Fraude al seguro

Detección del fraude en siniestros de automóvil mediante aprendizaje automático

Ainara Acha Sánchez

Grado de Ciencia de datos aplicada

Nombre Tutor/a de TF
Antonio Gutiérrez Blanco

Profesor/a responsable de la asignatura
Susana Acedo

Enero 2026

FICHA DEL TRABAJO FINAL

Título del trabajo:	Detección del fraude en siniestros de automóvil mediante aprendizaje automático
Nombre del autor:	*Ainara Acha Sánchez*
Nombre del director/a:	*Antonio Gutiérrez Blanco*
Nombre del PRA:	*Susana Acedo Nadal*
Fecha de entrega (mm/aaaa):	*01/2026*
Titulación o programa:	Grado Ciencia de Datos Aplicada
Área del Trabajo Final:	*Aprendizaje automático y modelos de lenguaje*
Idioma del trabajo:	*Castellano*
Palabras clave	*Detección de fraude, aprendizaje automático, procesamiento de lenguaje natural, inteligencia artificial explicable, sector asegurador, análisis predictivo*

Resumen del Trabajo

El fraude en siniestros de seguros de automóvil representa un problema relevante para el sector asegurador, debido a su impacto económico y a la dificultad de detección temprana en entornos con datos heterogéneos y altamente desbalanceados. En este trabajo se desarrolla y evalúa un sistema predictivo orientado a estimar la probabilidad de fraude en partes de siniestro, integrando técnicas de aprendizaje automático, métodos de explicabilidad y modelos de lenguaje natural.

La metodología propuesta incluye un análisis exploratorio del conjunto de datos, un proceso de preprocesado con codificación de variables categóricas y distintas estrategias para gestionar el desbalanceo de clases. Se comparan varios modelos de clasificación, prestando especial atención a *Random Forest*, que resulta ser el algoritmo con mejor equilibrio entre rendimiento y estabilidad tras aplicar sobremuestreo mediante *Random OverSampling* y optimización del umbral de decisión. El modelo final se optimiza mediante búsqueda de hiperparámetros y se evalúa utilizando métricas adecuadas para problemas de fraude, como *recall,* F1-score y ROC-AUC, así como mediante gráficos de *Gain* y *Lift* para analizar su capacidad de priorización.

Con el fin de garantizar la transparencia, se incorporan técnicas de explicabilidad como SHAP y LIME, y se desarrolla un agente basado en un modelo de lenguaje local que permite introducir descripciones de siniestros en lenguaje natural y generar explicaciones comprensibles para usuarios no técnicos. Los resultados muestran que el sistema es capaz de detectar aproximadamente la mitad de los casos fraudulentos y priorizar eficazmente los expedientes más sospechosos, constituyendo una herramienta útil para tareas de cribado inicial y apoyo a la toma de decisiones.

Abstract

Automobile insurance fraud represents a major challenge for the insurance sector due to its economic impact and the inherent difficulty of early detection in highly imbalanced and heterogeneous datasets. This work presents the design and evaluation of a predictive system aimed at estimating the probability of fraud in automobile insurance claims by combining machine learning techniques, model explainability methods, and natural language processing.

The proposed methodology includes exploratory data analysis, data preprocessing with categorical variable encoding, and the evaluation of several strategies to address class imbalance. Multiple supervised classification models are compared, with Random Forest achieving the best balance between performance and robustness after applying Random Oversampling and decision threshold optimization. The final model is further refined through hyperparameter tuning and evaluated using metrics suitable for fraud detection, such as recall, F1-score, and ROC-AUC, as well as Gain and Lift charts to assess its prioritization capability.

To enhance transparency and usability, explainability techniques such as SHAP and LIME are integrated, and a local large language model–based agent is developed to process insurance claim descriptions written in natural language. This agent converts textual input into structured features and generates human-readable explanations of the model's predictions. The results indicate that the proposed system can successfully identify a significant proportion of fraudulent claims and effectively prioritize high-risk cases, making it suitable as a decision-support tool for initial fraud screening processes.

Índice

Lista de figuras

1. Introducción

1.1. Contexto y justificación del Trabajo

El fraude en siniestros de automóviles constituye uno de los desafíos más relevantes para el sector asegurador, tanto por su frecuencia como por sus consecuencias económicas. Se estima que entre un 10% y un 20% de los siniestros contienen algún tipo de irregularidad, generando pérdidas significativas para las aseguradoras y afectando indirectamente al resto de los asegurados [1]. Este fenómeno, además de comprometer la sostenibilidad financiera del sistema, distorsiona los cálculos actuariales, eleva las primas de los clientes legítimos y debilita la confianza en las instituciones aseguradoras.

A pesar de los esfuerzos actuales, las técnicas tradicionales de detección de fraude basadas en reglas fijas y la revisión manual de expedientes o investigaciones del perito, han demostrado ser insuficientes frente a fraudes cada vez más sofisticados y difíciles de detectar. En respuesta a esta problemática, la inteligencia artificial y el análisis de grandes volúmenes de datos han surgido como soluciones eficaces para automatizar la identificación de patrones anómalos, reducir errores humanos y acelerar la toma de decisiones.

En este contexto, el presente trabajo tiene como finalidad diseñar, implementar y evaluar un sistema de detección de fraude apoyado en algoritmos de aprendizaje automático y en técnicas de procesamiento de lenguaje natural. El enfoque se centra no solo en lograr un buen rendimiento predictivo, sino también en garantizar la transparencia y explicabilidad del modelo, aspectos fundamentales para su aceptación e implementación en entornos reales que no requieran conocimientos técnicos para su utilización.

1.2. Objetivos del Trabajo

El objetivo general del presente trabajo es diseñar y evaluar un sistema predictivo capaz de estimar la probabilidad de que un parte de siniestro de automóvil sea fraudulento. Este sistema no solo busca alcanzar un rendimiento adecuado en términos de predicción, sino que además integra técnicas de explicabilidad y un

componente lingüístico basado en modelos de lenguaje, con el fin de facilitar su utilización por parte de usuarios no expertos. La finalidad última es proporcionar una herramienta clara, interpretable y robusta que pueda servir como apoyo en procesos reales de detección de fraude en el sector asegurador.

A partir de este propósito general se desarrollan diversos objetivos específicos. En primer lugar, se plantea la realización de un análisis exploratorio del conjunto de datos de siniestros, prestando especial atención a la identificación de errores de codificación, valores inconsistentes, categorías atípicas o variables con muy baja variabilidad. Este análisis es fundamental para comprender la estructura del *dataset*, detectar posibles fuentes de ruido y, en definitiva, preparar adecuadamente la información para su posterior modelización. Una parte relevante de este objetivo consiste también en estudiar el desbalanceo entre las clases de fraude y no fraude que constituye una de las principales dificultades del problema.

