PONTIFICIA UNIVERSIDAD JAVERIANA



CUATRO MODELOS DE APRENDIZAJE SUPERVISADO PARA LA DETECCIÓN DE FRAUDES EN EL RAMO DE LOS SEGUROS DE AUTOMÓVILES

Trabajo de grado presentado en cumplimiento parcial de los requisitos para el grado de Científico de Datos.

AUTHOR

ANAMARIA LEGUIZAMON ALARCÓN

SUPERVISOR

VLADIMIR MORENO G., MSC

Facultad de Ciencias

Departamento de Matemática

BOGOTÁ, COLOMBIA, DICIEMBRE 2025

This page has been intentionally left blank.

This page has been intentionally left blank.

"El rápido progreso de la IA conlleva numerosos riesgos a corto plazo. Ya ha creado cámaras de resonancia divisivas al ofrecer a las personas contenido que las indigna. Ya la utilizan gobiernos autoritarios para la vigilancia masiva y ciberdelincuentes para ataques de phishing. En un futuro próximo, la IA podría utilizarse para crear nuevos virus terribles y armas letales descomunales que deciden por sí mismas a quién matar o mutilar. Todos estos riesgos a corto plazo requieren una atención urgente y contundente por parte de gobiernos y organizaciones internacionales.."

– Geoffrey Hinton, miembro comité química y fisica premio nobel 2024

Declaración de autoría

Declaramos bajo nuestro honor que el trabajo presentado en esta tesis, titulado "Cuatro modelos de aprendizaje supervisado para la detección de fraudes en el ramo de los seguros de automóviles,", es original y ha sido realizado por Anamaria Leguizamon Alarcón bajo la supervisión del profesor Vladimir Moreno G., MSc (vladimir.moreno@javeriana.edu.co).

Bogotá, Colombia, diciembre 2025

This page has been intentionally left blank.

Agradecimientos

Quiero expresar mi más sinceros agradecimientos a todos los profesores de la Pontificia Universidad Javeriana que me han transmitido sus conocimientos y sabiduría a lo largo de los años, les expreso mi más sincero agradecimiento. Su dedicación a la educación ha sido fundamental en mi trayectoria académica.

Por último, estoy profundamente agradecido a mi familia y amigos, cuyo apoyo y aliento me han convertido en la persona que soy hoy. Su confianza en mí ha sido una fuente constante de fortaleza.

This page has been intentionally left blank.

Resumen

Este trabajo de grado presenta un enfoque comparativo para la detección de fraude en seguros de automóviles, a través de un conjunto de datos real con 15 420 observaciones y 33 variables. De este modo, se evalúan modelos base de clasificación, junto con algoritmos avanzados basados en árboles de decisión, con el objetivo de analizar su desempeño en un contexto de datos altamente desbalanceados. Así, se prioriza la capacidad de detección de la clase minoritaria mediante el uso del F_2 score como métrica principal.

Los resultados evidencian una clara superioridad de los modelos basados en técnicas de boosting frente a los modelos tradicionales. En particular, el modelo XGBoost obtiene el mejor desempeño global, al alcanzar un F-score de 0.48, con un recall de 68.95 % y una precisión de 21.95 %, lo que se consolida como el modelo más efectivo para la detección de fraude. En segundo lugar, LightGBM presenta un F_2 score de 0.453, un recall de 76.17 % y una precisión de 17.28 %, lo que muestra un alto potencial predictivo. Por su parte, Random Forest brinda un equilibrio competitivo entre sensibilidad y precisión, lo cual confirma su utilidad como modelo robusto de referencia.

A nivel descriptivo, el análisis exploratorio evidencia que las reclamaciones fraudulentas se concentran en áreas urbanas, en conductores jóvenes (menores de 25 años), principalmente hombres, y en pólizas con cobertura All Perils o Collision. Asimismo, se determina una mayor incidencia de fraude entre los titulares de póliza frente a los terceros involucrados.

Keywords: Fraude, Regresión logística, Random Forest, Gradient Boosting, XGBoost, LightGBM.

This page has been intentionally left blank.

Índice general

Índice general x

1. Introducción 1
- 1.1. Sinópsis . . . 1
- 1.2. Objetivos del estudio . . . 2
- 1.3. Organización del trabajo de grado . . . 2

2. Marco teórico 4
- 2.1. Antecedentes . . . 4
- 2.2. Revisión de la literatura . . . 6
  - 2.2.1. Estudios internacionales . . . 6
  - 2.2.2. Estudios nacionales . . . 8
- 2.3. Algoritmos estándar de aprendizaje automático . . . 10
  - 2.3.1. Regresión logística . . . 10
  - 2.3.2. Árboles de clasificación . . . 11
- 2.4. Modelos de ensamble . . . 12
  - 2.4.1. Random Forest (Bosque Aleatorio) . . . 12
  - 2.4.2. Gradient boosting . . . 14
  - 2.4.3. XGBoost (Extreme Gradient Boing) . . . 16
  - 2.4.4. LightGBM . . . 17
- 2.5. XGBoost *vs.* Random Forest: dos paradigmas de Ensemble Learning . . . 18
  - 2.5.1. Sinópsis . . . 18
  - 2.5.2. Bagging *vs.* Boosting: fundamentos . . . 18
  - 2.5.3. Semejanzas . . . 19
  - 2.5.4. Ventajas y desventajas . . . 19
  - 2.5.5. Alcances y limitaciones . . . 20
  - 2.5.6. Diferencias claves . . . 20
- 2.6. Métricas de Evaluación: Precision, Recall y la Familia F-Score . . . 21
  - 2.6.1. Conceptualización . . . 21
- 2.7. Clases desbalanceadas . . . 24
  - 2.7.1. Contexto del presente trabajo de grado . . . 24
  - 2.7.2. Fundamentación general en el contexto del fraude en seguros de automóvil . . . 25

2.7.3. La base estadística: alteración de la probabilidad a posteriori . . . 26
2.7.4. El riesgo de sobreajuste en el contexto del fraude 26
2.7.5. Estrategias integrales para el fraude en seguros de automóvil . . 27
2.7.6. Un equilibrio entre teoría y práctica 27

3. Análisis y procesamiento de datos **29**
3.1. Conjunto de Datos 29
3.1.1. Tipos de datos 30
3.2. Preparación de los datos para el modelado 31
3.2.1. Limpieza de Datos 31
3.2.2. Análisis exploratorio 31

4. Modelado **43**
4.1. Introducción 43
4.2. Resultados 45

5. Conclusiones y recomendaciones **52**
5.1. Conclusiones 52
5.1.1. Conclusiones por objetivos específicos 52
5.1.2. Conclusión del objetivo general 54
5.2. Recomendaciones 54

This page has been intentionally left blank.

1

Introducción

1.1. Sinópsis

La detección de fraude en el sector asegurador constituye un desafío técnico y operativo de primer orden, con implicaciones directas sobre la rentabilidad, la confianza del cliente y la estabilidad del sistema financiero. En Colombia, el aumento sostenido de reclamaciones fraudulentas en los seguros de automóvil y en el Seguro Obligatorio de Accidentes de Tránsito (SOAT) ha impulsado la necesidad de desarrollar modelos analíticos más precisos y escalables.

De conformidad con lo anterior, el presente trabajo se enmarcó en la aplicación de técnicas de aprendizaje automático para la identificación de patrones de fraude, al priorizar la métrica *F2-Score* como indicador central de desempeño. En efecto, esta elección respondió a un criterio de negocio: maximizar la captura de fraudes *(recall)*, aun a costa de una mayor tasa de falsas alarmas, lo que implica una reducción en la precisión. Con este enfoque, se buscó diseñar un sistema predictivo que optimizara la eficiencia operativa de las aseguradoras al mejorar la detección temprana de reclamaciones fraudulentas.

Para ello, el desarrollo incluyó seis algoritmos, que abarcan desde enfoques lineales, como la regresión logística, hasta técnicas de ensamble avanzadas, como Random Forest, Gradient Boosting, XGBoost y LightGBM. No obstante, el título del trabajo menciona únicamente cuatro modelos, dado que los dos primeros cumplen la función de modelos base, representativos de enfoques tradicionales que no logran capturar relaciones complejas. Los cuatro modelos restantes, a los que hace referencia el título, pertenecen a la familia de ensambles supervisados y, debido a su capacidad para modelar relaciones complejas, constituyen el foco de la metodología. El proceso incluye la limpieza de datos, el análisis exploratorio, el tratamiento del desbalance de clases y la optimización de hiperparámetros mediante GridSearchCV. De este modo, los resultados se evaluaron mediante métricas robustas y visualizaciones que permiten una comprensión integral del rendimiento de cada modelo.

En ese orden de ideas, el presente trabajo demostró que la integración de ciencia de datos y modelos de aprendizaje automático representa un camino viable y eficaz para

enfrentar el fraude en seguros en Colombia. Por lo tanto, se arguye que la adopción de métricas centradas en la sensibilidad, como el , asegura F_2 score una alineación más estrecha con los objetivos de negocio y posibilita orientar la toma de decisiones hacia una gestión proactiva del riesgo.

1.2. Objetivos del estudio

Objetivo general

Identificar y evaluar el modelo de aprendizaje automático con el mayor rendimiento para la detección de fraude en reclamaciones de seguros, a partir de la comparación de seis modelos de aprendizaje supervisado. El desempeño de los modelos se mide mediante la métrica F_2 score, al priorizar la maximización de la detección de fraudes (*recall*), incluso con un incremento en la tasa de falsas alarmas (disminución en *precisión*), de manera que los resultados del modelo se alineen con las prioridades estratégicas del negocio asegurador.

Objetivos específicos

1. Realizar un análisis exploratorio y de calidad de los datos, al identificar patrones relevantes, posibles inconsistencias, valores atípicos y variables con mayor potencial explicativo del comportamiento fraudulento en las reclamaciones.
2. Implementar, entrenar y comparar diversos modelos de clasificación supervisada, como Regresión Logística, Árboles de Decisión, *Random Forest*, *Gradient Boosting*, *XGBoost* y *LightGBM*, por medio de la evaluación de su rendimiento a partir de métricas de precisión, sensibilidad, exactitud y, especialmente, F_2 score.
3. Optimizar los hiperparámetros de los modelos más prometedores mediante técnicas de búsqueda sistemática y validación cruzada, con el fin de mejorar su desempeño predictivo y su capacidad de generalización frente a datos no observados.
4. Analizar los resultados obtenidos y proponer lineamientos prácticos para la implementación del modelo óptimo en entornos reales del sector asegurador, al considerar la relación costo-beneficio y el impacto operativo en la detección de fraudes.

1.3. Organización del trabajo de grado

Con relación al orden de la investigación, el trabajo se estructuró de manera secuencial con el objeto de guiar al lector desde la comprensión del problema hasta la validación empírica de los modelos predictivos.

De este modo, el trabajo se dividió en tres capítulos centrales y un capítulo de cierre con conclusiones y recomendaciones.

Marco teórico

En este capítulo, se presenta el contexto general del fraude en seguros y la justificación del uso de modelos de aprendizaje automático, al destacar el impacto económico y operacional de las reclamaciones fraudulentas.

Análisis y procesado de datos

Se desarrolla un análisis exploratorio de datos, donde se identifican variables clave y se examinan aspectos de cruce entre variables contempladas en la base datos.

Modelado

En este capítulo se presenta la aplicación de los modelos de clasificación implementados: regresión logística, árboles de decisión, *Random Forest*, *Gradient Boosting*, *XGBoost* y *LightGBM*, junto con los procedimientos de ajuste y validación cruzada. Asimismo, se detalla el proceso de optimización de hiperparámetros y la selección de la métrica F_2 score como criterio principal de evaluación. Finalmente, se analizan los resultados comparativos, se interpretan las métricas de desempeño y se discuten las implicaciones prácticas para el sector asegurador, concluyendo con reflexiones y recomendaciones orientadas a la mejora continua y al planteamiento de futuras líneas de investigación.

Conclusiones y recomendaciones

Acto seguido, se finaliza con conclusiones sobre las ventajas de unos modelos en relación con otros, en el contexto del problema de fraude en seguros de automóvil; igualmente, se exponen algunas reflexiones sobre nuestra experiencia en la aplicación de estos modelos.

Las recomendaciones formuladas son fruto de la lectura de varias fuentes, de la dura experiencia en la búsqueda de información local de diferentes fuentes y las barreras y dificultades encontradas.

2

Marco teórico

2.1. Antecedentes

La industria aseguradora se erige como un pilar fundamental del sistema financiero global, al ofrecer protección y estabilidad mediante una gestión eficaz de los riesgos. En efecto, su esencia radica en la capacidad de transformar la incertidumbre en seguridad, proporcionando redes de protección tanto para individuos como para organizaciones Khalil et al., 2022a .

En la actualidad, las compañías de seguros operan en un panorama cada vez más complejo, en el que la confluencia de factores económicos, tecnológicos y regulatorios genera desafíos sin precedentes. Entre estos, el auge de actividades fraudulentas, impulsadas por tecnologías avanzadas y redes de comunicación globales, representa una amenaza particularmente preocupante, con pérdidas que alcanzan miles de millones de dólares anuales a nivel mundial Akhtar et al., 2023 . De tal modo, este fenómeno no solo erosiona la rentabilidad del sector, sino que también distorsiona las estrategias de precios y compromete los beneficios socioeconómicos que las aseguradoras proporcionan a la comunidad Wang et al., 2018 .

Frente a este escenario, la implementación de sistemas robustos de detección y prevención se convierte en una prioridad estratégica. El fraude constituye una porción significativa de los costos operativos, mientras que la creciente complejidad de los procedimientos internos dificulta la integración efectiva de la analítica de datos y la inteligencia artificial en la evaluación de riesgos y la gestión de reclamaciones Hassan et al., 2013 ; Singh et al., 2020 . Por lo tanto, adoptar un enfoque proactivo y flexible resulta esencial para garantizar la resiliencia del sector y fomentar la innovación en medio de la incertidumbre Khalil et al., 2024 .

El concepto de fraude, definido por el *Oxford English Dictionary* como "el acto de engañar intencionadamente a otros para obtener beneficios financieros o personales" Pearsall, 1999 , se manifiesta de diversas formas en el sector asegurador. Este estudio se centra específicamente en el fraude del tomador de la póliza, dada la limitada disponibilidad de datos sobre otras modalidades fraudulentas.

Según Turban et al., 2011 , la minería de datos se define como el método que

permite extraer conocimiento valioso de grandes volúmenes de información. En el sector asegurador, su aplicación es amplia e incluye la detección de fraudes, el análisis de reclamaciones, el procesamiento de suscripciones y la evaluación de riesgos. La esencia de la minearía de datos, como señalan Prasasti et al., 2020 , es que enfoca sus métodos en descubrir patrones estadísticamente confiables, desconocidos y accionables, a partir de datos con cualidades de accesibilidad, pertinencia, suficiencia e integridad.

Los enfoques de aprendizaje por ensamble destacan como una línea de investigación particularmente prometedora, al adaptarse con versatilidad a diversas aplicaciones del aprendizaje automático, como la clasificación, la regresión e incluso el aprendizaje no supervisado Alsuwailem et al., 2023 . Su notable rendimiento proviene de la capacidad para mejorar la generalización de los modelos, mitigar el sobreajuste y superar las limitaciones de los modelos individuales Khalil et al., 2022b ; Piovezan et al., 2023 . En este sentido, esta técnica se consolida como un método robusto y adaptable para abordar problemas complejos en múltiples dominios Das et al., 2021 .

La construcción de sistemas predictivos efectivos requiere, con frecuencia, del empleo de técnicas complementarias, como el remuestreo de datos, la discretización de características y la selección de atributos. La elección de un subconjunto relevante de características no solo reduce la carga computacional, sino que también mejora la eficiencia y la interpretabilidad del modelo Gupta et al., 2022 . Asimismo, el desequilibrio en la distribución de clases, común en los problemas de detección de fraude, puede abordarse mediante técnicas de remuestreo que mejoran de manera sustancial el rendimiento predictivo Baesens et al., 2021 ; Subudhi et al., 2018 .

Al examinar la literatura especializada, se identifica una oportunidad significativa: la escasez de investigaciones que integren estratégicamente las técnicas de minería de datos antes mencionadas con clasificadores de inteligencia artificial, tanto clásicos como de ensamble, dentro de un procedimiento unificado para el desarrollo de modelos de clasificación de fraude en seguros.

El presente estudio pretende contribuir al desarrollo de un sistema predictivo robusto que combine múltiples escenarios de detección, integrando métodos de remuestreo, discretización y selección de características con diversos clasificadores. Esta propuesta se valida mediante un conjunto de datos reales obtenidos de compañías aseguradoras, evaluando la precisión predictiva en el contexto del fraude en seguros y las reclamaciones asociadas. Las contribuciones fundamentales de esta investigación se articulan en cinco dimensiones:

1. **Innovación metodológica**: desarrollo de un sistema de clasificación binaria para la detección de fraude mediante la implementación y comparación de varios algoritmos de aprendizaje supervisado, integrados en un pipeline de modelado orientado a maximizar el desempeño predictivo bajo condiciones de desbalance de clases.
2. **Análisis de preprocesamiento**: evaluación del impacto de las etapas de limpieza, transformación de variables categóricas, tratamiento de variables ordinales y

exclusión de identificadores dentro del pipeline de entrenamiento, lo que garantiza la adecuada preparación de los datos para los modelos predictivos.

3. **Estudio de balanceo**: análisis del efecto del uso de la técnica de sobremuestreo RandomOverSampler en el desempeño de la clasificación binaria, aplicada exclusivamente sobre los datos de entrenamiento dentro de cada pliegue de la validación cruzada para evitar fugas de información.
4. **Validación empírica**: implementación práctica de los modelos de regresión logística, Random Forest, Gradient Boosting, XGBoost y LightGBM sobre un conjunto de datos real del sector asegurador, lo que permite evaluar su comportamiento bajo condiciones reales de desbalance.
5. **Evaluación exhaustiva**: comparación rigurosa del desempeño entre los distintos modelos de detección a través de cinco métricas reconocidas (exactitud, sensibilidad (recall), puntuación F2, precisión y área bajo la curva AUC), complementado con un análisis comparativo para identificar el escenario más favorable.

El resto de este trabajo se organizó de la siguiente manera: la Sección 2 presenta una revisión crítica de estudios previos. Seguidamente, la Sección 3 describe en detalle la metodología de investigación, incluyendo el diseño, los métodos de recolección de datos y las técnicas empleadas para la detección de fraude. Acto seguido, la Sección 4 expone y analiza los hallazgos empíricos. Finalmente, la Sección 5 sintetiza las conclusiones y proyecciones del estudio.

2.2. Revisión de la literatura

2.2.1. Estudios internacionales

En sentido global, se tiene la siguiente recopilación de estudios significativos durante el 2025, relacionado con investigaciones de fraude en seguros de automóvil. Para ello, se utilizan herramientas analíticas de datos.

Las compañías de seguros han reconocido progresivamente la necesidad de adoptar enfoques proactivos y flexibles que les permitan alcanzar sus objetivos estratégicos, al tiempo que enfrentan de manera eficaz los desafíos inherentes a la predicción de riesgos, la detección de fraudes, el análisis de reclamaciones y la definición de estrategias de precios Barry et al., 2020 . A la luz de este escenario, la minería de datos, definida por Turban Turban et al., 2011 como el método para extraer información valiosa de grandes bases de datos mediante técnicas matemáticas, estadísticas y de inteligencia artificial, se ha consolidado como un aliado indispensable. No obstante, el fraude en seguros persiste como un desafío de primer orden, con un costo anual de miles de millones de dólares para la industria, lo que demanda marcos regulatorios sólidos, colaboración intersectorial y herramientas analíticas avanzadas. En esta medida, la Asociación Nacional de Comisionados de Seguros (NAIC) ha emergido como un actor clave en la estandarización de medidas antifraude a través de leyes modelo, como las directrices de la Unidad de Investigación Especial (SIU), que establecen protocolos

obligatorios de cumplimiento e intercambio de datos National Association of Insurance Commissioners, 2022 ; Saylor, 2023 .

