uc3m | Universidad Carlos III de Madrid

Máster en Ciencias Actuariales y Financieras
2022-2024

Trabajo Fin de Máster

"Reducción del sesgo en el proceso de tarificación no vida con técnicas de inteligencia artificial"

Tutor/es

Raquel Pérez Calderón

Madrid, julio 2024

DETECCIÓN DEL PLAGIO

La Universidad utiliza el programa **Turnitin Feedback Studio** para comparar la originalidad del trabajo entregado por cada estudiante con millones de recursos electrónicos y detecta aquellas partes del texto copiadas y pegadas. Copiar o plagiar en un TFM es considerado una **Falta Grave**, y puede conllevar la expulsión definitiva de la Universidad.

CLAUSULA DE RESPONABILIDAD

RESUMEN

El problema del sesgo en los modelos de tarificación de seguros no vida es una de las preocupaciones más crecientes en el sector asegurador. Este trabajo pretende analizar tanto las definiciones del sesgo como las diversas técnicas para mitigarlo, incluyendo enfoques tradicionales y algoritmos avanzados de aprendizaje automático, sin comprometer la capacidad predictiva de los modelos. Para demostrar la aplicabilidad de estas técnicas, se han presentado ejemplos prácticos desarrollados en *Python*, utilizando datos de frecuencia de siniestros de la cobertura de Responsabilidad Civil en el seguro de automóviles. La exploración e implementación de estos enfoques innovadores ha revelado un conjunto de herramientas efectivas que no solo logran reducir el sesgo, sino que también mejoran significativamente la capacidad predictiva de los modelos. Finalmente, se han discutido las implicaciones éticas y regulatorias de utilizar métodos de inteligencia artificial, subrayando la importancia de garantizar la interpretabilidad y equidad de los modelos finales aplicados en la tarificación.

PALABRAS CLAVE: tarificación, seguros, *machine learning*, inteligencia artificial, modelos lineales generalizados, equidad, sesgo, *Python*.

ABSTRACT

The issue of bias in non-life insurance pricing models is one of the growing concerns in the insurance sector. This paper aims to analyze both the definitions of bias and the various techniques to mitigate it, including traditional approaches and advanced machine learning algorithms, without compromising the predictive power of the models. To demonstrate the applicability of these techniques, practical examples developed in Python have been presented, using claims frequency data from automobile liability insurance. The exploration and implementation of these innovative approaches have revealed a set of effective tools that not only reduce bias but also significantly improve the predictive power of the models. Finally, the ethical and regulatory implications of using artificial intelligence methods have been discussed, emphasizing the importance of ensuring the interpretability and fairness of the final models applied in pricing.

KEY WORDS: non-life insurance pricing, machine learning, artificial intelligence, generalized lineal models, bias, equity, Python

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

El proceso de fijación de primas en el sector asegurador es una función fundamental en la operación de las compañías. Este proceso implica evaluar los riesgos potenciales asociados con cada póliza de seguro y establecer una tarifa que refleje adecuadamente esos riesgos. Es esencial que esta tarifa sea suficiente para cubrir los posibles siniestros futuros y, al mismo tiempo, sea competitiva en el mercado.

El cálculo de las primas se basa en la recopilación y análisis de datos, que para los seguros de autos pueden incluir, entre otros, la información sobre el asegurado, el historial siniestral, comportamientos de conducción e historial crediticio; así como detalles sobre el bien asegurado, su valor, ubicación y características específicas. Esta información se utiliza para evaluar el riesgo asociado con cada póliza y ajustar la prima en consecuencia.

Por otra parte, a pesar de la necesidad de segmentar las tarifas y contar con políticas de suscripción basadas en el riesgo, es importante evitar cualquier forma de discriminación injusta en el proceso de fijación de precios. La discriminación puede ocurrir cuando ciertos grupos de asegurados son tratados de manera desfavorable debido a características personales protegidas; como la edad, el género o el estado de salud. Esto puede conducir a prácticas injustas y a una percepción negativa por parte de los consumidores, lo que a su vez puede minar la confianza en la industria aseguradora.

En este contexto, es fundamental para las aseguradoras comprender y abordar cualquier sesgo potencial en sus modelos de tarificación. Esto implica no solo cumplir con las regulaciones y leyes contra la discriminación, sino también adoptar enfoques proactivos para mitigar cualquier diferenciación no deseada en las primas de seguro.

El propósito del presente trabajo es abordar de manera exhaustiva las definiciones de discriminación en el ámbito de la fijación de tarifas de seguros, así como proponer estrategias y técnicas para mitigar el sesgo que puede surgir hacia variables protegidas. A través de un enfoque integral, que abarca desde métodos tradicionales hasta técnicas más innovadoras de inteligencia artificial, este trabajo tiene como objetivo identificar y mitigar el sesgo en los modelos de tarificación más utilizados por las compañías de seguros.

Para ello, se llevará a cabo un análisis detallado que involucra el uso de datos recogidos sobre la cobertura de responsabilidad civil obligatoria en España[1]. La elección de esta cobertura no es opcional, dado que los propietarios de los vehículos no pueden optar por el no aseguramiento, por lo que mitigar o eliminar el sesgo en los procesos de tarificación se vuelve una tarea esencial.

Este enfoque, basado en datos empíricos, permitirá examinar estadísticamente la relación entre las diferentes variables y evaluar si existen sesgos potenciales en los modelos de tarificación existentes. Además, se explorarán y aplicarán una variedad de técnicas con el fin de identificar y mitigar cualquier sesgo identificado. Esto permitirá, más allá de un marco teórico, abordar de manera práctica las diferentes metodologías que pueden ayudar a mejorar la equidad y la transparencia en la fijación de precios de seguros.

Al abordar este tema desde múltiples perspectivas y utilizando la herramienta accesible de libre difusión Python, este trabajo busca ofrecer una visión integral y pragmática de

[1] Real Decreto Legislativo 8/2004, BOE núm. 278, de 29 de octubre, por el que se aprueba el texto refundido de la Ley sobre responsabilidad civil y seguro en la circulación de vehículos a motor, 2004.

cómo mitigar el sesgo hacia diferentes variables protegidas en los modelos de tarificación de seguros.

2. ESTADO DEL ARTE

En la actualidad, la preocupación por la equidad y la justicia en la fijación de precios de seguros ha aumentado significativamente. Las regulaciones y las normativas legales continúan evolucionando para abordar estas preocupaciones y garantizar que las prácticas de fijación de precios sean transparentes, no discriminatorias y equitativas para todos los asegurados. Además, el avance tecnológico y el acceso a grandes cantidades de datos han permitido a las aseguradoras sofisticar sus modelos de tarificación y personalizar las primas según el riesgo individual.

El uso de datos personales y la aplicación de algoritmos en los procesos de fijación de precios también plantean desafíos éticos y legales. Existe el riesgo de que los modelos de tarificación basados en algoritmos puedan introducir sesgos no intencionales o perpetuar sesgos existentes en los datos históricos. Por ejemplo, si los datos históricos muestran una tendencia discriminatoria en la fijación de precios hacia ciertos grupos demográficos, los algoritmos pueden aprender y replicar estos sesgos, incluso si no son explícitamente programados para hacerlo.

Ante estos desafíos, es fundamental que las aseguradoras adopten un enfoque proactivo para evaluar y mitigar cualquier sesgo en sus modelos de tarificación. Esto implica no solo realizar análisis de datos exhaustivos y pruebas rigurosas de los modelos, sino también desarrollar políticas y procedimientos claros para garantizar la equidad y la transparencia en la fijación de precios.

La discriminación injusta en la fijación de precios puede tener consecuencias significativas, como excluir a ciertos grupos de la población de acceder a coberturas de seguros básicas. O puede resultar en primas excesivamente altas para estos grupos, lo que a su vez puede desequilibrar la distribución del riesgo dentro del mercado. El avance tecnológico anteriormente mencionado también impacta la manera en la que los asegurados y usuarios interactúan, principalmente en redes sociales, donde una mala opinión o práctica puede ser compartida con miles de usuarios, suponiendo un riesgo reputacional para la aseguradora responsable.

Por lo tanto, es fundamental que las aseguradoras implementen medidas efectivas para prevenir y abordar cualquier forma de discriminación en la fijación de precios. Esto incluye el desarrollo de políticas y procedimientos claros que prohíban explícitamente la discriminación injusta y el establecimiento de controles y mecanismos de supervisión adecuados para garantizar el cumplimiento de estas políticas; así como concienciarse, como se analizará más tarde en este trabajo, de que eliminar algunas variables protegidas no es suficiente para eliminar la discriminación de ciertos grupos. Por otra parte, las aseguradoras deben comprometerse activamente a promover la diversidad, la equidad y la inclusión en procedimientos internos y externos, desde la contratación y la formación de sus empleados, hasta la prestación de servicios a los asegurados.

En este contexto, el análisis de la discriminación en la fijación de precios de seguros se vuelve cada vez más relevante. Comprender las diferentes formas en que puede surgir la discriminación y desarrollar estrategias efectivas para identificar y abordar estos problemas, son pasos críticos para garantizar que la fijación de precios sea justa y equitativa para todos los asegurados. En las próximas secciones, se explorará en detalle las definiciones de discriminación en el contexto de la fijación de tarifas de seguros y se discutirán diversas técnicas para mitigar el sesgo hacia variables protegidas.

2.1. Diferenciación: definición y tipos

En el proceso de fijación de precios de seguros, la diferenciación es fundamental para evaluar adecuadamente los perfiles de riesgo y establecer primas acordes con el nivel de riesgo de cada asegurado. Sin embargo, el problema surge cuando esta diferenciación se basa en variables que están relacionadas con características personales protegidas, como el género. En tal caso, la diferenciación puede derivar en discriminación injusta, lo que afecta a la equidad y la justicia en el acceso a la protección del seguro.

El concepto de diferenciación en el contexto del seguro puede ser complejo de abordar, ya que es necesario distinguir entre la diferenciación legítima, basada en el riesgo, y la discriminación injusta. Para hacerlo, es importante reflexionar sobre los diferentes tipos de discriminación que pueden ocurrir a lo largo del ciclo de vida de una póliza. Esto incluye la discriminación en la suscripción de la póliza, en la fijación de precios, en la renovación de la póliza y en la gestión de siniestros o reclamaciones. Al comprender y reconocer estos diferentes tipos de discriminación, podemos desarrollar estrategias efectivas para identificar y abordar cualquier práctica discriminatoria que pueda surgir, en especial, en el proceso de suscripción y tarificación. Esto nos permite garantizar que la diferenciación en la tarificación de seguros se base en criterios legítimos y objetivos relacionados con el riesgo, en lugar de en variables poco éticas o discriminatorias.

2.1.1. Discriminación directa

El concepto de discriminación directa, según la definición que proporciona la Comisión Europea, hace mención a “una situación en la que una persona es tratada de manera menos favorable que otra que ha sido o sería tratada en una situación comparable, por motivos de origen racial o étnico”. Ley 15 de 2022. Por la cual se promulga la integral para la igualdad de trato y la no discriminación. BOE. Núm. 167.

Al examinar las características consideradas en el sector asegurador, la pregunta que surge es por qué algunas, como la edad, son aceptadas como factores discriminantes en productos como los seguros de vida, mientras que otras, como la raza o etnia, no lo son. Una primera reflexión lleva a clasificar estos factores en dos categorías: alterables o inmutables. Se podría argumentar que los factores sobre los cuales un individuo no tiene control alguno deberían estar prohibidos para evitar la discriminación. Además, diversos estudios, como *The Social Psychology of Stigma*, indican que “las personas tienden a

aceptar mejor la discriminación cuando perciben que existe algún tipo de elección en la característica discriminatoria, como puede ser el caso de la obesidad". (Major B & O'Brien LT, 2005).

Sin embargo, la presencia de elección no debería necesariamente justificar la discriminación desde un punto de vista normativo. Por ejemplo, aunque una característica pueda ser modificable, como la religión, sigue siendo un atributo protegido contra la discriminación. Del mismo modo, la imposibilidad de cambiar una característica no garantiza que la discriminación basada en ella sea justificada, como sucede con el uso de la edad como factor discriminante. La edad es un rasgo inmutable y no puede modificarse, pero eso no implica automáticamente que sea justo discriminar a las personas mayores en términos de acceso a los seguros de vida. La edad por sí sola no determina la idoneidad de una persona para obtener cobertura de seguro, ya que otros factores como la salud y el estilo de vida también influyen en el riesgo asegurable. Por lo tanto, la discriminación basada únicamente en la edad no sería justificable en este contexto.

Adicionalmente, para comprender si una característica es discriminante, se debe tener en cuenta si dicha característica permanece constante o si cambia a lo largo de la vida del asegurado. Recuperando de nuevo la edad del asegurado como ejemplo, parece razonable que se utilice como factor discriminante dado que todos los asegurados tienen la misma oportunidad de experimentar tanto sus ventajas como desventajas a lo largo de vida.

Por último, es importante analizar si la característica que se está evaluando constituye una causa del riesgo o si correlaciona con el mismo. Para los cálculos actuariales generalmente el hecho de que una característica esté correlacionada con el riesgo es suficiente para que se catalogue como factor a emplear en los modelos de tarificación.

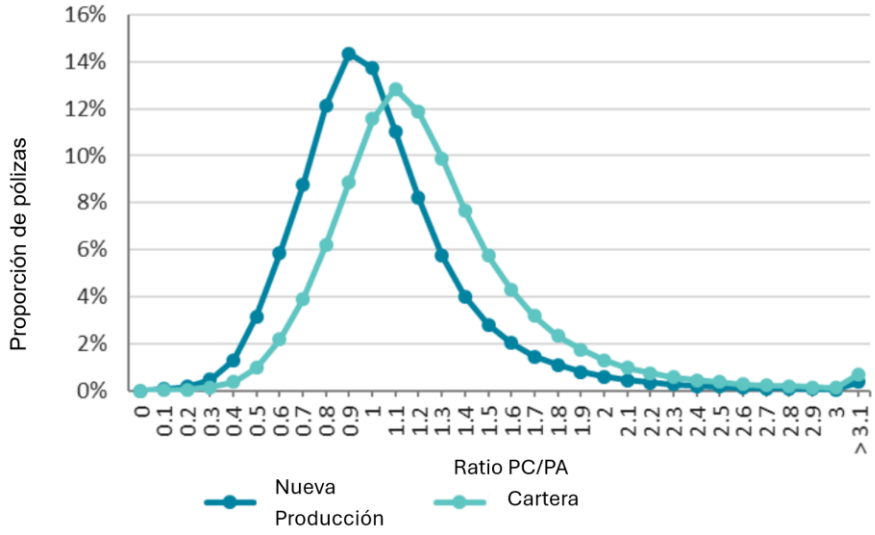
Si extrapolamos este concepto al ámbito asegurador, encontramos que esta situación puede manifestarse en varios momentos a lo largo del ciclo de vida de una póliza. En el momento inicial, durante la suscripción de una póliza, la aseguradora puede optar por rechazar la cobertura debido a ciertas características del asegurado. Un ejemplo de esta práctica es el denominado *redlining*, que constituye una práctica ilícita mediante la cual las aseguradoras deciden no ofrecer cobertura a individuos que residen en ciertas comunidades debido a su raza, color u origen nacional. El término *redlining* deriva de las marcas en rojo que se utilizaban en mapas codificados por colores para determinar qué áreas se consideraban más o menos seguras para otorgar préstamos hipotecarios. Las áreas marcadas en rojo, generalmente aquellas con una mayoría de población afrodescendiente, eran clasificadas como de mayor riesgo, lo que resultaba en la denegación sistemática de préstamos hipotecarios y otros servicios a sus residentes. El Movimiento por los Derechos Civiles, principalmente en Estadis Unidos, impulsó la promulgación de leyes *anti-redlining* que todavía hoy influyen en las prácticas del sector asegurador.

Asimismo, durante la vigencia de la cobertura, pueden producirse discriminaciones de tipo directo cuando las garantías se limitan para un asegurado debido a su estado de salud, como la presencia de enfermedades crónicas ya conocidas, como pueden ser asma o diabetes. Sin embargo, es en el proceso de tarificación donde más comúnmente se

observan prácticas discriminatorias. Aquí, se establecen primas más elevadas para ciertos asegurados en comparación con otros, incluso cuando los perfiles de riesgo son similares. Para ilustrar estos fenómenos, se examinarán más de cerca dos prácticas destacadas en el proceso de tarificación: el *dual pricing* y el *price walking*.

Con el término *dual pricing* nos referimos al hecho de diferenciar en la fijación de precios a los nuevos clientes y a los ya existentes en la cartera de la aseguradora, por motivos que no están relacionados con el riesgo ni el coste del servicio. Este fenómeno surge como estrategia de captación de clientes, en el cual se ofrece a los nuevos clientes una prima más baja, que a menudo se establece a un nivel que resulta en pérdidas (o, como mínimo, con un margen menor al del portafolio general). Esto se debe a que la siniestralidad y los gastos superan las primas recibidas en los nuevos negocios. Dichas pólizas son posteriormente empujadas hacía la rentabilidad mediante incrementos de prima, potencialmente a lo largo de varios años. El abuso de esta práctica llevó al Banco Central de Irlanda a realizar un estudio para analizar la variabilidad de las primas actuales en relación con esta técnica en los seguros de autos individuales y de hogar. Las conclusiones de dicho estudio se reflejan en el siguiente gráfico:

Figura 1. Distribución del ratio de la prima cobrada frente a la prima actuarial.



Fuente: Banco Centra de Irlanda, 2020

En la figura número 1 se observa en el eje vertical la proporción de pólizas y en el eje horizontal el ratio de la prima cobrada frente a la prima actuarial. La prima actuarial, comúnmente conocida por el término prima pura, es aquella que garantiza la suficiencia de una tarifa.

