar las inconsistencias y los errores detectados durante el análisis descriptivo, se garantiza que los resultados posteriores sean precisos y confiables, lo que proporciona una base sólida para la investigación y el análisis subsiguiente.

- Importe de Siniestro. Se procede a eliminar los siguientes 27 registros del conjunto de datos debido a que el importe del siniestro no puede ser negativo: 29099, 487503, 493803, 41590, 544692, 479970, 533553, 404576, 495576, 539106, 456085, 484762, 474672, 35749, 432577, 559810, 6796, 45469, 482169, 511528, 18183, 412403, 379993, 478405, 543284, 12794, 557233.
- Edad del conductor: Se procede a eliminar los siguientes 9 registros del conjunto de datos debido a que la edad del conductor es menor a 18 años, lo cual no es válido para seguros de coches: 509355, 543284, 458643, 429701, 65922, 218757, 403910, 147527, 206147.
- SiniestroId. La variable "siniestroid" se eliminará del conjunto de datos, ya que no es relevante para el estudio en cuestión. Esta variable representa simplemente un número de referencia para el siniestro y no proporciona información sustancial para el análisis de seguros de coches. Por lo tanto, su exclusión no afectará negativamente el análisis, pero ayudará a simplificar el conjunto de datos y a enfocar el estudio en las variables pertinentes.

  El proceso de eliminación de la variable "siniestroid" se lleva a cabo con el fin de mejorar la calidad y la utilidad del conjunto de datos para el análisis posterior. Al reducir la complejidad y el ruido innecesario en los datos, se facilita la identificación de patrones y tendencias relevantes que pueden ser de interés para el estudio de seguros de coches.

- Cilindrada: Se procede a eliminar los siguientes 26 registros del conjunto de datos debido a que la cilindrada es menor a 60cc, lo cual indica que se trata de motocicletas o ciclomotores en lugar de coches: 456085, 509355, 550887, 458643, 40648, 443347, 235374, 429701, 237078, 328310, 282905, 322048, 218757, 151029, 403910, 298243, 238313, 123698, 147527, 297659, 187322, 220418, 107738, 140381, 206147.
- Peso: Se han eliminado los siguientes 10 registros del conjunto de datos debido a que el peso de los vehículos supera los 3000 kg, lo cual es inusual y probablemente indica errores en los datos o la presencia de camiones: 48538, 484392, 28961, 285911, 417497, 382166, 350109, 149378, 163727, 416503.

  Los siguientes 57 registros se han eliminado del conjunto de datos debido a que el peso indicado es inferior a 500 kg, lo cual se considera poco probable y posiblemente sean motocicletas: 543284, 154231, 173201, 261063, 513944, 229004, 201355,

517413, 423098, 487630, 247773, 543695, 441536, 330977, 145860, 539158, 385655, 31321, 475483, 516483, 202952, 5177, 368903, 248403, 6796, 446032, 109477, 62922, 40648, 180539, 484197, 328310, 403910, 458643, 218757, 429701, 322048, 123698, 147527, 509355, 238313, 183130, 206147, 443347, 282905, 151029, 298243, 297659, 550887, 456085, 220418, 140381, 187322, 235374, 107738, 237078, 367891.

- Valor vehículo de fábrica Se ha decidido excluir tres conjuntos de datos de la referencia del valor del vehículo de fábrica. Estos conjuntos fueron eliminados debido a que el valor resultante era exorbitante. Además, al compararlo con el valor del vehículo en el mercado, se encontraron discrepancias significativas. Los valores excluidos son los siguientes: 424274, 346479, 330220.
- El proceso de obtención de la tabla se realizó utilizando el código *mine.plot(datos_ana)* en el entorno de programación R. En esta representación, cada dato faltante se muestra visualmente mediante una línea roja, lo que facilita la identificación de los valores ausentes en los datos.

- Como se puede observar en la tabla inferior, las variables "fecha de intervención", "importe siniestro", "importe invertido" y "importe ahorrado" tienen un porcentaje de datos faltantes del 80%, 70%, 70% y 70% respectivamente. Esta ausencia de datos representa un impedimento significativo, por lo que se ha optado por eliminar estas variables para el estudio del fraude.
- Por otro lado, la variable "provincia" presenta un porcentaje de datos faltantes del 36%, lo cual también es considerable. Sin embargo, por el momento no se eliminará esta variable. Posteriormente, una vez realizado el modelo estadístico, se determinará su relevancia. Dependiendo de los resultados, se decidirá si mantenerla en el análisis o descartarla.

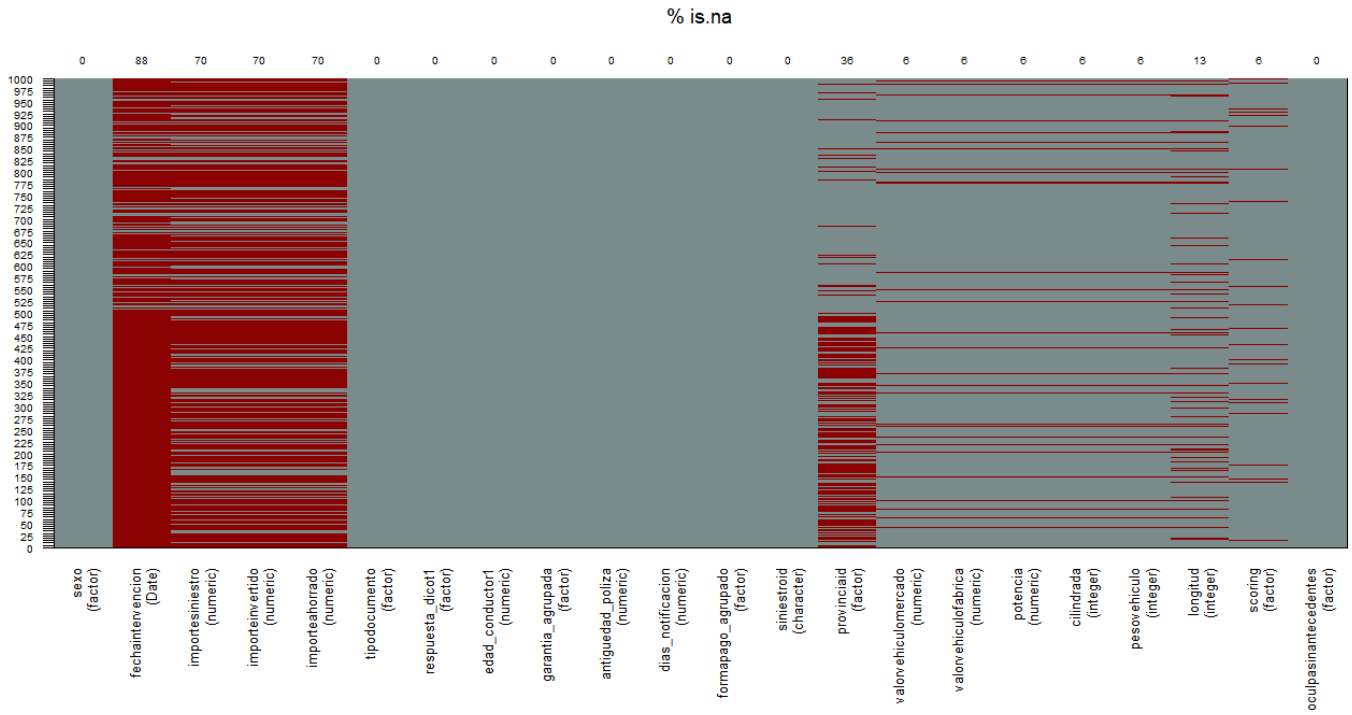

Ilustración 24. Gráfica de observaciones sin datos, resultante del código "mine.plot(datos_ana". Fuente: Elaboración propia)

En conclusión, luego de examinar los datos, se determina que se eliminarán cinco variables: siniestroid, fecha de intervención, importe siniestro, importe invertido e importe ahorrado. Además, se procederá a eliminar 99 observaciones del conjunto de datos: 5177, 6796, 12794, 18183, 28961, 29099, 31321, 35749, 40648, 41590, 45469, 48538, 62922, 65922, 107738, 109477, 123698, 140381, 145860, 147527, 149378, 151029, 154231, 163727, 173201, 180539, 183130, 187322, 201355, 202952, 206147, 218757, 220418, 229004, 235374, 237078, 238313, 247773, 248403, 261063, 282905, 285911, 297659, 298243, 322048, 328310, 330220, 330977, 346479, 350109, 367891, 368903, 379993, 382166, 385655, 403910, 404576, 412403, 416503, 417497, 423098, 424274, 429701, 432577, 441536, 443347, 446032, 456085, 458643, 474672, 475483, 478405, 479970, 482169, 484197, 484392, 484762, 487503, 487630, 493803, 495576, 509355, 511528, 513944, 516483, 517413, 533553, 539106, 539158, 543284, 543695, 544692, 550887, 557233, 559810.

Es importante destacar que, al revisar detenidamente el listado de observaciones a eliminar, se ha observado que muchas de estas observaciones no solo presentan errores en una variable, sino en múltiples variables. Este hallazgo refuerza aún más la decisión de eliminar estas observaciones, ya que sugiere que podrían haber ocurrido errores sistémicos o inconsistencias en la recopilación de datos en esas instancias.

La presencia de errores en varias variables de una misma observación puede tener un impacto significativo en la validez y la fiabilidad de los resultados obtenidos a partir de esos datos. Al eliminar estas observaciones, se asegura que el conjunto de datos resultante sea más coherente y representativo de los fenómenos o procesos que se están estudiando.

Por lo tanto, la decisión de eliminar estas observaciones con errores en múltiples variables se respalda no solo en la necesidad de mantener la integridad de los datos, sino también en la mejora de la calidad y la confiabilidad de los análisis posteriores que se realicen con estos datos. Esto subraya la importancia de realizar una limpieza exhaustiva de los datos antes de realizar cualquier análisis o interpretación.

Con la siguiente secuencia de códigos se eliminaron las observaciones comentadas anteriormente:

```
Base1<-datos_ana[datos_ana$edad_conductor1 >= 18, ]
Base2<-Base1[Base1$cilindrada >=60, ]
Base3<-Base2[Base2$pesovehiculo <=3000, ]
Base4<-Base3[Base3$pesovehiculo >=500, ]
Base5<-Base4[Base4$valorvehiculofabrica <=85000, ]
Base5<-Base5[Base5$importesiniestro >0, ]
```

Con el código `bananew<-Base5[,c(1,6:12,14:22)]` se han eliminado las variables que no son necesarias para el estudio.

4.1.2.1. Imputación de valores faltantes

En este apartado, nos enfrentamos al desafío de manejar datos faltantes, identificados comúnmente como "NA", en nuestro conjunto de observaciones. Para abordar esta situación, recurrimos a una técnica de imputación estocástica implementada en R a través de la biblioteca

"missForest". Este método, conocido como "MissForest", es una herramienta valiosa que preserva la variabilidad original de la variable al imputar los valores faltantes (Stekhoven, D. J. & Bühlmann, P., 2012).