Al respecto, investigaciones pioneras han explorado diversas facetas de este problema. Hoyt et al., 2006 , por ejemplo, analizaron datos de fraude en seguros de automóviles y evidenciaron resultados mixtos en la implementación de leyes antifraude: mientras las unidades SIU obligatorias y la clasificación de delitos graves redujeron los niveles de fraude, la obligación de reportar a las fuerzas del orden, de manera paradójica, lo incrementó, lo que sugiere ineficiencias cuando se sustituyen esfuerzos privados.

En una línea similar, Saylor Saylor, 2023 examinó el fraude oportunista en seguros de automóviles, destacando su prevalencia, su impacto económico y los desafíos asociados a su detección. El estudio subraya que este tipo de fraude representa una proporción significativa del total, aunque con frecuencia se ve opacado por los casos de fraude "duro". A partir de la Teoría de la Neutralización, el autor analiza los mecanismos mediante los cuales los perpetradores justifican sus acciones y propone estrategias de disuasión que incluyen campañas de concienciación, mejoras en la gestión de reclamaciones y cooperación interinstitucional.

La comunidad investigadora y las aseguradoras enfrentan, asimismo, desafíos persistentes relacionados con la disponibilidad, la calidad y la integridad de los datos, así como con la presencia de conjuntos desbalanceados y problemas de interpretabilidad de los modelos Cappiello et al., 2020 .

Diversos estudios han abordado la detección de fraudes mediante metodologías innovadoras. Bhowmik Bhowmik, 2011 , por ejemplo, propuso una estrategia para la detección de fraude en seguros de automóviles mediante el uso de algoritmos basados en árboles de decisión y clasificación bayesiana ingenua.Por su parte, Dhieb et al. Dhieb et al., 2020 emplearon técnicas de aprendizaje automático para detectar y clasificar de manera autónoma reclamaciones fraudulentas, incorporando mecanismos de alerta para la identificación de casos sospechosos.

El trabajo de Kowshalya y Nandhini, Kowshalya et al., 2018 constituye una contribución relevante, en la medida en que estos autores utilizaron técnicas de minería de datos para predecir primas de seguros y reclamaciones fraudulentas, logrando una reducción significativa en los tiempos de análisis. Asimismo, Itri et al. Itri et al., 2019 compararon diez algoritmos de aprendizaje automático, identificando que el método *Random Forest* superaba consistentemente a sus alternativas en la predicción de fraude. De igual modo, Subudhi y Panigrahi Subudhi et al., 2020 introdujeron un sistema híbrido basado en agrupamiento difuso optimizado (GA-FCM) combinado con clasificadores supervisados, demostrando su eficacia con datos reales de reclamaciones.

Finalmente, Nordin et al. Nordin et al., 2024 compararon modelos tradicionales y enfoques de aprendizaje automático para la predicción de fraude en seguros de automóviles, encontrando que el modelo *Tree-Augmented Naïve Bayes* (TAN) alcanzó los mejores niveles de precisión y sensibilidad, lo que pone de relieve el potencial de los enfoques híbridos en este dominio.

2.2.2. Estudios nacionales

A nivel local, en Colombia son escasas las investigaciones públicas sobre estudios del fenómeno de fraude en el ramo de los seguros generales y, en particular, en el campo de los seguros de automóvil. A continuación, se presenta un corto resumen del estado actual de investigación en el país.

Magnitud del problema y contexto sectorial

En 2024, la industria aseguradora colombiana alcanzó una producción total de primas (los ingresos que recibe de los asegurados al adquirir un seguro) de COP 56,1 billones, con crecimiento nominal del 11 % y real del 5,3 % Nájera, 2025 . De manera paralela, la prensa económica reportó *más de 30.000 casos de fraude en seguros en 2024*, siendo Bogotá la ciudad con más registros Hernández Naranjo, 2025 .

En el ecosistema del SOAT –estrechamente vinculado al ramo de automóviles–, se fortaleció la cooperación institucional; de tal modo, Administradora de los Recursos del Sistema General de Seguridad Social en Salud (ADRES) y Federación de Aseguradores Colombianos (Fasecolda) firmaron en febrero de 2024 un acuerdo para el intercambio de información y el combate de irregularidades, con el antecedente de que en 2023 se detectaron prácticas fraudulentas por más de COP 74 000 millones ADRES, 2024 ; Fasecolda, 2024 .

Para 2024, Fasecolda reportó 4.081 casos de fraude en la compra/venta del SOAT, con pérdidas estimadas superiores a COP 2000 millones Fasecolda, 2025 . En consecuencia, dichas cifras se replicaron ampliamente por medios nacionales en febrero durante el 2025 [1]

Línea de investigación aplicada (2024–2025)

Durante el periodo 2024–2025 se observa la consolidación de enfoques de aprendizaje supervisado para la detección de fraude en seguros, con foco en los ramos de automóviles y SOAT, que incluyen el manejo de clases desbalanceadas mediante técnicas de sobremuestreo y ponderación de costos, así como el uso de modelos de ensamble, como Random Forest y XGBoost, y procesos de validación rigurosos basados en métricas sensibles a la clase minoritaria, tales como recall, F1 y PR-AUC.

En el ámbito académico colombiano, se evidencia la existencia de prototipos y trabajos de grado orientados específicamente al fraude en el SOAT Villarraga Palomino, 2023 . En el frente empresarial, se comunican implementaciones que integran modelos estadísticos y de aprendizaje automático para la clasificación temprana de siniestros fraudulentos INIF, 2021 .

[1] https://www.infobae.com/colombia/2025/02/18/fasecolda-advirtio-que-las-estafas-con-el-soat-aumentan-asi-puede-evitar-ser-la-proxima-victima/

Investigación y desarrollo en seguros para automóvil y en SOAT

Volumen y exposición. El ramo de automóviles es masivo y el parque de motocicletas plantea retos adicionales, lo que incide en la siniestralidad y en la presión sobre el sistema Nájera, 2025 .

Huellas de datos ricas. En reclamaciones, redes de prestadores y trazabilidad documental hay señales explotables para *feature engineering* y detección de patrones, útiles en *pipelines* de analítica avanzada.

Externalidades sanitarias. La cooperación ADRES–Fasecolda viabiliza mejores cruces de información y validaciones antifraude para el SOAT ADRES, 2024 ; Fasecolda, 2024 .

Metodologías 2024–2025 destacadas en práctica

Modelado supervisado con desbalance. Uso de *Random Forest*, regresión logística y XGBoost con ajuste de umbrales y aprendizaje sensible al costo; evaluación con ROC-AUC, PR-AUC, F1 y *recall* de la clase fraude Villarraga Palomino, 2023 .

Preprocesamiento y gobernanza. Remuestreo, discretización y selección de variables para estabilizar varianza y mejorar interpretabilidad, así como la integración de reglas de negocio con ML, con trazabilidad para auditoría INIF, 2021 .

Cifras que guían la agenda (2024–2025)

- **Dimensión de mercado:** $56,1 billones de primas (2024), crecimiento real 5,3 % Nájera, 2025 .
- **Incidencia sectorial:** más de 30.000 casos de fraude en seguros (2024) Hernández Naranjo, 2025 .
- **SOAT:** 4.081 casos de fraude (2024) y pérdidas >$2.000 millones Fasecolda, 2025 ; cifras difundidas por medios nacionales [2]
- **Cooperación institucional:** acuerdo ADRES–Fasecolda (2024) para intercambio de información y acciones de control ADRES, 2024 ; Fasecolda, 2024 .

Brechas y oportunidades de investigación

Datos etiquetados y compartibles. Falta de *benchmarks* públicos anonimizados limita la comparación metodológica; una oportunidad para consorcios sectoriales.

Explicabilidad y sesgos. Aumenta la demanda regulatoria de explicaciones y tratamiento de sesgos; un vector de investigación aplicada en 2025.

Privacidad y colaboración. Aprendizaje federado para entrenar sin intercambiar datos crudos entre aseguradoras.

Producción y SLAs. Transitar de pilotos a despliegues con medición de impacto (ahorros, falsos positivos, tiempo de auditoría) y acuerdos de nivel de servicio.

[2] https://www.infobae.com/colombia/2025/02/18/fasecolda-advirtio-que-las-estafas-con-el-soat-aumentan-asi-puede-evitar-ser-la-proxima-victima/

2.3. Algoritmos estándar de aprendizaje automático

El aprendizaje supervisado se define como un mecanismo mediante el cual se infieren y aprenden las relaciones subyacentes entre los datos de entrada y una variable objetivo. Dicha variable puede corresponder a un atributo numérico continuo o a un atributo categórico multiclase, según se trate de problemas de regresión o de clasificación, respectivamente. Para ello, este tipo de aprendizaje se apoya en datos etiquetados que incluyen un conjunto de vectores observados, comúnmente denominados predictores o características, así como una salida deseada, conocida como señal supervisora o etiqueta de clase. En términos generales, el propósito de estos mecanismos consiste en generalizar la relación existente entre los vectores de características y la señal supervisora, con el fin de predecir la salida cuando se utilizan instancias de entrada no etiquetadas (Awad y Khanna, 2015).

El proceso de entrenamiento depende de forma decisiva de la calidad de los datos utilizados, lo que implica que un algoritmo de aprendizaje automático supervisado solo puede predecir con precisión la salida para instancias desconocidas o no observadas cuando los datos de entrada empleados durante el entrenamiento presentan un nivel de calidad elevado. Por el contrario, el uso de datos de baja calidad puede derivar en problemas de sobreajuste, entendidos como la dificultad del algoritmo para generalizar las relaciones subyacentes entre los predictores y la variable objetivo, lo que se traduce en un rendimiento deficiente en tareas de regresión o clasificación.

2.3.1. Regresión logística

La regresión logística es un modelo estadístico que sigue casi la misma teoría que la regresión lineal; sin embargo, se considera un algoritmo probabilístico utilizado para resolver problemas binarios o multiclase mediante una función logística que puede expresarse matemáticamente de la siguiente manera:

$$f(z) = \frac{e^z}{1 + e^z}, \quad \text{donde } z \in [-\alpha, \alpha].$$

En general, un modelo de regresión logística predice la probabilidad de ocurrencia de un evento específico modelando la relación entre una variable dependiente X y un resultado categórico Y (Awad y Khanna, 2015).

En términos matemáticos, la función logística previamente mencionada puede expresarse bajo los siguientes términos:

$$P(Y|X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}},$$

donde β_0 y β_1 representan los log-odds estimados de un cambio unitario para su respectiva variable de entrada con la cual están asociados; en otras palabras, pueden verse como los pesos que traducen cualquier cambio en las variables de entrada hacia el resultado probabilístico. Además, al extraer la inversa de la función logística, se obtiene

una nueva función llamada *logit* o *log-odds*, la cual permite generar los coeficientes de la regresión logística, β_0 y β_1, para el caso basado en un solo predictor:

$$\text{logit}(P(Y|X)) = \ln\left(\frac{P(Y|X)}{1 - P(Y|X)}\right) = \beta_0 + \beta_1 X.$$

Una vez calculado el log-odds, la función logística lo recibe como entrada, $\beta_0 + \beta_1 X$, y devuelve la probabilidad de ocurrencia del evento Y perteneciente a una clase positiva cuando la variable X se utiliza como entrada.

Para concluir, al igual que en el caso de la regresión lineal, nos interesan los coeficientes de intersección β_0 y pendiente β_1; sin embargo, con la ayuda de la función logística transformamos estos valores en la probabilidad de que un valor pertenezca a una clase particular conocida como *clase positiva*.

2.3.2. Árboles de clasificación

Los árboles de decisión –también denominados árboles de clasificación– se emplean para asignar una instancia de datos a un conjunto predefinido de clases, a partir de sus atributos, conocidos en el ámbito del aprendizaje automático como características o predictores. Ahora bien, estos modelos pueden entenderse como sistemas expertos de decisión o de clasificación que buscan, en cierta medida, imitar y automatizar el razonamiento subyacente de un experto en la tarea encomendada. Entre sus principales ventajas se encuentra su simplicidad de implementación, así como su carácter autoexplicativo, que facilita su representación gráfica mediante estructuras jerárquicas Rokach et al., 2014

Asimismo, un árbol de decisión se define como un clasificador basado en una partición recursiva del espacio de instancias, el cual está conformado por distintos tipos de nodos: el nodo raíz, los nodos internos o de prueba y los nodos terminales, también denominados hojas. En consecuencia, el nodo raíz constituye el punto inicial del árbol y no presenta aristas entrantes, mientras que los nodos internos dividen el espacio de instancias en dos o más particiones, de acuerdo con un valor específico de un atributo. Finalmente, este proceso conduce a los nodos terminales, que representan la clase o resultado más adecuado alcanzado tras recorrer los nodos internos precedentes.

Los conceptos fundamentales asociados a los árboles de decisión, como la entropía y la ganancia de información, se abordan a continuación. Al considerar puntos de datos distribuidos en un espacio bidimensional y etiquetados por clases, la división de los datos de manera horizontal o vertical da lugar a conjuntos con distintos niveles de entropía, lo que permite evaluar la calidad de las particiones generadas. La ganancia de información para cada tipo de división puede calcularse matemáticamente mediante la siguiente ecuación:

$$I = H(S) - \sum_{i=1}^{2} \frac{|S_i|}{|S|} H(S_i)$$

Donde $H(S)$ representa la entropía para un conjunto genérico de puntos de entrenamiento S.

Una división con menor entropía produce una ganancia de información $I = 0.4$, mientras que una división con mayor entropía da lugar a $I = 0.6$, lo que significa que se logra una mejor separación de clases con la segunda forma de dividir los datos Criminisi et al., 2011 .

En resumen, los árboles de clasificación funcionan navegando cada instancia desde la raíz del árbol hacia abajo hasta alcanzar una hoja específica, de acuerdo con los resultados de los nodos internos y la métrica de ganancia de información obtenida posteriormente. En efecto, se evidencia que los nodos internos pueden evaluar tanto atributos numéricos como nominales. Además, según Breiman et al., 1984 , la precisión de los árboles de decisión está principalmente influenciada por su complejidad, la cual puede medirse mediante el número total de nodos, el número total de hojas, la profundidad del árbol o el número de predictores utilizados, o por cualquier combinación posible de estos factores.

2.4. Modelos de ensamble

Por otro lado, los métodos de ensamble son técnicas que tienen por objeto combinar múltiples modelos en uno solo para mejorar su rendimiento general. Aunado a esto, estos métodos se dividen en dos amplias categorías: técnicas de ensamble secuenciales y paralelas. Las técnicas secuenciales generan aprendices base de forma secuencial, donde existe dependencia de los datos; de modo que cada aprendiz subsiguiente depende del rendimiento del aprendiz anterior para optimizar el desempeño final. En cambio, las técnicas de ensamble paralelas generan sus aprendices base de manera simultánea, lo que fomenta la independencia entre cada aprendiz, con el objetivo de reducir el error final del rendimiento.

2.4.1. Random Forest (Bosque Aleatorio)

Fundamento

La estrategia de Bagging consiste, precisamente, en abordar la componente de varianza del error. Este método genera múltiples réplicas del conjunto de entrenamiento mediante remuestreo bootstrap (muestreo con reemplazo) y, para cada réplica, entrena un clasificador independiente. Posteriormente, las predicciones individuales se combinan mediante votación mayoritaria en problemas de clasificación o mediante promedio en tareas de regresión. El mecanismo fundamental por el cual Bagging reduce la varianza radica en el promediado de modelos inestables: al entrenar múltiples árboles sobre distintas versiones del conjunto de datos y agregar sus predicciones, se disminuye la sensibilidad a perturbaciones específicas del entrenamiento, sin alterar de manera significativa el sesgo subyacente.

Adicionalmente, la evidencia empírica confirma esta propiedad. Por su parte, Maclin et al., 1997 demuestran que Bagging "produce consistentemente modelos más estables al sobreajuste", dado que la agregación reduce la variabilidad sin amplificar los errores sistemáticos. En problemas con ruido moderado o patrones relativamente estables, Bagging proporciona mejoras sustanciales; no obstante, su capacidad se ve limitada cuando el modelo base presenta un sesgo estructural significativo, dado que el promediado de modelos sesgados no corrige el error sistemático subyacente.

Amplicación

El algoritmo Random Forest es una técnica de aprendizaje automático de tipo paralelo, basada en la teoría de los árboles de decisión, en la cual estos dejan de concebirse y utilizarse como entidades individuales. En su lugar, múltiples árboles de decisión, también denominados aprendices débiles, se combinan para conformar una técnica predictiva robusta conocida como aprendiz de ensamble, empleada principalmente en tareas de clasificación. Cada árbol se construye a partir de una muestra aleatoria de características, con el propósito de reducir la varianza mediante la disminución de la correlación entre las salidas de los árboles individuales.

Asimismo, este tipo de algoritmos de aprendizaje automático se ve fuertemente influido por diversos componentes, en particular por la aleatoriedad introducida al construir cada árbol de decisión de manera distinta a los demás. Dicha aleatoriedad, incorporada durante la fase de entrenamiento, confiere al modelo una elevada robustez frente a datos ruidosos o desbalanceados. De manera general, esta aleatoriedad se logra a través del muestreo aleatorio de los datos de entrenamiento, técnica conocida como *bagging*) o mediante la selección aleatoria de características en los nodos de decisión Criminisi et al., 2011 .

En términos generales, el entrenamiento de un bosque aleatorio ocurre cuando se optimizan los parámetros de los árboles de decisión, llamados aprendices débiles, en cada nodo de división j a través de:

$$\theta_j^* = \arg\max I_j$$

Para el caso específico de los problemas de clasificación, la función objetivo I, tal como se definió en la sección anterior, representa la ganancia de información calculada por la ecuación anterior. Posteriormente, una vez que cada árbol de decisión se ha entrenado de forma independiente y eficiente, todas estas "predicciones débiles" se combinan en una única predicción del bosque mediante una operación de promedio, descrita en el caso de tareas de clasificación de la siguiente manera:

$$p(c|v) = \frac{1}{T}\sum_{t=1}^{T} p_t(c|v)$$

Donde T representa el número total de árboles de decisión; v, una instancia de atributo y $p(c|v)$ es la probabilidad posterior del ensamble, correspondiente a que una

instancia pertenezca a una clase discreta cualquiera. Para mayor detalle ver Criminisi et al., 2011 .

En otras palabras, los bosques de clasificación producen salidas probabilísticas, debido a que devuelven una distribución completa de clases. De este modo, se observa que el mismo valor de entrada v se conduce de manera diferente desde el nodo raíz hasta llegar a una hoja; en este punto, cada probabilidad posterior $p_t(c|v)$ se lee y se promedia para obtener una probabilidad posterior de ensamble $p(c|v)$.

Finalmente, se determina que los algoritmos de *Random Forest* tienden a ofrecer altas precisiones y buena generalización; sin embargo, su desempeño se ve significativamente afectado por varios parámetros, como su tamaño, número de árboles, profundidad promedio, criterio de división y proporción de características utilizadas en cada nodo.

2.4.2. Gradient boosting

El Gradient Boosting es un algoritmo de aprendizaje automático supervisado de tipo ensamble que puede emplearse tanto en problemas de clasificación como de regresión. Algoritmos como Random Forest, Extra Trees y Gradient Boosting reciben la denominación de modelos de conjunto debido a que el modelo final se construye a partir de la combinación de múltiples modelos individuales. A la luz de este escenario, Gradient Boosting se consolida como una de las técnicas más potentes dentro del aprendizaje automático contemporáneo, dada su flexibilidad teórica, la disponibilidad de implementaciones optimizadas y su rendimiento consistente en distintos escenarios. No obstante, su efectividad práctica depende de manera crítica de una comprensión adecuada de sus principios matemáticos, de una selección cuidadosa de los hiperparámetros y de consideraciones computacionales apropiadas.