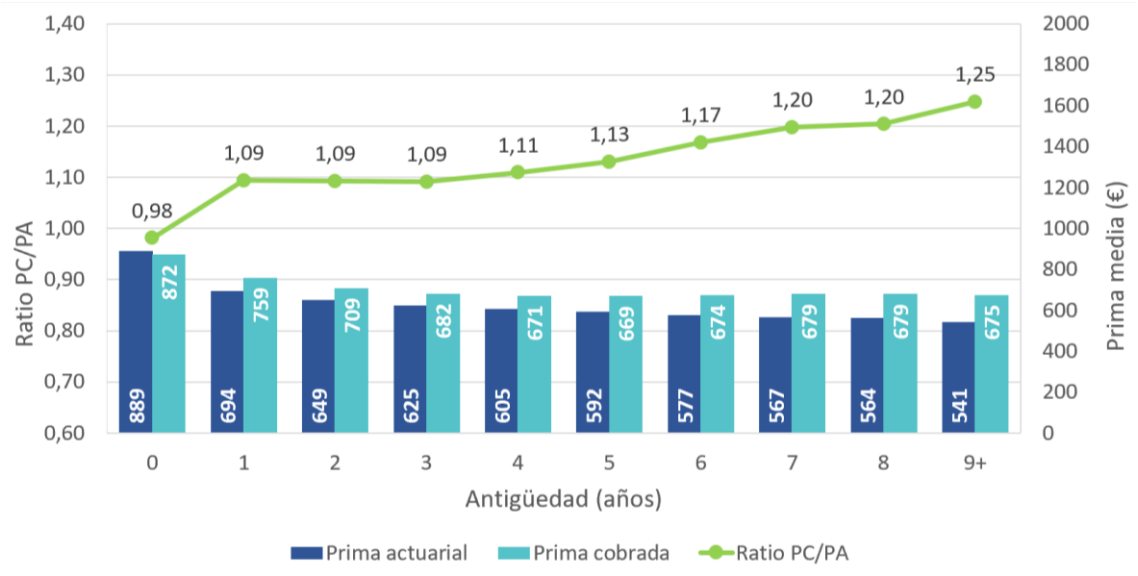
Con los resultados que se observan en la figura 1, se evidencia que la práctica de ofrecer descuentos para nuevos negocios es ampliamente utilizada en los mercados de seguros de automóviles privados y hogar. En promedio se observa que los clientes que renuevan pagan una prima más alta que los nuevos clientes en relación con los costes esperados de sus pólizas. Se observa que el 72% de los clientes que renuevan tienen una proporción

del ratio de prima cobrada y prima actuarial mayor que 1, en comparación con el 47% de los clientes nuevos.

Ante este tipo de resultados, el organismo EIOPA (European Insurance and Occupational Pensions Authority), ha emitido una serie de recomendaciones a los organismos nacionales competentes para supervisar que las prácticas de precios diferenciales no resulten en un trato injusto para los asegurados. Entre otros, estos organismos deben monitorear el mercado, revisar los procedimientos y materiales de marketing y cooperar con otras autoridades para proteger a los consumidores y asegurar la transparencia en el mercado.

Por otro lado, el fenómeno del *price walking* se manifiesta cuando se imponen primas más elevadas a los asegurados en relación con su coste esperado, a medida que prolongan su permanencia en la compañía. Este patrón implica un incremento gradual de las primas a lo largo del tiempo para los clientes existentes. En el mismo estudio citado anteriormente, se dedica un espacio para examinar esta práctica y se muestra el siguiente gráfico:

Figura 2. Ejemplo de price walkig



Fuente: Banco Centra de Irlanda, 2020

Las barras representan la prima media, mostrando en azul oscuro la prima actuarial y en azul claro la prima actual según la antigüedad de la póliza. La línea verde indica el ratio entre la prima actual y la prima técnica. Como se puede observar, aquellos asegurados que renuevan su póliza con el mismo asegurador durante nueve años o más, tienen un ratio promedio de 1,25; mientras que aquellos que lo hacen después de un año, tienen un ratio de 1,09. En otras palabras, el asegurado a largo plazo paga un 14% más en relación con los costes esperados de la póliza que el asegurado que renueva por primera vez.

Este patrón puede tener un impacto significativo en la percepción del consumidor sobre la lealtad hacia una compañía de seguros. Los clientes pueden sentirse injustamente tratados al enfrentar aumentos graduales en sus primas simplemente por mantener su

fidelidad a una aseguradora. Esta situación puede minar la confianza en la compañía y en el mercado de seguros en general. Cuando los consumidores interpretan que su lealtad es penalizada, es probable que cuestionen la validez de mantener relaciones a largo plazo con la misma aseguradora. En lugar de sentirse valorados y retribuidos por su fidelidad, pueden percibir que están siendo explotados o tratados injustamente. Como consecuencia, podrían optar por explorar otras opciones en el mercado, cambiando de aseguradora con una mayor frecuencia de lo que lo harían de otro modo.

Este incremento en la rotación de clientes puede acarrear consecuencias adversas para la estabilidad y la confianza en el mercado de seguros. Las compañías de seguros dependen en gran medida de la lealtad sostenida de sus clientes para asegurar su estabilidad financiera y operativa. La alta rotación de clientes genera incertidumbre y volatilidad en el mercado, lo cual puede afectar la capacidad de las compañías para proporcionar cobertura a precios competitivos, ya que adquirir nuevos clientes supone un coste adicional para las mismas.

2.1.2. Discriminación indirecta

La discriminación indirecta, también conocida como el efecto adverso, es un término que se remonta a 1971. Se empleó para ilustrar una instancia de discriminación laboral, el caso *Griggs v. Duke Power*, donde la empresa *Duke Power* requería un diploma de escuela secundaria y una serie de pruebas para el empleo en divisiones con salarios más altos dentro de la compañía. Dado que se determinó que estas pruebas no eran necesarias para los puestos de trabajo en cuestión y que las personas de raza negra tenían 10 veces menos probabilidades de cumplir con estos estándares, la Corte Suprema de Estados Unidos dictaminó que la política violaba la Ley de Derechos Civiles. Desde entonces, esta expresión ha sido ampliamente utilizada en distintos sectores, abarcando incluso la industria del seguro. Este tipo de discriminación hace referencia al efecto no intencionado que una práctica aparentemente neutral tiene sobre clases protegidas, entendiendo por clases protegidas el conjunto de individuos que comparten una característica común y que están protegidos legalmente contra la discriminación basada en dicha característica.

En el ámbito asegurador, la discriminación indirecta se utiliza para describir escenarios en los que una compañía de seguros emplea variables que parecen neutrales en su aplicación, pero que terminan afectado de manera no intencionada a las clases protegidas de forma negativa con la aplicación de tarifas más elevadas para esos grupos. La discriminación indirecta es una forma de discriminación que puede ser legalmente aceptada cuando las aseguradoras aplican políticas basadas en el perfil de riesgo de sus clientes. Aunque esto significa que pueden tratar de manera diferente a individuos basándose en aspectos como la ubicación o el historial de reclamaciones, esta práctica es vista como válida siempre que esté vinculada a una evaluación imparcial del riesgo. En contraste con la discriminación directa, que se basa en características protegidas sin una justificación objetiva, la discriminación indirecta puede ser defendible si está asociada con una evaluación razonable del riesgo.
Para que la discriminación indirecta sea considerada legalmente aceptable, deben cumplirse tres condiciones clave:

1. Justificación del objetivo: El motivo detrás de la discriminación indirecta debe ser válido y estar alineado con el propósito legítimo de la acción o práctica en cuestión. Un ejemplo de esto podría ser que una compañía de seguros aumente las tarifas en zonas geográficas específicas debido a un mayor nivel de riesgos, buscando así gestionar eficazmente los riesgos y mantener la solidez financiera de la empresa.
2. Adecuación de los medios utilizados: Los métodos empleados para implementar la discriminación indirecta deben ser adecuados y proporcionales al objetivo que se persigue. Esto significa que las variables utilizadas para diferenciar deben estar directamente relacionadas con el riesgo cubierto y no deberían ser intrínsecamente discriminatorias. Si una aseguradora usa el historial de reclamaciones como criterio de tarificación, debe asegurarse de que este sea un indicador confiable y relevante del riesgo asegurado.
3. Necesidad de los medios utilizados: Los medios empleados para la discriminación indirecta deben ser indispensables para alcanzar el objetivo legítimo y no deberían existir alternativas menos discriminatorias. Las aseguradoras deben demostrar que no hay otros métodos tan efectivos pero menos dañinos para lograr sus metas. Por ejemplo, si una aseguradora clasifica a los clientes según su ubicación geográfica, debe demostrar que no hay maneras menos discriminatorias de evaluar el riesgo en esas áreas específicas.

Sin embargo, en la práctica, estas condiciones no siempre se cumplen, dando lugar a situaciones alarmantes, especialmente en ramos como los de hogar y autos. En el ramo de hogar, por ejemplo, el uso de factores como la antigüedad o el valor de la vivienda para calcular las tarifas puede parecer justificado desde una perspectiva de riesgo. Sin embargo, estos factores han sido criticados por estar estrechamente relacionados con la raza, lo que podría dar lugar a una discriminación indirecta. A pesar de que estos factores son relevantes para la evaluación del riesgo, su uso puede no cumplir con los criterios mencionados, especialmente si se considera que no son necesarios ni proporcionales al objetivo de gestionar el riesgo de manera justa y equitativa.
Similarmente, en el ramo de autos, el uso del *scoring* de crédito como factor de riesgo ha sido objeto de controversia. Aunque esta variable muestra una alta correlación con la frecuencia de accidentes, su uso puede no cumplir con los criterios de necesidad y adecuación, especialmente si se considera que existen alternativas menos discriminatorias para evaluar el riesgo. Además, un informe de la Reserva Federal en 2007 señaló una posible relación entre la raza y las evaluaciones de seguro basadas en el crédito, lo que subraya la importancia de revisar cuidadosamente la práctica de utilizar el *scoring* de crédito como factor de riesgo.

Estos ejemplos ilustran cómo, a pesar de los criterios teóricos para justificar la discriminación indirecta, estas prácticas pueden no cumplir con los estándares requeridos, lo que pone en cuestión su justificación y plantea la necesidad de revisar y ajustar las políticas y prácticas de evaluación de riesgo en el sector de seguros.

2.1.3. Discriminación estadística

En *The Ethics of Statistical Discrimination*, Stephen Maitzen (1991) examina las cuestiones éticas que surgen al utilizar la discriminación estadística, una problemática relevante también en el análisis estadístico en el sector de seguros. Maitzen analiza cómo basar decisiones en datos estadísticos puede resultar problemático desde un punto de vista moral, especialmente cuando se evalúa a individuos según características de grupo.

El análisis estadístico en seguros permite a las empresas predecir riesgos futuros y tomar decisiones basadas en datos empíricos. No obstante, este enfoque puede resultar en discriminación estadística, tal como señala Maitzen. Esto sucede cuando las aseguradoras aplican patrones observados en grandes conjuntos de datos a individuos, resultando en tratamientos diferenciados basados en correlaciones que no necesariamente reflejan el riesgo individual. Maitzen sostiene que la discriminación estadística es problemática desde una perspectiva moral porque trata a los individuos no como entidades únicas, sino como miembros de grupos estadísticos.

Este problema se manifiesta en el sector de seguros cuando, por ejemplo, las compañías asignan tarifas más altas a ciertos grupos demográficos sin considerar las circunstancias individuales, perpetuando así injusticias y desigualdades. Un ejemplo práctico puede ser la aplicación de recargos a conductores noveles en el seguro de autos.

Además, la falta de innovación y adaptabilidad en las políticas de aseguramiento puede ser una consecuencia de depender excesivamente de modelos estadísticos tradicionales, un punto que también aborda Maitzen. Si las aseguradoras no incorporan nuevos métodos que consideren una gama más amplia de factores, pueden fallar en proporcionar una cobertura justa y equitativa.

Otros estudios, como el de Deborah Hellman en *When discrimination is wrong?*(2008), argumentan que la discriminación estadística puede consolidar desigualdades existentes al utilizar características de grupo para justificar el trato desigual. Hellman sugiere que este tipo de discriminación no solo perpetúa las desigualdades, sino que también puede crear nuevas formas de injusticia al basarse en datos históricos que reflejan prejuicios y desigualdades del pasado.

Finalmente, en *Fairness and Machine Learning* (Barocas et al., 2023) se explora cómo los algoritmos utilizados en la evaluación de riesgos pueden perpetuar sesgos si no se diseñan cuidadosamente. En este libro se argumenta que es crucial desarrollar algoritmos que no solo sean precisos, sino también equitativos, considerando factores éticos en su diseño y aplicación para evitar la perpetuación de sesgos y discriminación.

Estos estudios complementan la visión de Maitzen, destacando la importancia de diseñar políticas de seguro y algoritmos de evaluación de riesgos que sean tanto precisos como justos. Subrayan la necesidad de un enfoque más holístico e innovador en la evaluación

de riesgos, que incluya consideraciones éticas y sociales para evitar la discriminación y garantizar una cobertura más equitativa.

2.2. Marco regulatorio

Las actividades descritas previamente han provocado la intervención de supervisores y reguladores, quienes han implementado normativas en países como Alemania, Suiza, Irlanda, Italia y Estados Unidos. Este análisis se centrará en la regulación específica de Europa y el Reino Unido.

En Europa, la Dirección General de Competencia de la Comisión Europea ha emitido directrices para evitar la discriminación en el sector de seguros. Estas directrices promueven la transparencia en la fijación de precios y buscan eliminar prácticas que puedan resultar en tarifas más altas para ciertos grupos de la población sin una justificación sólida basada en la evaluación del riesgo. La meta es asegurar que las decisiones basadas en análisis estadísticos no perpetúen sesgos o desigualdades.

Además, la Autoridad Europea de Seguros y Pensiones de Jubilación (EIOPA) desempeña un papel vital en la regulación del sector de seguros en la Unión Europea. EIOPA ha lanzado varias iniciativas para enfrentar los desafíos en la fijación de precios de seguros y la discriminación estadística. Estas iniciativas incluyen orientaciones sobre el uso de datos y algoritmos en el sector asegurador, asegurándose de que las prácticas no conduzcan a la discriminación. También han implementado un marco de supervisión para que las aseguradoras evalúen el impacto de sus productos en diferentes segmentos de la población, revisando los criterios de fijación de precios e identificando posibles sesgos. Además, EIOPA trabaja para mejorar la protección del consumidor, promoviendo prácticas justas y transparentes, y obligando a las aseguradoras a proporcionar información clara y comprensible sobre la determinación de primas de seguros.

En el Reino Unido, la Autoridad de Conducta Financiera (FCA) ha implementado regulaciones similares para prevenir la discriminación indirecta en el sector de seguros. La FCA requiere que las aseguradoras documenten sus procesos de fijación de precios y justifiquen cualquier diferencia en las tarifas basadas en factores demográficos o de salud. Además, la FCA exige que las aseguradoras proporcionen explicaciones detalladas a los consumidores sobre cómo se calculan sus tarifas, permitiendo revisiones de precios si cambian los perfiles de riesgo de los clientes. Esta transparencia es crucial para que los consumidores comprendan las bases de las tarifas aplicadas y puedan tomar decisiones informadas.

En España, la Dirección General de Seguros y Fondos de Pensiones (DGSFP) es el organismo encargado de la supervisión y regulación del sector asegurador. La DGSFP ha implementado normativas que buscan proteger a los consumidores y asegurar la equidad en la fijación de precios de seguros. Estas normativas exigen a las aseguradoras que ofrezcan claridad sobre cómo se calculan las primas de seguros y que proporcionan justificaciones detalladas de las diferencias de precios entre los clientes. Además, la normativa española se centra en evitar la discriminación basada en factores como la edad, el género o la condición de salud, asegurando que las tarifas se basen en evaluaciones de riesgo justificadas y no en prejuicios estadísticos. La DGSFP también trabaja para

fortalecer la protección del consumidor, promoviendo la transparencia y la competencia justa en el mercado de seguros.

El Banco Central de Irlanda ha propuesto medidas para restringir el *"price walking"* y mejorar la transparencia en las renovaciones automáticas en el mercado de seguros no vida. El *"price walking"*, como se ha comentado anteriormente, se refiere a la práctica de incrementar gradualmente las primas de seguro de los clientes leales sin que estos se den cuenta. Al requerir la comunicación explícita de descuentos por adquisición de nuevo negocio y establecer requisitos claros para las renovaciones automáticas, estas propuestas buscan asegurar que los consumidores estén completamente informados y tengan opciones claras, contribuyendo así a un mercado de seguros más justo y transparente. Estas medidas no solo protegen a los consumidores, sino que también incentivan a las aseguradoras a competir de manera más equitativa.

Estas regulaciones buscan proteger a los consumidores contra prácticas discriminatorias, así como fomentar una competencia saludable en el sector de seguros, garantizando que las aseguradoras operen de manera justa y transparente. Al exigir que las aseguradoras expliquen claramente sus prácticas de fijación de precios y renovación de pólizas, se busca garantizar que los consumidores reciban un trato equitativo, sin importar su situación personal o demográfica. Esto es esencial para fomentar la confianza del consumidor y asegurar que todos los clientes tengan acceso a productos de seguros justos y razonables.

Además de las regulaciones que sirven como guía para abordar estos temas, numerosos autores se han pronunciado sobre la importancia de la transparencia en las prácticas de fijación de precios. Por ejemplo, en el estudio *Credit Card Pricing: The Persistence of Introductory Offers* (Bar-Gill & Bubb, 2012) se analiza cómo la falta de transparencia y las prácticas de fijación de precios engañosas pueden perjudicar a los consumidores. Aunque el estudio se centra en las tarjetas de crédito, los hallazgos son aplicables al sector de seguros, resaltando la importancia de regulaciones robustas para proteger a los consumidores y asegurar prácticas éticas y transparentes por parte de las empresas.