En segundo lugar, se establece como objetivo desarrollar un proceso de preprocesado adecuado a las características del *dataset*. Esto incluye el tratamiento de valores ausentes o erróneos, la codificación apropiada de variables categóricas tanto ordinales como nominales, y la posible reagrupación de categorías poco representadas para evitar que generen inestabilidad en los modelos. El preprocesado debe asegurar que los datos estén en un formato numérico y consistente que permita el entrenamiento fiable de determinados algoritmos de clasificación.

El objetivo esencial del preprocesado, tal y como se comentó anteriormente, consiste en gestionar el fuerte desbalanceo presente en la variable objetivo. Para ello, se ha estimado emplear varias estrategias complementarias de sobre muestreo (SMOTE, SMOTENC, RandomOverSampler [2]), el uso de pesos de clase en los modelos o el ajuste manual del umbral de decisión para optimizar métricas sensibles a la clase minoritaria. Estos ajustes buscan garantizar que los modelos sean capaces de detectar casos de fraude reales, evitando que el

desequilibrio en los datos lleve a predicciones sesgadas hacia la clase mayoritaria.

Tras la corrección del desbalanceo en la variable *target*, el siguiente objetivo se centra en el entrenamiento y la comparación de diversos modelos de clasificación supervisada. Entre ellos se consideran métodos lineales y no lineales como la regresión logística [3], los árboles de decisión [4], los bosques aleatorios (Random Forest) [5] y algoritmos basados en *boosting* como XGBoost [6]. Cada uno de estos modelos se evalúa mediante métricas adecuadas para escenarios desbalanceados, incluyendo *precisión* [7], *recall* [8], F1-score [9], ROC-AUC [10] y el análisis detallado de la matriz de confusión [11]. La finalidad de este apartado es determinar cuál de los modelos ofrece el mejor equilibrio entre rendimiento, interpretabilidad y estabilidad.

Una vez comparados los distintos modelos, el trabajo busca seleccionar y optimizar aquel que muestre un comportamiento superior. Esta optimización incluye la búsqueda de hiperparámetros, la evaluación de la estabilidad del modelo y el análisis de su capacidad para generalizar sin sobreajustarse. También se prioriza la función de interpretar sus predicciones, un aspecto fundamental cuando el modelo se utiliza en ámbitos con implicaciones económicas y legales, como la detección de fraude.

En relación con la interpretabilidad, se han aplicado técnicas de explicabilidad como SHAP [12] y LIME [13]. SHAP permite analizar la importancia global de las variables y explicar, de forma precisa, qué factores influyen en cada predicción realizada por el modelo. Por su parte, LIME facilita explicaciones locales adicionales que complementan la visión proporcionada por SHAP. Ambas técnicas tienen como propósito otorgar transparencia al modelo y permitir que sus decisiones puedan ser auditadas y justificadas.

Finalmente, el trabajo incorpora como objetivo el desarrollo de un agente basado en modelos de lenguaje LLM [14] capaz de recibir descripciones de siniestros en lenguaje natural, convertir este texto en las variables estructuradas que utiliza el

modelo predictivo y generar una explicación comprensible sobre la probabilidad estimada de fraude. Este componente lingüístico busca acercar la herramienta a usuarios no técnicos y demostrar cómo se puede mejorar la accesibilidad y la utilidad de sistemas avanzados de análisis.

El proyecto concluye con una evaluación de los resultados obtenidos y una reflexión sobre las limitaciones detectadas, los riesgos éticos asociados y las posibles líneas de trabajo futuro que podrían mejorar la solución propuesta.

1.3. Impacto en sostenibilidad, ético-social y de diversidad

La perspectiva ética-social constituye un reto, especialmente por tratarse de un sistema basado en inteligencia artificial aplicado a un ámbito sensible como la detección de fraude en seguros. Desde el inicio del proyecto se ha buscado garantizar que el desarrollo respete los principios de transparencia, responsabilidad y no discriminación.

En este sentido, se ha priorizado la creación de un modelo que no funcione como una “caja negra”, incorporando técnicas de explicabilidad como SHAP y LIME que permiten comprender qué factores influyen en cada decisión y facilitan la identificación de posibles sesgos o efectos desproporcionados sobre determinados colectivos.

Así mismo, el sistema se ha diseñado para alinearse con los principios de IA explicable (XAI) y con los Objetivos de Desarrollo Sostenible (ODS 9 y 11) [15] [16]. El ODS 9 promueve el uso responsable de la tecnología para mejorar la eficiencia industrial, mientras que el ODS 11 impulsa sistemas más seguros y sostenibles para los ciudadanos.

En el contexto asegurador, esto implica desarrollar herramientas que mejoren la toma de decisiones sin comprometer los derechos de los asegurados ni generar desigualdades. El presente trabajo contribuye a esa visión mediante un modelo

transparente, documentado y verificable, capaz de justificar sus predicciones y permitir auditorías internas y externas.

Finalmente, se ha prestado especial atención a evitar riesgos éticos asociados al uso de datos en el ámbito del cumplimiento del RGPD [17][18]. El *dataset* empleado está completamente anonimizado y no contiene identificadores personales. Los resultados se analizan únicamente desde una perspectiva técnica, sin realizar inferencias que puedan vincularse a individuos concretos.

1.4. Enfoque y método seguido

El trabajo se ha estructurado en fases iterativas desde la revisión bibliográfica, la validación, la explicabilidad del modelo final y la implementación del agente LLM. Primero se ha realizado una revisión exhaustiva sobre técnicas de detección de fraude en el sector asegurador, con foco en el aprendizaje automático y el procesamiento de lenguaje natural, para identificar retos y criterios de éxito.

Luego, se han tratado los datos de siniestros y se han entrenado varios modelos supervisados para evaluar su rendimiento y capacidad explicativa. Se ha complementado con un módulo de NLP que interpreta descripciones textuales y genera explicaciones automáticas sobre la probabilidad de fraude.

Así mismo, se han aplicado técnicas de explicabilidad y métricas de evaluación desde un enfoque cuantitativo y cualitativo que asegura la transparencia, reproducibilidad y ética en el sistema desarrollado.

1.5. Planificación del Trabajo

La planificación del proyecto se ha estructurado en fases consecutivas y parcialmente solapadas, abarcando desde octubre de 2025 hasta enero de 2026. El trabajo comenzó con una primera etapa dedicada a la revisión bibliográfica, donde se recopiló literatura especializada sobre fraude en seguros, aprendizaje

automático y técnicas de explicabilidad. Esta fase permitió establecer un marco teórico sólido y definir la metodología a seguir.