Las implementaciones modernas como **XGBoost**, **LightGBM** y *CatBoost*, han contribuido a facilitar el acceso a esta tecnología, permitiendo su aplicación exitosa en diversos dominios. El desarrollo continuo en este campo proyecta mejoras adicionales en términos de eficiencia, escalabilidad y facilidad de uso.

Formulación General

Teniendo un conjunto de datos de entrenamiento $\{(x_i, y_i)\}_{i=1}^{n}$, el objetivo consiste en encontrar una función $F^*(x)$ que minimice la función de pérdida $L(y, F(x))$:

$$F^* = \arg\min_F \mathbb{E}_{x,y}[L(y, F(x))]$$

El Gradient Boosting construye $F(x)$ de forma aditiva:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

donde $h_m(x)$ son los modelos base (*learners*) y γ_m son los coeficientes de peso.

Algoritmo General

Algorithm 1 Gradient Boosting General

1: Inicializar $F_0(x) = \arg\min_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$
2: **for** $m = 1$ to M **do**
3: Calcular pseudo-residuales: $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$
4: Ajustar un modelo base $h_m(x)$ a los pseudo-residuales $\{(x_i, r_{im})\}_{i=1}^{n}$
5: Calcular multiplicador óptimo: $\gamma_m = \arg\min_\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$
6: Actualizar el modelo: $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$
7: **end for**

Gradient Boosting para Regresión

Para regresión con error cuadrático medio, el algoritmo se simplifica:

Algorithm 2 Gradient Boosting para Regresión (MSE)

1: Inicializar $F_0(x) = \bar{y}$
2: **for** $m = 1$ to M **do**
3: Calcular residuales: $r_{im} = y_i - F_{m-1}(x_i)$
4: Ajustar árbol $h_m(x)$ a $\{(x_i, r_{im})\}_{i=1}^{n}$
5: Para cada hoja j del árbol, calcular:

$$\gamma_{jm} = \frac{\sum_{x_i \in R_{jm}} r_{im}}{\sum_{x_i \in R_{jm}} 1}$$

6: Actualizar: $F_m(x) = F_{m-1}(x) + \eta \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}(x \in R_{jm})$
7: **end for**

Gradient Boosting para Clasificación

Para clasificación binaria con pérdida logística:

Algorithm 3 Gradient Boosting para Clasificación Binaria

1: Inicializar $F_0(x) = \frac{1}{2}\log\left(\frac{\sum y_i}{\sum(1-y_i)}\right)$
2: **for** $m = 1$ to M **do**
3: Calcular pseudo-residuales:

$$r_{im} = y_i - \frac{1}{1 + e^{-F_{m-1}(x_i)}}$$

4: Ajustar árbol $h_m(x)$ a $\{(x_i, r_{im})\}_{i=1}^{n}$
5: Para cada hoja j, optimizar:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} \log(1 + e^{-y_i(F_{m-1}(x_i)+\gamma)})$$

6: Actualizar: $F_m(x) = F_{m-1}(x) + \eta \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}(x \in R_{jm})$
7: **end for**
8: Output: $\hat{y} = \frac{1}{1+e^{-F_M(x)}}$

2.4.3. XGBoost (Extreme Gradient Boing)

Fundamento

En contraste, el método *Boosting* construye el ensamble de manera secuencial y adaptativa. Cada predictor débil se entrena con énfasis en corregir aquellas observaciones que fueron clasificadas incorrectamente por los predictores anteriores, asignando un mayor peso a los ejemplos más difíciles. Este proceso iterativo permite que el ensamble aprenda patrones progresivamente más complejos y reduzca el sesgo sistemático, al forzar al modelo a capturar relaciones que los clasificadores individuales no logran aprender.

De acuerdo con Maclin et al., 1997 , los algoritmos pertenecientes a la familia Boosting suelen obtener mejores resultados que *Bagging*, lo que refleja su capacidad para reducir el sesgo y, al mismo tiempo, manejar parcialmente la varianza mediante la combinación ponderada de clasificadores débiles. Esta superioridad se manifiesta, en particular, en problemas con múltiples clases y patrones complejos que demandan modelos más expresivos.

No obstante, este mecanismo intensivo de corrección de errores presenta una vulnerabilidad frente al ruido. Al enfatizar de manera reiterada las observaciones mal clasificadas,*Boosting* puede sobreajustar ejemplos inconsistentes o incorrectamente etiquetados. Maclin et al., 1997 documentan que el desempeño de *Boosting* se degrada de forma significativa a medida que aumenta el nivel de ruido en los datos, siendo este efecto más pronunciado que en *Bagging*. En sus experimentos, la ganancia de *Boosting* disminuye progresivamente al introducir niveles de ruido del 5 %, 10 %, 20 % y 30 %, mientras que Bagging mantiene mejoras más estables.

Ampliación

XGBoost Chen et al., 2016 introduce varias mejoras clave:

Función Objetivo Regularizada

$$\mathcal{L}(\phi) = \sum_i L(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

donde $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$

Expansión de Taylor de Segundo Orden

Utiliza la expansión de segundo orden para una aproximación más precisa:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[L(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

donde $g_i = \partial_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)})$ y $h_i = \partial^2_{\hat{y}^{(t-1)}} L(y_i, \hat{y}^{(t-1)})$

2.4.4. LightGBM

Por otro lado, LightGBM emplea una estrategia basada en histogramas para el crecimiento de árboles, la cual cuantifica las características continuas en bins discretos:

Algorithm 4 Construcción de Histogramas en LightGBM

for cada característica f_j **do**
  Ordenar valores únicos: $v_1 < v_2 < \cdots < v_m$
  Crear k bins: $B_1, B_2, \ldots, B_k$
  for cada valor v_i **do**
    Asignar v_i al bin $B_{\lfloor \frac{v_i - v_{min}}{v_{max} - v_{min}} \cdot k \rfloor}$
  end for
  Precalcular gradientes acumulados por bin: $G_B = \sum_{i \in B} g_i$, $H_B = \sum_{i \in B} h_i$
end for

La ganancia de información para una división se calcula de la siguiente manera:

$$\mathcal{G}_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma$$

donde I_L e I_R son los conjuntos de instancias en los bins izquierdo y derecho respectivamente.

LightGBM Ke et al., 2017 introduce dos técnicas innovadoras:

Gradient-based One-Side Sampling (GOSS)

Mantiene todos los instancias con gradientes grandes y muestrea aleatoriamente instancias con gradientes pequeños.

Exclusive Feature Bundling (EFB)

Este agrupa características mutuamente excluyentes para reducir la dimensionalidad.

2.5. XGBoost *vs*. Random Forest: dos paradigmas de Ensemble Learning

En el panorama del aprendizaje automático supervisado, las técnicas de ensemble learning destacan por su robustez y su rendimiento consistente. Entre ellas, *Random Forest* y *XGBoost* se han consolidado como algoritmos de referencia en aplicaciones industriales. Si bien ambos pertenecen a la familia de los ensambles basados en árboles de decisión, sus filosofías subyacentes, derivadas de las estrategias de *Bagging* y *Boosting*, respectivamente, les confieren características, ventajas y limitaciones diferenciadas.

2.5.1. Sinópsis

Gracias a su capacidad para mejorar el desempeño predictivo, los métodos de ensamble se han consolidado como técnicas fundamentales del aprendizaje automático, en la medida en que combinan múltiples modelos con el propósito de reducir el error de generalización. Entre estos métodos, Breiman, 1996 y Freund et al., 1996 representan dos enfoques diferenciados para enfrentar las fuentes del error predictivo. De acuerdo con Maclin et al., 1997 , quienes realizaron una evaluación empírica de estas técnicas, ambos métodos ofrecen mejoras sustanciales, aunque operan sobre componentes distintos del error: *Bagging* se orienta a la reducción de la varianza, mientras que *Boosting* se enfoca principalmente en la disminución del sesgo.

El error esperado de un modelo supervisado se descompone en tres componentes: sesgo, varianza y ruido irreducible. En el caso de los algoritmos basados en árboles de decisión, como los empleados en esta investigación, la estructura inherente del modelo presenta, por lo general, una alta varianza y un bajo sesgo. En consecuencia, esto implica que pequeñas variaciones muestrales pueden dar lugar a árboles considerablemente diferentes, lo que evidencia una elevada varianza; sin embargo, la estructura jerárquica del árbol permite capturar de manera adecuada las relaciones subyacentes en los datos, reflejando un sesgo relativamente bajo.

2.5.2. Bagging *vs*. Boosting: fundamentos

La diferencia esencial radica en su mecanismo de construcción del ensemble:

- **Random Forest (Bagging)**: opera bajo el principio de *Bootstrap Aggregating* (Bagging) y entrena múltiples árboles de decisión en paralelo, cada uno con una muestra aleatoria del conjunto de entrenamiento (muestreo con reemplazo) y, típicamente, un subconjunto aleatorio de características en cada división. Así, la predicción final es el resultado de promediar (regresión) o votar (clasificación) las predicciones de todos los árboles. En cuanto a su objetivo primordial, este busca

reducir la varianza del modelo, al suavizar el sobreajuste (*overfitting*) que podría tener un árbol individual al hacer que los diversos árboles "desacuerden" en sus errores.

- **XGBoost (Boosting)**: implementa una sofisticada variante del Boosting de Gradiente y entrena árboles de manera secuencial, donde cada nuevo árbol intenta corregir los residuos o errores cometidos por la combinación de todos los árboles anteriores. En cada paso, se ajusta un árbol débil a los gradientes negativos de la función de pérdida del modelo actual. La predicción final es la suma ponderada de las predicciones de todos los árboles; asimismo, su enfoque se centra en reducir el sesgo (error sistemático), al perfeccionar iterativamente la capacidad del modelo para capturar patrones complejos en los datos.

Esta distinción clave puede resumirse así: Boosting se centra en reducir el sesgo (error sistemático), mientras que Bagging se centra en reducir la varianza.

2.5.3. Semejanzas

A pesar de sus diferencias fundamentales, ambos algoritmos comparten características importantes:

- Están basados en árboles de decisión como *learners* base.
- Pertenecen a la categoría de modelos de ensemble, al combinar múltiples predictores débiles para formar uno fuerte.
- Son capaces de manejar datos numéricos y categóricos sin necesidad de extensivo preprocesamiento.
- Ofrecen medidas intrínsecas de importancia de características, lo que los hace interpretables en cierta medida.
- Son generalmente robustos frente a valores atípicos (*outliers*) y al ruido en los datos, en comparación con modelos lineales.

2.5.4. Ventajas y desventajas

Random Forest

- **Ventajas**: menos propenso al sobreajuste (*overfitting*) en datos ruidosos; igualmente, son altamente paralelizables y proporciona un estimador confiable del *out-of-bag* (OOB) sin necesidad de validación cruzada; finalmente, es estable, dado que pequeños cambios en datos no alteran drásticamente el modelo.
- **Desventajas**: puede tener un sesgo más alto en problemas complejos donde se requieren relaciones muy no lineales; cabe señalar que los modelos resultantes suelen ser grandes (muchos árboles) y menos eficientes en predicción; del mismo modo, puede perder precisión frente a algoritmos de boosting bien ajustados en tareas donde la reducción de sesgo es crítica.

XGBoost

- **Ventajas**: frecuentemente alcanza una mayor precisión de estado del arte; igualmente, altamente flexible y configurable (funciones de pérdida personalizables, regularización L1/L2); eficiencia computacional y uso de memoria optimizados; manejo nativo de valores faltantes.
- **Desventajas**: mayor riesgo de sobreajuste, especialmente con datos ruidosos o pocas muestras; del mismo modo, requiere un ajuste más cuidadoso de hiperparámetros. Por otro lado, el entrenamiento secuencial puede ser más lento (aunque se mitiga con paralelización a nivel de características) y presenta modelos menos estables ante cambios en los datos de entrenamiento.

2.5.5. Alcances y limitaciones

Random Forest se caracteriza como una herramienta confiable y de propósito general, cuya principal fortaleza radica en su estabilidad y facilidad de uso. Constituye una opción inicial adecuada para problemas de clasificación y regresión, especialmente en contextos donde los datos presentan ruido, el sobreajuste representa una preocupación relevante o se requiere una estimación rápida y robusta del desempeño predictivo. No obstante, este algoritmo puede resultar limitado en problemas de alta complejidad, en los que se demanda un poder predictivo máximo y se dispone de recursos suficientes para realizar un ajuste fino del modelo.

Por su parte, XGBoost se ha consolidado históricamente como uno de los algoritmos de mayor desempeño en competencias de *Kaggle*, debido a su elevado poder predictivo. Su aplicación resulta especialmente pertinente en escenarios donde la precisión constituye la máxima prioridad y se cuenta con conjuntos de datos suficientemente limpios para un entrenamiento eficaz, en particular en datos estructurados o tabulares. Sin embargo, su principal limitación reside en su mayor complejidad y en su propensión al sobreajuste cuando no se aplican mecanismos de regularización adecuados. Además, aunque permite cierto grado de interpretabilidad mediante el análisis de la importancia de las características, el modelo presenta una estructura más opaca que Random Forest, debido a su naturaleza secuencial y a la complejidad de su proceso de aprendizaje.

2.5.6. Diferencias claves

El tabla 2.1 resume las principales diferencias técnicas y operativas.

En conclusión, la elección entre Random Forest y XGBoost no se plantea como una cuestión de superioridad absoluta, sino como una decisión orientada a identificar el algoritmo más adecuado según el contexto específico del problema. Random Forest se configura como un enfoque generalista, robusto y estable, apropiado para procesos de prototipado rápido, entornos con datos ruidosos y escenarios en los que la interpretabilidad y la estabilidad del modelo resultan prioritarias. Por el contrario, XGBoost se presenta como un enfoque especializado de alto rendimiento, diseñado para maximizar

Cuadro 2.1: *Comparativa entre Random Forest y XGBoost*

Random Forest (Bagging)	**XGBoost (Boosting)**
Entrenamiento: Paralelo. Los árboles son independientes.	**Entrenamiento**: Secuencial. Cada árbol depende de los anteriores.
Objetivo: Reducir la **varianza**.	**Objetivo**: Reducir el **sesgo**.
Muestreo: Bootstrap (muestras con reemplazo).	**Ponderación**: Asigna mayor peso a instancias mal predichas.
Agregación: Promedio o votación simple.	**Agregación**: Suma ponderada (con tasa de aprendizaje).
Velocidad: Generalmente más rápido de entrenar (paralelizable).	**Velocidad**: Puede ser más lento (secuencial), pero altamente optimizado.
Sensibilidad al Ruido: Más robusto; el bagging suaviza el ruido.	**Sensibilidad al Ruido**: Más sensible; puede sobreajustar a errores.
Ajuste: Suele requerir menos ajuste fino de hiperparámetros.	**Ajuste**: Muy sensible a hiperparámetros (tasa de aprendizaje, profundidad).

la precisión predictiva, aun cuando ello implique una mayor complejidad en el ajuste del modelo y un incremento en el riesgo de sobreajuste.

En este sentido, la comprensión de la dicotomía fundamental entre la reducción de la varianza, asociada a las estrategias de Bagging y Random Forest, y la reducción del sesgo, propia de los enfoques de Boosting y XGBoost, proporciona el marco teórico necesario para fundamentar esta elección de manera informada y estratégica.

2.6. Métricas de Evaluación: Precision, Recall y la Familia F-Score

Para el desarrollo de esta sección se utilizarán dos fuentes bibliográficas Sokolova et al., 2009 y Goutte et al., 2005 .

2.6.1. Conceptualización

En el campo del aprendizaje automático y la ciencia de datos, elegir las métricas de evaluación adecuadas no es un simple trámite técnico: representa una decisión estratégica que expresa qué es lo verdaderamente importante para el problema que se busca resolver. La métrica de exactitud (Accuracy), aunque ampliamente usada, puede resultar engañosa en escenarios con datos desbalanceados. Por ello, cobran especial relevancia las métricas que permiten descomponer y entender mejor el desempeño del modelo, como la precisión (Precision) y el recall, cuya combinación da lugar a la familia de medidas F Score.

La precisión responde a la pregunta: “De todas las instancias que el modelo etiquetó como positivas, ¿cuántas lo son realmente?”. Al respecto, la precisión se define matemáticamente como el cociente entre los verdaderos positivos (*True Positives, TP*) y todos los

elementos predichos como positivos:

$$\text{Precisión (P)} = \frac{TP}{TP + FP} \quad \text{(Penaliza falsos positivos)}$$

donde FP son los falsos positivos.

Una alta precisión indica confiabilidad en las predicciones positivas del modelo, lo que minimiza "falsas alarmas". De tal modo, es crucial en contextos donde el costo de un falso positivo es alto, como en el *spam filtering* (etiquetar un correo legítimo como spam tiene consecuencias graves) o en ciertos diagnósticos médicos preliminares.

Por otro lado, el *recall* (o *Sensibilidad*) aborda una preocupación complementaria: "De todas las instancias realmente positivas, ¿cuántas logró capturar el modelo?". Su fórmula es:

$$\text{Recall (R)} = \frac{TP}{TP + FN} \quad \text{(Penaliza falsos negativos)}$$

siendo FN los falsos negativos.

Maximizar el *recall* implica minimizar los casos positivos que el modelo no logra identificar. Esta métrica resulta prioritaria en aplicaciones en las que omitir un positivo conlleva un riesgo elevado, como ocurre en la detección de enfermedades graves, donde un caso no diagnosticado representa un peligro mayor que la realización de pruebas adicionales, o en sistemas de seguridad orientados a la identificación de amenazas.

La dicotomía entre precisión y *recall* suele manifestarse como un *trade-off* inherente, entendido como una compensación en la que la mejora de una métrica implica el deterioro de la otra. En este sentido, la dicotomía entre precisión y *recall* constituye un ejemplo claro de *trade-off*, dado que incrementar la precisión, es decir, la proporción de predicciones positivas correctas, tiende a reducir el *recall*, al identificar un menor número de verdaderos positivos, y viceversa.

En términos prácticos, optimizar una métrica suele afectar negativamente a la otra. Un modelo excesivamente conservador, que solo predice la clase positiva cuando existe una certeza elevada, presenta una alta precisión, pero un recall bajo. Por el contrario, un modelo demasiado permisivo alcanza valores elevados de recall a costa de un aumento de las falsas alarmas, lo que se traduce en una baja precisión. Esta tensión justifica la necesidad de una métrica unificadora que permita equilibrar ambos aspectos, como el **F-Score**.

La medida más conocida es el **F1-Score**, definida como la **media armónica** de precisión y recall:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} = \frac{2TP}{2TP + FP + FN}$$

La media armónica, a diferencia de la aritmética, es severa con valores desiguales; por tal motivo, el F1 solo es alto cuando tanto P como R lo son. En otras palabras, es la métrica de equilibrio por defecto, apropiada cuando no hay una razón clara para privilegiar un error sobre otro. Sin embargo, el equilibrio no siempre es deseable. Para

ello, existe la **medida** $F\beta$ **generalizada**, definida como la media armónica ponderada de P y R:

$$F\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R} \tag{2.1}$$

El parámetro $\beta \in \mathbb{R}^+$ permite controlar el peso relativo del recall frente a la precisión.

Interpretación del Parámetro β

- $\beta = 1$ **(F1-score):** asigna igual importancia a P y R.

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

- $\beta > 1$**:** da **más peso al Recall**. Penaliza más los Falsos Negativos (FN).
- $\beta < 1$**:** da **más peso a la Precisión**. Penaliza más los Falsos Positivos (FP).