Asimismo, en *Discrimination in the Age of Algorithms* (Kleinberg et al., 2019), se examina cómo los algoritmos en la fijación de precios de seguros pueden perpetuar sesgos si no son cuidadosamente diseñados y supervisados. Los autores abogan por una regulación que garantice la equidad de los algoritmos, evitando que reproduzcan injusticias sociales existentes, lo cual refuerza la necesidad de regulaciones que promuevan tanto la precisión como la justicia. La inclusión de hallazgos de otros estudios subraya la importancia de una regulación cuidadosa y adaptativa para enfrentar los desafíos éticos y prácticos en la industria de seguros.

3. METODOLOGÍAS PARA MITIGAR EL SESGO. UNA APLICACIÓN AL SEGURO DE AUTOS.

3.1. Modelización predictiva

La modelización predictiva es una metodología de análisis de datos que utiliza métodos estadísticos y de aprendizaje automático para predecir resultados futuros basándose en datos históricos, en nuestro caso, la experiencia siniestral representada en el número de siniestros. Si bien el objetivo de este trabajo no es principalmente este aspecto, supone la base inicial para la posterior aplicación de las técnicas de reducción del sesgo. Para ello, se han llevado a cabo los pasos más frecuentes de este proceso, descritos en los próximos puntos.

3.1.1. Análisis univariante

En este paso inicial del proceso de tarificación se pretende analizar y explorar todas las variables susceptibles de ser empleadas en el modelo predictivo. La finalidad de este paso es múltiple.

El objetivo es familiarizarse con los valores y distribuciones de las variables, que combinado con el juicio experto que se pueda tener, sea un paso adicional de validación de los datos. Por otra parte, mediante el análisis de estas variables, se pueden observar casos atípicos que puedan impactar negativamente en la estimación de los parámetros de los modelos. Es, por tanto, una fase crucial de cualquier proceso de modelización predictiva.

En esta parte, además, se representa cada variable para cada mes (o año si los datos incluyen varios años) con el fin de analizar su estacionalidad, distribución y estabilidad temporal. Esta última parte puede ser de alto interés ya que si empleamos variables que no son estables a lo largo del tiempo, esto podría impactar negativamente el modelo propuesto.

3.1.2. Análisis bivariante

En esta segunda fase del proceso de modelización, una vez analizadas las variables individualmente, se pretende analizar de manera bivariante las variables del modelo. Una aproximación puede ser enfrentar todas las variables de interés frente a la variable respuesta o variable a explicar. Este paso permite entender la relación que mantiene cada variable con la variable respuesta de manera visual. En este paso, podemos ya obtener algunas reflexiones acerca de qué variables pueden segmentar mejor nuestra variable de interés, o por el contrario, no mantener una relación directa con la misma a priori. Así mismo, y como se comentaba anteriormente, en esta etapa, el modelizador con juicio experto podrá relacionar los resultados observados con el conocimiento del negocio, con la finalidad de validar sus suposiciones o investigar en detalle cualquier anomalía o relación no esperada inicialmente.

Como segunda fase de este apartado, se pretende analizar la correlación lineal y no lineal entre las variables de interés. Es bien conocido que el uso de variables correlacionadas en modelos estadísticos puede causar problemas de multicolinealidad, lo que dificulta la interpretación y estabilidad de los coeficientes y consecuentemente los errores estándar asociados a los mismos. La consecuencia de este efecto puede conducir a la creación de modelos predictivos con menor capacidad predictiva y estabilidad, así como un mayor riesgo de sobreajuste y menor capacidad de generalización ante nuevos datos.

3.1.3. Selección de variables

La selección de variables en el proceso de modelización es una etapa crucial para mejorar la interpretabilidad, precisión y eficiencia. Este paso conlleva la identificación de aquellas variables que mejor ayudan a explicar la variable de interés, minimizando el riesgo de sobreajuste y ayudando a mejorar el comportamiento del modelo con datos futuros.

Las técnicas más habituales incluyen métodos de filtro (basados en correlación o información mutua), métodos de *wrapper* (como *forward selection* y *backward elimination*) y métodos basados en modelos (como regresión *Lasso y Ridge*).

En el presente trabajo se ha optado por la metodología *BorutaShap*, la cual combina el algoritmo *Boruta* y *SHAP* para identificar las variables más importantes, ofreciendo una selección robusta y explicable basada en métodos de aprendizaje automático y que resulta de gran utilidad para datos complejos y de alta dimensionalidad, ya que ofrece la posibilidad de seleccionar una muestra de la base de datos originales, reduciendo los tiempos de ejecución.

El algoritmo *Boruta* es una técnica de selección de variables propuesta por Miron B. Kursa y Witold R. Rudnicki en 2010. Este procedimiento nos permite conocer las variables más relevantes, habiendo excluido previamente las más correlacionadas y es independiente del resto de métodos. Se detalla a continuación el funcionamiento de este método:

1. Se incorporan nuevas variables aleatorias al conjunto de datos inicial denominadas variable *shadow* o variables sombra. La distribución de estas variables se basan en las variables originales presentes en la base de datos.
2. Entrenamiento del conjunto de datos transformados mediante un modelo de aprendizaje automático permitido por el paquete de *BorutaShap*. A partir del resultado de este modelo, se obtiene la importancia relativa de cada variable, descartando las no importantes e iterando sobre el resto de variables.
3. Si procede a analizar si la variables originales tienen mayor importancia que las variables sombra.
4. Las variables se clasifican en tres categorías: importantes, no importantes y no decisivas. Dicha clasificación depende de la importancia de las variables sombra con respecto a las variables originales, dado que si la variable sombra es más importante que la variable original, no tendrá sentido incluir esta variable, puesto que su aportación no es fiable.

3.1.4. Modelo inicial

El proceso comienza con la construcción de un modelo base que incorpora las variables más importantes para la predicción de la frecuencia de siniestros, incluida la variable protegida. El tipo de modelo que se va a utilizar es un modelo lineal generalizado (GLM), una técnica ampliamente empleada en la industria, para evaluar tanto el poder predictivo como el sesgo de este modelo inicial. Su popularidad de debe en parte a su flexibilidad al admitir varias distribuciones para la variable respuesta distintas a la distribución normal como podrían ser la distribución Poisson o Gamma.

La expresión de estos modelos viene dada a través de la siguiente expresión:

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

donde la $g(\mu_i)$ se refiere a la función de enlace que transforma μ_i y que es igual al intercepto (β_0) y una combinación lineal de los predictores denotados por $X_{i1}, \dots, X_{p1}$ y sus coeficientes $\beta_0, \dots, \beta_p$. Una vez se ha calculado el predictor lineal, se aplica la inversa de la función especificada en $g(.)$ para conocer el valor de interés μ_i.

A continuación, procederemos a comparar este modelo base con otras técnicas con el fin de identificar aquella que pueda reducir el sesgo mientras mantiene o mejora la capacidad predictiva del modelo inicial. Este enfoque nos permitirá seleccionar la estrategia más eficaz para desarrollar un modelo que sea tanto preciso como justo. Para ello, es preciso definir tanto una métrica que permita evaluar el sesgo de un modelo como una métrica que mida el poder predictivo de dicho modelo.

En cuanto al sesgo, se va a utilizar el *Disparate Impact Ratio (DIR)* que es una métrica que se utiliza mayormente en los análisis de equidad para medir la disparidad en los resultados entre diferentes grupos protegidos como podría ser el caso del género. La fórmula original es:

$$DIR = \frac{P(Y = 1 | G = grupo\ sensible\ o\ minoría)}{P(Y = 1 | G = grupo\ base\ o\ mayoría)}$$

Es decir, trata de calcular un ratio con la proporción del grupo no privilegiado que recibió un resultado positivo frente a la proporción del grupo privilegiado que recibió el mismo resultado positivo. Como en este caso de estudio, se trata de un problema de regresión en el que se pretende predecir la frecuencia del número de siniestros, se ha versionado dicho ratio calculándolo como la media de la predicción para el grupo no protegido sobre la media de la predicción para el grupo protegido. Si el ratio es mayor que 1, significa que, en promedio, las predicciones para el grupo minoritario son más altas que las predicciones para el grupo mayoritario. Es decir, el modelo no está capturando de manera equitativa la variabilidad de la variable respuesta. En el caso de que este ratio fuese menor a uno, indicaría un sesgo se da hacia el grupo mayoritario. Un valor de la métrica igual a 1 indicaría que no hay sesgo con respecto a la variable protegida en el modelo utilizado.

Por otro lado, para determinar el poder predictivo y así poder comparar las distintas técnicas que se van a comparar. En este caso de estudio se va a utilizar el *MAE* (*Mean Square Error*) que se define a partir de la siguiente fórmula como la distancia promedio entre el valor real y las predicciones del modelo:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

Valores altos del MAE indican aumento del error de predicción del modelo.

3.2. Técnicas de reducción del sesgo

Este análisis se ha dividido en tres etapas atendiendo al proceso de modelización: pre-procesamiento, procesamiento del modelo y post-procesamiento. Además, se aportarán ejemplo prácticos en cada una de estas fases para evaluar el impacto tanto en el poder predictivo como en el sesgo de cada una de estas técnicas

3.2.1. Pre-procesamiento

La creencia generalizada de que la precisión del modelo aumenta proporcionalmente con la cantidad de datos empleados, ha sido ampliamente corroborada en diversas aplicaciones tanto de técnicas de aprendizaje automático como con el uso de técnicas más tradicionales. Sin embargo, cuando se pretende abordar la equidad en el proceso de modelización, la calidad de los datos desempeña un papel aún más crucial que la cantidad de registros.

3.2.1.i. Omisión de la variable protegida (*unawarness principle*)

En una primera aproximación para reducir el sesgo, cabría pensar que al eliminar explícitamente la variable protegida del proceso de modelización cualquier decisión tomada por el modelo sería independiente de la variable protegida. Sin embargo, esa afirmación sólo es cierta si la característica que hemos eliminado del modelo es independiente de cualquier otra característica que esté incluida en el modelo. Esta situación se presenta escasamente en la práctica.

3.2.1.ii. Correlación mediante coeficientes

La correlación es una herramienta estadística muy útil que revela el grado de relación entre dos variables al mostrar cómo cambian conjuntamente a una tasa constante.

Con el fin de evitar que las variables correlacionadas con la variable protegida actúen como sustitutos de la misma, y por ende afecten indirectamente al criterio de independencia, se procede a eliminar del modelo tanto la variable protegida como aquellas variables correlacionadas con la misma para comprobar el efecto en el sesgo.

Para conocer las variables correlacionadas con la variable protegida se han aplicado dos técnicas de acuerdo a la naturaleza de las variables que se contrasten. En el caso de dos variables categóricas como por ejemplo la variable protegida y el tipo de agente, se ha utilizado el coeficiente de corrección V de Cramer basado en la prueba de Chi-cuadrado. Este coeficiente mide la fuerza de asociación entre dos variables categóricas, siendo el 1 el valor máximo y el 0 el valor mínimo que indica que las variables no están asociadas. La fórmula que se utiliza para obtener este coeficiente es la siguiente:

$$V = \sqrt{\frac{\chi^2}{N - m}}$$

donde N es el número total de observaciones de la tabla y m el valor mínimo (f-1, c-1) siendo f el valor de filas y c el valor de las columnas de la tabla de contingencia. El numerado que representa el índice 'Chi(Ji) cuadrado" de define como:

$$\chi^2 = \sum_{t=1}^{n} \frac{(f_e \ - f_t)^2}{f_t}$$

donde f_e hace referencia a la frecuencia empírica y f_t a la frecuencia teórica.

En el caso de contrastar la relación entre la variable protegida y una variable continua como podría ser la edad, se ha utilizado un caso especial del coeficiente de correlación de Pearson: el coeficiente de correlación biserial puntual (que viene dado por la siguiente expresión:

$$r_{p,b} = \frac{\tilde{X}_1 - \tilde{X}_0}{S_{1,0}} * \sqrt{\frac{n_1 * n_0}{n_{1,0}^2}}$$

donde

$\tilde{X}_1$: media de la muestra de respuestas del primer grupo,

$\tilde{X}_0$: media de la muestra de respuestas del segundo grupo,

$S_{1,0}$: desviación estándar de los valores de la muestra de la variable continua,

n_1: el número de respuesta del primer grupo,

n_0: el número de respuestas del segundo grupo,

$n_{1,0}$: número total de pólizas que integran la muestra.

3.2.1.iii.Correlación no lineal (*dependency analysis*)

En el análisis de conjuntos de datos complejos, la exploración de relaciones lineales entre las variables explicativas en un modelo puede ser limitada, y en ocasiones insuficiente, para comprender las interacciones presentes entres las distintas características de las pólizas. La realidad de muchos fenómenos observados en el contexto asegurador revela que las relaciones entre variables pueden ser intrínsecamente no lineales, lo que requiere un enfoque analítico más sofisticado para su comprensión. En este caso de estudio el análisis de dependencia se va a basar en analizar la importancia de las variables resultado de un modelo de *Gradient Boosting Machine (GBM)* donde la variable objetivo será la

variable protegida. De esta forma, si una variable explicativa en particular muestra una alta importancia en la predicción de la variable objetivo, esto indicará una dependencia entre ellas.

En el proceso de modelización tradicional la práctica más común es la creación de un único modelo predictivo que sea robusto para predecir la variable respuesta. No obstante, si se hace uso de las técnicas de *Machine Learning* o aprendizaje automático es posible cambiar dicho enfoque y utilizar un conjunto de modelos para abordar una tarea de aprendizaje específica. La familia de modelos de tipo *Boosting* se caracteriza por añadir nuevos modelos al conjunto utilizado de manera secuencial. Es decir, se parte de un primer árbol de decisión y se calcula el error residual de ese modelo como la diferencia entre el modelo y los valores reales. A continuación, se construye un segundo árbol de decisión que intenta corregir los errores residuales del modelo anterior. Este proceso se repite iterativamente añadiendo nuevos árboles de decisión hasta que se ha alcanzado un criterio de detención que se ha predefinido inicialmente. Este criterio de detención puede ser tanto el número máximo de iteraciones, un umbral para el error residual o el número de registros que deben estar en cada una de las hojas del árbol de decisión.

Al utilizar este algoritmo, uno de los hiper-parámetros a seleccionar por el analista es la función de pérdida. El avance de este tipo de técnicas permiten al analista seleccionar dicho hiperparamétro entre una amplia variedad de funciones de pérdida desarrolladas hasta la fecha y la posibilidad de implementar una función de pérdida específica para la tarea en cuestión. En el caso de que la función de error sea la pérdida cuadrática clásica, el procedimiento de aprendizaje resultaría en un ajuste secuencial de los errores. Es decir, se llevarían a cabo los siguientes pasos:

A. Inicialización

En el ejemplo

$$F_0(x) = \log\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} E_i}\right)$$

Donde:

- y_i es el numero de siniestros para la observación i
- E_i es la exposición para la observación i

B. Interacción

Para cada interacción de m, se realizan los siguientes pasos:

B.1. Cálculo de los residuos:

Los residuos, o técnicamente pseudo-residuos r_{im} se basan en la diferencia entre el número de siniestros observados y valor esperado por el modelo anteriormente indicado ajustado por la exposición en nuestro ejemplo:

$$r_{im} = y_i - E_i \exp\left(F_{m-1}(x_i)\right)$$

B.2 Ajuste de un nuevo modelo débil $h_m(x)$:

Se ajusta un nuevo modelo débil $h_m(x)$ a los pseudo-residuos r_{im}:

$$h_m(x) = \arg min_h \sum_{i=1}^{n} (\, r_{im} - \, E_i h\,(x_i))^2$$

B.3 Actualización del modelo global:

El modelo global se actualiza sumando el nuevo modelo ajustado multiplicado por la tasa de aprendizaje Л:

$$F_m(x) = \, F_{m-i}(x) - Л\mathrm{h}_m\,(x)$$

Matemáticamente, este tipo de modelo se construyen agregando sucesivamente modelos débiles, lo que permite capturar relaciones no lineales y complejas, que con un modelo de regresión lineal generalizado, no sería posible.

C. Predicción final

La predicción final de la tasa de siniestros se obtiene exponenciando la suma de los modelos ajustados:

$$\lambda(x) = \, \exp\,(F_M(x))$$

Donde $\lambda(x)$ es la tasa esperada de siniestros por unidad exposición.

Cabe recordar que el GBM es un modelo no paramétrico, presenta un número de parámetros que puede crecer con la cantidad y complejidad de los datos. Se trata de modelos mas flexibles y pueden modelizar relaciones complejas, si bien como se explicará mas tarde, son modelos menos interpretables.

3.2.1.iv. Análisis de residuos

Al emplear el enfoque de la correlación y excluir variables del modelo, estamos descartando parte de la información que estas variables aportan para explicar la variable de respuesta, en este caso, la frecuencia de siniestros. Es decir, este enfoque tiene limitaciones, ya que no logra capturar toda la complejidad de las relaciones entre las variables. Para abordar esta complejidad, recurrimos al método de análisis factorial.

El análisis factorial busca identificar factores latentes subyacentes que expliquen las correlaciones observadas entre las variables. Estos factores representan patrones más amplios y subyacentes en los datos, lo que proporciona una comprensión más profunda de las relaciones entre las variables. Por ejemplo, si tenemos un conjunto de variables relacionadas con las características de los asegurados, el análisis factorial podría identificar factores como "edad", "frecuencia de pagos" y "score crediticio", para explicar el perfil de los clientes de la cartera.

Sin embargo, aunque el análisis factorial proporciona una visión más completa de las relaciones entre las variables, sus resultados suelen ser abstractos y difíciles de interpretar directamente. Por esta razón, se propone una alternativa más práctica: utilizar los residuos del modelo de la variable protegida junto con las variables correlacionadas como predictores adicionales para predecir la frecuencia de siniestros. Este enfoque nos permite emplear la información proporcionada por las correlaciones sin depender exclusivamente

de los resultados del análisis factorial. Además, nos ofrece una herramienta más interpretable y práctica para el análisis y la toma de decisiones en el contexto de seguros.