Posteriormente, durante la segunda y tercera fase, se abordó el análisis exploratorio del *dataset*, el preprocesado de variables y el desarrollo inicial de modelos supervisados, comparando su rendimiento mediante métricas adecuadas para problemas con clases desbalanceadas.

A partir de la segunda mitad del proyecto se llevó a cabo la integración del módulo basado en modelos de lenguaje (LLM), encargado de interpretar descripciones textuales de siniestros y generar explicaciones automatizadas. De forma paralela se fueron realizando tareas de validación, elaboración de visualizaciones, interpretación de resultados y redacción progresiva de la memoria.

Finalmente, el trabajo culminó en enero con la revisión global del documento, la incorporación de observaciones del tutor y la preparación de la versión final del TFG, siguiendo la estructura y estándares formales requeridos. Esta planificación permitió avanzar de manera ordenada y coherente, asegurando la correcta conexión entre todas las fases del proyecto y garantizando la calidad de los resultados obtenidos.

Figura 1: Diagrama de Grantt

Fase / Duración estimada	Actividades principales
Fase 1. **Revisión bibliográfica y marco teórico** **(1–15 octubre 2025)**	- Actualización de fuentes académicas y oficiales. - Incorporación de literatura sobre *Explainable AI* y modelos LLM. - Redacción del borrador de la memoria. - Análisis exploratorio del dataset.
Fase 2. **Preparación y preprocesamiento de datos** **(15–31 octubre 2025)**	- Preprocesado del dataset: limpieza e imputación de valores ausentes. - Tratamiento de variables categóricas. - Generación y documentación del dataset final. - Implementación de modelos supervisados (Regresión Logística, Árboles, Random Forest, XGBoost).
Fase 3. **Desarrollo del modelo base de Machine Learning** **(1–15 noviembre 2025)**	- Ajuste de hiperparámetros y comparación de rendimiento. - Selección del modelo con mejor equilibrio entre precisión y explicabilidad. - Diseño del módulo para interpretar descripciones textuales de siniestros.
Fase 4. **Integración del módulo NLP / LLM** **(15–30 noviembre 2025)**	- Implementación de *prompts* generativos para explicar resultados. - Validación y análisis de coherencia de respuestas.
Fase 5. **Validación, visualización e interpretación de resultados** **(1–15 Diciembre 2025)**	- Validación cruzada y revisión de métricas. - Creación de visualizaciones interactivas (curvas ROC, SHAP, importancia de variables).
Fase 6. **Redacción y documentación progresiva del TFG** **(Octubre 2025 – Enero 2026)**	- Redacción paralela de capítulos conforme avanza el proyecto. - Integración de secciones (introducción, metodología, resultados, conclusiones). - Revisión continua de estilo, coherencia y citación académica (APA 7ª).
Fase 7. **Revisión final y entrega del TFG** **(Enero 2026)**	- Revisión integral con el tutor. - Incorporación de *feedback* y correcciones formales. - Entrega definitiva del documento.

Figura 2: Detalle de tareas de las fases

1.6. Sumario de productos obtenidos

Al finalizar el proyecto se ha obtenido un modelo predictivo que evalúa si el siniestro de automóvil es un posible fraude al seguro, mediante la introducción de la descripción del siniestro, devolviendo una respuesta en lenguaje natural que indique los motivos asociados.

A nivel técnico, se ha documentado el rendimiento de los distintos algoritmos y sus métricas identificando el más adecuado en base a sus resultados. Así mismo, desde una perspectiva ética, se garantiza que el trabajo contribuye al desarrollo de soluciones de inteligencia artificial transparente y responsable, alineadas con los principios de AI explicable y los Objetivos de Desarrollo Sostenible de la Agenda 2030 para fomentar la confianza en los sistemas automatizados en el sector asegurador.

Todo este proceso de desarrollo se documenta en la presente memoria, así como en el Plan de Trabajo que se detalla en el siguiente capítulo.

1.7. Descripción de capítulos

La memoria se organiza en varios capítulos que abarcan de forma progresiva todos los aspectos necesarios para comprender, desarrollar y evaluar el sistema de detección de fraude propuesto.

El capítulo 1 introduce el contexto del fraude en seguros de automóvil y justifica la relevancia del problema desde una perspectiva económica, tecnológica y social. También presenta los objetivos del trabajo, los impactos éticos y de sostenibilidad considerados y el enfoque metodológico seguido, junto con la planificación del proyecto y un resumen de los productos obtenidos.

El capítulo 2 desarrolla el marco conceptual y metodológico del estudio. Comienza con una revisión del marco teórico y del estado del arte en detección de fraude con inteligencia artificial, analizando los modelos y técnicas más

utilizadas en la literatura especializada. Posteriormente, se detallan los materiales y métodos empleados: el *dataset* utilizado, el preprocesado aplicado, las estrategias implementadas para abordar el desbalanceo de clases, los modelos evaluados, la optimización de hiperparámetros y las herramientas de explicabilidad. El capítulo concluye con la descripción del agente LLM y del entorno técnico utilizado.

El capítulo 3 presenta los resultados experimentales. En él se comparan distintos modelos y estrategias de sobremuestreo, se analiza la optimización del umbral y los hiperparámetros, y se expone el rendimiento del modelo final seleccionado. También incluye una evaluación de las capacidades del agente LLM, así como una discusión crítica que contextualiza los resultados obtenidos y analiza sus implicaciones prácticas.

El capítulo 4 recoge las conclusiones generales del trabajo, evaluando el grado de consecución de los objetivos planteados y valorando la metodología y la planificación empleadas. Se revisan los principales impactos éticos, sociales y de sostenibilidad asociados al desarrollo del sistema y se plantean diversas líneas de trabajo futuro orientadas a mejorar el rendimiento, la robustez y la aplicabilidad de la solución en entornos reales.

Finalmente, la memoria se completa con un glosario que define los términos técnicos utilizados, y con la bibliografía que recoge todas las fuentes consultadas.

2. Materiales y métodos

2.1. Marco teórico

En los últimos años, el uso de técnicas de aprendizaje automático (Machine Learning, ML) ha demostrado un alto potencial en tareas de clasificación binaria como la detección de fraude. Entre los modelos supervisados más empleados destacan los árboles de decisión y sus variantes ensambladas, como *Random Forest*, debido a su robustez, capacidad para manejar relaciones no lineales y buen comportamiento en presencia de variables mixtas (categóricas y numéricas). Además, los modelos de tipo árbol resultan especialmente adecuados en contextos donde la interpretabilidad resulta relevante para asistentes de tramitación y departamentos de auditoría.