La imputación estocástica con "MissForest" se distingue por su capacidad para predecir los valores faltantes basándose en los datos disponibles y en otras variables presentes en el conjunto de datos. En lugar de simplemente asignar valores promedio o mediana, "MissForest" utiliza un enfoque más sofisticado, empleando múltiples árboles de decisión ajustados a diferentes subconjuntos de datos. Esta estrategia permite capturar la complejidad y la estructura de los datos de manera más precisa, lo que resulta en una imputación más efectiva.

Una ventaja destacada de este método es su capacidad para manejar conjuntos de datos con una combinación de variables categóricas y numéricas, así como para abordar tanto patrones de falta de datos aleatorios como no aleatorios. Además, al preservar la variabilidad original de la variable, "MissForest" contribuye a mantener la integridad de los datos y a garantizar que las conclusiones derivadas del análisis posterior sean más sólidas y confiables.

Como se observó en la ilustración 32, se identificaron datos faltantes en varias variables del conjunto de datos. La variable "provinciasid" presentaba un 36% de valores faltantes, mientras que "Valor del vehículo en el mercado", "valor del vehículo en la fábrica", "potencia", "cilindrada", "peso del vehículo" y "scoring" tenían un 6% de valores ausentes cada una. Por su parte, la variable "longitud" tenía un 13% de datos faltantes. Ante esta situación, se procedió a realizar la imputación de datos utilizando el método "missForest" en R, como se muestra en la ilustración inferior. Este proceso permitió completar los valores faltantes en el conjunto de datos, asegurando así su integridad y preparándolos para análisis posteriores con datos completos y confiables.

```
> library(missForest)
> bbddimp<-missForest(bananew)
> summary(bbddimp)
         Length Class      Mode
ximp     17     data.frame list
OOBerror  2     -none-     numeric
> bbddimp_final<-(bbddimp$ximp)
> summary(bbddimp_final)
 sexo    tipodocumento  respuesta_dicot1 edad_conductor1      garantia_agrupada
 a: 34   cif    : 21   fraude   :788     Min.   :19.10   asistencia    : 70
 f: 56   dni/nie:880   no fraude:113     1st Qu.:37.97   daa?os propios: 12
 m:811                                   Median :43.39   incendio      :  0
                                         Mean   :43.18   lunas         : 15
                                         3rd Qu.:47.66   obligatorio_rc:801
                                         Max.   :78.67   robo          :  2
                                                         viajeros      :  1
 antiguedad_poliza dias_notificacion   formapago_agrupado  provinciaid  valorvehiculomercado
 Min.   : 0.000    Min.   :  0.000   domiciliado:483      3      :450   Min.   : 6870
 1st Qu.: 2.910    1st Qu.:  4.720   efectivo   :418      46     :331   1st Qu.:21244
 Median : 3.362    Median :  4.990                       30     : 49   Median :21264
 Mean   : 3.395    Mean   :  5.617                       29     : 25   Mean   :21430
 3rd Qu.: 4.397    3rd Qu.:  5.230                       2      : 10   3rd Qu.:21380
 Max.   :21.169    Max.   :206.000                       11     :  8   Max.   :76450
                                                         (Other): 28
 valorvehiculofabrica    potencia       cilindrada    pesovehiculo     longitud
 Min.   :    0        Min.   : 45.0   Min.   : 970   Min.   : 810   Min.   :3430
 1st Qu.:17078        1st Qu.:107.8   1st Qu.:1667   1st Qu.:1329   1st Qu.:4305
 Median :17103        Median :108.0   Median :1700   Median :1332   Median :4316
 Mean   :17184        Mean   :108.1   Mean   :1723   Mean   :1337   Mean   :4329
 3rd Qu.:17157        3rd Qu.:110.0   3rd Qu.:1756   3rd Qu.:1357   3rd Qu.:4345
 Max.   :55647        Max.   :340.0   Max.   :2993   Max.   :2165   Max.   :7345
```

Ilustración 25. Uso del código MissForest en R. Fuente: Elaboración propia

Después de aplicar el método "MissForest" en R, como se puede observar en las imágenes inferiores, se aprecia una notable transformación en los datos. Las observaciones que inicialmente contenían valores faltantes (representados como "NA") ahora muestran valores asignados. Este cambio se debe a que "MissForest" emplea una estrategia sofisticada para imputar los datos faltantes en cada registro.

Para cada variable con datos faltantes, "MissForest" utiliza el resto de las variables disponibles en el conjunto de datos para realizar predicciones y asignar valores a los registros afectados. Esto implica que, en lugar de simplemente asignar un valor promedio o mediano, el algoritmo de "MissForest" aprovecha la información disponible en las otras variables para realizar estimaciones más precisas y detalladas. Al utilizar un enfoque basado en múltiples árboles de decisión, "MissForest" logra capturar la complejidad y las interacciones entre las variables, lo que conduce a una imputación más efectiva y a una mejor preservación de la variabilidad original de los datos.

En resumen, la aplicación de "MissForest" ha permitido llenar los huecos causados por los datos faltantes de manera integral y precisa, utilizando la información disponible en el conjunto de datos para generar valores estimados que sean coherentes y relevantes para cada observación. Esto contribuye a una base de datos más completa y confiable, preparándola para el análisis posterior con mayor precisión y confianza, así poder obtener un modelo de regresión logística más sólido y robusto.

	sexo	tipodocumento	respuesta_dicot1	edad_conductor1	garantia_agrupada	antiguedad_poliza
NA	NA	NA	NA	NA	NA	NA
9699	m	dni/nie	no fraude	68.01095	obligatorio_rc	0.33675565
NA.1	NA	NA	NA	NA	NA	NA
NA.2	NA	NA	NA	NA	NA	NA
NA.3	NA	NA	NA	NA	NA	NA
NA.4	NA	NA	NA	NA	NA	NA
NA.5	NA	NA	NA	NA	NA	NA
NA.6	NA	NA	NA	NA	NA	NA
NA.7	NA	NA	NA	NA	NA	NA

Ilustración 26. Base de datos antes del uso de MissForest en R. Fuente: Elaboración propia

	sexo	tipodocumento	respuesta_dicot1	edad_conductor1	garantia_agrupada	antiguedad_poliza
NA.47	m	dni/nie	fraude	43.38570	obligatorio_rc	3.36234999
NA.48	m	dni/nie	fraude	47.65988	obligatorio_rc	4.71597536
NA.49	m	dni/nie	fraude	37.26831	obligatorio_rc	2.91009811
NA.50	m	dni/nie	fraude	47.65988	obligatorio_rc	4.71597536
46155	m	dni/nie	no fraude	40.87337	asistencia	0.06570842
474734	m	dni/nie	no fraude	44.95003	asistencia	0.54757016
537627	a	dni/nie	no fraude	54.09993	obligatorio_rc	2.48596851
NA.51	m	dni/nie	fraude	39.48131	obligatorio_rc	4.39682409
26078	m	dni/nie	no fraude	28.71458	asistencia	1.46748802

Ilustración 27. Base de datos después del uso de MissForest en R. Fuente: Elaboración propia

4.1.3. Variaciones de las Variables imputadas

Este análisis descriptivo se basa en un conjunto de datos que contiene información sobre siniestros de seguros de automóviles. El objetivo es presentar una descripción completa y precisa de las variables que componen el conjunto de datos, utilizando diferentes técnicas estadísticas y visualizaciones.

4.1.3.1. Variables Numéricas:

1. Valor vehículo mercado: Después de la imputación de datos, la media disminuyó ligeramente a aproximadamente 20,224.84€ desde los 20,935.62€ previos, indicando un ajuste en el valor promedio de los vehículos. Sin embargo, la desviación estándar aumentó a aproximadamente 12,832.60€ desde los 12,794.40€ anteriores, lo que sugiere una mayor dispersión de los precios alrededor de la nueva media.
2. Valor vehículo fabrica: La variable después de la imputación, la media se mantuvo prácticamente igual.
3. Potencia: Después de la imputación de datos la media aumentó significativamente a aproximadamente 109.505 unidades desde los 100.042 unidades anteriores. Además, la desviación estándar disminuyó considerablemente a aproximadamente 19.183 unidades.
4. Cilindrada: Después de la imputación de datos, la media experimentó una ligera disminución, pasando de aproximadamente 1.795,331cc a 1743.643cc. Además, la desviación estándar disminuyó notablemente de aproximadamente 1.081,833cc a 204.010cc, lo que sugiere una menor dispersión de los datos alrededor de la media después de la imputación.
5. Peso vehículo: Después de la imputación de datos, la media experimentó una ligera disminución, pasando de aproximadamente 1.329,610 Kg a 1.350,076 Kg, indicando un ajuste en el valor promedio del peso del vehículo. Además, la desviación estándar disminuyó significativamente de aproximadamente 968,228 Kg a 141.993 Kg, lo que sugiere una menor dispersión de los datos alrededor de la media después de la imputación.
6. Longitud: Después de la imputación de datos, la media aumentó ligeramente, indicando un cambio en el valor promedio de la longitud de los vehículos. Además, la desviación estándar disminuyó notablemente de aproximadamente 484.133 mm a 235.656 mm, lo que sugiere una reducción en la dispersión de los datos.

4.1.3.2. Variables Categóricas:

7. Provincia id: Antes de la imputación, se observa una distribución más equilibrada entre los casos de fraude y no fraude en la mayoría de las provincias. Por ejemplo, en la provincia con ID 3, se registran 139 casos de fraude y 72 de no fraude. Sin embargo, después de la imputación, hay un cambio significativo en la distribución, con una provincia (ID 46) mostrando una prevalencia mucho mayor de casos de fraude (671) en comparación con los 39 casos de no fraude.
8. Scoring: Antes de la imputación, la clase más común en la variable "scoring" era "grupo optima", que representaba aproximadamente el 53.2% de los registros, mientras que después de la imputación, esta clase se convirtió en la más dominante, representando el 87.9% de los datos. Además, la clase minoritaria, "grupo inicia", que constituía cerca del 1.3% antes de la imputación, experimentó una disminución significativa en representación, convirtiéndose en "grupo avanza", que constituye solo el 0.4% de los registros después de la imputación.

4.1.3.3. Comparación de las variables con la variable respuesta Fraude después de la imputación.

En este caso procederemos a comparar las variables con la variable respuesta fraude, observando por medio de las gráficas sus distribuciones y relaciones potenciales, lo que nos permitirá identificar patrones y posibles correlaciones con el comportamiento fraudulento.