Para aplicar esta metodología, se comienza construyendo un modelo donde la variable a predecir sea la variable protegida y la única variable independiente sea la más correlacionada, que en este caso es la edad. Una vez obtenido este modelo, se calculan los residuos. Más tarde, se repite el mismo proceso con la segunda variable más correlacionada, el modelo del vehículo, utilizando su respectivo modelo para obtener los residuos correspondientes. Finalmente, se emplean los residuos de ambos modelos como variables independientes en el modelo final.

3.2.1.v. Remuestreo (*re-sampling*)

Después de explorar en detalle las relaciones entre las variables explicativas en nuestro estudio, es crucial ahora dirigir nuestra atención hacia la distribución de la variable protegida. Como se ha podido comprobar anteriormente la distribución de la variable protegida en la base de datos de estudio es muy desbalanceada. Para mitigar este efecto, se ha realizado un exhaustivo análisis sobre la aplicación de la técnica de remuestreo como estrategia fundamental para mitigar los desafíos derivados del desbalance de clases en un conjunto de datos. A través del remuestreo, se busca equilibrar la representación de clases durante el entrenamiento del modelo, garantizando así una consideración justa de todas las muestras, independientemente de su clase.

El remuestreo ofrece una serie de ventajas significativas, tanto en términos generales como aplicadas específicamente a los Modelos Lineales Generalizados (GLM). En primer lugar, la capacidad del remuestreo para equilibrar la representación de clases es fundamental. Esta práctica asegura que cada clase esté representada de manera equitativa durante el proceso de entrenamiento del modelo, lo que ayuda a contrarrestar cualquier sesgo introducido por el desbalance de clases en los datos.
Además, el remuestreo contribuye a mejorar la precisión del modelo al proporcionar una representación más equitativa de las clases. Esto se traduce en predicciones más precisas y confiables en general, lo que es esencial para la toma de decisiones informadas y la generación de conclusiones precisas a partir de los resultados del modelo.

Otra ventaja significativa del remuestreo es su capacidad para reducir el riesgo de sobreajuste al modelo. Al garantizar una representación equitativa de todas las clases durante el entrenamiento, el remuestreo ayuda al modelo a aprender patrones generales en los datos, en lugar de simplemente memorizar las clases más comunes. Esto promueve la capacidad de generalización del modelo y reduce la probabilidad de que el modelo se ajuste demasiado a los datos de entrenamiento, lo que podría llevar a predicciones menos precisas en datos nuevos.

Al aplicar el remuestreo específicamente a los GLM, se observan varias ventajas adicionales. Por ejemplo, el remuestreo contribuye a mejorar la estimación de los parámetros del modelo al proporcionar una representación equitativa de todas las clases durante el proceso de entrenamiento. Esto es fundamental para realizar inferencias precisas sobre los datos y para garantizar que las conclusiones derivadas del modelo sean sólidas y confiables.
Además, el remuestreo promueve la equidad en las predicciones del modelo al mitigar el sesgo asociado al desbalance de clases. Al proporcionar una representación equitativa de todas las clases durante el entrenamiento, el remuestreo asegura que las predicciones del

modelo no estén sesgadas hacia ninguna clase en particular, lo que es esencial para garantizar la equidad y la imparcialidad en el análisis de los datos con GLM.

3.2.1.vi. Reponderación (*Re-weightning*)

Tras haber examinado detalladamente el proceso de remuestreo y su relevancia en la equidad y precisión de los modelos, resulta fundamental ahora adentrarnos en otro componente esencial del preprocesamiento de los datos: la reponderación.

El uso de reponderación mediante remuestreo con pesos, en contraposición a la técnica de remuestreo utilizada en el apartado anterior, se justifica por su capacidad para abordar de manera más efectiva el desbalance de clases y mitigar cualquier sesgo asociado en los datos. Mientras que el remuestreo tradicional selecciona muestras al azar, el remuestreo con pesos asigna pesos a las muestras de acuerdo con su clase, garantizando así una representación equitativa de todas las clases durante el entrenamiento del modelo. Esta asignación ponderada de pesos permite que el modelo tenga en cuenta de manera justa y equitativa las muestras de todas las clases, lo que contribuye a mejorar la precisión y equidad del modelo en general. Además, el remuestreo con pesos proporciona una mayor flexibilidad para ajustar la importancia relativa de las diferentes clases, lo que puede ser especialmente útil en situaciones donde ciertas clases son más importantes o representativas que otras.

La reponderación mediante remuestreo con pesos implica una serie de pasos críticos que contribuyen a corregir este desbalanceo. Para ello en primer lugar, se asignan pesos a cada muestra en función de su clase correspondiente. En el contexto de este estudio, los pesos se han determinado de manera que sean inversamente proporcionales al número de muestras en cada categoría de la variable protegida. Esta estrategia garantiza que las muestras de la clase minoritaria (en el caso de la variable “sexo” las mujeres) reciban un mayor peso, permitiendo así una representación más equitativa durante el proceso de entrenamiento del modelo.

Es importante destacar que la asignación de pesos es solo el primer paso en el proceso de reponderación. La técnica también implica el uso del remuestreo, donde las muestras se seleccionan con probabilidades proporcionales a los pesos asignados. Esto asegura que el modelo considere de manera justa y equitativa las muestras de ambas clases durante el proceso de entrenamiento, lo que a su vez conduce a una captura más precisa de la distribución subyacente de los datos.

En el contexto de los Modelos Lineales Generalizados (GLM), donde el sesgo puede tener un impacto significativo en la estimación de los parámetros del modelo y, consecuentemente, en las predicciones resultantes, la reponderación mediante remuestreo con pesos adquiere una importancia aún mayor. Al asignar pesos a las muestras en función de su clase y luego aplicar el remuestreo con probabilidades proporcionales a estos pesos, se asegura que el modelo considere de manera justa y equitativa todas las muestras durante el entrenamiento. Esto es crucial para contrarrestar el sesgo introducido por el desbalance de clases, lo que a su vez mejora la precisión y fiabilidad de las estimaciones de los parámetros del modelo y reduce el riesgo de predicciones sesgadas.

3.2.2. Procesamiento

Tras explorar las técnicas de preprocesamiento de datos, es relevante adentrarse en las técnicas de procesamiento interno (*in-processing*) que complementan y profundizan la capacidad de los modelos para abordar desafíos específicos de los datos. Las técnicas de procesamiento interno se refieren a métodos que se aplican durante el entrenamiento del modelo mismo, lo que les permite adaptarse y mejorar su capacidad de aprender y generalizar patrones en los datos.
En contraste con las técnicas de preprocesamiento, que se aplican antes de que los datos se introduzcan en el modelo, las técnicas de procesamiento interno trabajan directamente con el modelo durante su entrenamiento. En este caso de estudio se van a analizar cuatro técnicas distintas: el Disparate Impact Remover, la regresión con regularización, el Generative Conditional Adversarial Network (cGAN) y finalmente el Bayes Optimal Equalized Odds Predictor.

3.2.2.i.Disparate Impact Remover

Esta técnica, introducida en el estudio *Certifying and removing disparate impact* (Feldman et al., 2015), busca eliminar el impacto desigual en las predicciones de los modelos, asegurando un trato justo entre diferentes grupos categorizados. Aunque en el artículo se detalla una aplicación a un problema de clasificación, en el caso de estudio de ha adaptado la idea original al problema de regresión.

En el contexto de un problema de regresión, la técnica de Disparate Impact Remover se adapta para abordar las disparidades y sesgos que pueden surgir en los resultados de modelos de regresión. En lugar de separar clases o categorías, en la regresión, el límite de decisión se refiere al punto en el espacio de características donde se realiza la predicción. El Disparate Impact Remover ajusta este límite de decisión para reducir las disparidades en las predicciones de valores de salida. Por ejemplo, si el modelo tiende a sobreestimar o subestimar los valores de salida para ciertos grupos, el algoritmo ajustará el límite de decisión para mitigar estas disparidades y garantizar una predicción más equitativa en todo el conjunto de datos.

En la regresión, las probabilidades de predicción pueden interpretarse como la confianza del modelo en sus predicciones. El Disparate Impact Remover modifica estas probabilidades para lograr una distribución más equitativa de las predicciones de valores de salida entre diferentes grupos. Esto se logra recalibrando las probabilidades de predicción para que reflejen de manera más precisa la verdadera variabilidad en los datos, considerando la diversidad de los grupos y minimizando cualquier sesgo injusto en las predicciones.

Al aplicar el Disparate Impact Remover a un problema de regresión, se busca garantizar que el modelo produzca predicciones justas y equitativas en términos de los valores de salida, considerando las características sensibles o protegidas presentes en el conjunto de datos. Esto contribuye a mitigar cualquier sesgo o disparidad potencial en los resultados del modelo y promueve la equidad en las predicciones para todos los grupos involucrados.

3.2.2.ii. Regularización: Ridge regression

El modelo Ridge es una técnica de regresión lineal que destaca por su capacidad para abordar el problema de la multicolinealidad en conjuntos de datos donde las variables predictoras están altamente correlacionadas. Este enfoque se centra en reducir el sobreajuste (*overfitting)* al introducir una penalización sobre los coeficientes de regresión, lo que contribuye a la estabilidad y la generalización del modelo.
Una de las características principales del modelo Ridge es la incorporación de una penalización de regularización a la función de pérdida del modelo. Esta penalización, conocida como término de regularización L2, es controlada por un parámetro de regularización lambda (λ) como se puede ver en la siguiente expresión:

$$\hat{\beta}^{ridge} = argmin_{\beta \epsilon \mathbb{R}} ||y - X\beta||_2^2 + \lambda ||\beta||_2^2$$

El aumento del valor de λ incrementa la penalización, lo que resulta en coeficientes de regresión más pequeños y, por ende, en modelos más sencillos. Esta estrategia de regularización tiene el efecto de reducir la varianza de los coeficientes de regresión, haciendo que el modelo sea menos sensible a pequeñas variaciones en los datos de entrenamiento.

Desde el punto de vista de la eliminación del sesgo, el modelo Ridge puede contribuir indirectamente al fortalecimiento de la estabilidad y la generalización del modelo. Al reducir el sobreajuste, el modelo es menos propenso a capturar patrones espurios en los datos de entrenamiento que podrían estar sesgados. Además, al regularizar los coeficientes de regresión, el modelo puede mitigar la influencia de variables predictoras que podrían introducir sesgos indeseados en las predicciones.

En el contexto de la protección de variables sensibles o protegidas, el modelo Ridge puede ser particularmente útil. La regularización L2 en el modelo Ridge puede ayudar a prevenir el uso excesivo de estas variables en las predicciones, al reducir la magnitud de los coeficientes asociados a ellas. Esto se logra al penalizar los coeficientes de regresión, lo que disuade el modelo de depender demasiado de estas variables, contribuyendo así a la equidad en las predicciones.

Para analizar la efectividad de este método se ha definido una función de pérdida personalizada, la cual integra el error cuadrático medio (MSE) estándar con un término de regularización que penaliza la diferencia promedio en las predicciones entre grupos protegidos y no protegidos. Esta penalización se calcula ponderando la diferencia media en las predicciones entre estos grupos por un factor lambda (λ), que determina la intensidad de la penalización. Más tarde, se ajusta un modelo de regresión Ridge inicialmente utilizando los datos de entrenamiento sin considerar la función de pérdida personalizada. Posteriormente, se utilizan las predicciones de este modelo para calcular los pesos de muestra personalizados, los cuales se basan en la función de pérdida personalizada y se emplean para ajustar el modelo nuevamente. Este ajuste final del modelo Ridge utiliza los pesos de muestra personalizados, lo que prioriza las instancias que contribuyen a reducir la disparidad en las predicciones entre los grupos protegidos y no protegidos. Este enfoque garantiza que el modelo tome en cuenta la equidad y la justicia al hacer predicciones, ya que otorga mayor peso a las instancias que ayudan a minimizar las disparidades entre los grupos.

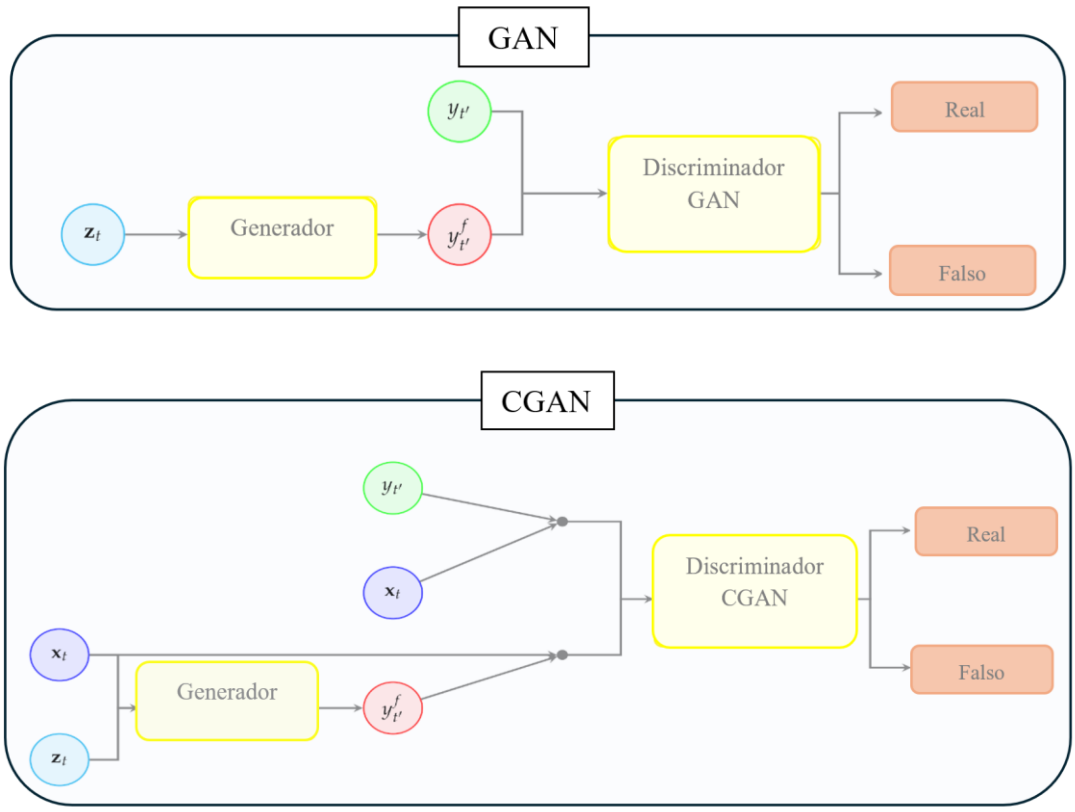
3.2.2.iii. Generative Conditional Adversarial Network (cGAN)

El aprendizaje automático (ML) se propone como un medio para emular la capacidad cognitiva humana al identificar y generalizar patrones en los datos. Este campo se encuentra en constante desarrollo y abarca una amplia gama de aplicaciones, desde el reconocimiento de patrones hasta el procesamiento del lenguaje natural y la detección de anomalías. Sin embargo, el éxito de los modelos de ML está intrínsecamente ligado a las características utilizadas durante el proceso de entrenamiento. Aunque tradicionalmente estas características son proporcionadas mediante técnicas de ingeniería de características (*feature engineering*), cuya implementación puede resultar una ardua tarea y, en ciertos contextos, poco factible de llevar a cabo manualmente, ha surgido un interés creciente en la capacidad de los modelos de ML para extraer características de manera autónoma. Este enfoque, conocido como aprendizaje de representaciones, ha adquirido particular relevancia en el ámbito del *Deep Learning*. Su objetivo es identificar y extraer información relevante de los datos para mejorar las predicciones o incluso generar nuevos datos. En este sentido, los modelos generativos representan una de las aproximaciones más prometedoras, ya que buscan capturar la distribución subyacente de los datos para poder generar ejemplos realistas a partir de esta distribución aprendida. En concreto, las redes generativas adversariales (GANs) han sido reconocidas como uno de los avances más significativos en el campo de la inteligencia artificial en los últimos años. Esta técnica se distingue por su capacidad para comprender la distribución de probabilidad de los datos y generar réplicas sintéticas que se asemejan a los datos reales. Entre estas aplicaciones se incluyen la ampliación de conjuntos de datos, la mitigación de problemas de desequilibrio de clases y el aprendizaje de representaciones justas.

Sin embargo, para el caso de estudio, aunque los Generative Adversarial Networks (GAN) son herramientas poderosas para generar datos sintéticos, en el contexto de muestras desbalanceadas al tener una variable protegida, como el sexo, se ha observado una preferencia por el uso de Conditional Generative Adversarial Networks (CGAN). Esta preferencia se debe a varias ventajas que ofrece CGAN en este escenario específico. En primer lugar, CGAN permite generar datos sintéticos condicionados a características específicas, como el sexo. Esto significa que se pueden controlar las características de los datos generados para cada grupo de interés, lo que facilita la generación de muestras equilibradas y representativas de todos los grupos, incluso aquellos que están subrepresentados en la muestra original.
Además, al condicionar la generación de datos al sexo, CGAN puede ayudar a comprender mejor cómo estas características afectan la distribución de los datos y cómo pueden influir en el rendimiento del modelo. Esto permite una mayor transparencia y comprensión del proceso de generación de datos, lo que puede ser crucial para identificar y mitigar posibles sesgos en el modelo generado.