Uno de los principales desafíos al trabajar con datos de fraude es la fuerte descompensación entre clases (fraude real vs. no fraude). Para abordar dicho problema existen técnicas de *resampling* como SMOTE, SMOTENC (orientado a datos con variables categóricas) y Random OverSampling (ROS), cuyo objetivo es equilibrar la clase minoritaria generando nuevas instancias sintéticas o réplicas. Estas técnicas permiten mejorar la sensibilidad del modelo sin alterar las distribuciones originales.

Para evaluar el rendimiento de los clasificadores se utilizan métricas como *Precision*, *Recall*, *F1-score* y *ROC-AUC*. Dado que los falsos negativos pueden representar un coste elevado en este contexto, resulta especialmente relevante optimizar el modelo en función del *F1-score* o mediante estrategias de ajuste de umbral (*threshold optimization*).

Respecto a la interpretabilidad, se incorporan dos metodologías ampliamente aceptados en el ámbito académico y empresarial: SHAP (SHapley Additive exPlanations), que proporciona explicaciones coherentes tanto globales como locales y LIME (Local Interpretable Model-Agnostic Explanations), que genera modelos lineales locales para explicar instancias concretas. Ambas técnicas

permiten identificar qué variables contribuyen más a la decisión del modelo, aspecto esencial en el análisis predictivo de reclamaciones.

Finalmente, se incorpora un agente basado en *Large Language Models* (LLM) para permitir la introducción de reclamaciones en lenguaje natural y generar un análisis automatizado. Este agente integra procesamiento de texto, extracción semántica de características, predicción del modelo entrenado y una explicación en lenguaje natural mediante un modelo local tipo “Mistral 7B Instruct” [19] ejecutado mediante *llama.cpp*, eliminando dependencias de servicios externos y evitando problemas de privacidad y coste.

2.2. Estado del arte en detección de fraude con inteligencia artificial

La detección automatizada de fraude en seguros ha experimentado una evolución significativa en las últimas décadas. Los primeros sistemas se basaban en enfoques expertos apoyados en reglas definidas manualmente por analistas, tales como la detección de reclamaciones repetidas, inconsistencias temporales o combinaciones específicas de variables. Aunque estos sistemas resultaron útiles en etapas iniciales, su rigidez y elevada dependencia del conocimiento experto limitaron su capacidad para adaptarse a nuevas tipologías de fraude, además de requerir un mantenimiento continuo y costoso.

Con la consolidación del aprendizaje automático, comenzaron a emplearse modelos estadísticos supervisados como la regresión logística, que ofrecían una base sólida para la clasificación binaria de siniestros fraudulentos. Sin embargo, la naturaleza principalmente lineal de estos modelos restringía su capacidad para capturar interacciones complejas entre variables. La introducción de algoritmos no lineales, como los árboles de decisión y, especialmente, los modelos de tipo ensemble como Random Forest, supuso un avance relevante. Estos modelos demostraron una mayor robustez frente a ruido, una mejor gestión de variables categóricas y una mayor capacidad para modelar relaciones

no lineales, lo que se tradujo en mejoras sustanciales en métricas como el recall y el F1-score en contextos de fraude altamente desbalanceados.

Posteriormente, los métodos de boosting, como Gradient Boosting y XGBoost, adquirieron protagonismo en la literatura académica y aplicada. Diversos estudios muestran que estos modelos logran resultados competitivos en la detección de fraude en seguros de automóvil, alcanzando valores de F1-score en el rango aproximado de 0,60–0,65 y recalls superiores al 70 % en datasets reales con una baja prevalencia de fraude. Un ejemplo representativo es el trabajo de Yankol-Schalck (2022) [20], que analiza la detección de fraude en seguros de automóvil utilizando Gradient Boosting sobre datos reales altamente desbalanceados, integrando además información textual procedente de descripciones de siniestros. En dicho estudio, la combinación de variables estructuradas y texto permite alcanzar un F1-score cercano a 0,64 y un ROC-AUC superior a 0,95, valores que se consideran una referencia relevante en el estado del arte.

De forma paralela al aumento del rendimiento predictivo, la investigación reciente ha puesto un énfasis creciente en la interpretabilidad de los modelos. En sectores regulados como el asegurador, la capacidad de explicar por qué un siniestro ha sido clasificado como fraudulento resulta fundamental tanto desde el punto de vista operativo como legal. En este contexto, técnicas de explicabilidad como SHAP y LIME se han consolidado como herramientas estándar para analizar la contribución de cada variable a las predicciones, permitiendo auditar el comportamiento del modelo y detectar posibles sesgos.

Más recientemente, la incorporación de técnicas de procesamiento de lenguaje natural ha ampliado el alcance de los sistemas de detección de fraude. Algunos trabajos integran descripciones textuales de los siniestros mediante representaciones basadas en n-gramas, TF-IDF o embeddings, mejorando la capacidad del modelo para capturar información semántica relevante. En esta línea, la aparición de modelos de lenguaje de gran escala (LLM) ha abierto

nuevas posibilidades para transformar texto libre en información estructurada y generar explicaciones automáticas en lenguaje natural.

El presente trabajo se alinea con estas tendencias recientes [21] [22], proponiendo una arquitectura híbrida que combina un modelo *Random Forest* optimizado, cuyos resultados se sitúan en un rango comparable a los reportados en la literatura, con técnicas de explicabilidad basadas en SHAP y LIME, y un agente LLM ejecutado de forma local para el procesamiento y la explicación de siniestros descritos en lenguaje natural.

Modelo / Estudio	Accuracy	Recall (fraude)	F1-score	ROC-AUC
Yankol-Schalck (2022)	—	≈ 0.75	≈ 0.64	≈ 0.95
Modelo propuesto	0.87	0.49	0.31	0.84

La comparación con el *baseline* de la literatura pone de manifiesto que, si bien existen enfoques que alcanzan valores superiores de *recall* y F1-score en entornos experimentales altamente especializados, el modelo desarrollado en este trabajo ofrece un equilibrio sólido entre capacidad predictiva, interpretabilidad y aplicabilidad práctica. En particular, los resultados obtenidos muestran que es posible alcanzar un rendimiento competitivo en términos de ROC-AUC utilizando modelos de complejidad moderada, siempre que se acompañen de estrategias adecuadas de balanceo de clases y optimización del umbral de decisión.