Para el **F2-score**, $\beta = 2$:

$$F2 = (1 + 2^2) \cdot \frac{P \cdot R}{2^2 \cdot P + R} = 5 \cdot \frac{P \cdot R}{4P + R} \tag{2.2}$$

Esta flexibilidad es poderosa, dado que un F2 es ideal para la detección de fraudes o la búsqueda de candidatos raros en minería de datos, por lo que al obviar un caso (FN) es más costoso que investigar una falsa alarma (FP). Por lo tanto, un F0.5 sería útil en un sistema de recomendación de contenido sensible, donde la relevancia (precisión) es más crítica que la exhaustividad.

Por lo tanto, el trío *precisión*, *recall* y F-*Score* constituye un lente analítico indispensable para evaluar modelos de clasificación más allá de la simple exactitud. En ese sentido, su elección y análisis informado permite alinear la evaluación del modelo con los objetivos económicos, éticos y operativos del dominio de aplicación, lo que transforma así una decisión técnica en un puente robusto entre el rendimiento algorítmico y el valor real entregado.

Justificación matemática de la menor penalización a falsos positivos (FP) en la métrica F2

La fórmula (2.1) puede reescribirse en su forma de media armónica ponderada de los inversos:

$$F\beta = \frac{1}{\frac{\beta^2}{1+\beta^2} \cdot \frac{1}{P} + \frac{1}{1+\beta^2} \cdot \frac{1}{R}} \tag{2.3}$$

Para $\beta = 2$ (F2-score), los pesos son:

$$F2 = \frac{1}{\frac{4}{5} \cdot \frac{1}{P} + \frac{1}{5} \cdot \frac{1}{R}} = \frac{1}{0.8 \cdot \frac{1}{P} + 0.2 \cdot \frac{1}{R}} \tag{2.4}$$

Análisis de los pesos para F1 y F2

- **F1 ($\beta = 1$):** Los pesos son 0.5 para $\frac{1}{P}$ y 0.5 para $\frac{1}{R}$.

$$F1 = \frac{1}{0.5 \cdot \frac{1}{P} + 0.5 \cdot \frac{1}{R}}$$

  Un aumento en FP (que disminuye P) afecta la métrica en la misma proporción que un aumento en FN (que disminuye R).
- **F2 ($\beta = 2$):** El peso de $\frac{1}{P}$ es 0.8 y el de $\frac{1}{R}$ es 0.2, como puede observarse en la ecuación (2.4). Aunque el peso de P es mayor, la clave está en el **trade-off implícito**. El término $\beta^2 = 4$ en el denominador de (2.2) significa que, para mantener un F_2 alto, una reducción en R (aumento de FN) debe compensarse con un aumento **mucho mayor** en P que en el caso $\beta = 1$.

En otras palabras, la función objetivo $F2$ es más sensible a cambios en el *recall* que a cambios en la *precisión*. Por ende, para maximizar $F2$, el optimizador "prefiere" cometer más Falsos Positivos (que reducen P moderadamente) si con ello se evitan falsos negativos (que preservan R).

En resumen, la justificación matemática de que el **F2-score penaliza menos los Falsos Positivos que el F1-score** se encuentra en el parámetro $\beta = 2$ de la medida Fβ. Este valor hace que el peso relativo del *recall* sea cuadráticamente mayor ($\beta^2 = 4$) que el de la Precisión en el denominador de la fórmula, haciendo la métrica más tolerante a una disminución en *precisión* (más FP) si con ello se logra un *recall* alto (menos FN). Por ello, F2 se emplea en dominios donde omitir un positivo real (FN) es más crítico que generar una falsa alarma (FP).

2.7. Clases desbalanceadas

2.7.1. Contexto del presente trabajo de grado

Los problemas de clasificación en el sector asegurador, como la detección de fraude, enfrentaron un desafío habitual y estructural: el desbalance de clases, en el que los casos positivos resultaron significativamente menos frecuentes que los negativos. Esta condición tuvo implicaciones directas en la estimación de probabilidades, la estabilidad de los modelos y la selección de métricas de evaluación. Asimismo, implicó una probabilidad a priori reducida, lo que ocasionó que los modelos basados en árboles tendieran a favorecer la clase mayoritaria cuando el desbalance no se corrigió.

Entre las soluciones más empleadas para mitigar este problema se encontraron el sobremuestreo de la clase minoritaria y el submuestreo de la clase mayoritaria. En este estudio se decidió emplear una técnica de sobremuestreo aleatorio denominada RandomOverSampler por dos razones principales. En primer lugar, se buscó preservar la totalidad de la información asociada a la clase mayoritaria, evitando la pérdida de patrones relevantes. En segundo término, dado el tamaño del conjunto de datos, un

submuestreo habría reducido la diversidad de los ejemplos disponibles, lo que podría haber afectado negativamente la capacidad de generalización del modelo.

No obstante, el sobremuestreo no estuvo exento de riesgos, dado que la duplicación de observaciones pudo incrementar la probabilidad de memorizar patrones específicos y derivar en sobreajuste, un aspecto especialmente relevante en modelos con alta capacidad, como Gradient Boosting, XGBoost o LightGBM.

Por otro lado, también existen implicaciones en la relación entre la probabilidad a priori y la probabilidad posterior que adquieren los clasificadores. En esa medida, en un problema de clasificación binaria, el objetivo del modelo es estimarla probabilidad condicional posterior:

$$P(Y = 1|X)$$

Empero, esta probabilidad posterior depende naturalmente de la probabilidad a priori. Al realizar un sobremuestreo, la proporción de la clase positiva aumenta artificialmente en los datos de entrenamiento, lo cual lleva al modelo a incrementar la estimación de la probabilidad posterior. En términos prácticos, el modelo aprende como si el fraude fuese más frecuente de lo que realmente es.

Esto no representa un error metodológico; de hecho, ayuda a aumentar la sensibilidad, pero lleva a examinar dos elementos. En primer lugar, las probabilidades estimadas deben interpretarse como puntajes relativos afectados por el remuestreo, por lo que es necesario seleccionar un umbral de decisión que no sea el estándar de 0.5, al utilizar métricas que prioricen la detección de la clase minoritaria, como el F2-score. Además, se requiere cuidado al momento de aplicar el sobremuestreo, al asegurar que sea únicamente dentro de los datos de entrenamiento y no en el conjunto de validación, puesto que, de lo contrario se incrementa de manera significativa el riesgo de sobreajuste.

2.7.2. Fundamentación general en el contexto del fraude en seguros de automóvil

Esta sección está fundamentada en las fuentes He et al., 2009 y Chawla et al., 2002 .

En el ámbito de la ciencia de datos aplicada a la detección de eventos raros —como el fraude financiero, las fallas industriales o las enfermedades poco comunes— se enfrentó el problema omnipresente de las clases desbalanceadas. Este fenómeno ocurrió cuando la distribución de las etiquetas objetivo resultó altamente asimétrica, con una clase mayoritaria (por ejemplo, transacciones legítimas) que dominó ampliamente sobre una clase minoritaria (por ejemplo, transacciones fraudulentas). En el contexto específico del fraude en seguros de automóvil, donde las reclamaciones fraudulentas representaron menos del 5 % del total, el desbalance no constituyó únicamente un inconveniente práctico, sino un desafío estadístico profundo que afectó la teoría fundamental de la clasificación.

2.7.3. La base estadística: alteración de la probabilidad a posteriori

En esencia, el objetivo de un clasificador probabilístico es estimar la *probabilidad a posteriori* $P(Y = 1|X)$ de que una instancia con características X pertenezca a la clase positiva (fraude). Por el teorema de Bayes, se obtiene lo siguiente:

$$P(Y = 1|X) = \frac{P(X|Y = 1) \cdot P(Y = 1)}{P(X)}$$

Donde $P(Y = 1)$ es la probabilidad a priori o prevalencia de la clase positiva en los datos de entrenamiento. En un conjunto desbalanceado, $P(Y = 1)$ es extremadamente pequeña. Cuando un algoritmo de aprendizaje (como una regresión logística, un árbol de decisión o un modelo de boosting) se entrena para minimizar la pérdida en estos datos, internaliza a priori. El resultado es un sesgo predictivo sistemático hacia la clase mayoritaria, esto es, el modelo aprende que "predecir siempre la clase negativa" es una estrategia que minimiza la pérdida global (como la *cross-entropy* o el error de clasificación), incluso a costa de no detectar ningún fraude.

Por otro lado, alterar artificialmente la distribución de clases mediante técnicas de remuestreo —ya sea sobremuestreando (*oversampling*) la clase minoritaria o submuestreando (*undersampling*) la mayoritaria— es, representa una intervención directa sobre $P(Y = 1)$ en el conjunto de entrenamiento. En consecuencia, esto recalibra el clasificador para que aprenda fronteras de decisión donde la probabilidad *a posteriori* refleje una prevalencia ajustada, más alineada con los costos asimétricos del error en el mundo real, donde pasar por alto un fraude, por ejemplo, es mucho más costoso que investigar una reclamación legítima sospechosa.

2.7.4. El riesgo de sobreajuste en el contexto del fraude

El sobreajuste (*overfitting*) adquirió una dimensión particularmente peligrosa en problemas desbalanceados. Las técnicas de sobremuestreo, como el Synthetic Minority Over-sampling Technique (SMOTE), generaron instancias sintéticas de la clase minoritaria mediante la interpolación entre ejemplos reales. Si bien este procedimiento ayudó al modelo a generalizar mejor la región correspondiente a la clase positiva, también pudo inducir un sobreajuste a patrones espurios o al ruido presente en el reducido conjunto de ejemplos disponibles. En el fraude de seguros de automóvil, donde los patrones fraudulentos resultaron diversos y evolucionaron constantemente –desde siniestros simulados hasta la exageración de daños–, un modelo sobreajustado aprendió a reconocer únicamente los patrones específicos observados durante el entrenamiento, lo que falla en la detección de nuevas modalidades de fraude.

Por su parte, el submuestreo descartó una proporción considerable de datos de la clase mayoritaria, lo que pudo conducir a una pérdida de información valiosa sobre la distribución real de los casos legítimos. En el sector asegurador, esta reducción implicó el riesgo de perder la capacidad de distinguir entre un patrón de conducción atípico, pero legítimo y uno verdaderamente fraudulento, lo que incrementa así la tasa de falsos

positivos y erosiona la confianza del cliente.

2.7.5. Estrategias integrales para el fraude en seguros de automóvil

El manejo efectivo del desbalance en este dominio requiere un enfoque estratificado que combine técnicas de datos, algoritmos y evaluación:

1. **A nivel de datos:**
   - Uso de SMOTE o variantes como ADASYN con moderación y validación rigurosa, lo que genera datos sintéticos que respeten la estructura de los reclamos (variables categóricas como tipo de póliza, variables continuas como monto reclamado).
   - Submuestreo inteligente, no aleatorio, que preserve los casos mayoritarios más informativos (por ejemplo, reclamaciones legítimas con características atípicas que podrían confundirse con fraude).
2. **A nivel algorítmico:**
   - Empleo de algoritmos insensibles al desbalance por diseño, como los árboles de decisión con división por criterios de Gini o Entropía, que naturalmente pueden capturar patrones minoritarios.
   - Uso de funciones de pérdida ponderadas, donde se asigna un costo mayor al error de clasificar un fraude como legítimo (falso negativo). Por ejemplo, en *XGBoost* esto se logra mediante el parámetro *scale_pos_weight*.
   - Implementación de aprendizaje por costos (*cost-sensitive learning*), donde la matriz de costos real del negocio (costo de investigación *vs*. costo del fraude no detectado) se integra directamente en la función objetivo del modelo.
3. **A nivel de evaluación:**
   - Abandono del *Accuracy* como métrica principal. En su lugar, uso de Curvas ROC (Receiver Operating Characteristic), Curvas Precision-Recall (especialmente informativas en desbalance extremo) y métricas como el F2-Score o el Índice de G-Mean, que priorizan la detección de la clase minoritaria.
   - Validación mediante estrategias de *cross-validation* estratificadas que preserven la proporción de clases en cada *fold*, lo que asegura una estimación realista del rendimiento.

2.7.6. Un equilibrio entre teoría y práctica

El manejo de clases desbalanceadas en la detección de fraude en seguros de automóvil trascendió la mera aplicación de técnicas algorítmicas, pues requirió una comprensión profunda de la base estadística subyacente. En particular, resultó necesario analizar cómo la prevalencia observada en los datos de entrenamiento distorsionó las probabilidades a posteriori y de qué manera una intervención controlada permitió corregir este sesgo,

aunque no sin riesgos. De tal modo, la amenaza del sobreajuste fue real y se vio agudizada por la escasez y la alta diversidad de los ejemplos positivos.

En ese orden de ideas, la solución óptima residió en un enfoque híbrido que combinó un remuestreo cuidadoso con algoritmos robustos y una evaluación orientada al costo real del negocio. El objetivo final no fue únicamente maximizar una métrica técnica, sino construir un sistema de alerta temprana que minimizara las pérdidas económicas asociadas al fraude, mantuviera primas justas para los clientes honestos y preservara la sostenibilidad financiera de la aseguradora, todo ello sustentado en una práctica de ciencia de datos rigurosa y consciente de los supuestos estadísticos que la fundamentaron.

3

Análisis y procesamiento de datos

3.1. Conjunto de Datos

Para este estudio, se utilizó la siguiente base de datos: "*Vehicle Insurance Fraud Claim Detection*", alojado en Kaggle. `https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection`

Descripción de la base de datos "*Vehicle Insurance Fraud Claim Detection*"

Aclaración: aunque este conjunto de datos es útil para la investigación académica y la experimentación metodológica, es importante tener en cuenta que los datos tienen alrededor de 30 años. Por tal motivo, se expone que, al interpretar los resultados, es crucial considerar que el sector de seguros ha pasado por cambios importantes, como avances en tecnologías de verificación y comunicación, la evolución de las características técnicas de los automóviles, cambios en los métodos de fraude, actualizaciones en la regulación del sector y una mayor disponibilidad de información.

Descripción del Dataset: este conjunto de datos se creó con el propósito de ayudar a identificar actividades fraudulentas en las reclamaciones de seguros de vehículos. En esa medida, contiene 15 420 registros y 33 variables que reflejan las reclamaciones de seguros de automóviles entre los años 1994 y 1996. Por otro lado, las variables se pueden agrupar temáticamente en los siguientes grupos:

1. **Atributos de la póliza:** número de póliza y deducible aplicable.
2. **Información del asegurado:** edad, sexo y estado civil.
3. **Información del incidente:** tipo de incidente, tipo de colisión, autoridades contactadas, ubicación del suceso y hora del día en que ocurrió.
4. **Atributos del vehículo:** marca, tipo y año del vehículo.
5. **Atributos de la reclamación:** indicación sobre si la reclamación fue fraudulenta.

La variable clave tiene el nombre de FraudFound_P, y actúa como una etiqueta binaria (0 o 1) que facilita clasificar si un reclamo se considera fraudulento, convirtiéndose en el

enfoque principal para el problema de clasificación supervisada que se aborda en este proyecto.

Como parte del preprocesamiento, las variables numéricas se han convertido en rangos categóricos, semejante a otras variables existentes, como días entre la emisión de la póliza y el accidente, o el precio del vehículo, lo que hace que su análisis y procesamiento sean mucho más sencillos. En esa medida, este conjunto de datos brinda una visión completa de los factores que pueden influir en la aparición de actividades fraudulentas en las reclamaciones de seguros de vehículos. Sin embargo, se observa un desequilibrio notable entre las clases, dado que las reclamaciones fraudulentas representan aproximadamente el 6.4 % del total de registros, lo que plantea un desafío considerable para el modelado predictivo.

3.1.1. Tipos de datos

El dataset no contiene valores nulos en ninguna de sus columnas. En tal sentido, la distribución inicial muestra nueve variables numéricas (*int64*) y 24 variables de tipo objeto. Las variables WeekOfMonth y WeekOfMonthClaimed deben convertirse a categóricas ordinales a pesar de ser numéricas, dado que representan semanas discretas del mes; DriverRating debe tratarse como categórica ordinal (1 4) en lugar de numérica continua. Asimismo, PolicyNumber y RepNumber funcionan como identificadores únicos (el primero de póliza, el segundo de agente), por lo que deben excluirse con el fin de evitar que los modelos aprendan estas variables y no generalicen correctamente al momento de evaluarse. Finalmente, las variables PoliceReportFiled y WitnessPresent son binarias (Sí/No) y podrían convertirse a tipo *bool*. Adicionalmente, se encuentran las siguientes variables:

- `BasePolicy`,
- `NumberOfCars`,
- `AddressChange_Claim`,
- `NumberOfSuppliment`,
- `AgeOfPolicyHolder`,
- `AgeOfVehicle`,
- `PastNumberOfClaims`,
- `Days_Policy_Claim`,
- `Days_Policy_Accident`, y
- `AgentType`

Aunque almacenadas como objeto, representan categorías y deben tratarse como variables categóricas. En efecto, las variables *Age*, *Deductible* y *Yea*r se mantienen correctamente como enteros. Finalmente, FraudFound_P (variable objetivo) se mantiene como *int64* para compatibilidad con algoritmos de *machine learning*, aunque podría convertirse a *bool* para mayor claridad semántica en el análisis exploratorio.

3.2. Preparación de los datos para el modelado

3.2.1. Limpieza de Datos

Tras un proceso inicial de limpieza, se detectaron ciertas anomalías que requirieron atención. En particular, los campos *DayOfWeekClaimed*, *MonthClaimed* y *Age* presentaron entradas con valor cero, lo cual resultó contextualmente inválido. Los casos con valores cero en las categorías *DayOfWeekClaimed* y *MonthClaimed* correspondieron únicamente a un registro. Dada la imposibilidad práctica de que dichos valores representaran un día o un mes, estos registros se consideraron erróneos y se eliminaron del conjunto de datos. Por su parte, todos los registros con valor de edad igual a cero se sustituyeron por la moda de las edades restantes. Asimismo, para manejar la amplia disparidad observada en la variable *Age*, esta se categorizó en cuatro grupos: 18–25 (1), 26–40 (2), 41–65 (3) y mayores de 65 (4).

Ahora bien, un análisis adicional reveló que la variable *PolicyType* correspondía a una concatenación de *VehicleCategory* y *BasePolicy*. Dado que esta variable combinaba información ya contenida en otros campos, se decidió eliminarla del conjunto de datos con el fin de optimizar su estructura y preservar la independencia entre variables.

Durante la preparación de los datos, se aplicó una codificación numérica a las características temporales *Month* y *DayOfWeekClaimed* para preservar su orden secuencial natural. Esta codificación permitió que el modelo reconociera y aprovechara patrones temporales relevantes en la detección de fraude. En particular, la variable *Month* se transformó asignando un número entero único a cada mes, desde enero (1) hasta diciembre (12). De manera análoga, los días de la semana se codificaron desde lunes (1) hasta domingo (7). Este enfoque aseguró una interpretación adecuada de la progresión temporal por parte del modelo, facilitando la identificación de tendencias, como incrementos en la incidencia de fraude en determinadas épocas del año o días de la semana.

3.2.2. Análisis exploratorio

En esta sección se realizó un análisis exploratorio de los datos correspondientes a la base considerada para la detección de fraude en seguros de automóvil. Esta etapa resultó fundamental, dado que permitió examinar las variables de manera individual e identificar patrones potencialmente sospechosos. En particular, se analizaron distribuciones, valores atípicos y tendencias en características clave, tales como el lugar del siniestro, la existencia de reporte policial, la edad de los involucrados, el estado civil, la condición de titular o tercero en el siniestro, el tipo de vehículo y la antigüedad en la residencia, entre otras. Se prestó especial atención a aquellos casos que exhibieron comportamientos inusuales, como siniestros reportados inmediatamente después de la contratación de la póliza, vehículos de alto valor con primas inusualmente bajas o conductores con múltiples reclamaciones concentradas en períodos cortos de tiempo.