Para una comprensión más completa de los beneficios mencionados, es crucial profundizar en el funcionamiento de esta técnica. Para este propósito, se recurrirá a los siguiente esquemas:

Figura 3. Estructura Gan y cGan



Fuente: elaboración propia

El propósito fundamental del modelo GAN es desarrollar un modelo generativo mediante un proceso adversarial. Este proceso implica el entrenamiento simultáneo de dos modelos distintos. En primer lugar, se entrena un modelo generativo, representado por G (generador), que, en el contexto de la predicción de datos, aprende los patrones previos en los datos y realiza inferencias sobre los valores predictivos. Por otro lado, se entrena un modelo discriminatorio, denotado como D (discriminador), que evalúa la probabilidad de que una muestra dada provenga de los datos de entrenamiento reales en comparación con ser generada por el modelo generador. Como resultado, el generador se enfrenta a un adversario, el discriminador, cuyo objetivo radica en detectar diferencias entre muestras de datos reales y generadas. Ambos componentes del modelo se entrenan mediante un proceso de optimización que busca maximizar la dificultad del discriminador para distinguir entre muestras reales y generadas. En general, tanto el generador como el discriminador se implementan como perceptrones multicapa, empleando métodos de gradiente estocástico para su optimización.

Esto se refiere a la arquitectura de los modelos generador y discriminador en un GAN. Los perceptrones multicapa son una forma de red neuronal artificial que consta de múltiples capas de neuronas, cada una conectada a la siguiente en secuencia. En el contexto de un GAN, tanto el generador como el discriminador se implementan utilizando esta estructura de red neuronal multicapa. El generador toma una entrada de ruido aleatorio y la proyecta a través de múltiples capas ocultas para generar datos sintéticos que se asemejen a los datos reales. Por otro lado, el discriminador recibe como entrada tanto datos reales como datos generados y los procesa a través de múltiples capas ocultas para distinguir entre ellos, es decir, para determinar si una muestra dada es real o generada.

Para optimizar los parámetros de estos modelos, se utilizan métodos de gradiente estocástico, que ajustan los pesos de las conexiones neuronales en función del gradiente de una función de pérdida. Este proceso de optimización se lleva a cabo iterativamente durante el entrenamiento del GAN, con el objetivo de mejorar continuamente el rendimiento del generador y del discriminador para que el generador pueda generar datos más realistas y el discriminador pueda discernir de manera más efectiva entre datos reales y generados.

Por otro lado, en el caso de un modelo GAN no condicionado, no existe control sobre el proceso de generación de datos, lo que implica que el modelo genera datos de manera completamente independiente. Sin embargo, en el modelo CGAN, se incorpora información adicional para condicionar el proceso de generación de datos. Esta información adicional permite dirigir el proceso de generación hacia un objetivo específico o un conjunto particular de características deseables en los datos generados. En resumen, mientras que el modelo GAN no condicionado genera datos de manera indiscriminada, el modelo CGAN proporciona un mayor control sobre el proceso de generación al permitir la incorporación de información adicional para guiar la producción de datos.

3.2.2.iv. Bayes Optimal Equalized Odds Predictor (BOEO)

Después de explorar las aplicaciones y capacidades de las Conditional Generative Adversarial Networks (cGAN) en la generación de datos sintéticos para mejorar la diversidad y calidad de los conjuntos de datos, surge la necesidad de abordar la equidad en el proceso de modelado predictivo. Una técnica que se centra en esta equidad es el Bayes Optimal Equalized Odds Predictor (BOEO), el cual busca garantizar que las predicciones de los modelos sean equitativas y justas para todos los grupos, independientemente de sus características protegidas.

La idea central del predictor Bayes Optimal Equalized Odds (BOEO) radica en su enfoque de equidad al abordar la disparidad en las predicciones de un modelo predictivo entre grupos protegidos y no protegidos. Para lograr esto, el BOEO sigue un enfoque fundamentalmente diferente al entrenamiento convencional de modelos predictivos.

En lugar de entrenar un solo modelo predictivo para todo el conjunto de datos, el BOEO opta por entrenar modelos predictivos separados para cada grupo protegido y no protegido. Esto significa que se entrena un modelo para el grupo privilegiado y otro para el grupo no privilegiado, reconociendo y tratando las posibles disparidades inherentes a cada grupo.

Una vez entrenados los modelos predictivos separados, el BOEO procede a ajustar las predicciones de cada modelo de tal manera que las tasas de error condicionales sean iguales para ambos grupos. En otras palabras, busca equilibrar las tasas de error condicionales entre el grupo protegido y el no protegido, asegurando que las predicciones del modelo sean justas e imparciales independientemente del grupo al que pertenezca el individuo.

Este ajuste se realiza mediante la optimización de las predicciones basadas en el teorema de Bayes. El teorema de Bayes es un concepto fundamental en la teoría de la probabilidad que establece cómo actualizar nuestras creencias sobre la ocurrencia de un evento en función de la evidencia observada. Matemáticamente, se expresa como:

$$P(B) = \frac{P(B|A)P(A)}{P(B)}$$

Donde:

$P(B)$ supone la probabilidad de que el evento A ocurra dado que el evento B ha ocurrido,

$P(B|A)$ es la probabilidad de que el evento B ocurra dado que el evento A ha ocurrido,

$P(A)$ y $P(B)$ son las probabilidades marginales de los eventos A y B respectivamente.

En el contexto del predictor Bayes Optimal Equalized Odds (BOEO), este marco probabilístico se utiliza para ajustar las predicciones del modelo de manera que reflejen de manera equitativa las probabilidades condicionales de error para ambos grupos protegidos y no protegidos. Esto implica que, al considerar la evidencia observada (es decir, los datos del conjunto de entrenamiento), el BOEO ajusta las predicciones del modelo de manera que las tasas de error condicionales sean iguales para ambos grupos.

En otras palabras, el BOEO busca equilibrar las probabilidades de cometer errores para cada grupo, lo que garantiza que las predicciones del modelo sean justas e imparciales independientemente del grupo al que pertenezca el individuo. Este proceso de ajuste se realiza iterativamente durante el entrenamiento del modelo, utilizando el teorema de Bayes como un marco probabilístico para guiar la optimización de las predicciones del modelo hacia un estado equitativo y justo para todos los grupos.

3.2.3. Postprocesamiento

Después de explorar las técnicas de in-processing para abordar la equidad en el aprendizaje automático, es fundamental considerar también las estrategias de post-procesamiento, que ofrecen otra perspectiva para mitigar los sesgos y garantizar la equidad en los modelos predictivos. Mientras que las técnicas de in-processing se centran en modificar el proceso de entrenamiento del modelo para incorporar la equidad desde el principio, las estrategias de post-procesamiento intervienen después de que el modelo ya ha sido entrenado, ajustando sus predicciones para lograr resultados más equitativos.

Para mostrar los efectos del post-procesamiento se ha utilizado la técnica de *Equalized Odds Postprocessing* (Procesamiento de Equidad en los Odds). La idea principal detrás de esta técnica es garantizar que las tasas de error condicionales sean comparables entre los grupos protegidos y no protegidos, lo que promueve la equidad en las decisiones del modelo. Esta iniciativa es similar a la que veíamos en el Bayes Optimal Equalized Odds (BOEO) sin embargo, La principal diferencia radica en el momento en que se aplican estas técnicas y en su enfoque. Mientras que el BOEO es un enfoque de in-process que ajusta la clasificación directamente durante el entrenamiento del modelo, *Equalized Odds Postprocessing* es una técnica de post-procesamiento que se aplica después de que el modelo ya ha sido entrenado. Equalized Odds Postprocessing ajusta las predicciones del modelo sin modificar los parámetros internos del modelo, lo que lo hace más flexible y aplicable a una variedad de modelos pre-entrenados. Es decir, aunque comparten el objetivo de garantizar la equidad en las predicciones del modelo, Equalized Odds Postprocessing y BOEO difieren en su implementación y momento de aplicación.

Si se analiza en mayor profundidad esta técnica tiene varios pasos. Primero, se comienzan con las predicciones del modelo sobre el conjunto de datos de prueba. Luego, se identifican las instancias en el conjunto de datos que pertenecen a grupos protegidos y no protegidos, basados en atributos como género, raza u otras características sensibles.

Después, se calculan las tasas de error condicionales para cada grupo protegido y no protegido. Estas tasas de error representan la proporción de predicciones incorrectas en cada grupo en relación con las instancias que deberían haber sido clasificadas correctamente.

El siguiente paso implica ajustar las predicciones del modelo para igualar las tasas de error entre los grupos protegidos y no protegidos. Si la tasa de error en un grupo protegido es mayor que en el grupo no protegido, se modifican las predicciones para reducir las predicciones incorrectas en el grupo protegido y viceversa. El objetivo es garantizar que las tasas de error condicionales sean comparables entre ambos grupos.

Finalmente, se evalúa el modelo ajustado para verificar si se han logrado las metas de equidad esperadas. Esto implica examinar las tasas de error condicionales después del ajuste y asegurarse de que sean comparables entre los grupos protegidos y no protegidos.

4. APLICACIÓN PRÁCTICA DE LA METODOLOGÍA PROPUESTA

4.1. Datos empleados

La puesta en práctica de las técnicas mencionadas en el apartado anterior se va a hacer a partir de una base de datos que contiene más de medio millón de pólizas (696.788) y cuya variable de interés es la frecuencia, es decir, el número de siniestros ponderado por la exposición de cada una de las pólizas de la base de datos. Los datos recopilados contienen información sobre las pólizas de cobertura obligatoria de Responsabilidad Civil, que abarcan tanto daños materiales como daños personales. El periodo de exposición de la base de datos corresponde al año 2019. Se ha destinado los primeros diez meses para entrenar el modelo, mientras que los dos últimos meses se reservaron para evaluar el rendimiento del modelo en datos previamente no observados.
Las variables predictoras que conforman el resto de la base de datos y que van a ayudar a predecir la frecuencia del número de siniestros se presentan a continuación:

- Tipo de carrocería: deportivo, convencional, monovolumen, todo terreno o derivado del turismo.
- Tipo de agente intermediario
- Antigüedad de la póliza: indica la antigüedad del contrato representados en valores absolutos.
- Estado de la póliza (anulación, sumplemento…)
- Actividad: uso del vehículo que puede ser clasificarse en la siguientes categorias: particular, actividades comerciales, reparto, autoescuela o renting.
- Ámbito en el que se hace el uso del coche: urbano, nacional, regional, UE o recintos portuarios.
- Antigüedad del vehículo: años transcurridos desde la fabricación del vehículo.
- Bonus-malus del cliente: variable que interacciona el número de años asegurado y el número de siniestros RC reportados.
- Cilindrara del coche

- Tipo de combustible: eléctrico, etanol, diesel, híbrido enchufable o no enchufable.
- Edad del asegurado
- Forma de pago
- Uso del vehículo: público o privado.
- Ocasional: indica si existe un conductor ocasional en la póliza
- Plazas: número de plazas del vehículo
- Potencia del vehículo
- Score crediticio:
- Género: mujer ó hombre.
- Tipo de vehículo: turismo, derivado de turismo, remolque, semirremolque o ciclo motor.
- Suma asegurada (valor del vehículo)
- Zona: provincia dentro del territorio nacional por la que circula el vehículo.
- Marca del vehículo
- Modelo del vehículo

4.2. Resultados de técnicas del reducción del sesgo

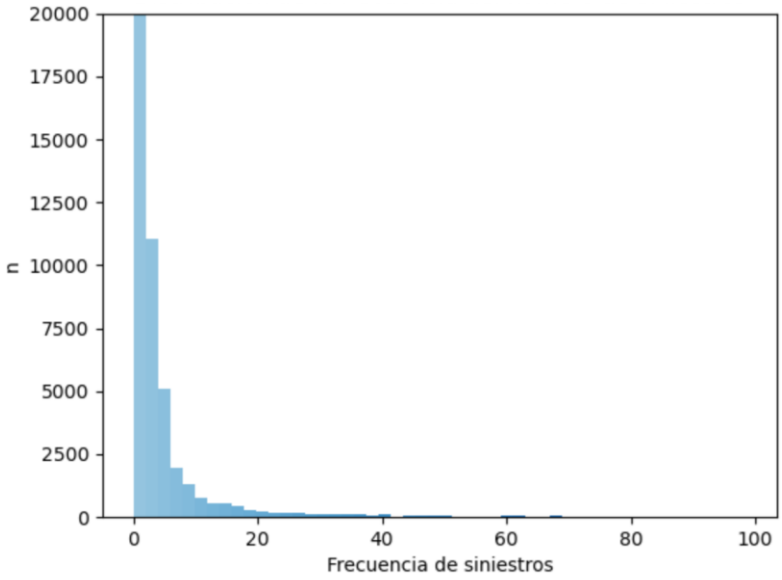
Después de haber examinado detenidamente las definiciones y conceptos relevantes relacionados con la diferenciación en el proceso de tarificación, se plantea la exploración de diferentes técnicas que permitan mitigar el sesgo relacionado con la variable protegida manteniendo al mismo tiempo la capacidad predictiva del modelo. Se han seleccionado dos casos prácticos para ilustrar este enfoque. En el primer ejemplo, se incluye el género como variable protegida debido a su relevancia y aplicación discutida en el sector. Se busca comprobar si simplemente eliminar esta variable es efectivo para eliminar el sesgo. En el segundo ejemplo, se adopta la antigüedad de la póliza (ya sea cartera o nueva producción) como variable protegida, abordando temas tan actuales como el *dual pricing*, que se discutió en detalle en la sección anterior.

4.2.1.Metodología predictiva

4.2.1.1.Análisis univariante

Antes de aplicar cualquier técnica, se realizó un análisis exploratorio inicial para identificar posibles problemas en los datos, como la presencia de valores faltantes o atípicos. Además, se examinó detalladamente la distribución de la variable respuesta. Este paso es crucial, ya que necesitamos comprender cómo se distribuyen los datos para asegurarnos de que la información sea correcta, especialmente dado que se utilizarán modelos lineales generalizados, los cuales son paramétricos y requieren conocer la distribución de la variable. La representación visual de esta variable se muestra en la siguiente figura:

Figura 4. Distribución de la frecuencia del número de siniestros

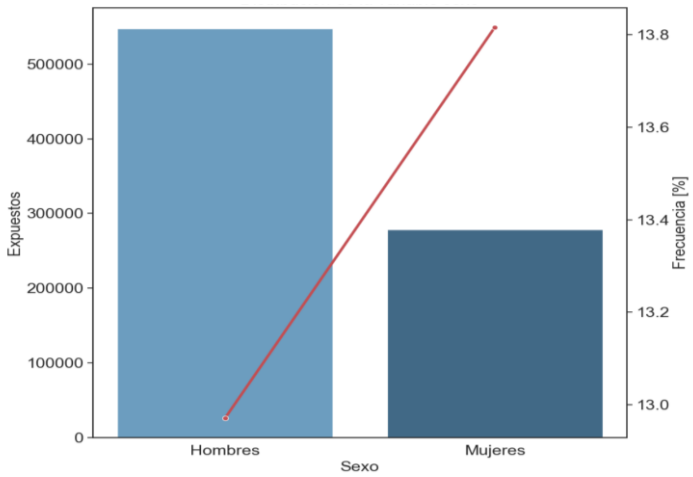


Fuente: elaboración propia

La hipótesis de que la frecuencia de siniestros sigue una distribución de Poisson es una suposición generalmente aceptada. Con base en el gráfico presentado, la distribución de la frecuencia del número de siniestros parece coincidir con una distribución de Poisson sobredispersa caracterizada por el parámetro λ. Sin embargo, para confirmar esta observación de manera más rigurosa, se aplica el test de bondad de ajuste de Kolmogorov-Smirnov. Esta prueba, que es no paramétrica, sirve para determinar si los datos de la muestra concuerdan con una distribución teórica específica. En este escenario, el test de Kolmogorov-Smirnov nos ayuda a verificar si los datos recopilados cumplen con la distribución de Poisson. Los resultados obtenidos a través de este test validan la distribución inicialmente asumida.

Para conocer la distribución de los grupos de cada una de las variables protegidas se presentas las siguientes figuras:

Figura 5. Distribución de la variable sexo

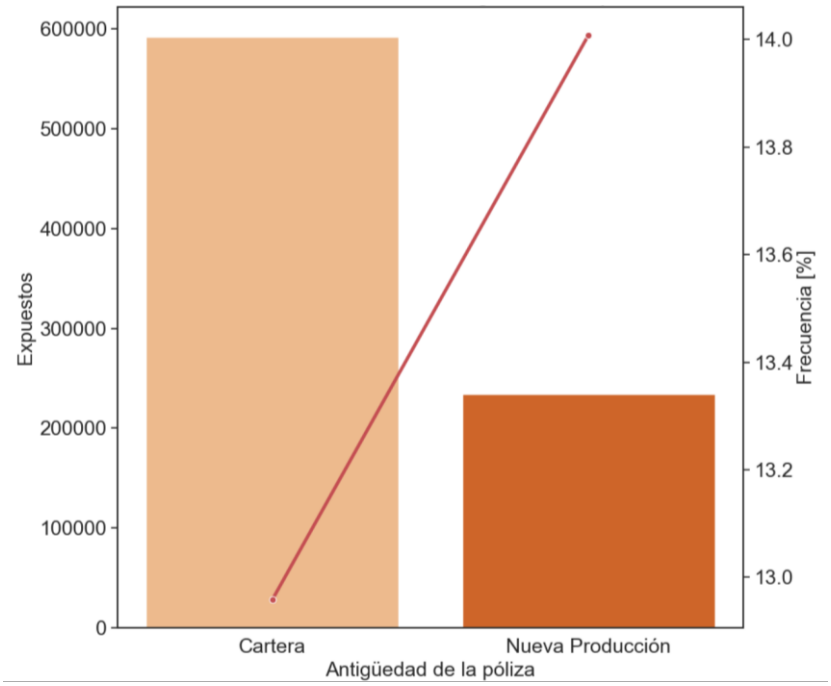


Fuente: elaboración propia

Como se puede ver en la figura anterior, se observa un desequilibrio importante en la distribución de la variable género donde la proporción de hombres sobrepasa considerablemente a la de las mujeres. En cuanto a la frecuencia media del número de siniestros encontramos que la frecuencia en los hombres es algo inferior a la de las mujeres, 12.07% frente al 12.88% en mujeres. En cuanto a la exposición concluimos que la clase minoritaria o sensible la forman las mujeres y la clase mayoritaria los hombres.