2.3. Datos

El conjunto de datos utilizado corresponde al “Vehicle Insurance Claim” de Oracle extraído de Kaggle, ampliamente empleado en estudios de detección de fraude. El *dataset* contiene 32 variables predictoras y una variable objetivo binaria (*FraudFound_P*) que indica si la reclamación fue clasificada como fraudulenta.

2.3.1. Variable objetivo (FraudFound_P)

La variable objetivo del presente estudio es *FraudFound_P*, una variable binaria que indica si una reclamación de seguro ha sido clasificada como fraudulenta (valor 1) o no fraudulenta (valor 0). Esta variable constituye el eje central del problema de clasificación abordado y condiciona de manera directa tanto la metodología empleada como las métricas de evaluación seleccionadas.

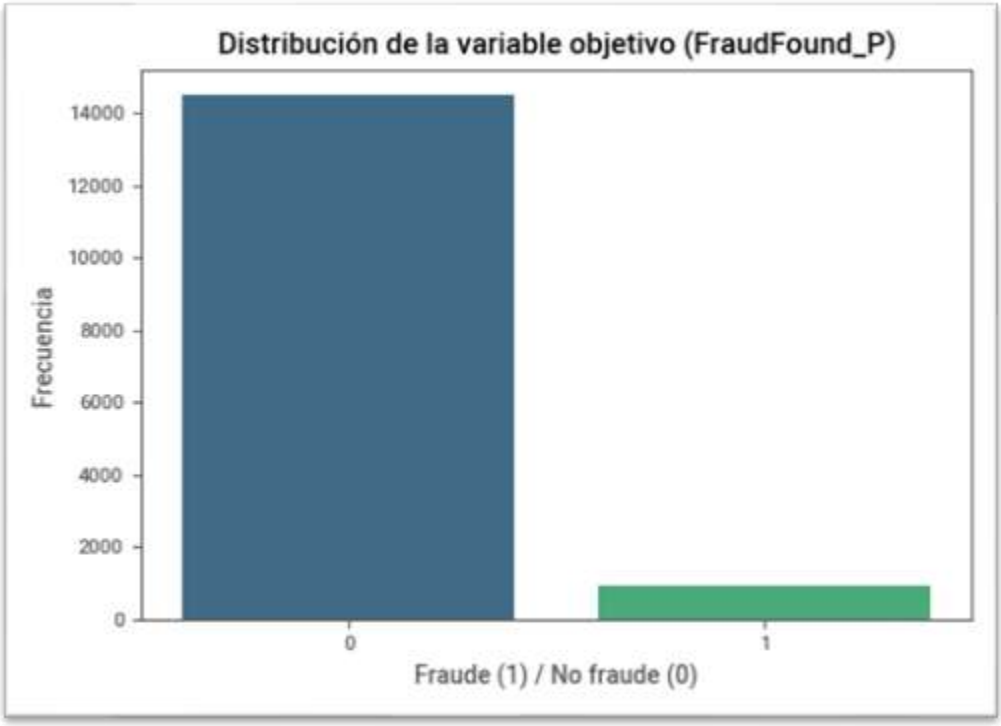

Figura 3: Distribución de variable objetivo

El análisis exploratorio inicial revela una fuerte descompensación entre clases, siendo los casos de fraude una proporción muy reducida del total de siniestros registrados. Tal como se muestra en la Figura 4, los siniestros no fraudulentos representan la inmensa mayoría del conjunto de datos, mientras que los casos etiquetados como fraude suponen menos del 10% del total. Este desequilibrio es característico en problemas reales de detección de fraude en seguros y plantea un desafío para los modelos de aprendizaje automático, que tienden a favorecer la clase mayoritaria si no se aplican técnicas específicas de mitigación.

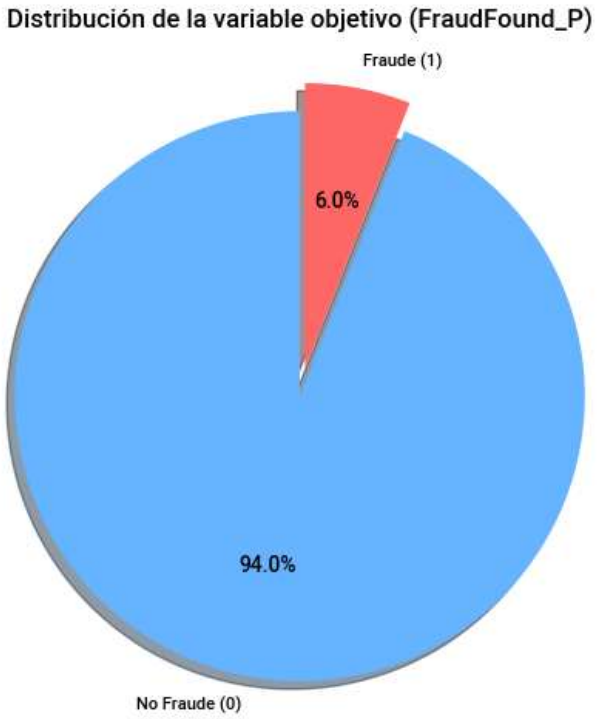

Figura 4: Distribución de porcentajes de variable objetivo

La presencia de este desbalanceo tiene implicaciones directas en la interpretación de las métricas de rendimiento. En este contexto, métricas globales como *accuracy* pueden resultar engañosas, ya que un modelo que clasifique todos los casos como “no fraude” obtendría una exactitud elevada sin aportar valor práctico. Por este motivo, el análisis posterior prioriza métricas sensibles a la clase minoritaria, como el *recall*, el F1-score y la curva ROC-AUC, así como estrategias específicas de balanceo y ajuste del umbral de decisión.

En consecuencia, el comportamiento observado en la variable objetivo justifica plenamente la adopción de técnicas de sobremuestreo, la evaluación comparativa de distintos enfoques de balanceo y la optimización del umbral de clasificación, aspectos que se desarrollan en los apartados posteriores.

2.3.2. Variables del asegurado

Las variables asociadas al asegurado recogen información demográfica y administrativa relevante para la detección de patrones de fraude. En este grupo se recogen, entre otras, las variables *Age*, *AgeOfPolicyHolder*, *Sex*, *MaritalStatus*, *NumberOfCars*, *PastNumberOfClaims* y *DriverRating*. El análisis exploratorio de estas variables permite identificar posibles concentraciones anómalas o comportamientos atípicos asociados a reclamaciones fraudulentas.