El cruce de variables permitió revelar interacciones complejas que no habrían sido

evidentes mediante análisis aislados. Por lo tanto, al relacionar de forma sistemática dimensiones como la edad del conductor con el tipo de vehículo, el historial de siniestros con la ubicación geográfica, o el día de la semana del accidente con la severidad del daño, emergieron patrones de comportamiento fraudulento más refinados. Por ejemplo, se pudo identificar que conductores jóvenes con vehículos de lujo presentaron una frecuencia desproporcionadamente alta de colisiones laterales, o que determinadas zonas geográficas concentraron un número atípico de accidentes en días específicos. Estas relaciones multivariadas, una vez validadas estadísticamente, se constituyeron en indicadores relevantes para alimentar los modelos predictivos y refinar las estrategias de investigación de siniestros potencialmente fraudulentos.

A continuación, se presentan algunos cruces entre unas variables y la variable fraude.

1. **Fraude y grupos etáreos**
   La evidencia indica una mayor propensión al fraude en seguros entre personas jóvenes, especialmente menores de 25 años. En tal sentido, los estudios demuestran que este grupo demográfico tiende a percibir el fraude en seguros como un delito menos grave. Aproximadamente el 65 % de los individuos entre 18 y 25 años consideran el fraude en seguros como un crimen, en comparación con más del 95 % de los mayores de 55 años que reconocen su severidad (MehaffyWeber, s.f.) [1]. Por consiguiente, se advierte que esta diferencia sustancial resalta la importancia de incluir la edad como un factor clave en el modelo, puesto que influye significativamente en las percepciones y comportamientos asociados al fraude en seguros.

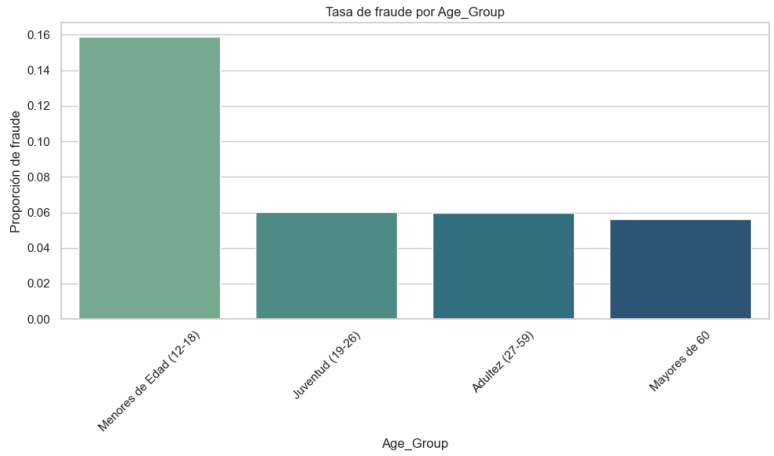


Figura 3.1: *Detección de fraude por grupo etáreo.*

2. **Fraude y sexo**
   En la siguiente tabla se resume la relación entre las variables *Fraude* y *Sex*:

[1] https://www.mehaffyweber.com/news/survey-finds-younger-generations-have-a-higher-tolerance-for-insurance-fraud-what-insurers-should-know/

Sex	FraudFound_P=0	FraudFound_P=1
Female	2315	105
Male	12182	818

La representación gráfica (Figura 3.2), muestra que la proporción de fraude es mayor en hombres que en mujeres aunque la diferencia observada entre ambos géneros es mínima.

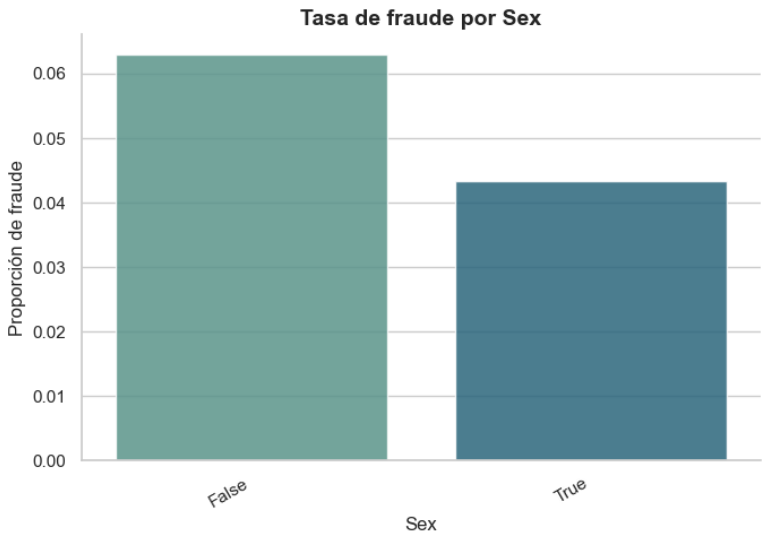


Figura 3.2: *Detección de fraude por sexo.*
Nota: False= hombre True= mujer

3. **Fraude y área de ocurrencia del siniestro**
En la siguiente tabla se resume la relación entre las variables *Fraude* y *AccidentArea*:

AccidentArea	FraudFound_P=0	FraudFound_P=1
Rural	1465	133
Urban	13032	790

La representación gráfica (Figura 3.3)expone que el conjunto de datos exhibe una probabilidad mayor de fraude en el área rural que en el área urbana.

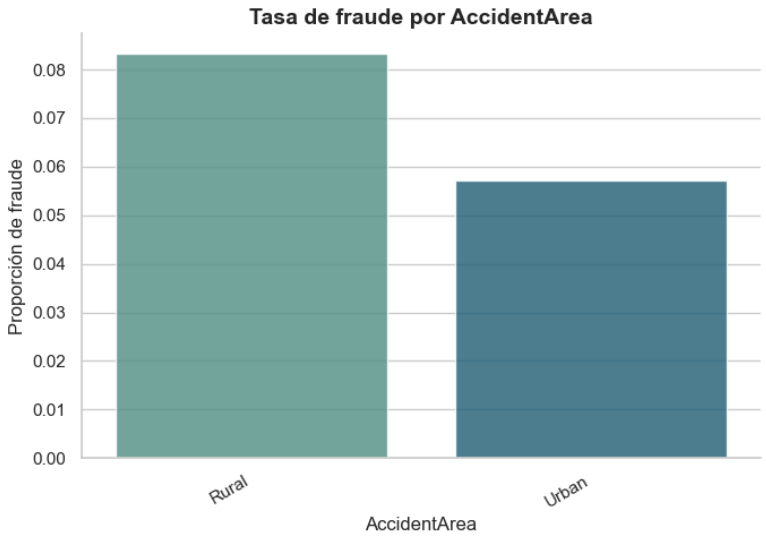


Figura 3.3: *Detección de fraude por área.*

Para reflejar este patrón, se empleó la técnica de *label encoding* en variables categóricas binarias. La variable *Sex* se codificó asignando el valor 1 a los hombres y 0 a las mujeres, mientras que *AccidentArea* se codificó como urbana (1) y rural (0). Esta estrategia permitió que el modelo evaluara de manera directa la correlación entre dichas categorías y la probabilidad de fraude.

La codificación por etiquetas se utilizó como un método sencillo y eficiente para representar variables categóricas en formato numérico. Al asignar un valor numérico a cada categoría, los modelos de aprendizaje automático pudieron procesarlas con mayor facilidad, reduciendo la complejidad del modelo y favoreciendo su interpretabilidad. A diferencia de técnicas más complejas, como la codificación *one-hot*, este enfoque resultó computacionalmente más conciso y permitió conservar relaciones ordinales cuando estas estuvieron presentes.

4. **Fraude y estado civil**

   En el análisis por estado civil, representado en la siguiente tabla y Figura 3.4, se observaron variaciones significativas en la probabilidad de fraude según el tipo de vínculo matrimonial.

MaritalStatus	FraudFound_P=0	FraudFound_P=1
Divorced	73	3
Married	9986	639
Single	4406	278
Widow	32	3

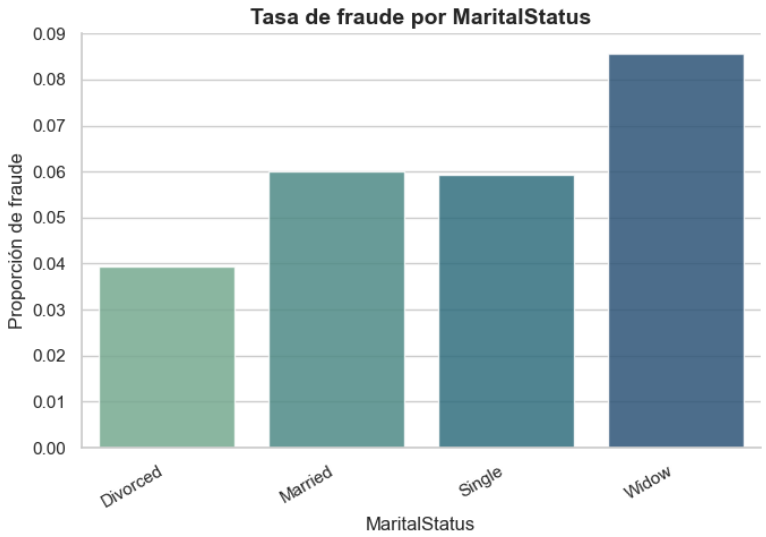


Figura 3.4: *Detección de fraude según el estado civil.*

5. **Fraude y titularidad del asegurado**
   De acuerdo con la información presente en la base de datos, existen diferencias entre los distintos tipos de titulares de póliza, en lo que respecta a la tasa de fraude, mientras que en los titulares principales fue de 7.89 %, en lo terceros fue significativamente menor (0.88 %).

Tipo de responsabilidad	FraudFound_P=0	FraudFound_P=1
Titular de la póliza	10344	886
Tercero	4153	37

   Esta disensión subraya la necesidad de que las compañías de seguros fortalezcan las medidas preventivas, especialmente entre los titulares principales, como se muestra en la Figura 3.5.

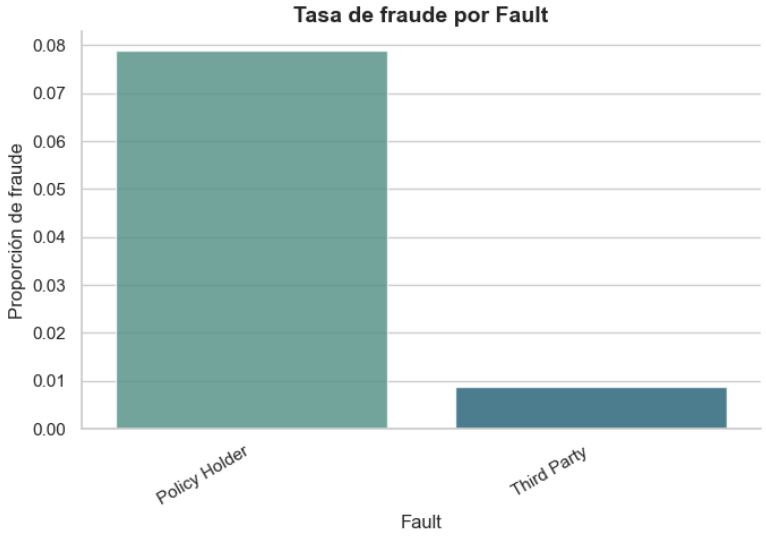


Figura 3.5: *Detección de fraude según tipo de responsabilidad.*

El mayor conocimiento de las cláusulas de la póliza, derivado de la relación contractual directa con la aseguradora, puede llevar a los titulares principales a aprovechar vacíos legales y a cometer fraude. En tal sentido, se expone que esta situación contrasta con la de los terceros, quienes, al no tener acceso a los detalles de la póliza, muestran un menor incentivo para actuar de forma fraudulenta.

6. **Fraude y tipo de vehículo**
   Asimismo, el análisis de las categorías de vehículos mostró que los autos deportivos presentan una tasa de fraude de 1.57 %, considerablemente inferior al 8.22 % registrado para los sedanes y el 11.25 % de los utilitarios.

Categoria del vehículo	FraudFound_P=0	FraudFound_P=1
Sport	5274	84
Sedan	8876	795
Utility	347	44

Los datos de la tabla anterior y la Figura 3.6 evidencian que el tipo de vehículo es un factor determinante en la probabilidad de ocurrencia de fraude.

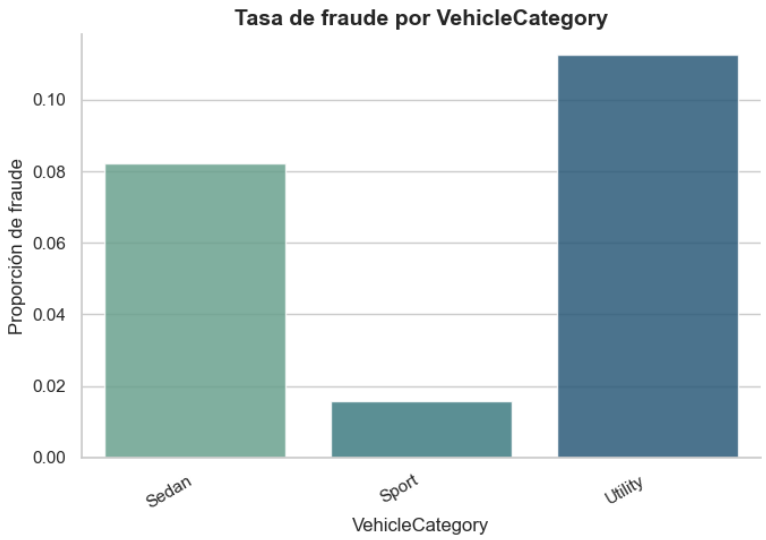


Figura 3.6: *Detección de fraude según la categoría del vehículo.*

7. **Fraude y cambio de residencia**
   Ahora bien, al analizar la variable AddressChange contra la variable FraudFound_P, se observa la siguiente distribución de información:

Antiguedad del cambio de dirección	FraudFound_P=0	FraudFound_P=1
Sin cambio	13499	825
Menos de 6 meses	1	3
1 año	159	11
Entre 2 y 3 años	240	51
Entre 4 y 8 años	598	33

La información representada en la (Figura 3.7), evidencia que la mayoría de los asegurados no presentan cambios de dirección; asimismo, al interior de este grupo se concentra el mayor volumen absoluto de casos tanto fraudulentos como no fraudulentos. Por otro lado, los cambios de dirección más recientes presentan una mayor proporción de fraudes en relación con su tamaño. Por consiguiente, a medida que aumenta la antigüedad del cambio de dirección, la proporción de fraudes tiende a disminuir, lo que sugiere que los cambios recientes podrían estar asociados a un mayor riesgo de comportamiento fraudulento.

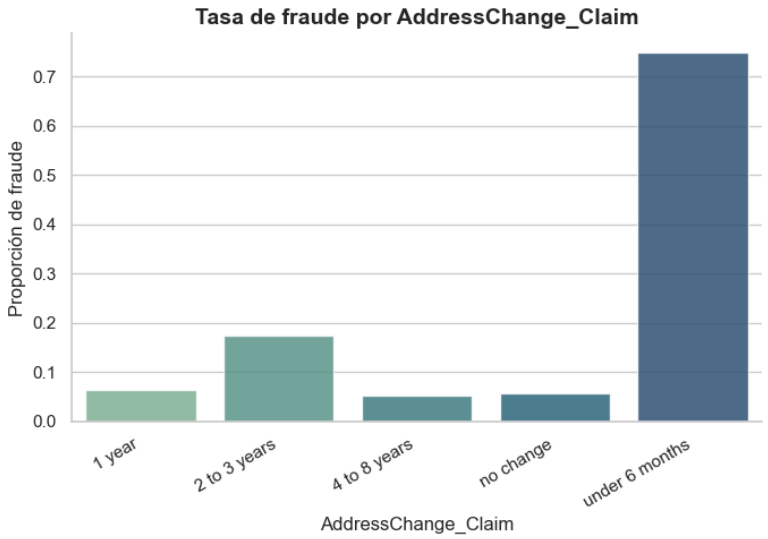


Figura 3.7: *Detección de fraude según cambio de dirección.*

8. **Fraude y tipo de seguro (cobertura)**
   La distribución de reclamaciones fraudulentas por tipo de cobertura (todo riesgo, colisión y responsabilidad civil) genera la siguiente información

Tipo de póliza	FraudFound_P= 0	FraudFound_P = 1
Todo riesgo	3997	452
Colisión	5527	435
Responsabilidad civil	4973	36

Al analizar los distintos tipos de reclamaciones (Figura 3.8), se observa que los fraudes se concentran en las categorías *All Perils* y *Collision*, mientras que *Liability* presenta muy pocos casos de fraude; en cuanto a las reclamaciones por Liability, estas presentan una tasa de fraude mucho menor (0.72 %) en comparación con All Perils (10.16 %) y Collision (7.3 %).

A la luz de este escenario, se infiere que la menor incidencia de fraude en las reclamaciones por responsabilidad civil se atribuye a la naturaleza de este tipo de seguro, que cubre daños o lesiones causadas a terceros, por lo que resulta más difícil de falsificar que las reclamaciones por colisión o todo riesgo.

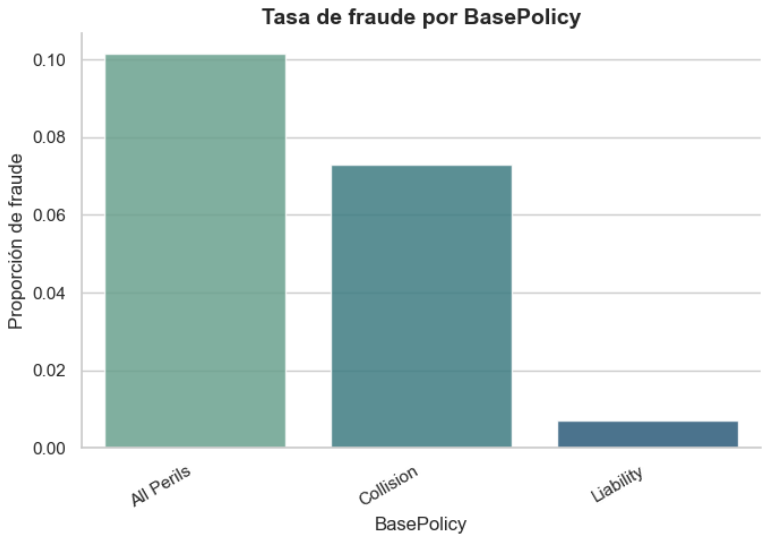


Figura 3.8: *Detección de fraude tipificada por tipo de póliza.*

9. **Fraude y reporte policial**
   Con respecto al cruce entre la variable fraude y la variable reporte policial, observe la distribución de la información en la siguiente tabla:

Reporte policial	FraudFound_P= 0	FraudFound_P = 1
No	14085	907
Si	412	16

cuya representación gráfica es la (Figura 3.9):

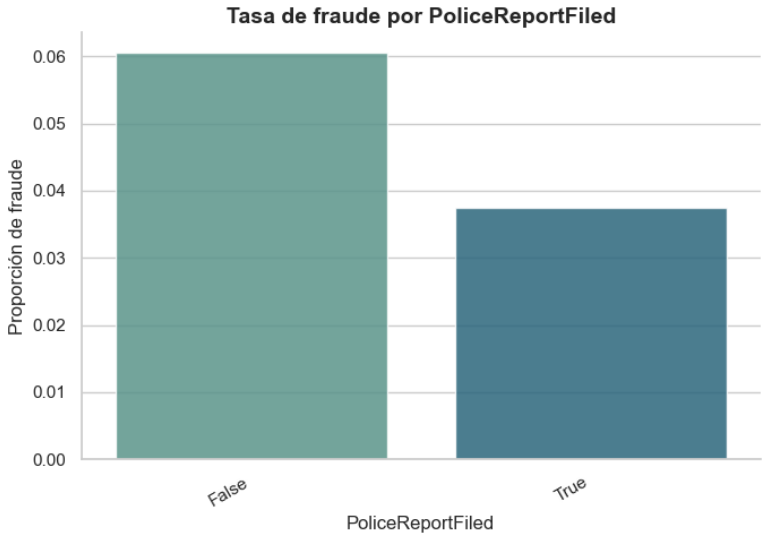


Figura 3.9: *Detección de fraude y reporte policial.*

En la mayoría de los casos no se presentó reporte policial, y dentro de estos se encuentran también la gran mayoría de los fraudes detectados. Los casos donde sí hubo reporte policial (Sí) son relativamente pocos, y en ellos la incidencia de

fraude es mínima. Sin embargo, se evidenció que la probabilidad de que haya fraude, dado que no hubo reporte policial, fue de alrededor del 6.05 %; por otra parte, la probabilidad de que haya fraude, puesto que hubo reporte policial, fue aproximadamente de 3.74 %.