Por otro lado, si se representa la distribución de la variable antigüedad de póliza (o negocio), que sería la segunda variable protegida a analizar, se obtiene el siguiente gráfico:

Figura 6. Distribución de la variable antigüedad de póliza



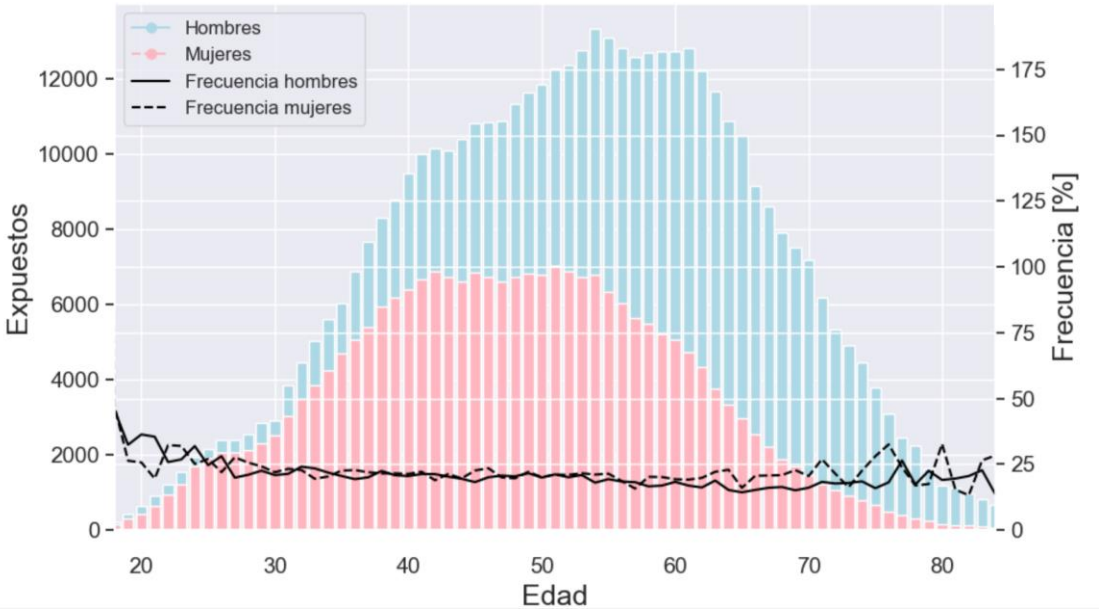
Fuente: elaboración propia

En la figura anterior, como era de esperar, se puede apreciar que el número de expuestos en cartera es muy superior al nuevo negocio. Lo contrario ocurre con la variable frecuencia que es superior en la nueva producción donde alcanza un porcentaje del 14% mientras que en cartera este porcentaje disminuye al 12.95%.

4.2.1.2.Análisis bivariante

Tras haber identificado el desequilibrio en la variable respuesta, es relevante explorar más a fondo la dinámica entre género y la edad para conocer algo más la muestra que se está tratando en este estudio. Para ello se presenta el siguiente histograma:

Figura 7. Distribución de la edad por sexo y frecuencia de siniestros



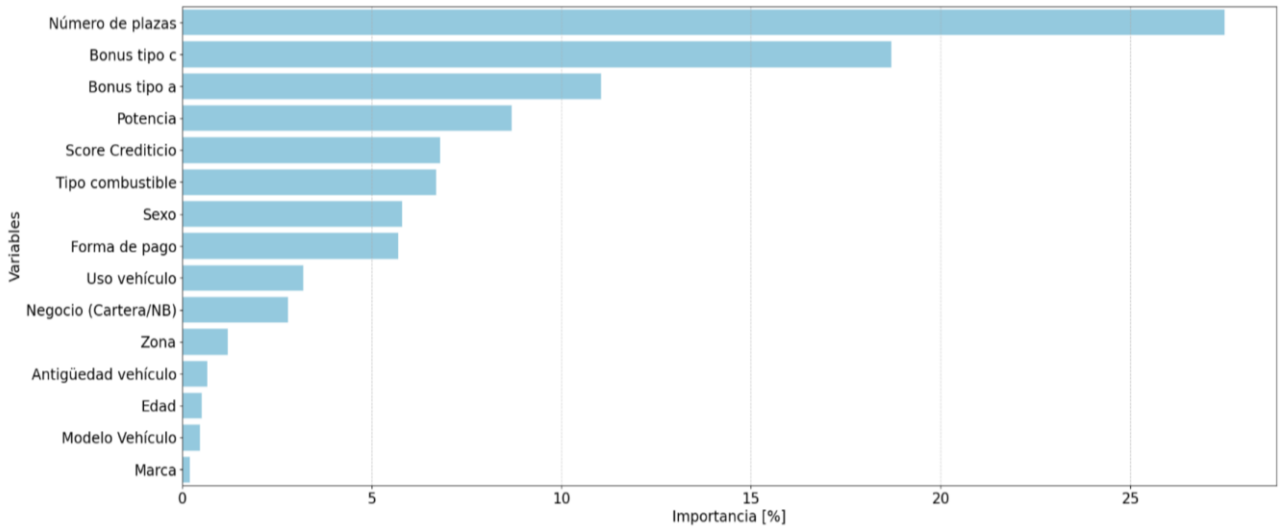
Fuente: elaboración propia

Como se puede apreciar en la figura anterior, tanto la distribución de los hombres como la de las mujeres son bimodales, mostrando modas alrededor de los 40 y 52 años para las mujeres, y alrededor de los 45 y 55 años para los hombres. Sin embargo, cabe destacar que la distribución de la frecuencia media de siniestros difiere entre ambos grupos. En general, la frecuencia media de siniestros tiende a ser menor en mujeres en comparación con hombres, a lo largo de la mayoría de los tramos de edad. No obstante, es interesante observar que estas curvas comienzan a igualarse a partir de los 58 años.

4.2.1.3. Selección de variables

Para identificar las variables más relevantes para el modelo que se va a constriu inicialmente, y que va a servir de base para su comparación con el resto de técnicas, se ha empleado el algoritmo que se ha mencionado en el marco teórico cuya base es *Gradient Boosting Machine (GBM).* Esta técnica se describe en mayor detalle en la sección 1.3. Como resultado de aplicar dicha técnica se obtiene el siguiente gráfico:

Figura 8. Importancia de las variables



Fuente: elaboración propia

Las variables más importantes y que compondrán los modelos son: número de plazas, zona, edad, bonus tipo c, bonus tipo a, edad, potencia del vehículo, sexo y negocio.

En esta sección se examinarán los resultados obtenidos de aplicar las técnicas mencionadas en la sección del marco teórico. Para ello, se evaluarán el sesgo así como el poder productivo con las métricas descritas anteriormente (Disparate Impact Rate y MAE) para cada una de las fases propuestas: pre-procesamiento, procesamiento y post-procesamiento.

4.2.2. Técnicas de reducción del sesgo

4.2.2.1.Pre-procesamiento

4.2.2.1.i.Omisión de la variable protegida (*unawarness principle*)

La técnica de eliminación de la variable protegida supone el primer paso en la fase de pre-procesamiento. Con ello se quiere verificar si es suficiente este paso para eliminar el sesgo del modelo.

A continuación, se presenta la tabla de resultados después de aplicar esa técnica.

Tabla 1. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Modelo sin la variable género	24.15%	1.007

Fuente: elaboración propia

En la tabla anterior se puede ver que al eliminar la variable protegida del modelo se consigue reducir el sesgo pero no se erradica completamente. En cuanto al poder predictivo, el modelo predictivo es ligeramente peor que el modelo inicial ya que el valor del MAE es ligeramente superior.

Tabla 2. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
Modelo sin la variable negocio	23.97%	1.005

Fuente: elaboración propia

En el segundo ejemplo donde la variable protegida es el negocio, las conclusiones a las que se llegan son muy similares a las obtenidas en el ejemplo inmediatamente superior, la eliminación de la variable protegida no supone una erradicación del sesgo.

4.2.2.1.ii.Correlación mediante coeficientes

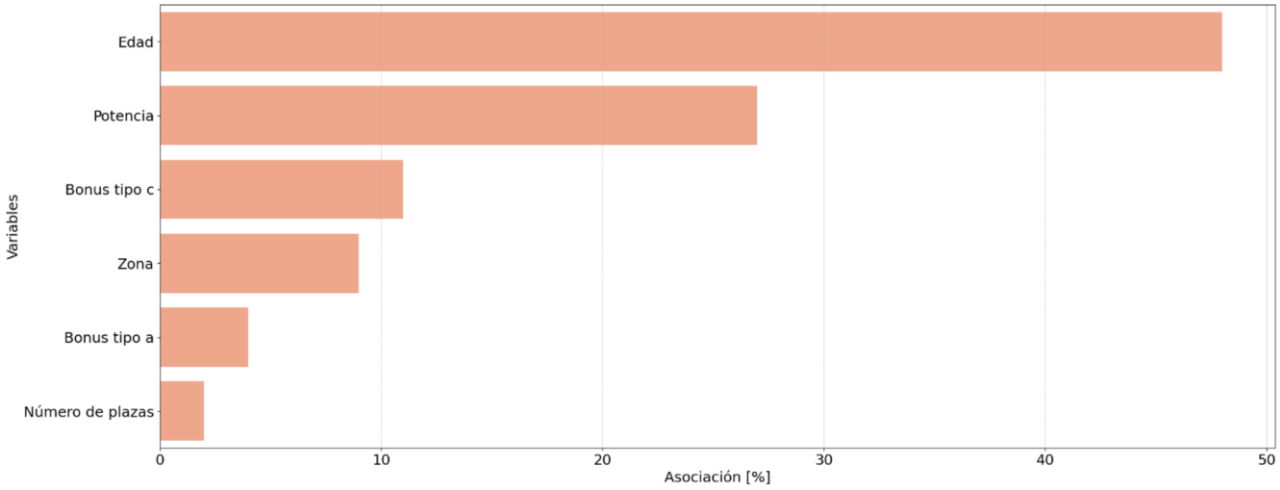
Puesto que en el apartado anterior se ha demostrado que la eliminación de la variable protegida no es suficiente para erradicar el sesgo del modelo, con esta técnica se propone no sólo eliminar la variable protegida sino también aquellas variable asociadas con la misma para evitar que estas actúen como sustitutos de la misma.

Como resultado de aplicar los coeficientes de Pearson y Cramer , según corresponda como se ha detallado en la sección anterior, se ha obtenido el valor del estadístico, que representa la fuerza de asociación de cada variable con la variable protegida así como el p-valor, indicando la probabilidad de que la asociación observada sea resultado del azar. Para facilitar la interpretación de estos estadísticos se ha elaborado un gráfico que expresa la asociación de la variable en forma de porcentaje. A continuación se presentan los resultados de dichos gráficos.

- Resultados para la variable protegida “Sexo”:

Al notar un salto significativo en los valores del estadístico entre las seis primeras variables y el resto de variables, se ha decidido decide considerar únicamente las siguientes seis variables en el siguiente gráfico:

Figura 9. Porcentaje de asociación de las variables con la variable protegida “sexo”



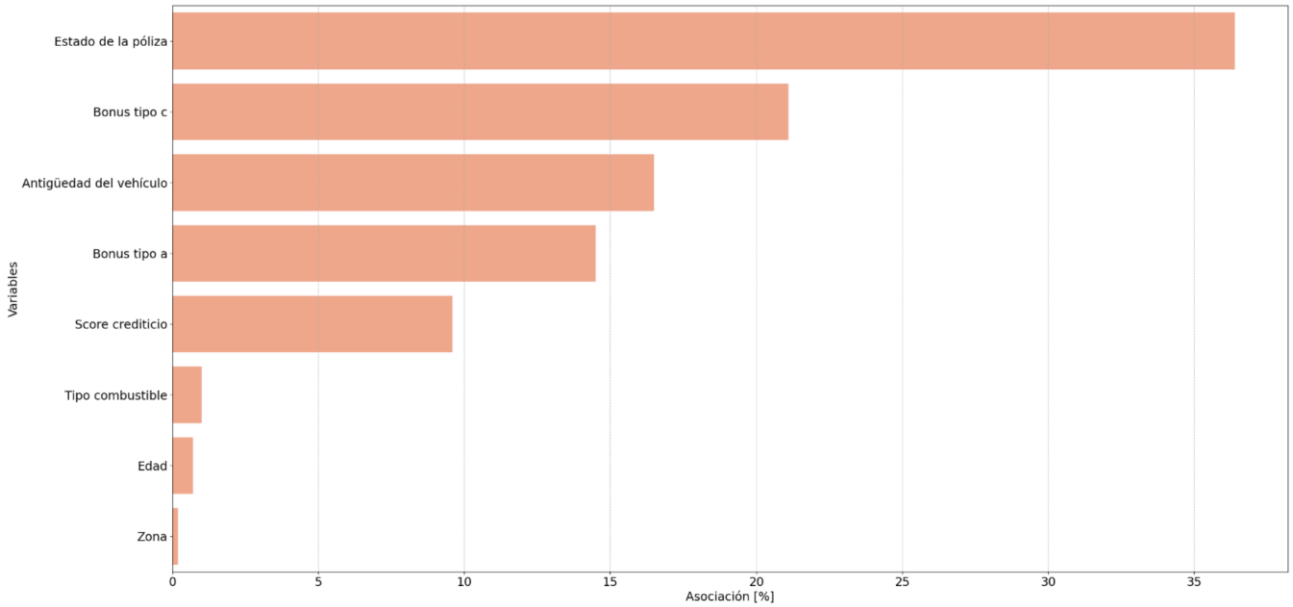
Fuente: elaboración propia

En el gráfico anterior, se observa un salto significativo entre las dos primeras variables y el resto de variables. Las variables más correlacionadas con la variable sexo y que por tanto se van a eliminar del modelo son la edad y la potencia.

- Resultados para la variable protegida “Negocio”:

Para el caso de esta variable se ha seguido el mismo procedimiento, graficando únicamente aquellas variables con valores significativos del estadístico. Como resultado se obtiene el siguiente gráfico.

Figura 10.Porcentaje de asociación de las variables con la variable protegida “negocio"



Fuente: elaboración propia

El gráfico muestra que las variables más correlacionadas con el negocio son el estado de la póliza y el bonus-malus de tipo c. Se eliminarán ambas variables para ver su efecto en el sesgo.

Con esta acción los resultados que se obtiene se encuentran en la siguiente tabla:

Tabla 3. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Modelo sin la variable género ni las variables correlacionadas	31.15%	1.000

Fuente: elaboración propia

Al eliminar las variables correlación se consigue un valor de la métrica del sesgo igual a 1 lo que supone que se ha eliminado completamente el sesgo del modelo. Sin embargo, esta mejora en el sesgo conlleva un deterioro significativo en la capacidad predictiva del modelo.

Tabla 4. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
Modelo sin la variable negocio ni las variables correlacionadas	30.67%	1.000

Fuente: elaboración propia

En la tabla inmediatamente anterior, de nuevo se puede apreciar que con esta técnica se consigue erradicar el sesgo pero el poder predictivo aumenta considerablemente poniendo en duda el uso de esta técnica.

4.2.2.1.iii.Correlación no lineal (dependency analysis)

Esta técnica es una alternativa al método anterior que explora las relaciones no lineales de la variable protegida en cuestión y el resto de variables. En el marco teórico se ha propuesto una metodología que consiste en predecir la variable protegida utilizando el resto de variables como variables explicativas mediante el uso de un modelo de *Gradient Boosting Machine (GBM).* El resultado de este algoritmo. De esta forma, si una variable explicativa muestra una alta importancia en la predicción de la variable objetivo, esto indica una dependencia con la misma.

Para examinar los resultados, se ha impuesto un punto de corte en la importancia de las variables de un 10% y poder ver así las variables más importantes resumidas en las siguientes tablas:

Tabla 5. Resultados para la variable protegida "sexo"

Variable respuesta	Variable explicatoria	Importancia de la variable (%)
Género	Edad	42%
Género	Modelo	27%
Género	Potencia	13%

Fuente: elaboración propia

La metodología ha identificado que las variables más útiles para predecir el género son la Edad, el Modelo y la Potencia.

En el caso donde la variable protegida es el negocio los resultados fueron los siguientes:

Tabla 6. Resultados para la variable protegida "negocio"

Variable respuesta	Variable explicatoria	Importancia de la variable (%)
Negocio	Bonus de tipo c	71%
Negocio	Bonus de tipo a	26%
Negocio	Zona	8%

Fuente: elaboración propia

Las variables que más afectan la predicción de la variable negocio son el tipo de bonus c y a y la zona de residencia del asegurado.

A continuación, se extraejorn estas variables junto a la variable protegida en cuestión y se compararon con el modelo inicial ofreciendo los siguientes resultados:

Tabla 7. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Modelo sin la variable género ni las variables correlacionadas no linealmente	32.19%	1.0048

Fuente: elaboración propia

El resultado de aplicar esta métrica es una mejora del sesgo en comparación con el modelo original pero de nuevo, ello conlleva un deterioro del poder predictivo del modelo.

Tabla 8. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	0.2397	1.0869

Modelo sin la variable género ni las variables correlacionadas no linealmente	0.3154	1.006

Fuente: elaboración propia

Los resultados de la tabla anterior muestran una notable similitud con los obtenidos mediante el método anterior, que implicaba la captura de coeficientes para evaluar la relación con la variable protegida. Esta consistencia en los hallazgos sugiere que la influencia de estas variables se mantiene coherente en diferentes análisis. Sin embargo, nuevamente se observa que, aunque este método puede eliminar casi por completo el sesgo, no logra mejorar la capacidad predictiva del modelo.