5. Provincia ID:

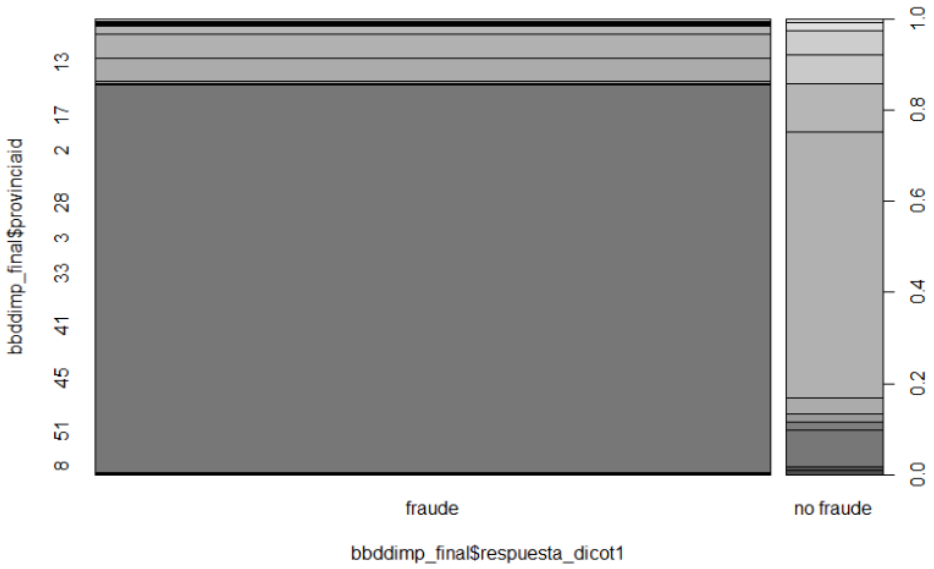

Ilustración 28. Gráfico de mosaicos para la variable independiente y la variable Provincia después de la imputación de datos. Fuente: Elaboración propia en programa R

Después de la imputación, se observa una reorganización significativa en la distribución de casos clasificados como fraude y no fraude en diversas provincias de España. Mientras que antes de la imputación se destacaban casos dispares entre provincias, como los 41 casos de fraude en la provincia 29, después de la imputación se identifican nuevas tendencias. Por ejemplo, la provincia 3 emerge como la que presenta la mayor cantidad de casos, con 108 en total, mientras que varias provincias, como las provincias 14, 16, 17, 33 y 43, no registran casos ni de fraude ni de no fraude después de la imputación. Además, se destaca la presencia de varias provincias con un número bajo de casos, entre 1 y 4, como las provincias 5, 7, 8, 21, 23, 30, 41, 44 y 51.

ID Provincia	No fraude	Fraude
11	5	1
12	1	2
13	1	0
14	0	0
16	0	0
17	0	0
18	2	6
2	2	7
21	1	0
23	1	0
28	0	0
29	13	12
3	42	66
30	40	4
33	0	0
39	0	0
4	3	2
41	3	0
43	0	0
44	0	0
45	0	2
46	671	9
5	0	0
21	0	0
7	2	1
8	1	1

Tabla 8.Tabla de contingencia para la variable independiente y la variable provincia después de la imputación de datos. Fuente: Elaboración propia

6. Valor del vehículo de mercado:

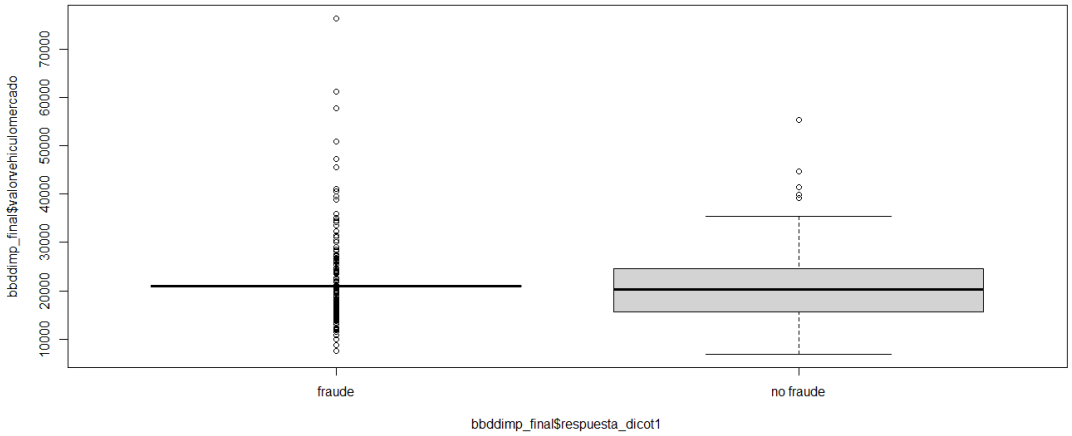

Ilustración 29. Diagrama de cajas para la variable independiente y la variable Valor del vehículo de mercado después de la imputación de datos. Fuente: Elaboración propia en programa R

Después del proceso de imputación, el aumento en el IQR para el grupo de fraude indica una mayor dispersión en los datos de valor del vehículo en este grupo después de la imputación. Esto implica que hay más variabilidad en el precio de mercado de los vehículos fraudulentos después del proceso de imputación.

7. Valor del vehículo de fábrica:

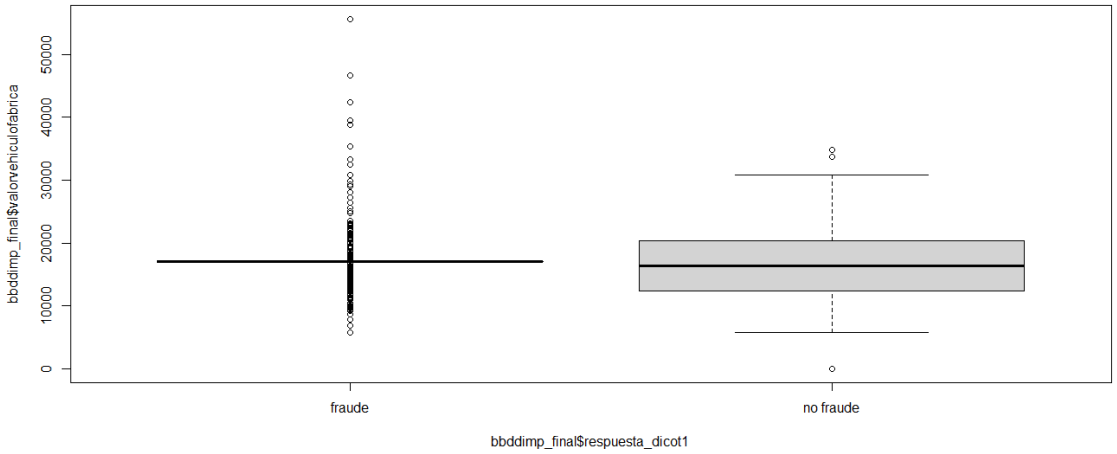

Ilustración 30. Diagrama de cajas para la variable independiente y la variable Valor del vehículo de fábrica después de la imputación de datos. Fuente: Elaboración propia en programa R

Después de la imputación, se registra un cambio en la dispersión de los datos, con un aumento en el IQR para el grupo de fraude, lo que indica una mayor variabilidad en el precio inicial de los vehículos fraudulentos después de la imputación. Además, se nota una continuación en la frecuencia de valores atípicos en ambos grupos, con una mayor incidencia en el grupo de fraude después de la imputación. Esto sugiere que los casos de fraude pueden implicar vehículos con valores iniciales más extremos o errores en la medición del valor después del proceso de imputación. A pesar de estas diferencias, la superposición entre las distribuciones de ambos grupos sigue siendo considerable, lo que subraya la limitación de utilizar únicamente el valor del vehículo en fábrica para distinguir entre casos de fraude y no fraude con precisión.

8. **Potencia**:

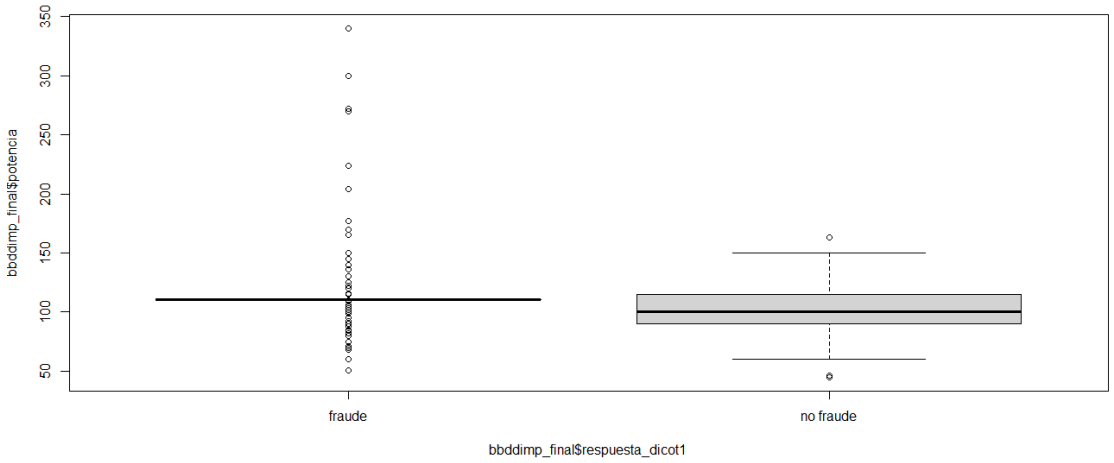

Ilustración 31. Diagrama de cajas para la variable independiente y la variable Potencia después de la imputación de datos. Fuente: Elaboración propia en programa R

Después de la imputación, la principal diferencia en la interpretación radica en la evidencia visual proporcionada por la gráfica de dispersión que muestra la relación entre la potencia del vehículo y la respuesta dicotómica de fraude o no fraude. Se observa una tendencia general representada por una línea de regresión lineal que indica que, a medida que la potencia del vehículo aumenta, la probabilidad de fraude disminuye. Esto contrasta con el análisis anterior, que se basaba en medidas descriptivas como la mediana y el IQR, donde se encontró una falta de diferencia significativa en la potencia entre los grupos de fraude y no fraude. Además, se señala que la relación no es perfectamente lineal, lo que sugiere la influencia de otros factores en la probabilidad de fraude además de la potencia del vehículo.

9. **Cilindrada**:

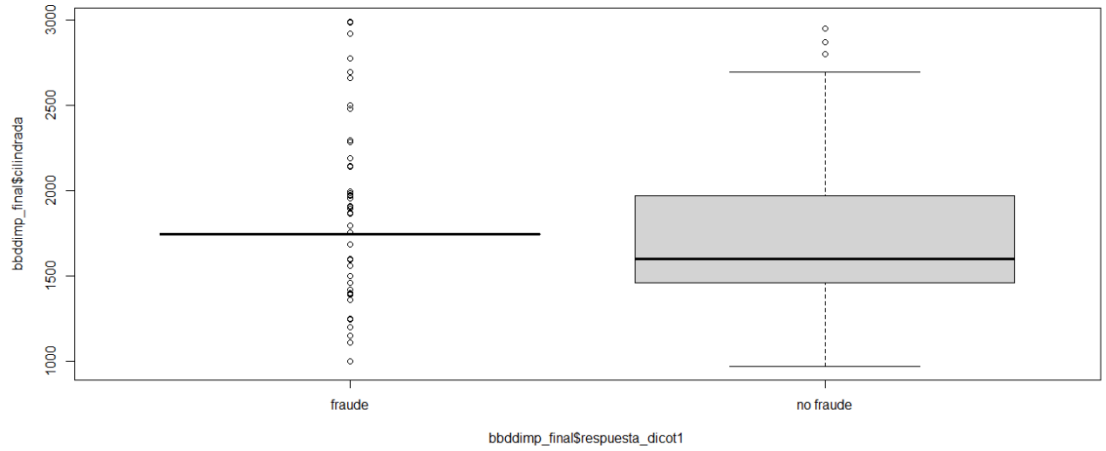

Ilustración 32. Diagrama de cajas para la variable independiente y la variable Cilindrada después de la imputación de datos. Fuente: Elaboración propia en programa R

Después de la imputación, se registra un cambio en la dispersión de los datos, con un aumento en el IQR para el grupo de fraude. Esto indica una mayor variabilidad en el tamaño del motor de los vehículos fraudulentos después de la imputación. Además, se nota una continuación en la frecuencia de valores atípicos en ambos grupos, con una mayor incidencia en el grupo de fraude después de la imputación. Esto sugiere que los casos de fraude pueden implicar vehículos con cilindradas más.