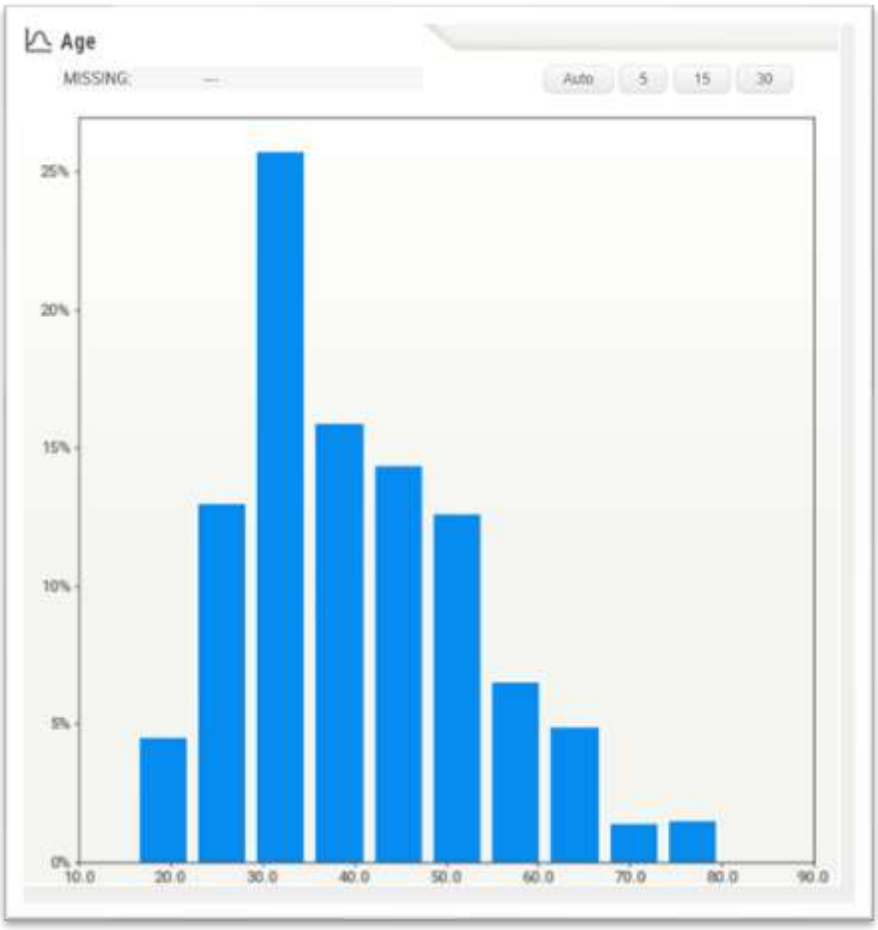

Figura 5: Distribución variable edad

En primer lugar, el análisis de la edad del asegurado y de la edad del titular de la póliza muestra que la mayoría de los siniestros se concentran en rangos de

edad intermedios, especialmente entre los 26 y 50 años, tal y como se aprecia en la Figura 5.

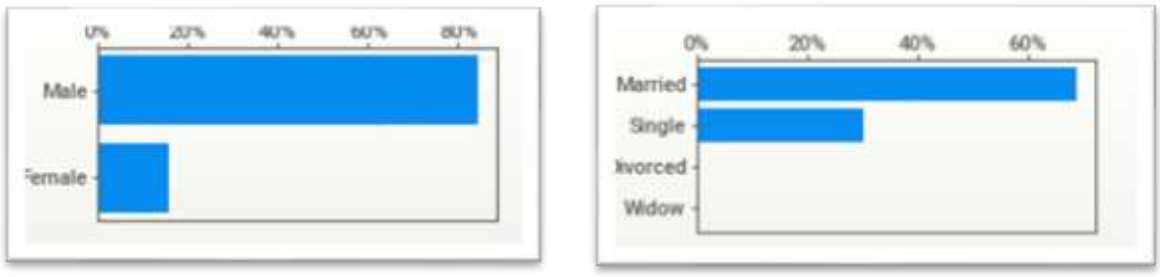

Figura 6: Distribución variables sexo y estado civil

Respecto al sexo y el estado civil, el conjunto de datos presenta una distribución sustancialmente desequilibrada entre hombres y mujeres, así como entre personas solteras y casadas (Figura 6).

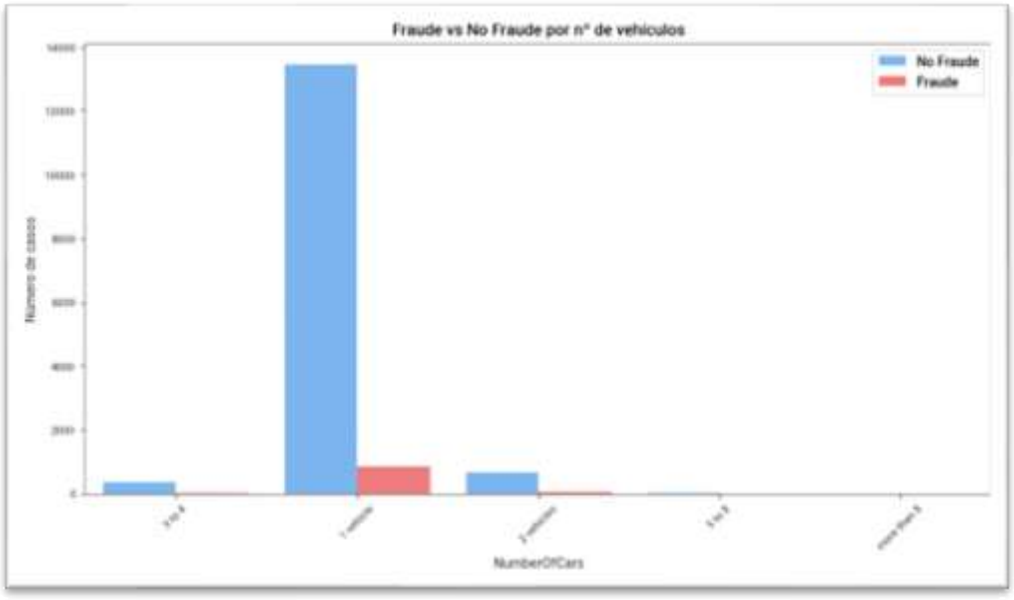

Figura 7: Distribución fraude por nº vehículos implicados

Por último, variables como *NumberOfCars* aportan información adicional sobre el perfil del asegurado. En general, los casos fraudulentos tienden a concentrarse en perfiles con un solo vehículo, aunque estas relaciones no son estrictamente lineales. Estas variables, combinadas con el resto de información disponible,

contribuyen a enriquecer la capacidad del modelo para capturar interacciones complejas entre el perfil del asegurado y el riesgo de fraude.