4.2.2.1.iv.Análisis de residuos

Con este enfoque se pretende rescatar aquella parte de las variable explicativas que sirve para predecir la variable respuesta y que no se encuentra correlacionada con la variable protegida. Para ello, se hace uso de los residuos de los modelos donde la variabe explicativa es la más correlacionada con la variable protegida como se ha mencionado en el marco teórico. Los resultados que se han obtenido se ven reflejado en las figuras siguientes:

Tabla 9. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Modelo sin la variable género ni las variables correlacionadas no linealmente + residuos	37.11%	0.9921

Fuente: elaboración propia

La tabla anterior muestra una ligera reducción en el error predictivo del modelo, aunque el ratio para medir el sesgo apenas presenta diferencias en comparación con los métodos anteriores.

Tabla 10. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	0.2397	1.0869
Modelo sin la variable género ni las variables correlacionadas no linealmente + residuos	0.3512	0.9979

Fuente: elaboración propia

Los resultados donde se ha tenido en cuenta el negocio como variable portegida muestran que este método no reduce el sesgo en comparación con los métodos anteriores y, además, empeora el MAE. Esto sugiere que los residuos de las variables correlacionadas con las variables protegidas no tienen una contribución significativa en la predicción de la frecuencia de siniestros.

4.2.2.1.v.Remuestreo (*re-sampling*)

Como se ha comentado anteriormente, se utiliza esta técnica con el fin de equilibrar la proporción d los datos de entrada del modelo con el objetivo de que cada clase esté representada de manera equitativa durante el proceso de entrenamiento del modelo, lo que ayuda a contrarrestar cualquier sesgo introducido por el desbalance de clases en los datos.

Los resultados obtenido tras aplicar esta técnica son los que se muestra a continuación.

Tabla 11. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Modelo con remuestreo	32.53%	0.9951

Fuente: elaboración propia

Al aplicar el remuestreo se puede observar que el valor de la métrica que se obtiene es cercano al modelo insesgado (1) pero que nuevamente esta mejora conlleva un valor del MAE mayor y con ello una pérdida de la capacidad predictiva.

Tabla 12. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
Modelo con remuestreo	33.78%	0.9851

Fuente: elaboración propia

En el caso en el que la variable protegida es el negocio, se aprecia de nuevo que aunque que este método logra una posición muy cercana a la de un modelo sin sesgo, donde la métrica es igual a 1 el poder predictivo del modelo se deteriora considerablemente.

4.2.2.1.vi.Reponderación (*Re-weightning*)

Esta técnica en contraposición con la utilizada anteriormente, como se ha descrito detalladamente en el marco teórico, no realiza una selección al azar de las muestras sino que asigna pesos a las muestras de acuerdo con su clase, garantizando así una representación equitativa de todas las clases durante el entrenamiento del modelo. Al aplicar esta técnica se han obtenido los siguientes resultados.

Tabla 13. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Modelo con reponderación	32.52%	0.9951

Fuente: elaboración propia

Con este método se observa que se consigue un ligero deterioro del poder predictivo, además el sesgo parece inclinarse hacia la clase no protegida (hombres).

Tabla 14. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
Modelo con reponderación	34.03%	0.9814

Fuente: elaboración propia

Este método de resmuestreo, al igual que el método anterior de reponderación, se pueden apreciar valores del Disparate Impact Ratio cercanas a uno pero el poder predictivo medido por el MAE sigue sin ser mejor al modelo base.

Antes de evaluar el rendimiento de las técnicas en la siguiente fase, es útil revisar los resultados de esta primera fase. Esto nos ayudará a identificar el mejor método de las técnicas in-processing y establecer las métricas que necesitamos superar en la próxima fase. A continuación, se presentan las tablas 15 y 16, que muestran las métricas analizadas en cada caso:

Tabla 15. Comparativa resultados pre-procesamiento variable "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Modelo sin la variable género	24.15%	1.007
Modelo sin la variable género ni las variables correlacionadas	31.15%	1.000
Modelo sin la variable género ni las variables correlacionadas no linealmente	32.19%	1.0048
Modelo sin la variable género ni las variables correlacionadas no linealmente + residuos	37.11%	0.9921
Modelo con remuestreo	32.53%	0.9951

Modelo con reponderación	34.03%	0.9814

Fuente: elaboración propia

Tabla 16. Comparativa resultados pre-procesamiento variable "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
Modelo sin la variable negocio	23.97%	1.005
Modelo sin la variable negocio ni las variables correlacionadas	30.67%	1.000
Modelo sin la variable género ni las variables correlacionadas no linealmente	31.54%	1.006
Modelo sin la variable género ni las variables correlacionadas no linealmente + residuos	35.12%	0.9979
Modelo con remuestreo	33.78%	0.9851
Modelo con reponderación	32.52%	0.9951

Fuente: elaboración propia

El método que ha ofrecido los mejores resultados en ambos análisis es aquel que no incluye la variable protegida, logrando un equilibrio entre la reducción del sesgo y la capacidad predictiva. Sin embargo, aún hay margen de mejora, ya que no elimina totalmente el sesgo y el poder predictivo es similar al del modelo inicial.

4.2.2.2.Procesamiento

Tras explorar las técnicas de preprocesamiento de datos, es relevante adentrarse en las técnicas de procesamiento interno (*in-processing*) que complementan y profundizan la capacidad de los modelos para abordar desafíos específicos de los datos. Las técnicas de procesamiento interno se refieren a métodos que se aplican durante el entrenamiento del modelo mismo, lo que les permite adaptarse y mejorar su capacidad de aprender y generalizar patrones en los datos.
En contraste con las técnicas de preprocesamiento, que se aplican antes de que los datos se introduzcan en el modelo, las técnicas de procesamiento interno trabajan directamente con el modelo durante su entrenamiento. En este caso se estudio se van a analizar cuatro técnicas distintas: el Adversarial Debias, la regresión con regularización, el Disparate Impact Remover y finalmente el Bayes Optimal Equalized Odds Predictor.

4.2.2.2.i.Disparate Impact Remover

Esta técnica que está orientada a modificar las probabilidades de predicción para captar la variabilidad de los datos minimizando el sesgo de la variable protegida ofrece los siguientes resultados.

Tabla 17. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Disparate Impact Remover	1.83%	1.0050

Fuente: elaboración propia

Se observa un reducción considerable del error del modelo al mismo tiempo que se consigue controlar el sesgo.

Tabla 18. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
Modelo con reponderación	1.83%	1.0030

Fuente: elaboración propia

En este segundo ejemplo también es visible que los resultados de esta técnica son notorios que los anteriores tanto en términos de capacidad predictiva, con un MAE considerablemente reducido en comparación con el resto de técnicas utilizadas hasta el momento, como en términos de sesgo. Estas mejoras se observan de manera consistente en ambos análisis, lo que respalda la eficacia del método.

4.2.2.2.ii.Regularización: Ridge regression

En el caso de técnica de penalización que utiliza el parámetro lambda (λ) para mejorar la estabilidad del modelo y reducir el sesgo de variables que podrían inducir sesgo, se ha hecho una modificación de su función de pérdida para reducir las disparidades entre los grupos protegidas y los no protegidos de la variable objetivo en cuestión.

El resultado dela aplicación de esta técnica se puede consultar en la siguiente tabla:

Tabla 19. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Regularización	14.49%	0.9126

Fuente: elaboración propia

Se puede ver un aumento considerable del error de predicción del modelo y un sesgo que se inclina cada vez más hacia la clase protegida.

Tabla 20. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
Regularización	14.36%	1.125

Fuente: elaboración propia

Los resultado para la segunda variable protegida negocio, muestran que aunque la técnica es capaz de mejorar el rendimiento del modelo, exhibe valores de sesgo superiores al modelo inicial que incluye la variable. Esta discrepancia puede atribuirse, en parte, al diseño original del método, concebido para reducir el sesgo global del modelo. En este caso de estudio, se ha modificado el método original para adaptarlo a las variables protegidas. Sin embargo, según los resultados presentados en las tablas anteriores, esta adaptación no ha resultado exitosa.

4.2.2.2.iii.Generative Conditional Adversarial Network (cGAN)

Este método de *deep learning* capaz de generar datos sintéticos para equilibrar las muestras y facilitar la transparencia del modelo ha arrojado los siguientes resultados.

Tabla 21. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
GAN	11.76%	1.0000

Fuente: elaboración propia

Esta metodología es capaz de acercarse al ratio objetivo de 1 entre la dos clases aunque ello supone un deterioro en el poder predictivo del modelo.

Tabla 22. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
GAN	12.48%	1.001

Fuente: elaboración propia

En este segundo ejemplo se confirma que hasta el momento, la técnica de Generative Adversarial Networks (GAN) ha demostrado ser la más efectiva en términos de rendimiento, ya que logra eliminar el sesgo y mejora considerablemente el poder predictivo del modelo. En comparación con el modelo inicial, el uso de GAN reduce a la mitad el error de predicción, lo cual representa una mejora sustancial en la capacidad del modelo para generalizar y predecir con precisión nuevos datos. Este avance sugiere que las GAN ofrecen una solución prometedora para abordar el sesgo y mejorar el rendimiento en tareas de modelización.

4.2.2.2.iv.Bayes Optimal Equalized Odds Predictor (BOEO)

Esta técnica es capaz de asegurar que las predicciones sean justas al equilibrar las probabilidades de error entre los grupos. Durante el proceso de entrenamiento hace uso del teorema de Bayes para guiar la optimización hacia la equidad. El resultado de este proceso queda reflejado en las siguientes figuras.

Tabla 23. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
BOEO	32.89%	0.9564

Fuente: elaboración propia

Esta técnica se aleja del ratio 1 y se inclina hacia el sesgo de la clase no protegida. Así mismo, no consigue mejorar el poder predictivo del modelo.

Tabla 24. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
BOEO	33.18%	1.134

Fuente: elaboración propia

La aplicación del método Bayes ofrece resultados contradictorios en relación al sesgo. En el primer análisis, donde la variable protegida es el sexo del asegurado, el Disparate Impact Ratio es menor a uno, lo que indica un sesgo hacia la clase mayoritaria (hombres). Por otro lado, en el segundo análisis, el valor de esta métrica es superior a 1, mostrando un sesgo hacia la clase minoritaria (el negocio de la nueva producción).

En cuanto al poder predictivo del modelo, se observa que es considerablemente inferior al del modelo inicial en ambos análisis. Esta disminución en el poder predictivo sugiere que la aplicación del método Bayes no mejora la capacidad del modelo para generalizar y predecir con precisión nuevos datos, lo que invita a descartar el uso de esta técnica en este contexto específico.

A continuación de ofrece un resumen de esta segunda fase para cada unos de los análisis con el objetivo de establecer una comprensión clara de las técnicas utilizadas.

Tabla 25. Comparativa resultados procesamiento variable "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Disparate Impact Remover	1.83%	1.0050
Regularización	14.49%	0.9126
GAN	11.76%	1.0000
BOEO	32.89%	0.9564

Fuente: elaboración propia

Tabla 26. Comparativa resultados procesamiento variable "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
Disparate Impact Remover	1.83%	1.0030
Regulzarización	14.36%	1.125
GAN	12.48%	1.001
BOEO	33.18%	1.134

Fuente: elaboración propia

Los resultados obtenidos en esta fase representan una mejora considerable con respecto a los obtenidos en la primera fase. Se observa que el método que ofrece una mejora sustancial tanto en el sesgo como en el poder predictivo es el Generative Adversarial Networks (GAN). Además, la técnica de Disparate Impact Remover arroja resultados prometedores en términos de capacidad predictiva, aunque no logra eliminar completamente el sesgo.

Por otro lado, se esperaba que la técnica de regularización proporcionara mejores resultados. Sin embargo, ambos análisis confirman que esta técnica no logra reducir el sesgo a un nivel cercano a 1, lo que desalienta su uso en este contexto.

4.2.2.3. Postprocesamiento

En esta última fase se busca ajustar las predicciones una vez el modelo ha sido entrenado. La técnica de *Equalized Odds Postprocessing*garantiza que las tasas de error condicionales sean similares entre grupos protegidos y no protegidos induciendo así la equidad sin altera kis parámetros internos de los modelos.

A continuación se presentan los resultados obtenidos con esta técnica:

Tabla 27. Resultados para la variable protegida "sexo"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	24.08%	1.0699
Postprocessing	1.38%	0.9480

Fuente: elaboración propia

Esta técnica ofrece una disminución del error predictivo del modelo muy notable pero no consigue erradicar el sesgo del modelo.

Tabla 28. Resultados para la variable protegida "negocio"

Método	Poder predictivo (MAE)	Sesgo (DIR)
Modelo inicial con todas las variables	23.97%	1.0869
Postprocessing	9.8%	1.030

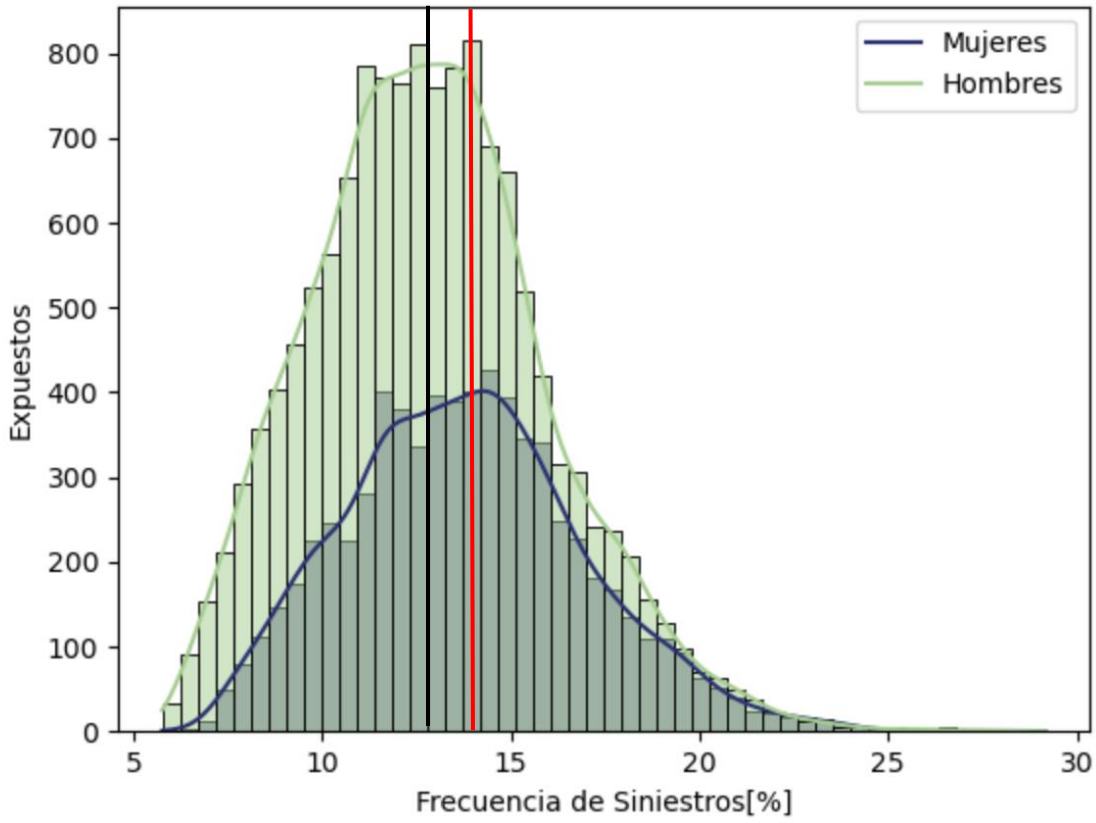
Fuente: elaboración propia

En este segundo ejemplo, se observa de nuevo que se ha logrado reducir notablemente el error predictivo del modelo, aunque no se ha logrado controlar el sesgo de la misma manera que en otros métodos utilizados.