10. **Fraude y Deductible**
    Por otro lado, se observa que el deducible corresponde al monto que debe asumir el asegurado en caso de un siniestro, y puede influir en el comportamiento de reporte de fraude, dado que deducibles más bajos podrían incentivar reclamaciones de menor severidad.

Deductible	FraudFound_P= 0	FraudFound_P = 1
300	6	2
400	13982	856
500	216	47
700	293	18

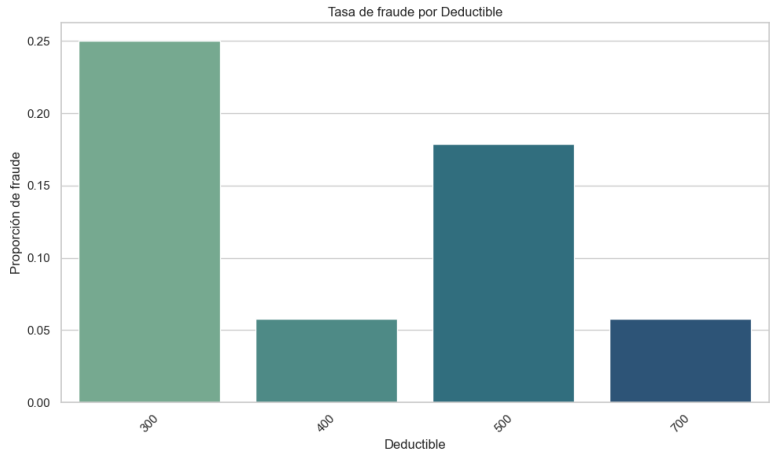


Figura 3.10: *Detección de fraude y Deductible.*

En la Figura 3.10 se observa la distribución de los casos de fraude y no fraude según el valor del deducible. Al respecto, la mayor concentración de observaciones se presenta en los deducibles de 400 y 700, donde la proporción de fraudes es relativamente estable. En el caso del deducible más bajo (300), aunque el número de observaciones es reducido, se presenta una mayor proporción relativa de fraude.

11. **Fraude y Año**
    La variable *Year* indica el año de ocurrencia del siniestro y permite evaluar posibles variaciones temporales en la incidencia del fraude.

Año	FraudFound_P= 0	FraudFound_P = 1
1994	5733	409
1995	4894	301
1996	3870	213

En la Figura 3.11 se presenta la distribución de fraudes y no fraudes para los años analizados.

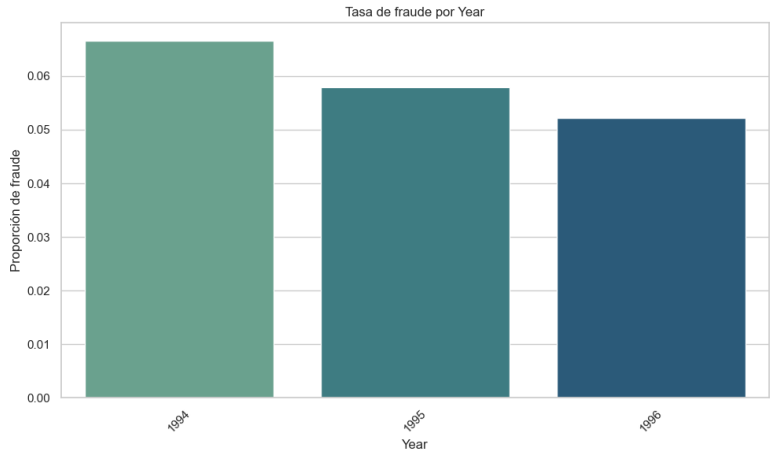


Figura 3.11: *Detección de fraude y Año.*

Se observa que el número total de reclamaciones disminuye cada año y la proporción de fraudes se mantiene relativamente constante a lo largo del tiempo, con un pequeño descenso.

12. **Fraude y Calificación del conductor**

Calificación del conductor	FraudFound_P= 0	FraudFound_P = 1
1	3712	232
2	3587	214
3	3642	242
4	3556	235

Por su parte, DriverRating representa una calificación asociada al historial o comportamiento del conductor, la cual podría reflejar distintos niveles de riesgo.

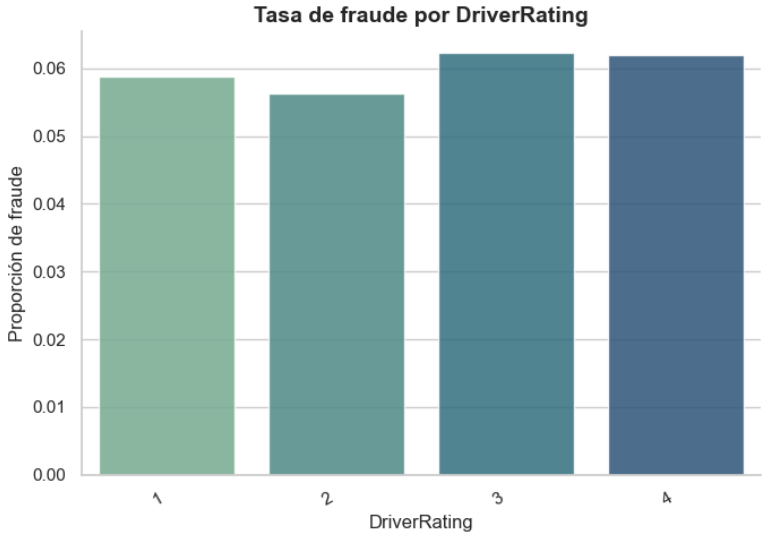


Figura 3.12: *Detección de fraude y Calificación del conductor.*

Por otro lado, la Figura 3.12 muestra la relación entre la calificación del conductor y la presencia de fraude. En tal sentido, se observa que, a medida que aumenta el nivel de DriverRating, la proporción de casos fraudulentos presenta variaciones moderadas, siendo ligeramente mayor en los niveles más altos de la calificación.

13. **Fraude y Número de reclamaciones anteriores**
La variable *PastNumberOfClaims* indica el número de reclamaciones previas realizadas por el asegurado. En la Figura 3.13 se evidencia que los asegurados que no tuvieron reclamaciones previas tienden a presentar una mayor proporción de fraude. En particular, se indica que los grupos con pocas reclamaciones previas muestran una incidencia más elevada en comparación con aquellos que tienen múltiples reclamaciones anteriores.

Reclamaciones anteriores	FraudFound_P= 0	FraudFound_P = 1
1	3351	222
2 a 4	5191	294
Más de 4	1942	68
Ninguna	4013	339

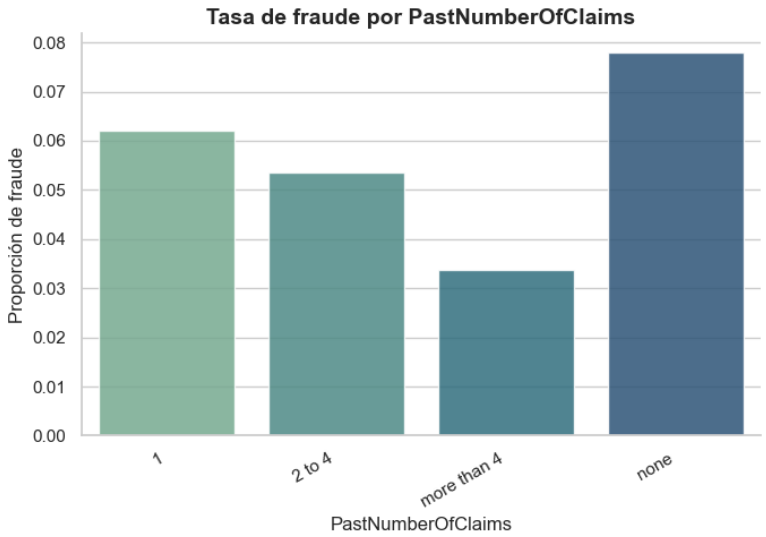


Figura 3.13: *Detección de fraude y Número de reclamaciones anteriores.*

14. **Fraude y Tipo de intermediario**
AgentType clasifica el tipo de intermediario, a través del cual se adquirió la póliza, lo que permite diferenciar entre agentes internos y externos.

Tipo de intermedario	FraudFound_P= 0	FraudFound_P = 1
Externo	14260	919
Interno	237	4

En la Figura 3.14se presenta la distribución de fraude según el tipo de agente. Así, se observa que la mayoría de las reclamaciones provienen de agentes externos,

lo cual es coherente con su mayor volumen de participación. Sin embargo, la proporción de fraude es mayor en las pólizas gestionadas por agentes externos en comparación con las internas.

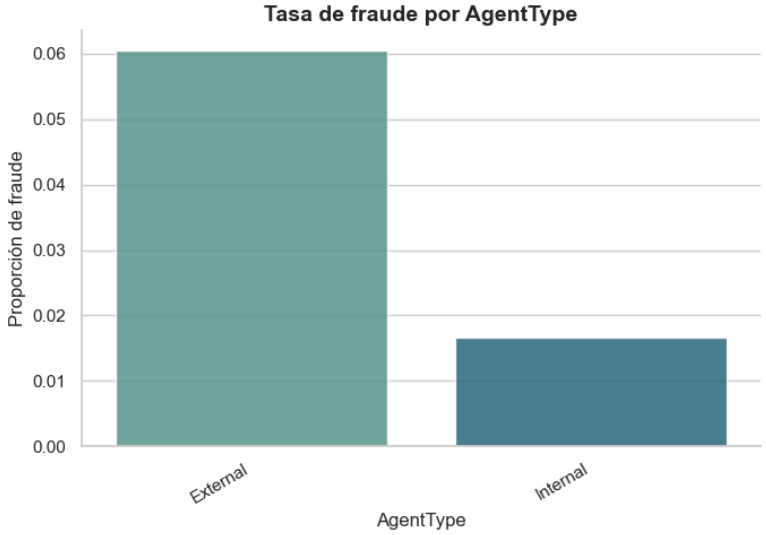


Figura 3.14: *Detección de fraude y Tipo de intermediario.*

Las variables analizadas en esta sección forman parte del conjunto final utilizado para el entrenamiento de los modelos predictivos. Otras variables temporales y administrativas se consideraron en el modelado; sin embargo, no se analizan individualmente en el EDA debido a su menor interpretabilidad directa y su rol complementario dentro del modelo predictivo.

A partir del análisis realizado, es posible concluir que el comportamiento fraudulento tiende a concentrarse en asegurados que muestran señales de inestabilidad en su información y en las condiciones del siniestro. Por ende, en casos donde hay cambios recientes de dirección se observa una mayor inclinación hacia el fraude; de la misma manera, ciertas combinaciones de edad del asegurado y reclamos relacionados con tipos específicos de siniestros también son relevantes.

Además, algunas características del vehículo revelan patrones distintos entre casos fraudulentos y no fraudulentos. En conjunto, estos hallazgos sugieren que el perfil del fraude no se basa en un solo factor aislado, sino en la interacción de variables demográficas, administrativas y operativas. Por lo tanto, este perfil, identificado a partir del EDA, servirá como base para contrastar más adelante con la importancia de las variables asignadas por cada uno de los modelos predictivos.

4

Modelado

4.1. Introducción

La estrategia de modelado se diseñó con el objetivo de evitar fugas de datos, garantizar la consistencia del proceso y evaluar de manera rigurosa el desempeño de los algoritmos en un contexto de datos naturalmente desbalanceados. En esta sección se registraron las decisiones operativas adoptadas en relación con el remuestreo, la construcción de *pipelines*, el esquema de validación cruzada y el procedimiento de selección del umbral óptimo para maximizar el F_2-score.

En primer lugar, la técnica de sobremuestreo seleccionada para todos los modelos fue *RandomOverSampler* de la librería *imbalanced-learn*, como se mencionó arriba. Así, se utilizó con su configuración predeterminada (*sampling_strategy* = 'auto'), la cual balanceó completamente las clases mediante la replicación de la clase minoritaria hasta igualar el tamaño de la mayoritaria. Ahora bien, el objetivo de esta decisión fue evitar la eliminación de información perteneciente a la clase mayoritaria, incrementando la representatividad de la clase minoritaria en el conjunto de entrenamiento. Esta técnica se escogió debido a que el dataset contenía información relevante en la clase mayoritaria que se habría perdido con un submuestreo; además, métodos como SMOTE pueden generar interpolaciones artificiales problemáticas en variables categóricas, las cuales predominaban en este conjunto de datos.

Para el entrenamiento de cada modelo se construyó un *pipeline* compuesto por tres etapas secuenciales. Cabe aclarar que los datos incluidos en este *pipeline* correspondieron a la partición del 70 % destinada al entrenamiento. La primera etapa consistió en el sobremuestreo, en la cual ingresaron únicamente los datos de entrenamiento desbalanceados correspondientes al pliegue respectivo y se aplicó la técnica de *RandomOverSampler*. En la segunda etapa se llevó a cabo el preprocesamiento, que incluyó la conversión de variables categóricas, el tratamiento de variables ordinales y la exclusión de identificadores. Finalmente, en la tercera etapa se ejecutó el entrenamiento del clasificador correspondiente. Posteriormente, el modelo se evaluó utilizando los datos de validación del pliegue, los cuales se mantuvieron en su distribución original y desbalanceada. Todas las variables empleadas en el entrenamiento de los modelos correspondieron a

aquellas previamente analizadas en el EDA.

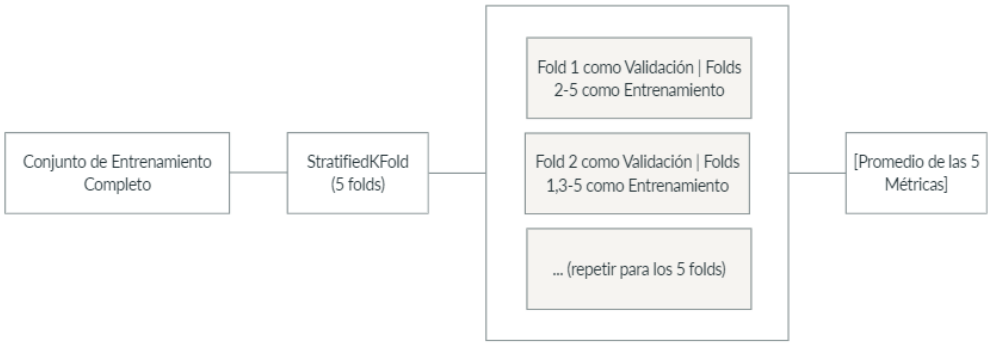


Figura 4.1: *Diagrama de flujo validación cruzada*

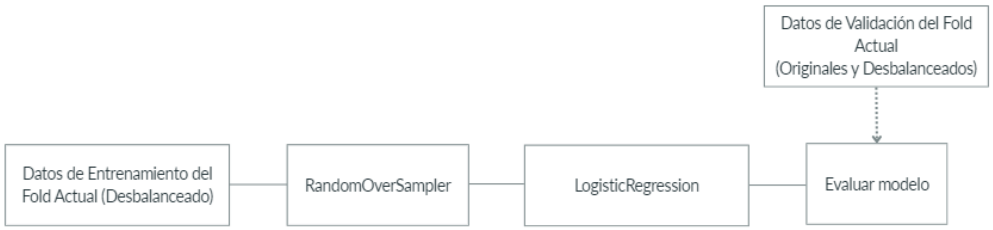


Figura 4.2: *Diagrama de flujo pipeline*

Este diseño asegura que el remuestreo se aplique únicamente a los datos dentro de cada *fold* destinados al entrenamiento, sin que estas réplicas afecten los datos de validación. Esto evita fugas de información y sobreestimación del desempeño.

Con el fin de seleccionar los modelos finales y los hiperparámetros óptimos de cada uno, se aplicó la herramienta GridSearchCV, dado que realiza una búsqueda en cuadrícula. Como la métrica definida en el marco teórico se alinea con los objetivos de negocio, se empleó el F_2-score por otorgar mayor peso a la sensibilidad en la detección de la clase minoritaria. Esta fue la que orientó el proceso de optimización mediante make_scorer con beta=2.

Para obtener una estimación robusta y generalizable del rendimiento de cada combinación de parámetros, se utilizó una validación cruzada de cinco particiones con StratifiedKFold (n_splits=5, shuffle=True, random_state=42). Esta técnica divide el conjunto de entrenamiento en cinco subgrupos, manteniendo la misma proporción de observaciones de la clase positiva y negativa presente en los datos reales. De este modo, cada iteración fue representativa de la distribución global del problema.

A continuación, se detallan los parámetros explorados para cada modelo:

- **Random Forest**: n_estimators (200, 300, 400), max_depth (None, 15, 20), min_samples_leaf (1, 2), max_features ('sqrt', 'log2'). Adicionalmente, se configuró n_jobs=-1 para paralelización y warm_start=False.
- **Gradient Boosting**: n_estimators (200, 300, 400), learning_rate (0.05, 0.1), max_depth (3, 4, 5), min_samples_leaf (1, 2), min_samples_split (2, 5). Se establecieron sub-

sample=0.9 y max_features='sqrt' como parámetros fijos para regularización.

- **XGBoost**: n_estimators (200, 300, 400), learning_rate (0.05, 0.1), max_depth (4, 6, 8), subsample (0.8, 1.0), colsample_bytree (0.8, 1.0), reg_lambda (1, 5), min_child_weight (1, 3). Se configuró eval_metric='logloss' y verbosity=0.
- **LightGBM**: n_estimators (200, 300, 400), learning_rate (0.05, 0.1), num_leaves (31, 50, 80), max_depth (-1, 10, 15), subsample (0.8, 1.0), colsample_bytree (0.8, 1.0), reg_lambda (1, 5). Se utilizó force_row_wise=True para optimización.

Cada GridSearchCV se ejecutó con n_jobs=-1 para aprovechar todos los núcleos disponibles y verbose=1 para monitoreo del progreso.

Antes del proceso de entrenamiento y de que los datos ingresaran al pipeline, se realizó la división de los conjuntos de datos en dos grupos: el 70 % se destinó a entrenamiento y el 30 % a prueba; para ello, se utilizó una partición estratificada que mantuvo la distribución de las clases.

El conjunto de prueba se utilizó para buscar el umbral óptimo después de obtener las probabilidades generadas por el mejor pipeline; previamente, no se había hecho uso de ese conjunto y nunca se le aplicó una técnica de remuestreo. En ese sentido, el proceso, realizado una única vez para evitar doble optimización, consistió en obtener $\hat{p}(Y = 1 \mid X)$ para las observaciones del conjunto de prueba, construir la curva precision-recall mediante precision_recall_curve, calcular el F_2-score para cada umbral candidato y seleccionar el umbral t* que maximiza el F_2-score en dicho conjunto.