En conjunto, el análisis exploratorio de las variables del asegurado pone de manifiesto que, aunque ninguna de ellas resulta concluyente de manera aislada, su combinación proporciona informaciones relevantes que justifican su inclusión en el modelo predictivo y su posterior análisis mediante técnicas de explicabilidad.

2.3.3. Variables del vehículo

Las variables relacionadas con el vehículo aportan información relevante sobre las características materiales del siniestro y tienen un peso significativo en la detección de fraude. Entre ellas destacan la marca del vehículo (*Make*), la categoría del vehículo (*VehicleCategory*), el rango de precio (*VehiclePrice*) y la antigüedad del vehículo (*AgeOfVehicle*).

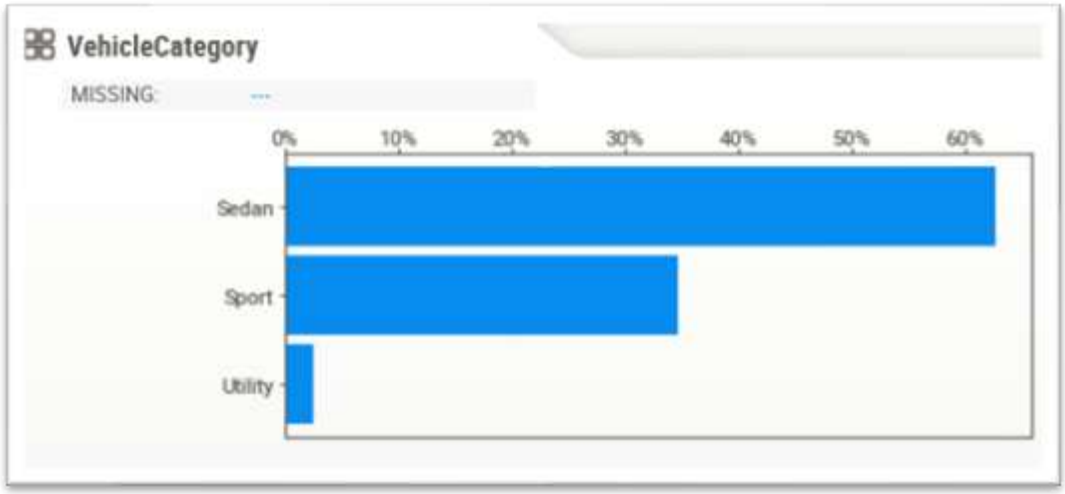

Figura 8: Distribución categoría vehículo

En primer lugar, la variable *VehicleCategory* muestra una clara predominancia de vehículos de tipo sedán y sport frente a categorías menos frecuentes como utilitarios (Figura 8). Esta distribución es coherente con la composición habitual del parque automovilístico, pero resulta relevante desde el punto de vista del fraude, ya que determinadas categorías pueden concentrar un mayor riesgo debido a su coste de reparación o perfil de uso.

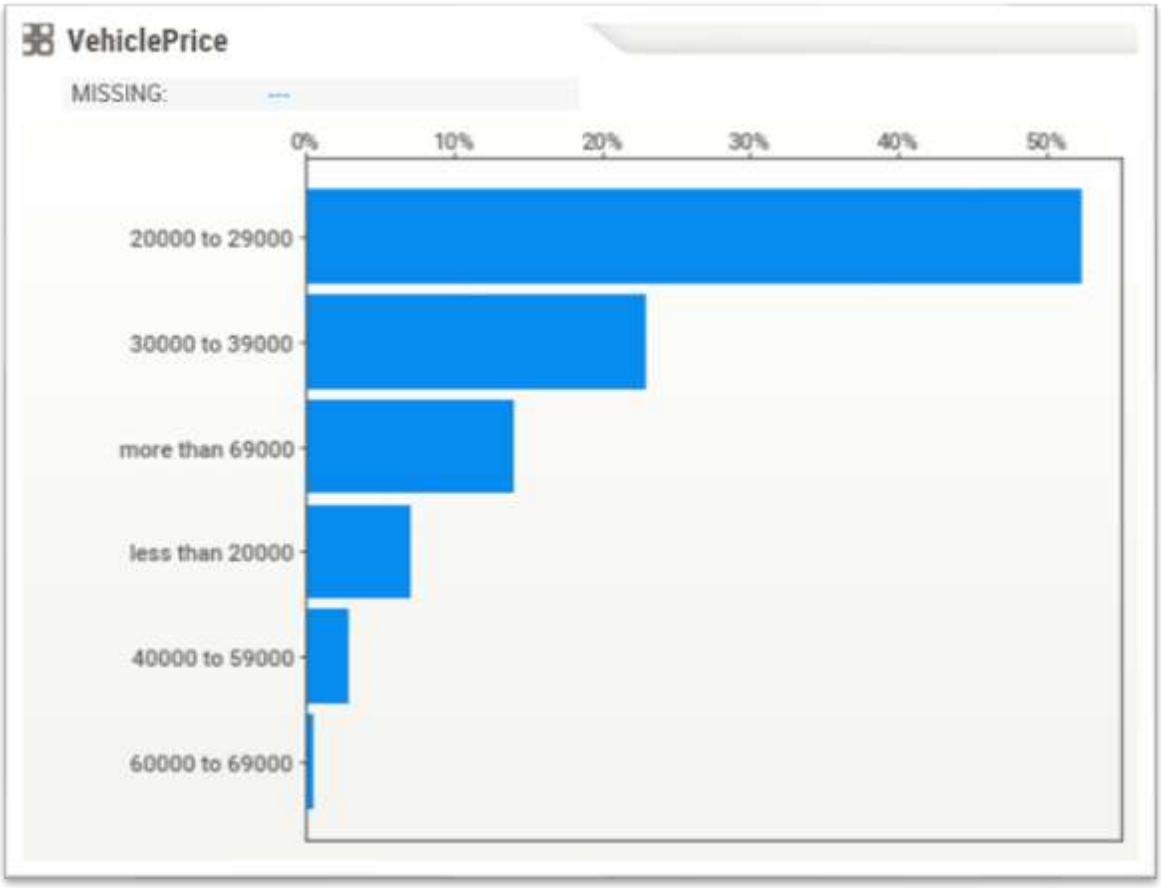

Figura 9: Distribución precio vehículos

En relación con el valor del vehículo, la variable *VehiclePrice* presenta una mayor concentración en rangos de precio medios, mientras que los vehículos de gama alta aparecen con menor frecuencia (Figura 9). No obstante, estos rangos superiores resultan especialmente interesantes para el análisis, ya que el valor económico del vehículo puede actuar como incentivo para la comisión de fraudes, tal y como se observa posteriormente en los análisis de explicabilidad.