4.2.3. Conclusiones de la aplicación práctica

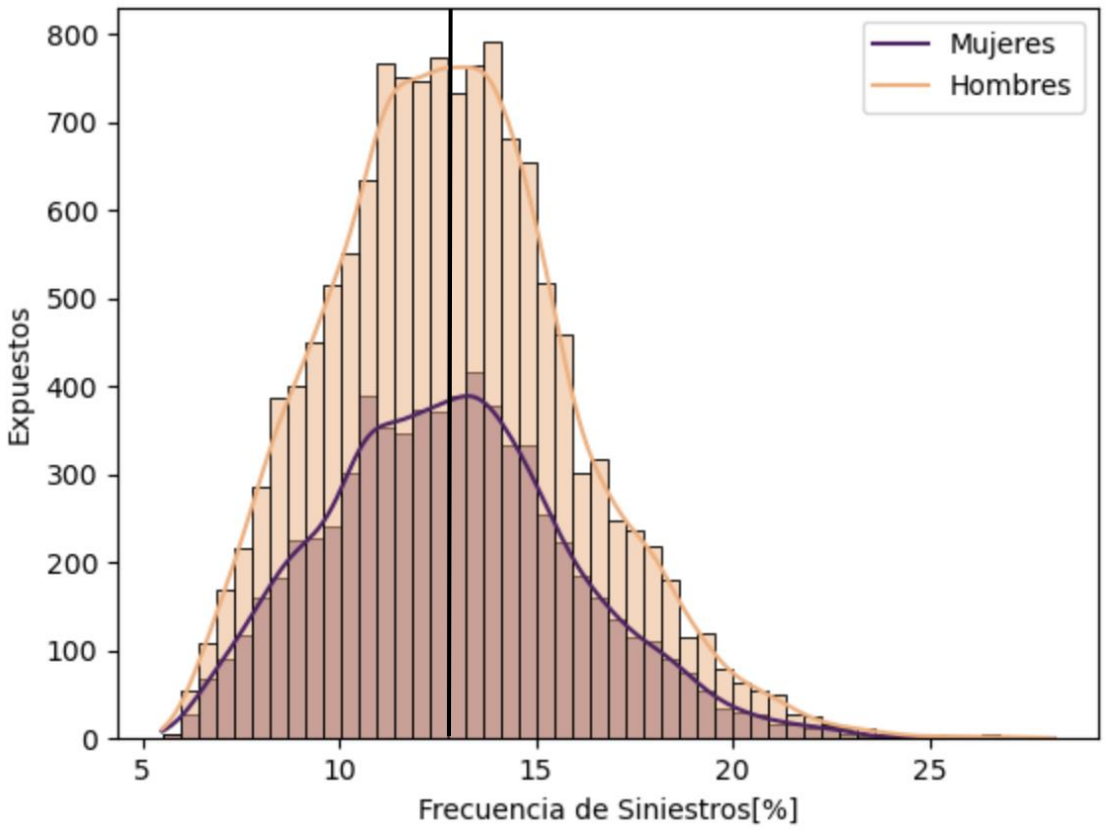
Después de explorar todas las técnicas definidas en secciones anteriores, se llega a la conclusión de que la metodología más adecuada para cumplir con el objetivo inicial de eliminar el sesgo y mantener o mejorar el poder predictivo del modelo es la técnica cGAN. En cuanto al sesgo, la métrica del *Disparate Impact Ratio* proporciona una indicación del sesgo, pero resulta útil complementar esta medida con un análisis visual de las distribuciones. Para ello, es recomendable analizar la distribución de la frecuencia de siniestros predicha en la variable protegida, como el género, comparando el modelo base con el modelo más efectivo, que en este caso es el cGAN. A continuación, se presentan dos gráficos que ilustran estas distribuciones.

Figura 11. Distribución de las predicciones para la variable protegida "sexo" utilizando el modelo base



Fuente: elaboración propia

Figura 12. Distribución de las predicciones para la variable protegida "sexo" utilizando la técnica cGAN



Fuente: elaboración propia

A través de los gráficos anteriores es posible observar que la distribución de las predicciones para las mujeres en el primer gráfico, en el que se ha usado el modelo base, es distinta a la distribución de los hombres presentando una media ligeramente superior (14.8 frente a 13.2) mientras que en el segundo gráfico las medias son prácticamente la misma. Estos gráficos son una prueba más de la eficacia del modelo cGAN para mitigar el sesgo en las variables protegidas.

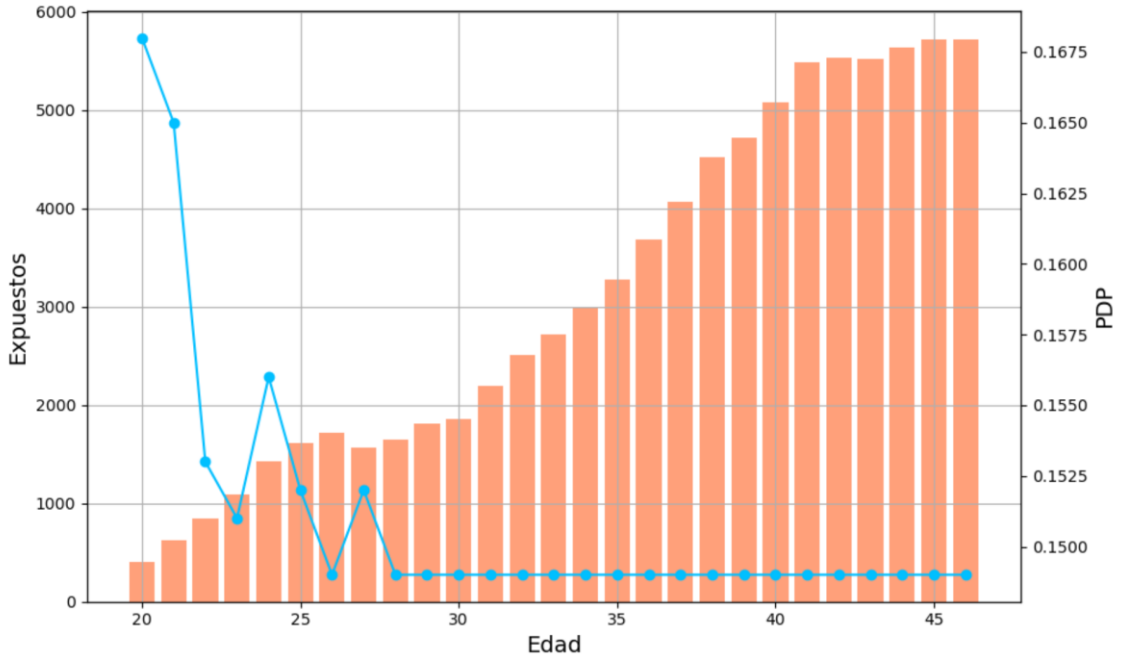
4.2.4. Limitaciones

La conclusión del apartado anterior es la idoneidad del algoritmo cGAN para mitigar el sesgo y mejorar el poder predictivo. Sin embargo, cuando se examinan las técnicas de Deep Learning como es el caso de la técnica cGAN, emerge una preocupación fundamental: la falta de interpretabilidad o explicabilidad inherente a estas herramientas de aprendizaje automático. La importancia de tratar la interpretabilidad en los modelos predictivos radica en su condición de componente esencial para garantizar la justicia algorítmica y la transparencia en el uso de la inteligencia artificial. La justicia algorítmica se enfoca en asegurar que los modelos predictivos sean justos y no discriminativos, mientras que la transparencia implica que los procesos y resultados de los modelos sean accesibles y comprensibles para todos los interesados. La interpretabilidad, en este contexto, se refiere a la capacidad de explicar claramente cómo funciona un modelo, lo que permite a los usuarios entender y verificar si el modelo está haciendo predicciones justas y equitativas. Esto es crucial para mantener la confianza en los sistemas de IA y evitar decisiones injustas o discriminatorias.

La falta de claridad sobre cómo funcionan los modelos predictivos puede conducir a malentendidos y decisiones basadas en información incorrecta. En respuesta a este desafío, ha surgido un campo en constante crecimiento dentro del ámbito del aprendizaje automático, conocido como Explicabilidad o Interpretabilidad de modelos. El propósito primordial de este campo es desarrollar herramientas que permitan examinar y comprender los modelos de "caja negra", con el fin de entender mejor el proceso de toma de decisiones y detectar posibles sesgos o deficiencias en los modelos de aprendizaje automático.

Uno de los métodos más ampliamente empleados en este contexto son los Gráficos de Dependencia Parcial (PDP por sus siglas en inglés). Estos gráficos están diseñados para ilustrar los efectos marginales de cada variable sobre las predicciones del modelo, lo que permite abordar preguntas específicas, como los cambios en la predicción cuando varía el valor de una variable, y si esta variación conlleva un aumento o una disminución en la predicción. La elaboración de estos gráficos requiere fijar el valor de la variable en cuestión al mismo nivel para todos los registros, calcular las predicciones y obtener un promedio ponderado. Repitiendo este proceso para todos los valores relevantes, se puede observar el efecto marginal de esa variable. Para entender mejor la interpretación del gráfico de dependencia parcial, se puede tomar como ejemplo la variable edad, centrándose en el rango que abarca la transición de conductor joven a conductor de mediana edad, momento en el cual se observan cambios significativos en la frecuencia.

Figura 13. Gráfico parcial de dependencia



Fuente: elaboración propia

En el gráfico previo se observa una reducción gradual en la frecuencia predicha a medida que el conductor progresa hacia la edad adulta.

Además de esta técnica visual, existe la posibilidad de realizar un estudio más profundo en las predicciones realizadas por el modelo. Con las últimas avances en el campo de la explicabilidad de modelos, es posible investigar cómo el modelo genera predicciones para situaciones particulares. Este tipo de análisis en profundidad de las predicciones dentro del marco del aprendizaje automático es crucial para entender la capacidad de explicación de los modelos. Mediante el empleo de métodos avanzados y herramientas especializadas, tenemos la posibilidad de estudiar cómo los modelos de aprendizaje automático deciden en casos individuales.

En términos simples, este análisis consiste en elegir un caso específico dentro de los datos, como un cliente de seguro, y examinar exhaustivamente toda la información relevante asociada a dicho cliente. Esto abarca aspectos como la edad, el género, el historial de seguros y otros elementos significativos que fueron tenidos en cuenta durante el entrenamiento del modelo.

Una vez que hemos escogido este caso particular, podemos observar cómo el modelo emite predicciones para este individuo en particular. Partimos de un punto de partida, generalmente la probabilidad media de que ocurra el evento que el modelo está pronosticando, calculada teniendo en cuenta las características medias de todos los clientes en el conjunto de datos.

A medida que incorporamos variables específicas de ese individuo en el modelo, podemos apreciar cómo cambia la predicción del modelo. Por ejemplo, si estamos estimando la probabilidad de que este individuo tenga un accidente de coche en el próximo año, podemos analizar cómo esa probabilidad fluctúa a medida que añadimos

información adicional, como el historial de conducción o la cantidad de kilómetros recorridos semanalmente.

Este análisis ofrece una visión detallada del efecto que cada característica individual tiene en la probabilidad prevista del evento. Nos ayuda a comprender mejor cómo el modelo utiliza cada característica para tomar decisiones y emitir predicciones. Además, este procedimiento nos permite detectar posibles sesgos o fallos en el modelo, lo que contribuye a la transparencia y confiabilidad de los modelos de aprendizaje automático.

5. CONCLUSIONES

La evolución del rol del actuario en la era actual refleja un complejo proceso de adaptación a nuevos avances metodológicos y tecnológicos. Durante la fase inicial de este estudio, se han exploraron diversas definiciones éticas que se consolidan a través de las regulaciones impuestas por los organismos supervisores. Este contexto subraya la importancia ética del actuario, requiriendo no solo la adhesión a definiciones éticas y el cumplimiento de regulaciones gubernamentales sino también un compromiso intrínseco con la integridad y la transparencia en todas sus actividades profesionales.

Una de las cuestiones fundamentales en este proceso radica en determinar si es viable mitigar el sesgo asociado a ciertas variables protegidas sin sacrificar la precisión y la eficacia de los modelos predictivos empleados en la práctica actuarial. Tras un meticuloso análisis, en el cual se han examinado diversas técnicas en cada etapa de un proceso de modelización habitual —incluyendo el preprocesamiento, el procesamiento y el postprocesamiento— se ha concluido que es factible alcanzar dicho equilibrio.

Durante el desarrollo de este estudio, se ha demostrado que la eliminación de las variables protegidas del modelo, aunque comúnmente practicada, no es suficiente para eliminar completamente el sesgo en los modelos predictivos. Esta constatación llevó a la búsqueda de soluciones alternativas que pudieran mitigar el sesgo sin comprometer la precisión y eficacia de los modelos. La exploración y la implementación de técnicas y enfoques innovadores han revelado un conjunto de herramientas efectivas que no solo logran reducir el sesgo, sino que también mejoran significativamente la capacidad predictiva de los modelos. Estas técnicas, aplicadas durante la fase de procesamiento del modelo y específicamente para los datos frecuencia de siniestros de la cobertura de Responsabilidad Civil en el seguro de automóviles, han demostrado ser particularmente efectivas, destacándose las técnicas de Disparate Impact Remover y Conditional Generative Adversarial Networks (cGAN).

El algoritmo cGAN, en particular, ha mostrado ser el más exitoso debido a su notable reducción del error del modelo y su habilidad para gestionar el sesgo. Este éxito se debe en parte a su capacidad para generar datos realistas que pueden ser utilizados para entrenar modelos predictivos, permitiendo así una mayor flexibilidad y precisión en las predicciones. Además, el cGAN ofrece una forma innovadora de abordar el problema del sesgo, ya que puede aprender a distribuir los datos de manera que minimice el sesgo inherente, manteniendo al mismo tiempo la capacidad predictiva del modelo.

Además de mejorar la capacidad predictiva, un aspecto crucial de estos estudios es la consideración de la interpretabilidad y transparencia de los métodos empleados. La falta de claridad acerca de cómo operan los modelos predictivos basado en inteligencia artificial puede dar lugar a malinterpretaciones y a la toma de decisiones informadas incorrectamente. Para abordar esto, se han introducido técnicas como la visualización mediante gráficos de dependencia parcial y el análisis de predicciones, con el objetivo de incrementar la comprensión del modelo. Estas herramientas facilitan a los usuarios comprender mejor cómo los modelos derivan sus conclusiones, lo que a su vez fortalece la confianza en las predicciones emitidas.

Estos descubrimientos abren nuevas posibilidades en el campo de la ciencia actuarial, concretamente, en todas aquellas funciones donde se recurre al uso de modelos predictivos, proporcionando un camino hacia una práctica más justa y efectiva. En este marco, conceptos como la justicia algorítmica y la transparencia en el uso de datos surgen como principios fundamentales para guiar las acciones profesionales hacia un futuro en el que la equidad y la precisión sean los impulsores principales de la industria aseguradora.

6. BIBLIOGRAFÍA

Awasthi, P., Kleindessner, M., & Morgenstern, J. (2016). Equalized odds postprocessing under imperfect group information.

Bar-Gill, O., & Bubb, R. (2012). Credit card pricing: The CARD Act and beyond. *Cornell Law Review, 97*(5), 967-1024. Disponible en http://scholarship.law.cornell.edu/clr/vol97/iss5/1.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities* (Adaptive Computation and Machine Learning series). Cambridge, MA: MIT Press.

Central Bank of Ireland. (2021). *Review of differential pricing in the private car and home insurance markets: Final report and public consultation*. Central Bank of Ireland.

Central Bank of Ireland. (2021, July). *Review of Differential Pricing in the Private Car and Home Insurance Markets: Final Report and Public Consultation*.

Chibanda, K. F. (2022). *Defining discrimination in insurance*. CAS Research Paper Series on Race and Insurance Pricing.

Crugnola-Humbert, J., Kivisaari, E., Leitner, M., & Zach, V. (2024). *Social sustainability in insurance: What, who and how* (Discussion Paper). Actuarial Association of Europe.

EIOPA. (2022, February 22). *Supervisory statement on differential pricing practices in non-life insurance lines of business*.

EIOPA. (2022, July 11). *CONSULTATION PAPER on Supervisory statement on differential pricing practices in non-life insurance lines of business*.

EIOPA. (2022, July 11). *Impact Assessment of Supervisory Statement on Differential Pricing Practices in Non-Life Insurance*.

Fafalios, S., Charonyktakis, P., & Tsamardinos, I. (2020, April). *Gradient boosting trees*. Gnosis Data Analysis PC.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259-268. https://doi.org/10.1145/2783258.2783311

Feldman, M., Scheidegger, C., Friedler, S. A., Moeller, J., & Venkatasubramanian, S. (2015, July 16). *Certifying and removing disparate impact*.

Financial Conduct Authority. (2020). *General insurance pricing practices: Final report market study MS18/1.3*. (Updated December 2020).

Frees, E. W., & Huang, F. (2023). The discriminating (pricing) actuary. *North American Actuarial Journal, 27*(1), 2-24. https://doi.org/10.1080/10920277.2021.195129

Gohar, U., Biswas, S., & Rajan, H. (2022, diciembre). Towards understanding fairness and its composition in ensemble machine learning.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014, June 10). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hellman, D. (2008). *When discrimination is wrong*. Cambridge, MA: Harvard University Press.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the age of algorithms. *NBER Working Paper, No. 25548*. National Bureau of Economic Research. https://doi.org/10.3386/w25548

Lima Rego, M. (2014). *Statistics as a basis for discrimination in the insurance business*. NOVA Law School, Universidade Nova de Lisboa.

Lippert-Rasmussen, K. (Ed.). (2018). *The Routledge Handbook of the Ethics of Discrimination*.

Madani, A. (2023). *Debugging Machine Learning Models with Python*.

Major, B., & O'Brien, L. T. (2005). The social psychology of stigma. *Annual Review of Psychology, 56*, 393-421. https://doi.org/10.1146/annurev.psych.56.091103.070137

Maitzen, S. (1991). The ethics of statistical discrimination. *Social Theory and Practice, 17*(1), 23–45. http://www.jstor.org/stable/23557057

Mirza, M., & Osindero, S. (2014, November). *Conditional Generative Adversarial Nets*.

Mosley, R., & Wenman, R. (2022). *Methods for quantifying discriminatory effects on protected classes in insurance*. CAS Research Paper Series on Race and Insurance Pricing. Casualty Actuarial Society.

Statutory Instruments. (2022). *S.I. No. 126 of 2022: Central Bank (Supervision and Enforcement) Act 2013 (Section 48(1)) (Insurance Requirements) Regulations 2022*.

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence* (SHS/BIO/REC-AIETHICS/2021). Recuperado de https://unesdoc.unesco.org/ark:/48223/pf0000377897

Van den Boom, F. (s.f.). *The Insurance Industry: insights into challenges for indirect discrimination*. Recuperado de https://techethicslab.nd.edu/assets/517823/the_insurance_industry_insights_into_the_challenges_for_indirect_discrimination.pdf

Van Wieringen, W. N. (2023, junio 27). *Lecture notes on ridge regression*.

Zand, J., & Roberts, S. (2021, septiembre 1). *Mixture Density Conditional Generative Adversarial Network Models (MD-CGAN)*.

Zeng, X., Dobriban, E., & Cheng, G. (2022, junio 3). FAIR BAYES-OPTIMAL CLASSIFIERS UNDER PREDICTIVE PARITY. Shenzhen Research Institute of Big Data.

Zhang, B. H. (2018, febrero 2–3). *Mitigating Unwanted Biases with Adversarial Learning*. New Orleans, LA, USA.