10. Peso del vehículo:

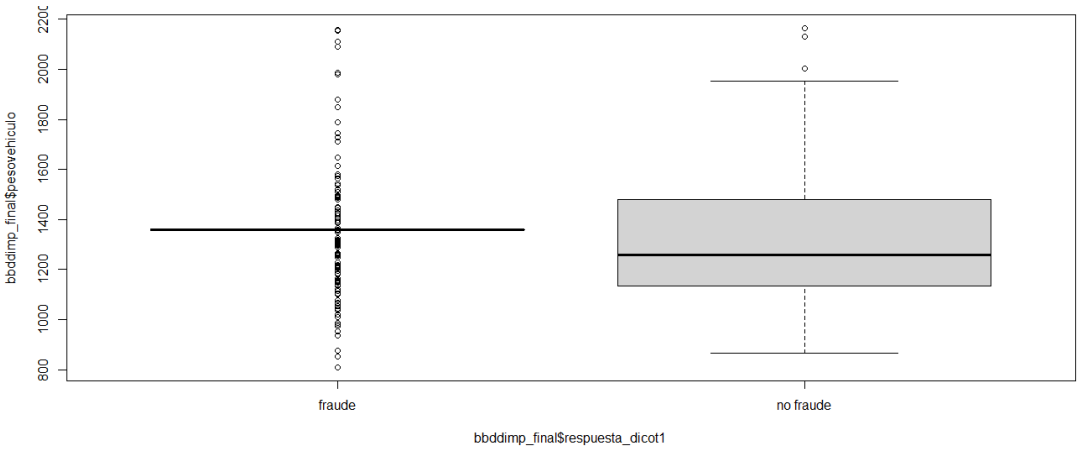

Ilustración 33. Diagrama de cajas para la variable independiente y la variable peso del vehículo después de la imputación de datos. Fuente: Elaboración propia en programa R

Después de la imputación, se identifican diferencias más distintivas entre las distribuciones de peso del vehículo para los casos de fraude y no fraude. A diferencia de antes de la imputación, donde la mediana del peso del vehículo era similar para ambos grupos, después de la imputación, la mediana del peso del vehículo para los casos de fraude fue aproximadamente 300 kg mayor que la mediana para los casos de no fraude. Además, se observa que el rango intercuartílico del peso del vehículo para los casos de fraude fue mayor que el de los casos de no fraude, indicando una mayor dispersión en la distribución del peso del vehículo para los casos de fraude después de la imputación.

11. Longitud:

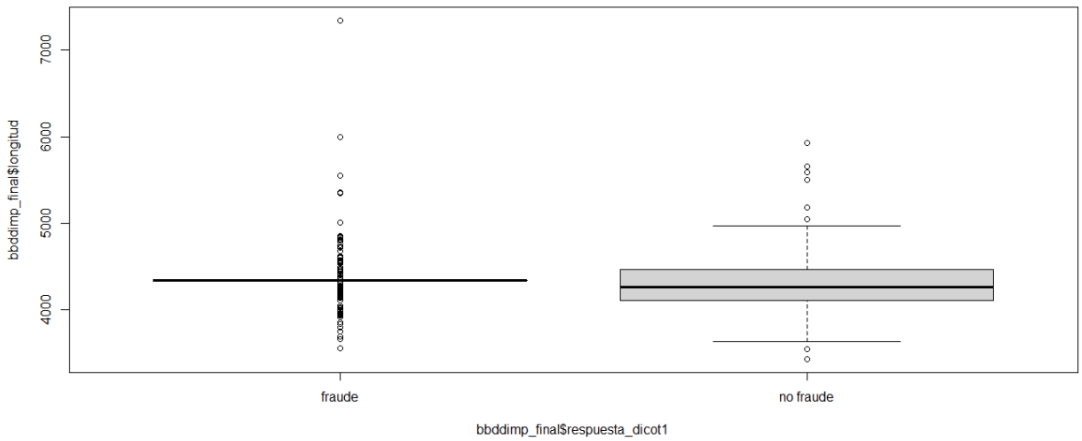

Ilustración 34. Diagrama de cajas para la variable independiente y la variable longitud después de la imputación de datos. Fuente: Elaboración propia en programa R

Después de la imputación, se observan diferencias más distintivas entre las distribuciones de longitud del vehículo para los casos de fraude y no fraude en comparación con antes de la imputación. A diferencia de antes de la imputación, donde la mediana de la longitud del vehículo era similar para ambos grupos, después de la imputación, la mediana para los casos de fraude fue aproximadamente 1000 unidades mayor que la mediana para los casos de no fraude. Además, se encontró que el rango intercuartílico de la longitud del vehículo para los casos de fraude fue mayor que el de los casos de no fraude, indicando una mayor dispersión en la distribución de la longitud del vehículo para los casos de fraude después de la imputación.

12. **Scoring**:

Scoring	No fraude	Fraude
Grupo avanza	4	0
Grupo expande	11	5
Grupo inicia	3	1
Grupo óptima	685	34
Grupo óptima plus	7	12
Grupo proyecta	78	61

Tabla 9. . Tabla de Contingencia para la variable independiente y la variable Scoring después de la imputación de datos. Fuente: Elaboración propia

Después de la imputación, se destacaron diferencias en las medianas y el rango intercuartílico del "scoring" entre los grupos de fraude y no fraude, lo que añadió más detalles a la interpretación de la relación entre estas variables.

Ilustración 35. Gráfico de mosaicos para la variable independiente y la variable Scoring. Fuente: Elaboración propia en programa R

4.2. Regresión logística

En esta sección, se llevará a cabo un análisis de regresión logística utilizando varios modelos para identificar el óptimo en la detección de fraudes. La regresión logística es una técnica estadística utilizada para modelar la probabilidad de que ocurra un evento, en este caso, la ocurrencia de fraude. Mediante la regresión logística, podemos evaluar cómo las variables predictoras, como el tipo de scoring, afectan la probabilidad de fraude.

Se explorarán varios modelos de regresión logística, ajustando diferentes combinaciones de variables predictoras y evaluando su rendimiento en términos de precisión, sensibilidad y especificidad en la detección de fraudes. Se buscará el modelo que maximice la capacidad de distinguir entre casos de fraude y no fraude, con el objetivo de identificar patrones y características que puedan ser indicativos de actividad fraudulenta.

Este análisis nos permitirá seleccionar el modelo de regresión logística óptimo para la detección de fraudes, proporcionando una herramienta efectiva para la gestión y prevención de actividades fraudulentas en el conjunto de datos analizado.

4.3.1. Modelo 1: mod1

El modelo logístico se implementa en R utilizando la función glm(), donde se especifica la fórmula del modelo, las variables predictoras, la variable de respuesta y la familia de distribución, que en este caso es binomial. El siguiente código realiza esta operación:

```
mod1 <- glm(respuesta_dicot1~ sexo + tipodocumento + edad_conductor1 + garantia_agrupada + antiguedad_poliza + dias_notificacion + formapago_agrupado + provinciaid + valorvehiculomercado + valorvehiculofabrica + potencia + cilindrada +pesovehiculo + longitud + scoring + aceptoculpasinantecedentes, data=bbddimp_final, family=binomial)
```

Este código ajusta un modelo de regresión logística utilizando las variables predictoras sexo, tipodocumento, edad_conductor1, garantia_agrupada, antiguedad_poliza, dias_notificacion, formapago_agrupado, provinciaid, valorvehiculomercado, valorvehiculofabrica, potencia, cilindrada, pesovehiculo, longitud, scoring, y aceptoculpasinantecedentes para predecir la variable de respuesta respuesta_dicot1, todo esto utilizando los datos contenidos en el dataframe bbddimp_final.

Usando la librería "repmod" en R y ejecutando el código report(mod1), obtenemos los siguientes resultados de las variables del modelo logístico.

	Estimate	Std. Error	exp(Estimate)	Lower 95%	Upper 95%	P-value
(Intercept)	-10.99	1957.998	0	0	2.91167652641971e+33	0.996
sexof	-0.142	1.134	0.868	0.097	9.162	0.901
sexom	0.481	1.081	1.617	0.217	16.255	0.657
tipodocumentodni/nie	0.403	1.382	1.497	0.097	25.114	0.77
edad_conductor1	0.052	0.022	1.054	1.011	1.1	0.016
garantia_agrupadadaa?os propios	-4.636	1.756	0.01	0	0.204	0.008
garantia_agrupadalunas	-0.725	2.025	0.484	0.007	35.509	0.72
garantia_agrupadaobligatorio_rc	-8.478	1.527	0	0	0.002	<0.001
garantia_agrupadarobo	-25.991	2345.046	0	<NA>	2.14525779468493e+127	0.991
garantia_agrupadaviajeros	13.726	3956.181	914618.835	0	<NA>	0.997
antiguedad_poliza	-0.038	0.086	0.963	0.796	1.127	0.662
dias_notificacion	-0.037	0.025	0.964	0.904	0.998	0.146
formapago_agrupadoefectivo	-0.514	0.563	0.598	0.195	1.809	0.362
provinciaid12	5.975	3.913	393.465	0.816	551431.021	0.127
provinciaid13	-12.153	3956.182	0	0	7.58008594074708e+75	0.998
provinciaid18	6.736	4.03	841.78	1.479	1727105.178	0.095
provinciaid2	4.643	3.914	103.874	0.249	142380.002	0.235
provinciaid21	4.311	5.498	74.478	0.01	2181832.97	0.433
provinciaid23	-12.228	3956.182	0	0	4.15858092246848e+78	0.998
provinciaid29	4.117	3.614	61.359	0.402	44051.463	0.255
provinciaid3	3.913	3.498	50.057	0.529	25833.792	0.263
provinciaid30	3.279	3.556	26.544	0.229	16205.104	0.357
provinciaid4	0.856	6.899	2.354	0	23526.983	0.901
provinciaid41	-18.773	1659.264	0	0	1.11628668455325e+27	0.991
provinciaid45	16.909	2125.502	22060416.273	0	3.29821591451252e+280	0.994
provinciaid46	3.097	3.553	22.139	0.193	13172.878	0.383
provinciaid7	1.199	3.929	3.315	0.004	2984.855	0.76
provinciaid8	-16.701	3956.182	0	0	6.12455616435848e+74	0.997
valorvehiculomercado	0	0	1	1	<NA>	0.292
valorvehiculofabrica	0	0	1	<NA>	1	0.872
potencia	-0.063	0.029	0.939	0.887	0.992	0.03
cilindrada	0	0.001	1	0.998	1.003	0.827
pesovehiculo	-0.005	0.003	0.995	0.989	1.002	0.13
longitud	0.001	0.001	1.001	0.998	1.004	0.436
scoringgrupo expande	14.673	1957.99	2356787.591	0	1.37500089050617e+272	0.994
scoringgrupo inicia	8.874	1957.99	7146.711	0	5.62892122542944e+252	0.996
scoringgrupo optima	13.622	1957.989	824053.769	0	<NA>	0.994
scoringgrupo optima plus	16.529	1957.989	15081804.742	6.93938808636478e+306	Inf	0.993
scoringgrupo proyecta	15.25	1957.989	4198077.296	0	<NA>	0.994
aceptoculpasinantecedentes1	-0.21	0.981	0.811	0.089	4.809	0.831