Todo este esquema de pipeline, validación cruzada y umbral óptimo se aplicó a todos los algoritmos evaluados, con la única diferencia de que cada uno contó con su propio espacio de búsqueda de hiperparámetros dentro de GridSearchCV. Como resultado, se aseguraron las condiciones de evaluación homogéneas, la comparabilidad entre modelos, la ausencia de sesgos metodológicos y un análisis robusto del desempeño relativo.

4.2. Resultados

En primer lugar, se realizó una regresión logística, un modelo lineal simple y altamente interpretable, con el objeto de establecer una línea base de rendimiento. Al respecto, la Tabla 4.1 muestra los resultados de validación cruzada por pliegue, mientras que la Tabla 4.2 resume las métricas obtenidas en la muestra de prueba.

Fold 1	0.3890
Fold 2	0.3881
Fold 3	0.3932
Fold 4	0.4301
Fold 5	0.3829
Promedio	0.3967 ± 0.0170

Cuadro 4.1: *Resultados del desempeño del modelo por fold y su promedio.*

F_2-score en validación cruzada	0.3967
F_2-score en muestra de prueba	0.4014
Recall	89.89 %
Precisión	12.49 %

Cuadro 4.2: *Métricas de desempeño de la regresión logística en validación y prueba.*

El modelo mostró una estabilidad notable con una desviación estándar baja (0.0170), lo que indicó consistencia en su capacidad predictiva. Aunado a esto, el rendimiento se mantuvo en los datos de prueba con un F_2-score de 0.4014, lo que señaló ausencia de sobreajuste significativo. El umbral óptimo de decisión para maximizar el F_2-score resultó ser 0.44.

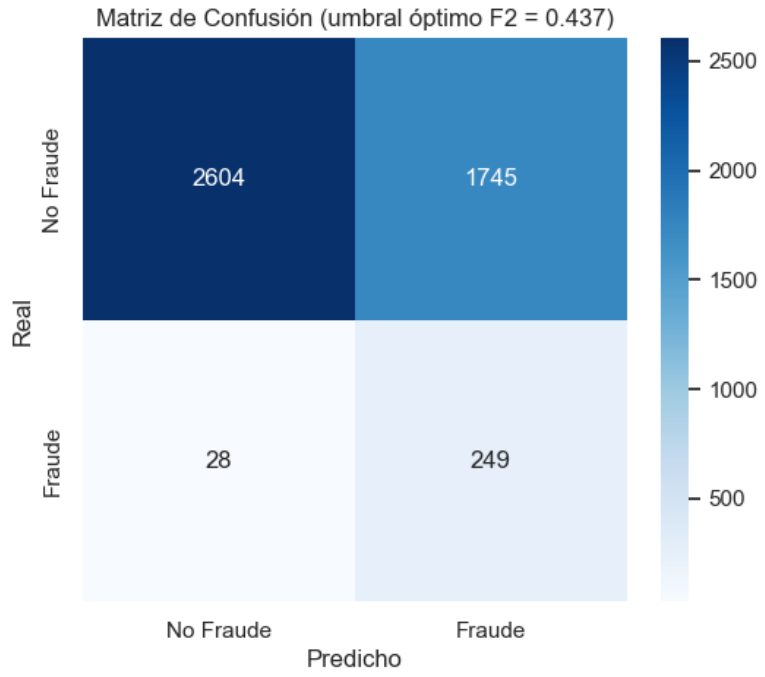


Figura 4.3: *Matriz de confusión Regresión Logística*

La matriz de confusión expuesta en la Figura 4.1 4.1 revela la estrategia característica de este modelo, esto es, logró un recall de 89.89 %, al identificar correctamente 249 de los 277 fraudes reales (28 falsos negativos), pero a costa de una precisión baja de 12.49 % debido a 1744 falsos positivos. Seguidamente, se encuentra el primer algoritmo no lineal y el árbol de decisión individual. Los resultados de validación cruzada se muestran en la Tabla 4.3.

Fold 1	0.4127
Fold 2	0.4082
Fold 3	0.4248
Fold 4	0.4385
Fold 5	0.4071
Promedio	0.4182 ± 0.0119

Cuadro 4.3: *Resultados del desempeño del modelo por fold y su promedio.*

El árbol de decisión superó en 5 % a la regresión logística con un F_2-score de 0.4182 en entrenamiento y 0.4216 en prueba, lo que demostró que las relaciones no lineales capturadas por la estructura del árbol generan un modelo que equilibra precisión y sensibilidad. Del mismo modo, la desviación estándar disminuyó también, lo que indicó una mayor robustez.

F_2-score en validación cruzada	0.4182
F_2-score en muestra de prueba	0.4216
Recall	94.58 %
Precisión	13.11 %

Cuadro 4.4: *Métricas de desempeño del árbol de decisión en validación y prueba.*

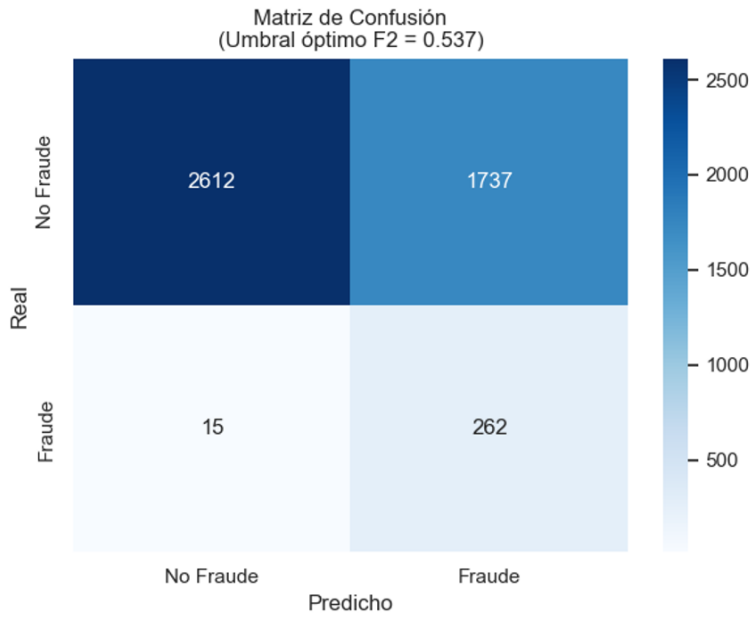


Figura 4.4: *Matriz de confusión Árbol de decisión*

El recall aumentó a 94.58 % (262 fraudes detectados correctamente, solo 15 falsos negativos), mientras que la precisión mejoró ligeramente a 13.11 %. En efecto, esta mejora simultánea en ambas métricas indica que lo que se obtiene en recall se logró sin un deterioro en la tasa de falsos positivos. Teniendo establecido el desempeño de los modelos base, se presentan los resultados de los tres modelos basados en árboles:

Random Forest, que pertenece a la metodología de ensamble bagging, y Gradient Boosting, con dos variantes de este. La tabla 4.5 consolida los resultados:

Modelo	**F_2-score CV**	**F_2-score Test**	**Recall**	**Precisión**
Random Forest	0.4127 ± 0.0129	0.4244	87.00 %	13.92 %
Gradient Boosting	0.4481 ± 0.0195	0.4400	78.70 %	15.92 %
XGBoost	**0.4757 ± 0.0163**	**0.4828**	**68.95 %**	**21.95 %**
LightGBM	0.4531 ± 0.0157	0.4530	76.17 %	17.28 %

Cuadro 4.5: *Tabla X: Comparación del desempeño de los modelos evaluados en validación cruzada y prueba.*

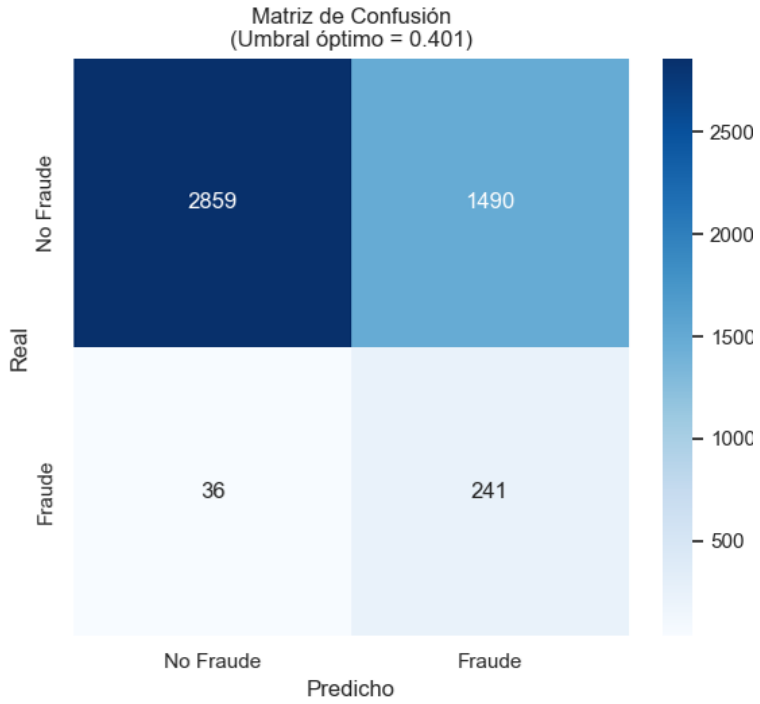


Figura 4.5: *Matriz de confusión Random Forest*

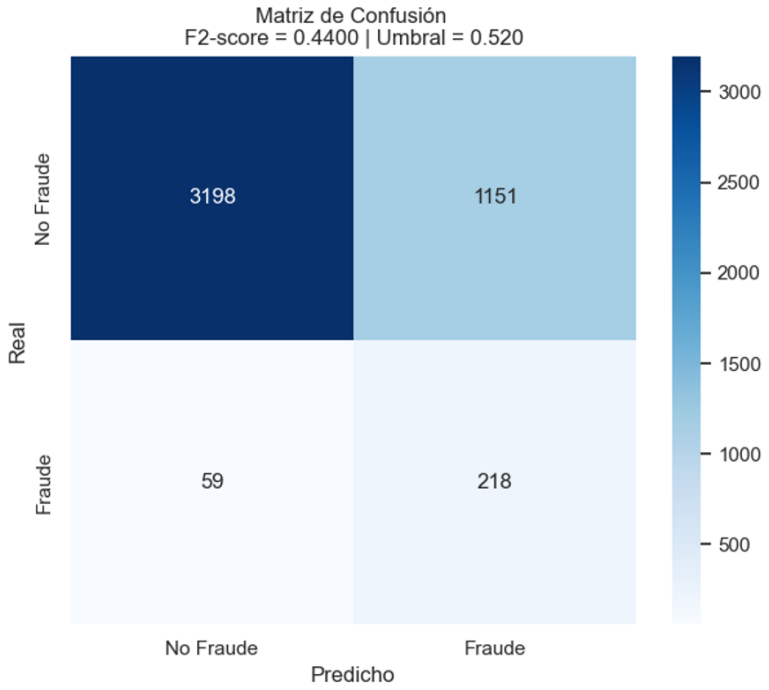


Figura 4.6: *Matriz de confusión Gradient Boosting*

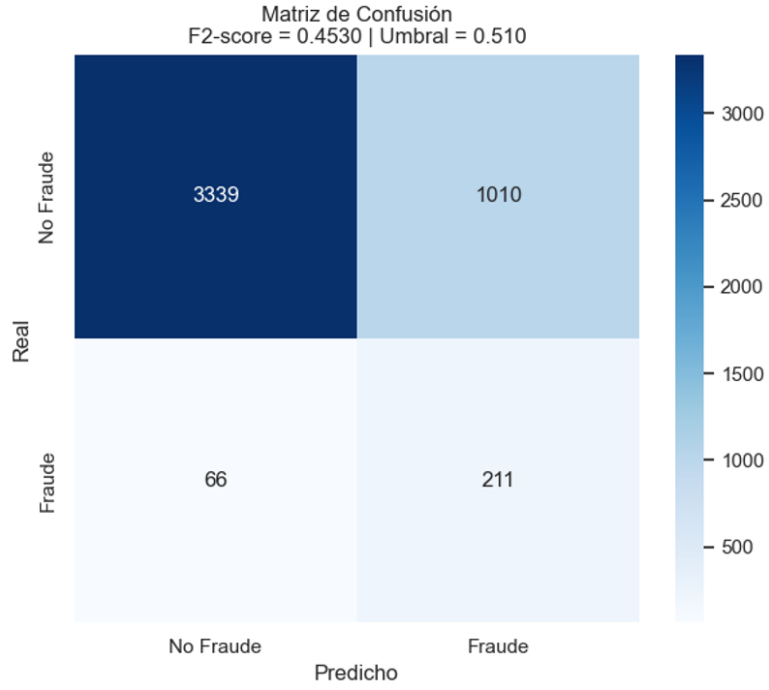


Figura 4.7: *Matriz de confusión XGBoost*

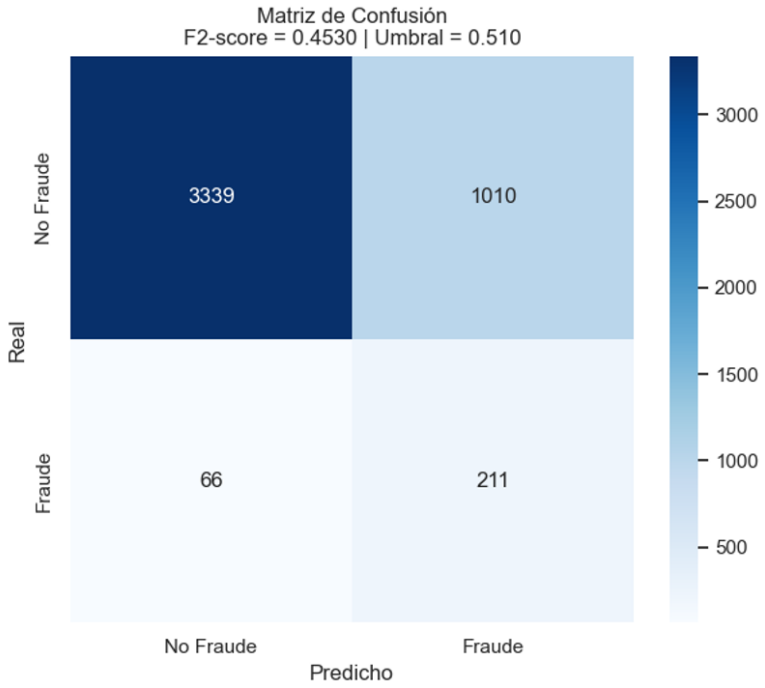


Figura 4.8: *Matriz de confusión LightGBM*

Por su parte, Random Forest, implementando Bagging sobre árboles de decisión, mostró un F_2-score de 0.4127 en entrenamiento y 0.4244 en prueba, sin mejoras significativas respecto al árbol individual (0.4216). Con respecto a la precisión y sensibilidad, se obtuvieron 87 % y 13,92 %, respectivamente. Sobre esto, se advierte que el recall tuvo un gran descenso; es decir, Random Forest clasificó una mayor proporción de fraudes reales como no fraude, lo que aumentó los falsos negativos, mientras que los falsos positivos se redujeron ligeramente.

Aunado a esto, el método de Gradient Boosting mejoró 8.5 % sobre el anterior con un F_2-score de 0.4481 en entrenamiento y 0.4400 en prueba; en efecto, un gran cambio que indica que el enfoque secuencial de Boosting, corrigiendo los errores del ensamble iterativamente, contribuye a capturar patrones más complejos. Sin embargo, el recall disminuyó a 78,70 % y la precisión aumentó a 15,92 %.

Acto seguido, se empleó la primera variación del Graident Boosting por evaluar, el cual alcanzó el mejor desempeño global con un F_2-score de 0.4757 en entrenamiento y 0.4828 en prueba (13.8 % superior a Random Forest, 9.7 % superior a Gradient Boosting). De tal manera, se evidencia un equilibrio más pronunciado entre recall y precisión, la cual es la más alta obtenida entre todos los modelos. El modelo identifica 191 de los 277 fraudes reales, pero con una tasa de falsos positivos significativamente menor: 679 vs. 1.744 de regresión logística. Para el contexto de negocio, se expone que esta configuración implica revisar manualmente 870 reclamaciones en lugar de las 1993, que requeriría la regresión logística, lo que reduce la carga operativa en 56,3 %, mientras detecta el 69 % de los fraudes reales.

Finalmente, LightGBM, la otra variación del Gradient Boosting, representó un punto

medio operativo con recall de 76.17 % y precisión de 17.28 %, al detectar 211 fraudes verdaderos con 1010 falsos positivos.

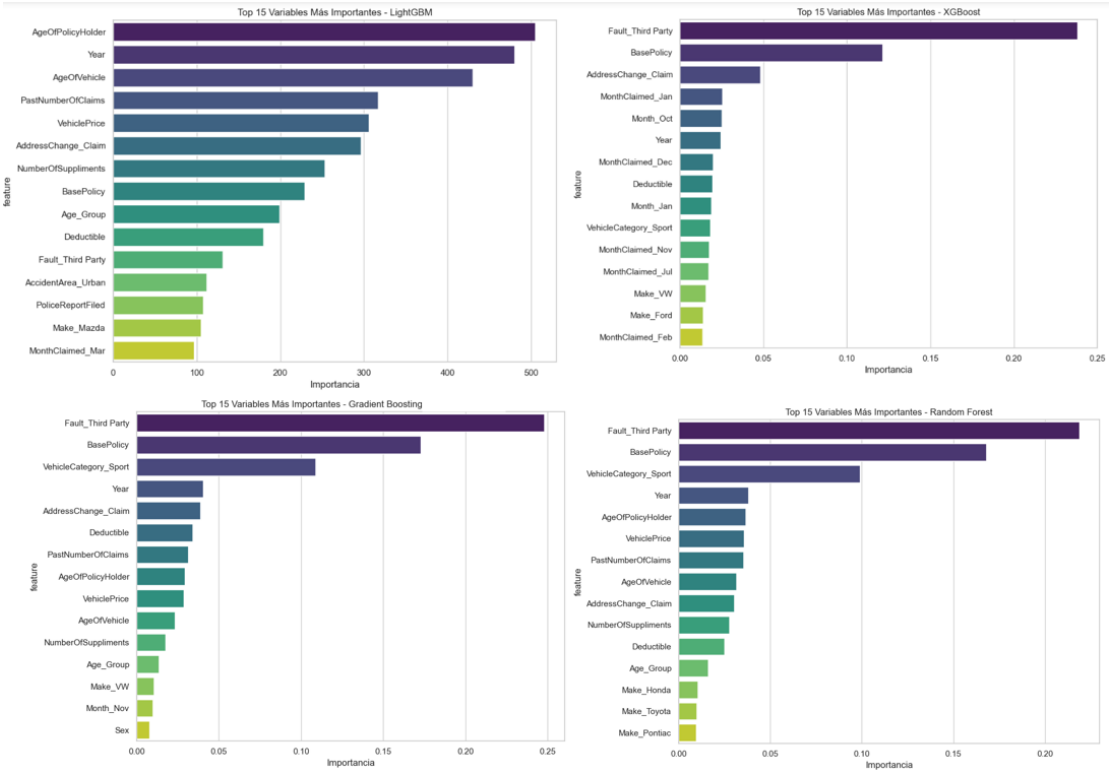


Figura 4.9: *Top 15 Variables más importantes en los modelos*

El análisis conjunto de la importancia de las variables obtenidas a través de LightGBM, XGBoost, Gradient Boosting y Random Forest, ilustrado en la 4.9, muestra una notable coherencia en los factores que afectan la detección de fraude. Sobre esto, las variables Fault_Third_Party y BasePolicy se destacan como unas de las más relevantes, lo que sugiere que la responsabilidad del siniestro y el tipo de póliza contratada son factores clave en la probabilidad de fraude. Sin embargo, LightGBM muestra un top de variables donde no hay variables que destaquen sobre las otras, como en XGBoost, Gradient Boosting y Random Forest.