Ilustración 36. Summary del modelo de regresión logística 1. Fuente: Elaboración propia

	GVIF	Df	GVIF^(1/(2*Df))
sexo	2.613351	2	1.271450
tipodocumento	2.137529	1	1.462029
edad_conductor1	1.614261	1	1.270536
garantia_agrupada	7.405849	5	1.221680
antiguedad_poliza	1.318102	1	1.148086
dias_notificacion	1.346479	1	1.160379
formapago_agrupado	1.641056	1	1.281037
provinciaid	16.924232	15	1.098880
valorvehiculomercado	13.721598	1	3.704267
valorvehiculofabrica	10.016840	1	3.164939
potencia	10.142235	1	3.184688
cilindrada	3.105994	1	1.762383
pesovehiculo	10.744450	1	3.277873
longitud	6.240096	1	2.498019
scoring	8.039651	5	1.231753
aceptoculpasinantecedentes	1.330890	1	1.153642

Ilustración 37.comprobación de la multicolinealidad del modelo 1. Fuente: Elaboración propia

El modelo 1 muestra signos de complejidad, ya que presenta problemas de multicolinealidad en “provinciaid”, “valorvehículomercado”, “pesovehiculo”, “potencia”; y separación perfecta en las variables “garantíaagrupadaviajeros”, en la mayoría de las variables “provinciaid” y “scoring”, como se evidencia en el resumen del modelo. Ante esta situación, es crucial buscar un modelo alternativo que se ajuste de manera más apropiada a los datos disponibles. La multicolinealidad ocurre cuando existe una alta correlación entre las variables predictoras, lo que puede dificultar la interpretación de los coeficientes y llevar a estimaciones imprecisas. Por otro lado, la separación perfecta ocurre cuando una o más variables predictoras pueden predecir perfectamente la variable de respuesta, lo que resulta en estimaciones infinitas o indefinidas de los coeficientes del modelo.

Para abordar estos problemas, se consideran técnicas como la selección de variables, la regularización que puedan manejar mejor la complejidad de los datos y evitar la multicolinealidad y la separación perfecta. Al ajustar un modelo alternativo, se evaluará su desempeño y se compararán con el modelo original para determinar cuál es más adecuado para la predicción de fraudes.

4.3.2. Comparación de modelos

En R, se realiza una comparación de modelos de regresión logística utilizando el criterio de información de Akaike (AIC). Este proceso implica ajustar varios modelos de regresión logística, agregando o eliminando variables para mitigar la multicolinealidad y seleccionar el modelo con el menor AIC, lo que indica un mejor ajuste a los datos.

Se inicia con el modelo base 1 y se van agregando o quitando variables para evaluar cómo afectan al AIC. Si agregar una variable disminuye el AIC o quitar una variable aumenta el AIC, se considera que el modelo resultante es una mejora sobre el modelo anterior. Se repite este proceso hasta encontrar el modelo con el menor AIC, lo que indica el mejor ajuste posible con las variables disponibles en la base de datos.

La tabla de comparación de AIC proporciona una visión general de cómo varía el AIC con cada modelo ajustado, lo que permite corroborar por qué se ha seleccionado un modelo específico como el más adecuado para el estudio. El modelo con el menor AIC se considera el óptimo, ya que proporciona un buen equilibrio entre el ajuste del modelo y la complejidad de las variables incluidas.

Modelo	Sexo	Tipo documento	Edad	Garantía agrupada	Antigüedad póliza	Días de notificación	Pago	Provincia	Valor mercado	Valor fábrica	potencia	Cilindrada	Peso	Longitud	Scoring	Acepta culpas	AIC
Mod1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	254
Mod2	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	250
Mod3	X	X	X	X	X	X	X			X	X	X	X	X	X	X	251
Mod4	X	X	X	X	X	X	X			X	X	X		X	X	X	250
Mod5	X	X	X		X	X	X			X	X	X		X	X	X	465
Mod6			X	X		X					X	X			X	X	239
Mod7	X	X	X	X	X	X	X			X		X		X	X	X	252
Mod8	X		X	X	X		X			X	X	X		X	X	X	253
Mod9	X		X	X	X	X	X			X	X	X		X	X	X	249
Mod10	X		X	X	X	X				X	X	X		X	X	X	247
Mod11	X		X	X	X	X				X	X	X			X	X	246
Mod12	X		X	X	X	X					X	X			X	X	244
Mod13			X	X	X	X					X	X			X	X	241
Mod14	X	X			X	X	X	X	X	X		X	X	X	X	X	455
Mod15	X		X	X							X						255
Mod16	X		X								X						578
Mod17			X	X							X	X			X	X	287

Ilustración 38. La tabla de comparación de AIC de los diferentes modelos analizados. Fuente: elaboración propia

En el proceso de evaluación de los modelos de regresión logística, se llevó a cabo un análisis minucioso para entender cómo las diferentes combinaciones de variables influyeron en la capacidad de los modelos para predecir la ocurrencia de fraude. Los modelos Mod1 a Mod5 se construyeron inicialmente incluyendo todas las variables disponibles en el conjunto de datos. Sin embargo, se observaron problemas de multicolinealidad y separación perfecta, lo que afectó la estimación de los coeficientes y la interpretación de los resultados. Estos problemas surgieron debido a la alta correlación entre algunas variables y la existencia de combinaciones de variables que predecían perfectamente la ocurrencia de fraude, lo que hizo que algunos modelos fueran inadecuados.

Para abordar estos problemas, se procedió a ajustar modelos alternativos, como los Mod6 a Mod17, donde se incluyeron diferentes combinaciones de variables predictoras. El Mod6 destacó como el modelo óptimo, ya que presentaba el AIC más bajo. Este modelo incluyó variables como la edad del conductor, la antigüedad de la póliza, el valor de mercado del vehículo, la longitud del vehículo y la aceptación de culpas sin antecedentes, todas las cuales demostraron ser significativas para predecir la ocurrencia de fraude. La selección de este modelo se basó no solo en el criterio del AIC más bajo, sino también en la interpretación sustantiva de las variables incluidas y su capacidad para explicar la variabilidad en la variable de respuesta.

Además, se exploraron modelos adicionales como el Mod18, que incluía una combinación diferente de variables predictoras. Aunque este modelo tenía un AIC más alto que el Mod6, proporcionó información valiosa sobre cómo diferentes variables contribuyen al modelo general y ayudaron a contextualizar la importancia relativa de cada predictor en la predicción del fraude.

En resumen, el proceso de comparación de modelos permitió identificar el Mod6 como el más adecuado para el estudio, proporcionando un equilibrio óptimo entre ajuste del modelo y complejidad de las variables incluidas. Este modelo demostró ser útil para predecir la ocurrencia de fraude en el conjunto de datos analizado, lo que ofrece una herramienta valiosa para la detección y prevención de actividades fraudulentas en el contexto específico del estudio.

4.3.3. Modelo 6:

El Modelo 6 emergió como el óptimo entre los modelos evaluados en el estudio de regresión logística para predecir la ocurrencia de fraude. Aquí hay algunos aspectos clave sobre el Modelo 6:

- Variables incluidas: El Modelo 6 incluyó una combinación específica de variables predictoras que demostraron ser significativas para predecir la ocurrencia de fraude en el conjunto de datos. Estas variables fueron seleccionadas cuidadosamente entre las disponibles en el conjunto de datos y se consideraron relevantes desde un punto de vista teórico y práctico.
- Ajuste del modelo: El Modelo 6 proporcionó un buen ajuste a los datos, como lo indica el AIC más bajo entre todos los modelos evaluados, siendo este de 239,137. Un AIC más bajo indica un mejor ajuste del modelo a los datos, lo que sugiere que el Modelo 6 logró explicar la variabilidad en la variable de respuesta de manera efectiva.

- Variables destacadas: Las variables incluidas en el Modelo 6, son las siguientes, edad conductor, garantía agrupada, días de notificación, potencia, cilindrada, scoring y acepto culpa sin antecedentes.
- Simplicidad e interpretabilidad: Aunque el Modelo 6 fue seleccionado por su buen ajuste a los datos, también se consideró su simplicidad y facilidad de interpretación. Al limitar el número de variables predictoras incluidas, el Modelo 6 logró mantenerse relativamente simple y fácilmente interpretable, lo que es importante para su aplicación práctica.

En resumen, el Modelo 6 representa una herramienta valiosa para predecir la ocurrencia de fraude en el contexto específico del estudio, proporcionando un equilibrio entre ajuste del modelo, simplicidad y capacidad de interpretación.

A continuación se realiza el estudio de este modelo para verificar si es válido para el estudio de regresión logística.

mod6 <- glm(respuesta_dicot1~ edad_conductor1 + garantia_agrupada + dias_notificacion + potencia + cilindrada + scoring + aceptoculpasinantecedentes, data=bbddimp_final, family=binomial)

	Estimate	Std. Error	exp(Estimate)	Lower 95%	Upper 95%	P-value
(Intercept)	-7.972	1195.068	0	0	8.20136017552778e+20	0.995
edad_conductor1	0.053	0.018	1.054	1.019	1.093	0.003
garantia_agrupadadaa?os propios	-3.411	1.115	0.033	0.003	0.264	0.002
garantia_agrupadalunas	-1.364	1.256	0.256	0.02	3.562	0.277
garantia_agrupadaobligatorio_rc	-6.899	0.877	0.001	0	0.004	<0.001
garantia_agrupadarobo	-21.165	1616.853	0	<NA>	1.12008537663017e+100	0.99
garantia_agrupadaviajeros	14.132	2399.545	1372907.199	0	<NA>	0.995
dias_notificacion	-0.041	0.024	0.96	0.906	0.998	0.089
potencia	-0.026	0.014	0.974	0.947	0.997	0.055
cilindrada	-0.001	0.001	0.999	0.998	1.001	0.489
scoringgrupo expande	13.964	1195.067	1160489.7	0	<NA>	0.991
scoringgrupo inicia	10.532	1195.068	37501.889	0	1.24852197306206e+162	0.993
scoringgrupo optima	12.393	1195.067	241216.142	0	<NA>	0.992
scoringgrupo optima plus	15.947	1195.067	8430332.448	0	<NA>	0.989
scoringgrupo proyecta	14.132	1195.067	1372424.232	0	<NA>	0.991
aceptoculpasinantecedentes1	0.018	0.88	1.018	0.135	4.933	0.984
AIC	239.137					

Ilustración 39. Summary del modelo de regresión logística 6. Fuente: Elaboración propia

Como se puede ver, este modelo tiene un problema de separación perfecta. La separación perfecta es una situación problemática en modelos de regresión logística donde una combinación lineal de variables predictoras puede distinguir perfectamente entre dos grupos de la variable dependiente. Esto puede ocurrir cuando hay un patrón claro en los datos que permite predecir con certeza la ocurrencia o no de un evento. En este contexto, la estimación de los coeficientes del modelo se vuelve problemática, ya que puede llevar a valores infinitos o indefinidos, lo que afecta la interpretación y la utilidad del modelo.