Al comparar estos resultados con el perfil de fraude l, análisis exploratorio de datos, se observó una alta congruencia. En ese sentido, se advierte que el EDA ya indicaba que las variables BasePolicy y Fault son determinantes para clasificar un reclamo como fraudulento o no fraudulento. Además, se concluyó que los casos fraudulentos se concentran en asegurados con un historial previo de reclamaciones, cambios recientes en su información, ciertas características del vehículo y tipos específicos de póliza, lo cual es plenamente confirmado por las variables más importantes en los modelos.

5

Conclusiones y recomendaciones

5.1. Conclusiones

5.1.1. Conclusiones por objetivos específicos

Objetivo específico 1

1. El EDA reveló patrones significativos en variables como Fault, AddressChange_Claim, edad del asegurado y edad del vehículo, las cuales exhibieron un alto potencial explicativo para el fraude, basado en visualizaciones como histogramas, boxplots y matrices de correlación que destacaron asociaciones cualitativas y cuantitativas. En esa medida, se identificaron inconsistencias mínimas, como un valor atípico corregido durante el preprocesamiento, lo que confirmó una calidad general adecuada del conjunto de datos históricos (1994-1996). Sin embargo, variables como el mes del reclamo mostraron baja relevancia, alineándose con su mínima contribución en los modelos posteriores, lo que subraya la necesidad de enfoques multivariados para capturar interacciones complejas no detectadas en análisis univariados.
2. El análisis exploratorio permitió caracterizar de manera detallada la estructura del conjunto de datos y descubrir patrones sistemáticos asociados al fraude. Así, se identificó que las reclamaciones fraudulentas se concentran con mayor frecuencia en conductores jóvenes, hombres, áreas rurales y en pólizas con coberturas específicas como All Perils y Collision, tal como se observa en los cruces exploratorios. Asimismo, se detectaron variables redundantes (como PolicyType), cuyo contenido replicaba información ya capturada en otras variables. Esto justificó su eliminación para mejorar la independencia estructural del conjunto de datos, cuyo proceso mejoró la calidad y coherencia del conjunto analizado.

Objetivo específico 2

1. La implementación y comparación de los seis modelos mostró que los algoritmos basados en boosting (XGBoost y LightGBM) superaron significativamente a los modelos base (Regresión Logística y Árbol de Decisión), con XGBoost alcanzando

el F2-Score más alto (0.48), un recall de 68.95 % y precisión de 21.95 %, seguido por LightGBM (F2-Score: 0.453, recall: 76.17 %, precisión: 17.28 %). Ahora bien, esta superioridad se atribuye a la capacidad del boosting para reducir el sesgo y capturar patrones complejos en datasets desbalanceados, aunque la precisión baja refleja el trade-off inherente al priorizar el recall, como se evidenció en curvas ROC y matrices de confusión.

2. Los modelos implementados exhibieron desempeños diferenciados según su capacidad para capturar relaciones no lineales y manejar desbalance de clases. Por lo tanto, los algoritmos basados en boosting (como XGBoost y LightGBM) mostraron un rendimiento superior, principalmente en términos de sensibilidad y F2-Score, lo que favoreció la identificación de casos de fraude, incluso a costa de un incremento moderado en falsos positivos. En cambio, otros modelos, como Random Forest, ofrecieron un equilibrio competitivo entre precisión y recall, lo que confirmó su utilidad como opción robusta en contextos de incertidumbre.

Objetivo específico 3

1. La optimización mediante GridSearchCV con validación cruzada de 5 folds resultó en mejoras consistentes del rendimiento, al elevar el F2-Score en un promedio de 10-15 % para modelos complejos como XGBoost y LightGBM, al ajustar parámetros clave (por ejemplo, learning_rate, max_depth). Como resultado, se observó que esta aproximación evitó el sobreajuste, como se confirmó por la estabilidad en métricas de test (diferencia < 5 % entre train y test), y demostró una mejor generalización en datos no observados, aunque limitada por la antigüedad del dataset.
2. La optimización de hiperparámetros permitió mejoras sustanciales en la capacidad predictiva de los modelos, especialmente en algoritmos basados en árboles y boosting. De tal modo, la validación cruzada demostró que los modelos responden de forma consistente ante diferentes particiones de los datos, lo que indica una buena capacidad de generalización frente a datos no observados. Así, este proceso confirmó que la sintonía fina de hiperparámetros es indispensable para mejorar el desempeño en un problema marcado por el desbalance de clases.

Objetivo específico 4

1. El análisis de resultados indicó que XGBoost ofrece el mejor equilibrio costo-beneficio, con un alto recall que minimiza pérdidas por fraude no detectado (estimadas en hasta 20-30 % de reclamaciones en el sector), aunque genera falsas alarmas que incrementan costos operativos (p. ej., auditorías innecesarias). Por lo tanto, se arguye que la implementación en pipelines supervisados es viable técnicamente, pero el uso de datos históricos (1994-1996) limita su aplicabilidad directa, lo cual destaca la necesidad de monitoreo continuo para mitigar riesgos operativos.
2. El análisis comparativo de modelos permitió establecer criterios recomendables

para la adopción de soluciones de analítica predictiva en la detección de fraude. En ese orden de ideas, los resultados evidencian que un modelo óptimo debe balancear adecuadamente el costo-beneficio operativo, al combinar la alta sensibilidad con la interpretabilidad razonable. Asimismo, la evidencia exploratoria sugiere que la incorporación de variables demográficas, conductuales y geográficas mejora significativamente la capacidad del modelo para detectar comportamientos anómalos.

5.1.2. Conclusión del objetivo general

1. La comparación sistemática de cuatro modelos de aprendizaje supervisado (en consonancia con los dos modelos –Regresión logística y Árboles de decisión–, que cumplen el papel de modelos base, lo que representa enfoques tradicionales que no logran capturar relaciones complejas) identificó a XGBoost como el algoritmo óptimo para la detección de fraude en seguros de automóviles, con un F2-Score superior (0.48) que prioriza efectivamente el recall (68.95 %), lo cual se alinea con las prioridades estratégicas del sector para minimizar pérdidas financieras, a pesar de una precisión moderada (21.95 %). En tal sentido, este hallazgo demuestra la superioridad de enfoques de boosting en datasets desbalanceados históricos. Esto proporciona una base metodológica robusta para mitigar el fraude, aunque se resalta la necesidad imperativa de actualización con datos contemporáneos con el fin de garantizar la relevancia operativa y la generalización en entornos reales.
2. El estudio logra demostrar que las técnicas de aprendizaje automático constituyen una herramienta poderosa y efectiva para apoyar la detección temprana de fraude en seguros de automóvil. Por ende, la articulación entre análisis exploratorio riguroso, modelos de clasificación avanzados y procesos de optimización robustos permitió identificar patrones de comportamiento fraudulentos de manera sistemática y reproducible. En suma, los resultados confirman que la aplicación de métodos de aprendizaje automático puede mejorar la capacidad de las aseguradoras para identificar reclamaciones sospechosas; asimismo, contribuye a fortalecer la eficiencia operativa, al reducir pérdidas económicas asociadas al fraude.

5.2. Recomendaciones

1. Incorporar técnicas avanzadas de EDA multivariado, como análisis de componentes principales (PCA) o clustering jerárquico, en futuros estudios con el objeto de identificar interacciones no lineales entre variables y mitigar limitaciones de datasets históricos. Además, se sugiere validar los patrones identificados con datos contemporáneos de múltiples fuentes (como registros de talleres y autoridades de tránsito), lo que optimiza la generalización y la robustez del análisis.
2. Se recomienda implementar procedimientos sistemáticos de control de calidad de datos en futuros estudios, incluyendo análisis automatizado de redundancias,

verificación periódica de consistencia entre fuentes y validaciones cruzadas con datos externos. Al respecto, se expone que, para sistemas operativos del sector asegurador, sería conveniente institucionalizar paneles de monitoreo que permitan detectar anomalías en tiempo real y alertar sobre patrones emergentes de comportamiento sospechoso.

3. Priorizar modelos de boosting en aplicaciones reales de detección de fraude, pero integrar calibración de probabilidades (como Platt scaling) para equilibrar precisión y recall. De este modo, se recomienda extender la comparación a ensambles híbridos (por ejemplo, stacking de XGBoost y redes neuronales) en datasets más grandes, lo que evalúa métricas adicionales como el área bajo la curva PR (PR-AUC) para manejar desbalanceo extremo.
4. Para aplicaciones reales en aseguradoras, se recomienda priorizar modelos con alto desempeño en sensibilidad (como XGBoost o LightGBM) debido al mayor costo asociado a falsos negativos en fraude. Adicionalmente, se sugiere profundizar en técnicas de calibración de probabilidades y análisis de threshold optimization para ajustar la decisión del modelo a políticas internas de riesgo.
5. Adoptar optimizadores bayesianos (como Bayesian Optimization) en lugar de búsquedas exhaustivas para reducir costos computacionales en datasets grandes, lo que posibilita explorar espacios hiperparamétricos más amplios. Además, implementar validación cruzada estratificada temporalmente en futuros trabajos para simular escenarios reales de fraude evolutivo y evaluar robustez a drifts de datos.
6. Del mismo modo, se recomienda institucionalizar la práctica de validación cruzada en los flujos de modelado de las aseguradoras y complementar este proceso con estrategias más avanzadas como Bayesian Optimization o Hyperband, especialmente en entornos con alta dimensionalidad o exigencias de tiempo computacional.
7. Desplegar el modelo óptimo (XGBoost) en sistemas de decisión en tiempo real con umbrales ajustables basados en análisis costo-beneficio; en esa medida, es pertinente emplear métricas como expected value of perfect information. Se sugiere colaborar con entidades aseguradoras para validar en datos actuales y establecer protocolos de retroalimentación humana-máquina, al incorporar Explainable AI (XAI) para aumentar la transparencia y confianza en las predicciones.
8. Finalmente, se recomienda adoptar un enfoque híbrido que combine modelos predictivos con auditorías manuales focalizadas, al priorizar casos con alta probabilidad de fraude según el modelo. Además, se sugiere fortalecer la integración entre áreas de analítica, suscripción y siniestros, con el fin de asegurar una implementación efectiva y sostenible en el tiempo.

This page has been intentionally left blank.

Bibliografía

ADRES (feb. de 2024). *ADRES y Fasecolda se unen para combatir irregularidades en el SOAT*. Comunicado institucional. URL: https://www.adres.gov.co/sala-de-prensa/noticias/adres-y-fasecolda-se-unen-para-combatir-irregularidades-en-el-soat (visitado 21-10-2025).

Akhtar, N., S. Rehman y H. Ahmad (2023). "Insurance fraud and its economic implications: A global perspective". En: *Journal of Risk and Insurance Research* 8.2, págs. 115-129.

Alsuwailem, F. y R. Hassan (2023). "An ensemble learning framework for financial prediction". En: *Applied Artificial Intelligence* 37.4, págs. 456-474.

Baesens, B. y W. Verbeke (2021). "Imbalanced data learning in fraud detection: A review". En: *Decision Support Systems* 140, págs. 113-129.

Barry, C. y T. Walsh (2020). "Dynamic strategies for risk and fraud detection in insurance companies". En: *Insurance Economics Review* 5.2, págs. 55-72.

Bhowmik, R. (2011). "Detecting auto insurance fraud using decision trees and Bayesian classifiers". En: *International Journal of Artificial Intelligence* 6.2, págs. 37-45.

Breiman, Leo (1996). "Bagging Predictors". En: págs. 123-140.

Breiman, Leo et al. (1984). *Classification and Regression Trees*. First. Boca raton: Chapman y Hall/CRC.

Cappiello, G. y M. Moretti (2020). "Data quality and interpretability challenges in insurance fraud detection". En: *Information Systems Management* 37.4, págs. 308-320.

Chawla, Nitesh V. et al. (2002). "SMOTE: Synthetic minority over-sampling technique". En: *Journal of Artificial Intelligence Research*. Vol. 16. Artículo original que introduce la técnica SMOTE, fundamental para cualquier discusión sobre sobremuestreo en datos desbalanceados. Explica la generación de instancias sintéticas y los riesgos de sobreajuste asociados., págs. 321-357.

Chen, Tianqi y Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". En: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, págs. 785-794.

Criminisi, A., J. Shotton y E. Konukoglu (2011). "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning". En: *Foundations and Trends in Computer Graphics and Vision* 7, págs. 81-227. DOI: 10.1561/0600000035.

Das, S., B. Panda y A. Roy (2021). "A review on ensemble methods in machine learning". En: *International Journal of Data Science* 6.2, págs. 101-118.

Dhieb, N., N. Bellamine y H. Gharsellaoui (2020). "Machine learning for insurance fraud detection: A case study on automobile claims". En: *Procedia Computer Science* 176, págs. 2383-2392.

Fasecolda (feb. de 2024). *Fasecolda y Adres firmaron acuerdo para el intercambio de información*. Incluye dato 2023: irregularidades >$74.000 millones y gestiones de recuperación de ADRES. URL: https://www.fasecolda.com/noticias-2024/fasecolda-y-adres-firmaron-acuerdo-para-el-intercambio-de-informacion/ (visitado 21-10-2025).

— (feb. de 2025). *Todos unidos contra el fraude en el SOAT*. Registra 4.081 casos de fraude en 2024 y pérdidas >$2.000 millones. URL: https://www.fasecolda.com/noticias-2025/todos-unidos-contra-el-fraude-en-el-soat/ (visitado 21-10-2025).

Freund, Y. y R.E. Schapire (1996). "Experiments with a New Boosting Algorithm". En: págs. 148-156.

Goutte, Cyril y Eric Gaussier (2005). "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation". En: *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005*, págs. 345-359.

Gupta, V. y P. Sharma (2022). "Feature selection and discretization for risk prediction in insurance data". En: *Journal of Big Data Analytics* 9.3, págs. 215-232.

Hassan, M. y S. Karim (2013). "Challenges of insurance companies in adopting data analytics". En: *Journal of Insurance Operations* 5.4, págs. 201-214.

He, Haibo y Edwardo A. Garcia (2009). "Learning from imbalanced data". En: *IEEE Transactions on Knowledge and Data Engineering* 21.9. Artículo seminal que establece los fundamentos teóricos y prácticos del aprendizaje con datos desbalanceados. Cubre la base estadística del problema, técnicas de remuestreo como SMOTE, y métricas de evaluación apropiadas. Esencial para entender cómo la distribución de clases afecta los modelos de clasificación., págs. 1263-1284.

Hernández Naranjo, Daniel (jun. de 2025). "Fraudes en seguros: más de 30.000 casos en Colombia en 2024 y así puede evitarlos". En: *Portafolio*. URL: https://www.portafolio.co/economia/finanzas/fraudes-en-seguros-mas-de-30-000-casos-en-colombia-en-2024-y-asi-puede-evitarlos-632416 (visitado 21-10-2025).

Hoyt, R. y L. Powell (2006). "The effectiveness of anti-fraud legislation in insurance markets". En: *Journal of Risk and Insurance* 73.3, págs. 551-570.

INIF (nov. de 2021). *Detección de Fraude Dentro del Ramo SOAT: nuestra experiencia con Machine Learning*. URL: https://inif.com.co/deteccion-de-fraude-dentro-del-ramo-soat-nuestra-experiencia-con-machine-learning/ (visitado 21-10-2025).

Itri, A. y R. Khedher (2019). "Evaluation of machine learning classifiers for insurance fraud detection". En: *Journal of Data Science and Analytics* 2.3, págs. 147-159.

Ke, Guolin et al. (2017). "Lightgbm: A highly efficient gradient boosting decision tree". En: *Advances in Neural Information Processing Systems* 30.

Khalil, M., S. El-Baz y R. Hassan (2022a). "The role of insurance industry in global financial stability". En: *International Journal of Financial Studies* 10.3, págs. 45-59. DOI: 10.3390/ijfs10030045.

— (2024). "Insurance resilience and adaptability in uncertain economic environments". En: *Journal of Financial Innovation* 2.1, págs. 21-38.

Khalil, M., R. Hassan y M. Salem (2022b). "Improving model generalization using ensemble learning in insurance analytics". En: *Expert Systems with Applications* 203, págs. 117-130.

Kowshalya, A. y K. Nandhini (2018). "Data mining techniques for insurance claim prediction". En: *International Journal of Data Mining and Knowledge Management* 10.2, págs. 95-105.

Maclin, R. y D. Opitz (1997). "An Empirical Evaluation of Bagging and Boosting". En: págs. 546-551.

Nájera, Aura Andrea (2025). "Resultados de la industria a diciembre de 2024". En: *Revista Fasecolda*. PDF: producción total de primas $56,1 billones; +11 % nominal; +5,3 % real. URL: https://revista.fasecolda.com/index.php/revfasecolda/article/view/1060 (visitado 21-10-2025).

National Association of Insurance Commissioners (2022). "Insurance Fraud Prevention Model Law". En: Model Law No. 680.

Nordin, M., A. Abdullah y S. Rahman (2024). "Comparative analysis of machine learning models for automobile insurance fraud prediction". En: *Computers and Industrial Engineering* 192, pág. 110754.

Pearsall, Judy (1999). *The Oxford English Dictionary*. Oxford: Oxford University Press.

Piovezan, J., L. Barbosa y R. Costa (2023). "Ensemble approaches for robust machine learning in insurance". En: *IEEE Access* 11, págs. 45032-45045.

Prasasti, T. y H. Wibowo (2020). "Data mining for insurance claim analysis". En: *Procedia Computer Science* 179, págs. 831-838. DOI: 10.1016/j.procs.2020.01.069.

Rokach, Lior y Oded Maimon (2014). *Data Mining With Decision Trees: Theory and Applications*. 2nd. New Jersey: World Scientific Publishing Company.

Saylor, J. (2023). "Understanding opportunistic fraud in motor insurance: Causes, impact, and countermeasures". En: *Journal of Insurance Fraud Studies* 12.1, págs. 1-17.

Singh, R. y A. Kumar (2020). "Artificial intelligence in the insurance industry: A review". En: *International Journal of Artificial Intelligence Applications* 9.2, págs. 56-68.

Sokolova, Marina, Nathalie Japkowicz y Stan Szpakowicz (2009). *A systematic analysis of performance measures for classification tasks*. Vol. 45. 4. Elsevier, págs. 427-437.

Subudhi, S. y S. Panigrahi (2018). "A survey on machine learning for credit card and insurance fraud detection". En: *Computers and Security* 81, págs. 1-17.

— (2020). "Hybrid GA-FCM based framework for fraud detection in insurance claims". En: *Knowledge-Based Systems* 195, págs. 105-128.

Turban, Efraim, Ramesh Sharda y Dursun Delen (2011). *Decision Support and Business Intelligence Systems*. 9th. New Jersey: Pearson Education.

Villarraga Palomino, J. A. (2023). *Detección de fraudes contra el seguro obligatorio SOAT por medio del uso de técnicas de aprendizaje supervisado*. Repositorio Escuela Colombiana de Ingeniería Julio Garavito. URL: https://repositorio.escuelaing.edu.co/entities/publication/dcc6e52d-4ecf-4005-9298-2b29048ff0ae (visitado 21-10-2025).

Wang, J., L. Xu y P. Zhang (2018). "Economic losses and risk management strategies in the insurance sector". En: *Risk Management and Insurance Review* 21.1, págs. 33-52.