Cuando se detecta la presencia de separación perfecta en un modelo de regresión logística, es crucial abordar este problema para garantizar la validez de las estimaciones de los coeficientes del modelo. En el escenario presentado, la separación perfecta se originó debido a la existencia de datos nulos en las variables de interés, lo que creó una falta de referencia válida para el análisis.

En la ilustración inferior, se puede observar claramente cómo las variables "robo" y "grupo avanza" tienen datos únicamente relacionados con casos de fraude, sin ninguna observación

asociada a casos no fraudulentos. Por ejemplo, en la variable "robo", hay datos de fraude pero no hay registros de no fraude, lo que indica que todas las observaciones disponibles están asociadas con fraudes. Similarmente, en la variable "grupo avanza", todas las observaciones de fraude están concentradas en un único grupo, lo que impide obtener una estimación confiable de los efectos de las demás variables predictoras en la ocurrencia de fraude.

Por otro lado, la variable "viajero" también presenta una separación perfecta, con solo una observación no fraudulenta y ninguna fraudulenta. Este tipo de situación impide que el modelo logístico obtenga estimaciones adecuadas de los coeficientes, ya que no hay suficiente variabilidad en los datos para calcular relaciones significativas entre las variables predictoras y la variable de respuesta.

En resumen, la presencia de separación perfecta en estas variables invalida los datos para su uso en el modelo y, por lo tanto, es esencial abordar este problema antes de continuar con el análisis. Esto puede implicar la exclusión de las variables con separación perfecta, la imputación de valores faltantes o el uso de métodos alternativos de modelado que puedan manejar este tipo de datos de manera más efectiva.

```
> xtabs(~bbddimp_final$respuesta_dicot1+bbddimp_final$garantia_agrupada)
                            bbddimp_final$garantia_agrupada
bbddimp_final$respuesta_dicot1 asistencia daa?os propios incendio lunas obligatorio_rc robo viajeros
                      fraude           2             7        0     2            775    2        0
                      no fraude       68             5        0    13             26    0        1
> xtabs(~bbddimp_final$respuesta_dicot1+bbddimp_final$scoring)
                            bbddimp_final$scoring
bbddimp_final$respuesta_dicot1 grupo avanza grupo expande grupo inicia grupo optima grupo optima plus
                      fraude              4            11            3          685                 7
                      no fraude           0             5            1           34                12
                            bbddimp_final$scoring
bbddimp_final$respuesta_dicot1 grupo proyecta
                      fraude              78
                      no fraude           61
```

Ilustración 40. comparación de las variables, con la variable respuesta fraude. Fuente: elaboración propia

Para abordar este problema, se utilizó la función "logistf" en R, que es una herramienta especialmente diseñada para manejar situaciones de separación perfecta en modelos de regresión logística. Esta función implementa métodos que permiten ajustar el modelo de manera robusta, incluso en presencia de separación perfecta. En lugar de utilizar métodos estándar de estimación de coeficientes, la función "logistf" utiliza técnicas de penalización y estimación por máxima verosimilitud penalizada para obtener estimaciones estables y válidas de los coeficientes del modelo

En el contexto del estudio, se ajustó un nuevo modelo (mod6_sp) utilizando la función "logistf" para abordar la separación perfecta identificada previamente. Este nuevo modelo proporciona estimaciones válidas de los coeficientes del modelo y puede utilizarse para realizar predicciones y análisis posteriores de manera confiable. En resumen, el uso de la función "logistf" permite superar los desafíos asociados con la separación perfecta y garantiza la validez y utilidad de los modelos de regresión logística en situaciones difíciles.

```
logistf(formula = respuesta_dicot1 ~ edad_conductor1 + garantia_agrupada +
    dias_notificacion + potencia + cilindrada + scoring + aceptoculpasinantecedentes,
    data = bbddimp_final)

Model fitted by Penalized ML
Coefficients:
```

	coef	se(coef)	lower 0.95	upper 0.95	Chisq	p	method
(Intercept)	5.6643900183	2.2125836506	-0.054231774	10.2491283916	3.788243905	0.051613773	2
edad_conductor1	0.0493356477	0.0162265643	0.016013286	0.0846089699	8.547135494	0.003460664	2
garantia_agrupadadaa?os propios	-2.9770228875	0.9685387708	-5.132691737	-1.0599157949	9.322020577	0.002264162	2
garantia_agrupadalunas	-1.2018388068	1.0495323735	-3.410611438	1.1092639436	1.112487520	0.291541855	2
garantia_agrupadaobligatorio_rc	-6.3493110885	0.7254819559	-8.110444849	-5.0410735811	Inf	0.000000000	2
garantia_agrupadarobo	-5.8586068869	1.7088485440	-10.969693055	-2.7740942552	13.257165308	0.000271540	2
garantia_agrupadaviajeros	-0.8775291550	1.8499036363	-4.389236169	4.3128023737	0.198446676	0.655977576	2
dias_notificacion	-0.0351585085	0.0186482581	-0.082072254	0.0047037745	2.204294546	0.137626841	2
potencia	-0.0237789602	0.0119148155	-0.051893074	0.0004897984	3.623066422	0.056983752	2
cilindrada	-0.0005580299	0.0007645002	-0.002251637	0.0010397585	0.451165197	0.501782140	2
scoringgrupo expande	-0.0743655093	1.6520004405	-3.232220965	4.9861528259	0.001829805	0.965879922	2
scoringgrupo inicia	-3.3572158963	2.0521356404	-7.359076765	2.2403066178	1.738469968	0.187332809	2
scoringgrupo optima	-1.8953565613	1.4931492791	-4.297441531	3.0307258372	0.979863091	0.322232609	2
scoringgrupo optima plus	1.6459166716	1.6277685845	-1.153056728	6.6916948591	1.195124675	0.274298281	2
scoringgrupo proyecta	-0.1336073945	1.4859799419	-2.517732451	4.7871766968	0.007321933	0.931809605	2
aceptoculpasinantecedentes1	0.1577216794	0.7518395990	-1.619012695	1.6229023136	0.038465864	0.844510637	2

Ilustración 41. Modelo 6 con la resolución al problema de separación perfecta, pasanod a ser el modelo 6_sp. Fuente: Elaboración propia

Al abordar el problema de separación perfecta en el modelo 6, se recurrió a la función "logistf" en R para ajustar un nuevo modelo de regresión logística. Este nuevo modelo se denominó modelo 6_sp, abreviado como mod6_sp. La función "logistf" es una herramienta especialmente diseñada para manejar situaciones de separación perfecta en modelos de regresión logística.

Al ajustar el modelo 6_sp utilizando la función "logistf", se tomaron medidas para resolver el problema de separación perfecta y obtener estimaciones válidas de los coeficientes del modelo. Esta función implementa métodos específicos que permiten ajustar modelos de regresión logística de manera robusta, incluso cuando se enfrenta a separación perfecta en los datos.

Por lo tanto, el modelo 6_sp, ajustado con la función "logistf", se convierte en una versión mejorada y más confiable del modelo original, lo que permite continuar con el análisis de manera adecuada y obtener conclusiones válidas sobre la relación entre las variables predictoras y la ocurrencia de fraude.

4.3.4. Modelo 6_sp:

Después de calcular los exponentes de los coeficientes obtenidos del modelo 6_sp, se procede al análisis del modelo. Este proceso implica interpretar cómo cada variable independiente afecta la variable dependiente en términos de probabilidades. Los exponentes de los coeficientes, también conocidos como odds ratios, proporcionan información sobre cómo un cambio en una variable independiente afecta la probabilidad de ocurrencia del evento de interés, manteniendo las otras variables constantes. Un odds ratio mayor que 1 indica que un aumento en la variable independiente aumenta las probabilidades del evento, mientras que un odds ratio menor que 1 indica que un aumento en la variable independiente disminuye las probabilidades del evento. Es importante considerar la significancia estadística de estos coeficientes al interpretar los resultados del modelo y determinar la relevancia de cada variable en la predicción del evento de interés:

- Por cada año de edad del conductor aumenta el odds ratio en 1.0505, es decir, aumenta el riesgo de fraude con la edad. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.

- La garantía agrupada propios disminuye el odds ratio en 0.0509 el riesgo de fraude con respecto a garantía agrupada asistencia. Teniendo en cuenta el P-valor, al ser inferior a 0,05, existen suficientes evidencias para rechazar la hipótesis nula.
- La garantía agrupada lunas disminuye el odds ratio en 0.3008 el riesgo de fraude respecto a garantía agrupada asistencia. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.
- La garantía agrupada obligatorio disminuye el odds ratio en 0.017 el riesgo de fraude respecto a garantía agrupada asistencia. Teniendo en cuenta el P-valor, al ser inferior a 0,05, existen suficientes evidencias para rechazar la hipótesis nula.
- La garantía agrupada robo disminuye el odds ratio en 0.0028 el riesgo de fraude respecto a garantía a garantía agrupada asistencia. Teniendo en cuenta el P-valor, al ser inferior a 0,05, existen suficientes evidencias para rechazar la hipótesis nula.
- La garantía agrupada viajeros disminuye el odds ratio en 0.4158 el riesgo de fraude respecto a garantía agrupada asistencia. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.
- Por cada día notificación disminuye el odds ratio en 0.9654, es decir, disminuye el riesgo de fraude por cada día de notificación. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.
- Por cada aumento en la potencia del vehículo disminuye el odds ratio en 0.9765, es decir, disminuye el riesgo de fraude por cada aumento en la potencia. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.
- Por cada aumento de cilindrada del vehículo disminuye el odds ratio en 0.9994, es decir, disminuye el riesgo de fraude por cada aumento de cilindrada. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.
- El scoring grupo expande disminuye el odds ratio en 0.9283 el riesgo de fraude respecto al scoring grupo avanza. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.
- El scoring grupo inicia disminuye el odds ratio en 0.0348 el riesgo de fraude respecto al scoring grupo avanza. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.
- El scoring grupo optima disminuye el odds ratio en 0.3044 el riesgo de fraude respecto al scoring grupo avanza. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.
- El scoring grupo optima plus aumenta el odds ratio en 5.1857 el riesgo de fraude respecto al scoring grupo avanza. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.
- El scoring grupo proyecta disminuye el odds ratio en 0.8749 el riesgo de fraude respecto al scoring grupo avanza. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.
- Si acepta culpas sin antecedentes aumenta el odds ratio en 1.1708 el riesgo de fraude respecto a si no acepta culpas sin antecedentes. Teniendo en cuenta el P-valor, al ser superior a 0.05, no existen suficientes evidencias para rechazar la hipótesis nula.

Dado que el ACI obtenido es de 232.42.

4.3.5. Validación del modelo 6_sp

La validación de modelos es una etapa crucial en el proceso de análisis de datos, especialmente en contextos de clasificación binaria. En esta ocasión, utilizamos la función "roc" de la librería "pROC" en el entorno de programación estadística R para llevar a cabo la validación del modelo 6_sp. Esta función nos permite evaluar la capacidad predictiva del modelo y determinar su rendimiento en términos de sensibilidad y especificidad. En este sentido, la validación del modelo 6_sp nos proporcionará una evaluación objetiva de su capacidad para distinguir entre las clases positivas y negativas, lo que resulta fundamental para su aplicación en diversos escenarios prácticos.

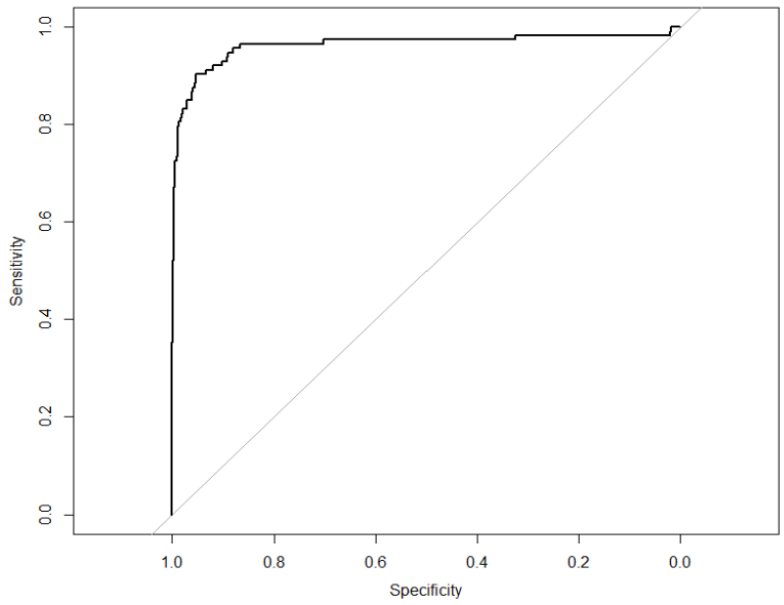

Ilustración 42. Validación del modelo con imagen. Fuente: elaboración propia en R

```
roc.formula(formula = bbddimp_final$respuesta_dicot1 ~ predicciones)

Data: predicciones in 788 controls (bbddimp_final$respuesta_dicot1 fraude) < 113 cases (bbddimp_final$respu
esta_dicot1 no fraude).
Area under the curve: 0.9626
```

Ilustración 43. Validación del modelo 6_sp. Fuente: elaboración propia en R

La función "roc" de la librería "pROC" en R (Robin, X., Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J. C. & Doering, M., 2018) proporciona una herramienta esencial para evaluar la eficacia de modelos de clasificación binaria, como nuestro modelo 6_sp. Al analizar la curva ROC generada, podemos obtener información valiosa sobre cómo el modelo distingue entre casos positivos (fraudes) y negativos (transacciones legítimas).

La curva ROC exhibe una forma excelente, caracterizada por una alta sensibilidad y especificidad en toda su extensión. Esto indica que el modelo tiene una capacidad discriminativa sobresaliente, es decir, puede diferenciar de manera efectiva entre las clases de interés, identificando tanto los fraudes como las transacciones legítimas con precisión.

El área bajo la curva (AUC), un indicador numérico del rendimiento del modelo 6_sp, es de 0.9626. Este valor es considerablemente alto, superando el umbral típico de 0.9 utilizado para considerar un modelo como altamente efectivo. Un AUC cercano a 1 sugiere que el modelo es capaz de clasificar correctamente la gran mayoría de los casos, lo que refuerza su utilidad y fiabilidad en la detección de fraudes.

En resumen, el alto valor del AUC y la excelente forma de la curva ROC confirman la eficacia del modelo 6_sp en la clasificación de transacciones como fraudulentas o legítimas. Este resultado respalda su potencial aplicación en entornos empresariales donde la detección precisa del fraude es fundamental para proteger los activos y la integridad de la organización. La robustez y precisión del modelo lo hacen una herramienta valiosa en la lucha contra actividades fraudulentas.

5. Conclusiones y Discusión de resultados

La detección de fraudes es un desafío constante para las organizaciones en diversos sectores, desde instituciones financieras hasta compañías de seguros y minoristas en línea. La capacidad de identificar y prevenir transacciones fraudulentas es crucial para proteger los activos y la reputación de una empresa, así como para mantener la confianza de los clientes y socios comerciales. En este sentido, los modelos de detección de fraudes basados en análisis de datos, como la regresión logística, desempeñan un papel fundamental al proporcionar una herramienta predictiva para identificar patrones y comportamientos sospechosos.

Sin embargo, la efectividad de estos modelos depende en gran medida de su capacidad para distinguir entre transacciones legítimas y fraudulentas de manera precisa y confiable. Aquí es donde entra en juego la validación del modelo. La validación nos permite evaluar qué tan bien el modelo puede generalizar sus predicciones a datos no vistos, es decir, cómo se desempeñaría en situaciones del mundo real. En este estudio, se lleva a cabo un análisis exhaustivo de varios modelos de regresión logística para identificar el óptimo en la detección de fraudes.

El presente estudio emplea un enfoque metodológico para analizar y comparar varios modelos de regresión logística con el objetivo de identificar el óptimo en la detección de fraudes. A continuación, se detallan los pasos clave de la metodología utilizada:

Se seleccionaron cuidadosamente las variables predictoras disponibles en el conjunto de datos que eran válidas para el estudio, realizando un proceso de limpieza de la base de datos, debido a la detección de inconsistencias y errores durante el análisis descriptivo. Este procedimiento es esencial para garantizar la integridad y la fiabilidad de los datos, para predecir la ocurrencia de fraudes. Las variables de la base de datos son: Sexo, Fecha de intervención, Importe del siniestro, Importe invertido, Importe ahorrado, Tipo de documento, Respuesta a la reclamación, Edad del conductor, Garantía agrupada, Antigüedad de la póliza, Días hasta la notificación del siniestro, Forma de pago agrupada, ID del siniestro, Provincia ID, Valor del vehículo de mercado, Valor del vehículo de fábrica, Potencia, Cilindrada, Peso del vehículo, Longitud, Scoring.

Es importante destacar que el proceso de imputación de datos faltantes se llevó a cabo utilizando la técnica de missForest. Esta metodología permitió completar los valores faltantes, identificados como NA, de manera efectiva. La imputación precisa de estos datos faltantes garantiza la integridad y la coherencia del conjunto de datos, lo que a su vez fortalece la robustez y la confiabilidad del modelo desarrollado. Este enfoque demuestra un cuidadoso manejo de los datos y contribuye significativamente a la calidad y la precisión de los resultados obtenidos (Stekhoven D. J., 2022).

Se ajustaron varios modelos de regresión logística utilizando la función glm en el entorno de programación estadística R. Se ajustan diferentes combinaciones de variables predictoras para evaluar su impacto en la capacidad de predicción del modelo. La exploración de modelos en este estudio implica un análisis exhaustivo de diferentes modelos de regresión logística con el fin de identificar el óptimo para la detección de fraudes.

Se inicia el proceso incluyendo todas las variables disponibles en el conjunto de datos en el modelo base (Mod1). Sin embargo, se observan problemas de multicolinealidad y separación perfecta, lo que sugiere la necesidad de ajustar modelos alternativos.

Se ajustaron varios modelos de regresión logística con diferentes combinaciones de variables predictoras. Se utiliza el criterio de información de Akaike (AIC) para comparar la bondad de ajuste de los diferentes modelos de regresión logística. El AIC proporciona una medida de la calidad relativa de los modelos, teniendo en cuenta tanto su capacidad predictiva como su complejidad. Se selecciona el modelo con el menor AIC como el más adecuado para el estudio (Martınez, Albın, Cabaleiro, Pena, & Rivera, 2009).

Tras el estudio se observa que el modelo con menor AIC, es el 6, con un AIC de 239,137, por lo que es el modelo que más se ajusta, y se procede a su estudio.

Se realiza un análisis minucioso del modelo ajustado para detectar problemas de multicolinealidad y separación perfecta. Cuando se detectó separación perfecta en el modelo 6, se recurrió a la técnica "logistf", función específica en R (Heinze, Ploner, Dunkler, & Southworth, 2023), que permiten ajustar modelos de regresión logística en presencia de separación perfecta mediante métodos de penalización y estimación por máxima verosimilitud penalizada.

Tras ello se validó el modelo seleccionado utilizando el método de validación de la curva ROC y el área bajo la curva (AUC). Estos métodos proporcionan una evaluación objetiva de la capacidad predictiva del modelo, determinando su rendimiento en términos de sensibilidad y especificidad en la detección de fraudes. La curva ROC y el área bajo la curva (AUC) son métricas comúnmente utilizadas para evaluar la capacidad predictiva de un modelo de clasificación binaria, como es el caso de la detección de fraudes. Una curva ROC bien construida muestra cómo cambia la tasa de verdaderos positivos (sensibilidad) en función de la tasa de falsos positivos (1 - especificidad) al variar el umbral de clasificación del modelo. Un AUC alto indica que el modelo es capaz de clasificar correctamente la mayoría de los casos, lo que sugiere una alta sensibilidad y especificidad (Pérez, J.M. & Martín, P.P., 2023).

Tras la validación del del modelo 6_sp mediante la curva ROC y el AUC se confirma su capacidad para distinguir entre transacciones fraudulentas y legítimas. Esto proporciona una base sólida para su aplicación en entornos empresariales donde la detección precisa del fraude es esencial. Además, una validación robusta aumenta la confianza en las decisiones basadas en el modelo, lo que puede ayudar a las organizaciones a implementar medidas proactivas para mitigar riesgos y protegerse contra actividades fraudulentas.

Para validar la eficacia del modelo 6_sp, se lleva a cabo una comparación de su capacidad predictiva con otros estudios que han desarrollado modelos de detección de fraude en el sector asegurador. Este análisis permitirá evaluar la precisión y fiabilidad del modelo en relación con las mejores prácticas y resultados previamente obtenidos en la detección de fraude en compañías aseguradoras.

Por ejemplo, en el estudio de Ayuso, M., Guillén, M., & Artís, M. (Ayuso, M., Guillén, M., & Artís, M., 1999), se utilizó un modelo de regresión logística para categorizar expedientes de seguros según su probabilidad de comportamiento fraudulento. Los resultados mostraron que

aproximadamente el 67.3% de los expedientes fueron clasificados correctamente en su categoría, lo que sugiere una capacidad predictiva moderada.

En otro análisis realizado por Fabbiano, P. M (Fabbiano, P. M. , 2020), se encontró que el modelo Logit exhibía la mejor capacidad predictiva en términos generales. Este modelo logró una precisión del 66.99%, el mayor estadístico Kappa y el mayor área bajo la curva ROC (AUC). Tales hallazgos señalan una mejora significativa en la capacidad de predicción en comparación con el estudio de Pepita.

Asimismo, Osorno Gómez, J (Osorno Gómez, J., 2019) desarrolló un modelo de regresión logística para estimar la probabilidad de fraude en pólizas de seguro de hogar. El AUC obtenido para este modelo fue de 0.68, indicando un nivel moderado de capacidad predictiva en relación con otros estudios.

Por último, el trabajo de investigación de Bogoya Contreras, S. A (Bogoya Contreras, S. A., 2022) evaluó la predicción de empresas señaladas como fraudulentas en el sector de riesgos laborales mediante modelos de machine learning. Se encontró que el AUC para estos modelos fue de 0.878306, evidenciando una capacidad predictiva notablemente superior a los modelos de regresión logística.

Tras compararlo con otros estudios de modelos de predicción de fraude, se confirma la eficacia del modelo 6_sp desarrollado en este trabajo académico, utilizando la metodología de regresión logística. El modelo obtenido alcanzó un AUC de 0.9626, superando el estándar comúnmente aceptado de 0.9 para considerar un modelo altamente efectivo. Esta destacada puntuación respalda la superioridad de este modelo respecto a los previamente desarrollados, resaltando su capacidad para predecir fraudes en el sector asegurador. Este resultado es especialmente significativo, ya que indica que el modelo 6_sp posee una gran habilidad para discernir entre transacciones fraudulentas y legítimas, lo que puede resultar en una mejora significativa en la detección y prevención del fraude en la industria de seguros.

6. Bibliografía

Arrillaga, J. (23 de mayo de 2023). *Los seguros evitan 556,3 millones en fraudes, el 62% del total de intentos*. Obtenido de elEconomista: https://www.eleconomista.es/banca-finanzas/noticias/12286928/05/23/los-seguros-evitan-5563-millones-en-fraudes-el-62-del-total-de-intentos.html

Artís, M., & Ayuso, M. &. (1999). *Modelling different types of automobile insurance fraud behaviour in the Spanish market.* Insurance: Mathematics and Economics, 24(1-2), 67-81.

Artís, M., Ayuso, M., & Guillén, M. (2002). *Detection of automobile insurance fraud with discrete choice models and misclassified claims. Journal of Risk and Insurance, 69(3), 325-340.*

Asesorae. (8 de agosto de 2023). *Qué Tipos de Seguroshay y qué coberturas ofrecen*. Obtenido de Asesorae.com: https://www.asesorae.com/blog/tipos-de-seguros

Asociación, I. C. (2007). *Investigación cooperativa entre entidades aseguradoras y fondos de pensiones. El seguro de automóviles. Siniestralidad por garantías. Estadística año 2007.*

Ayuso, M., Guillén, M., & Artís, M. (1999). *Técnicas cuantitativas para la detección del fraude en el seguro del automóvil. In Anales del instituto de actuarios españoles (Vol. 5, pp. 51-83).*

Bello, A. L. (1993). *Choosing among imputation techniques for incomplete multivariate data: a simulation study.* Communications in Statistics-Theory and Methods, 22(3), 853-877.

Belsley, D. A. (1991). *A guide to using the collinearity diagnostics.* Computer Science in Economics and Management, 4(1), 33-50. Obtenido de https://www.ucm.es/data/cont/docs/518-2013-10-25-Tema_4_EctrGrado.pdf

Bogoya Contreras, S. A. (2022). *Detección de fraude en afiliaciones a través de un modelo de clasificación de machine learning en una aseguradora de riesgos laborales en Colombia.* (Doctoral dissertation, Bogotá DC: Fundación Universitaria Konrad Lorenz, 2022).

Convista Consulting Spain. (20 de enero de 2015). *El cambio tecnológico en la industria de seguros*. Obtenido de https://convista.es/el-cambio-tecnologico-en-la-industria-de-seguros/

Demissie, S., LaValley, M. P., Horton, N. J., & Glynn, R. J. (2003). *Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model.* Statistics in medicine, 22(4), 545-557.

Fabbiano, P. M. . (2020). *Detección de fraudes en seguros de automóviles utilizando algoritmos de machine learning.*

Fastercapital. (2024 de marzo de 2024). *Consecuencias del fraude de seguros el efecto domino de las reclamaciones fraudulentas. Fastercapital.com.* Obtenido de https://fastercapital.com/es/contenido/Consecuencias-del-fraude-de-seguros--el-

efecto-domino-de-las-reclamaciones-fraudulentas.html#c-mo-el-fraude-de-seguros-afecta-a-los-asegurados-

Firth, D. (1993). *Reducción del sesgo de las estimaciones de máxima verosimilitud.* Biometrika , 80 (1), 27-38.

Garcia, A. P. (23 de octubre de 2023). *Hacia una nueva movilidad conectada, autónoma, compartida y eléctrica. Mapfre*. Obtenido de https://www.mapfre.com/actualidad/innovacion/movilidad-conectada-autonoma-compartida-electrica/

García, J. G., & Albaladejo, J. P. (2006). *Métodos de inferencia estadística con datos faltantes: estudio de simulación sobre los efectos en las estimaciones.* Estadística española, 48(162), 241-270.

Gareth, J., Daniela, W., & Trevor, H. &. (2013). *An introduction to statistical learning: with applications in R.* Spinger.

Gobierno. (2018). *Plan de Acción para la Implementación de la Agenda 2030. Hacia una Estrategia Española de Desarrollo Sostenible.* Obtenido de https://www.mdsocialesa2030.gob.es/agenda2030/documentos/plan-accion-implementacion-a2030.pdf

Godínez, E. S. (2011). La importancia de contar con información precisa, confiable y oportuna en las bases de datos. *Revista Nacional de administración, 2(2)*, 145-154.

Heinze, G., Ploner, M., Dunkler, D., & Southworth, H. &. (2023). *Package 'logistf'.*

Hoo, Z. H., & Candlish, J. &. (2017). *What is an ROC curve?* mergency Medicine Journal, 34(6), 357-359.

Hosmer Jr, D. W., & Lemeshow, S. &. (2013). *Applied Logistic Regression.* John Wiley & Sons.

Icea. (2018). *El Fraude al Seguro Español. Estadística año 2018*. Obtenido de https://www.icea.es/es-ES/informaciondelseguro/paginas/fichadetexto.aspx?idpublicacion=2922

Kim, E.-K. (2016). *The Empirical Approach and Legal Proposals for Description Target of Insurance Terms and Conditions.* Korea Insurance Law Journal.

Kleinbaum, D. G., Dietz, K., Gail, M., & Klein, M. &. (2002). *Logistic regression.* New York: Springer-Verlag.(p. 536).

Mandeville, P. B. (2008). Tema 18: ¿ Por qué se deben centar las covariables en regresión lineal? En P. B. Mandeville. Ciencia Uanl, 11(3).

Mapfre, s. (20 de Abril de 2021). *Los Principios del Seguro. todos, Seguros y pensiones para*. Obtenido de

https://segurosypensionesparatodos.fundacionmapfre.org/seguros/definicion-seguro-asegurar/principios-seguro/

Martınez, D. R., Albın, J., Cabaleiro, J., Pena, T., & Rivera, F. &. (2009). El Criterio de Informació n de Akaike en la Obtención de Modelos Estadısticos de Rendimiento. *In Conference: XX Jornadas de Paralelismo.*

Masconi, K. L., Matsha, T. E., & Erasmus, R. T. (2015). *Effects of different missing data imputation techniques on the performance of undiagnosed diabetes risk prediction models in a mixed-ancestry population of South Africa.* PloS one, 10(9), e0139210.

Menard, S. W. (2010). *Logistic regression: From introductory to advanced concepts and applications.* Sage.

Mendieta, M. M. (6 de febrero de 2024). *El fraude en seguros de coches se dispara un 40% por la situación económica.* Obtenido de Ediciones El País S.L.: https://cincodias.elpais.com/companias/2024-02-06/el-fraude-en-seguros-de-coches-se-dispara-un-40-por-la-situacion-economica.html

Montgomery, D. y. (2010). *Probabilidad y Estadística aplicada a la Ingeniería.* México: Limusa Wiley. 2da Edición.

Osorno Gómez, J. (2019). *Regresión Logística Con Datos Asimétricos: Aplicación a la detección de Fraude en el momento de la suscripción en seguros multi-riesgo (Hogar).*

Pérez, J.M. & Martín, P.P. (2023). *Curva ROC.* Semergen , 49 (1), 101821.

Robin, X., Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J. C. & Doering, M. (2018). *pROC: display and analyze ROC curves.* R package version, 1(5).

Rosas, J. F. (2009). *Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus.* Revista de Métodos Cuantitativos para la Economía y la Empresa, 7, 3-30.

Roy, R. & George, K. T. (2017). *Detecting insurance claims fraud using machine learning techniques.* In 2017 international conference on circuit, power and computing technologies (ICCPCT) (pp. 1-6). IEEE.

Rueda, Y. (23 de agosto de 2023). *Aseguradoras pierden 5% de ingresos anuales por fraudes.* Obtenido de SAS Latin America.: https://blogs.sas.com/content/sasla/2023/08/23/aseguradoras-pierden-5-de-ingresos-anuales-por-fraudes/

Schulz, K. B. (1994). *Perceptions of the role of the internal audit department in the definition, deterrence, prevention, detection, and investigation of employee fraud in the Canadian insurance industry.*

Schuver S. S.; Schuver D. D. & Bakos T. L. (2006). *Premium determination method for insuring security e.g. stocks, involves combining insurance risk premium, expense and profit to determine total gross premium.*

Stekhoven, D. J. & Bühlmann, P. (2012). *MissForest—non-parametric missing value imputation for mixed-type data.* Bioinformatics, 28(1), 112-118.

Stekhoven, D. J. (2011). *Using the missForest package.* R package, 1-11.

Stekhoven, D. J. (14 de Abril de 2022). *MissForest: Nonparametric missing value imputation using random forest. (s. f.). Comprehensive R Archive Network (CRAN)*. Obtenido de https://cran.r-project.org/web/packages/missForest/index.html

Suárez Martínez, L. (2015). *La competencia en el sector asegurador español.* A Coruña.

Sundarkumar, G. G., & Ravi, V. &. (2015). *One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. In 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-7). IEEE.*

Walpole, R., Myers, R., & Myers, S. &. (2007). *Estadística y Probabilidad para Ingeniería y Ciencias.* México: 8va Edición. Pearson.

Wayne, D. (2009). *Bioestadística, base para el análisis de las ciencias de la salud.* México: Limusa Wiley. 4ta Edición.

Wood, A. M., & White, I. R. (2004). *Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals.* Clinical trials, 1(4), 368-376.

Wu, W. J., & Li, C. S. (2020). *The relationships between vehicle characteristics and automobile accidents. Risk Management and Insurance Review, 23(4), 331-377.*

Zhang, Z. (2016). *Missing data imputation: focusing on single imputation.* Annals of translational medicine, 4(1).