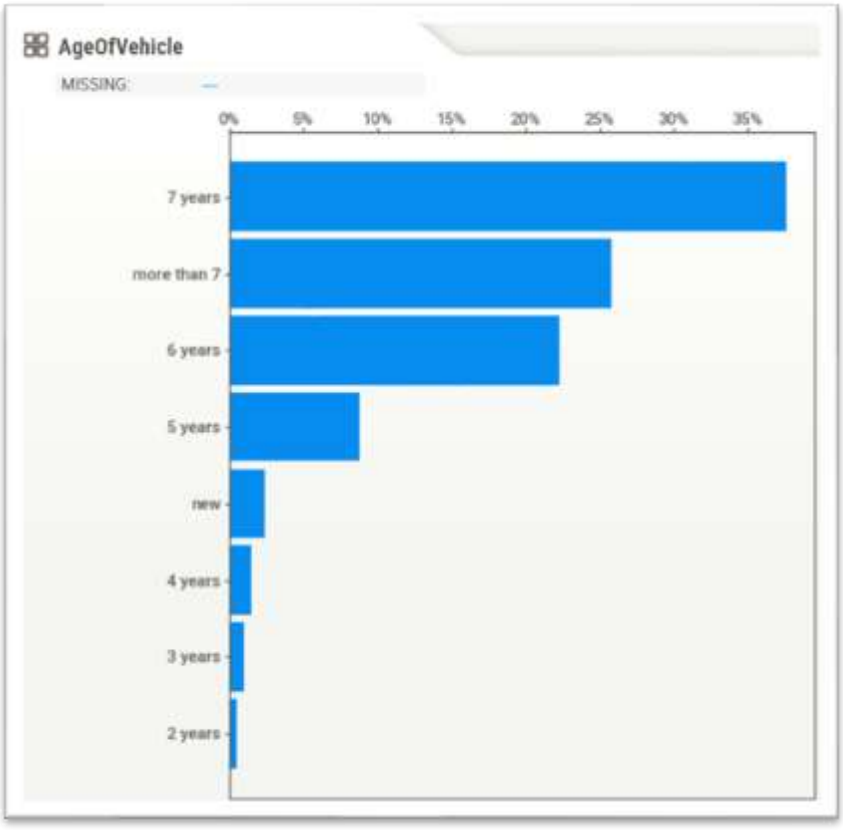

Figura 10: Distribución antigüedad vehículo

Por otro lado, la antigüedad del vehículo (*AgeOfVehicle*) muestra una distribución heterogénea de vehículos relativamente nuevos y de vehículos con varios años de antigüedad (Figura 10). Esta variable resulta especialmente relevante, ya que puede influir tanto en la probabilidad de sufrir un siniestro como en el coste asociado a la reparación o indemnización.

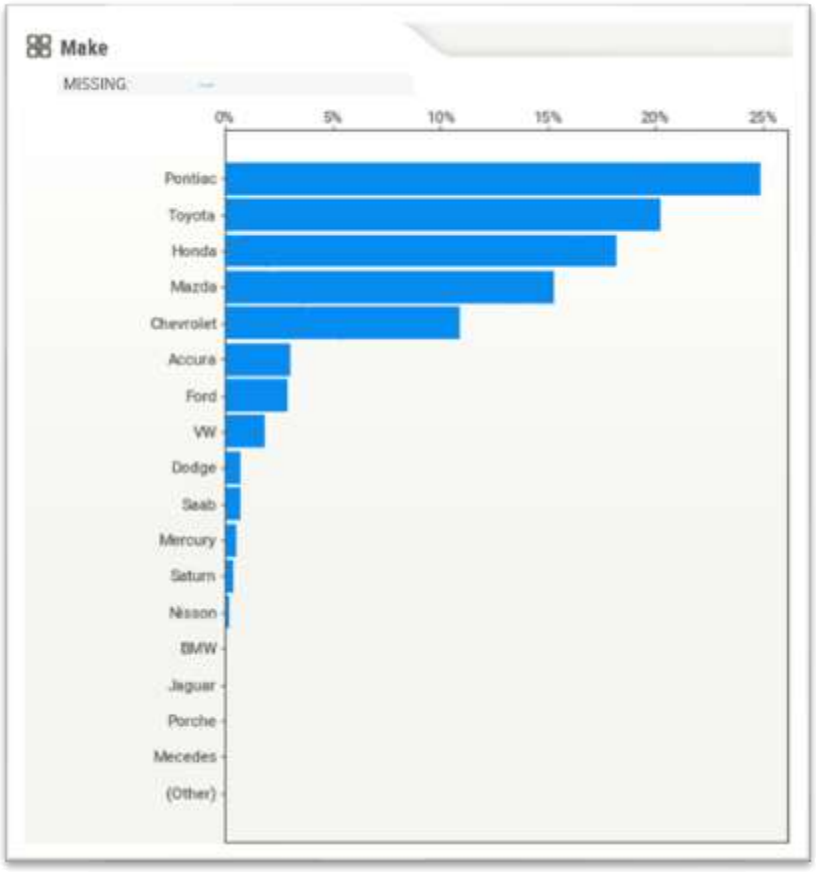

Figura 11: Distribución variable marcas

En relación con la variable *Make*, que identifica la marca del vehículo implicado en el siniestro, se observa una distribución claramente desigual entre las distintas categorías (Figura 11). Un número reducido de marcas concentra la mayor parte de las reclamaciones, destacando especialmente Pontiac, Toyota y Honda, que en conjunto representan una proporción significativa del total. El resto de marcas presentan frecuencias mucho menores, algunas de ellas testimoniales. Esta concentración sugiere que la variable *Make* refleja, en gran medida, la composición del parque automovilístico presente en el conjunto de datos más que un patrón directo de fraude. No obstante, su inclusión resulta relevante, ya que determinadas marcas pueden correlacionar indirectamente con otros factores de riesgo, como el valor del vehículo, la antigüedad o el tipo de póliza contratada.

El análisis de las variables del vehículo pone de manifiesto la existencia de patrones estructurales coherentes con la realidad dentro del sector asegurador, así como su potencial influencia en la detección de comportamientos fraudulentos.

2.3.4. Variables del siniestro

Las variables asociadas directamente al siniestro aportan información fundamental sobre el contexto en el que se produce la reclamación y permiten analizar patrones temporales, espaciales y de responsabilidad que pueden estar vinculados a comportamientos fraudulentos. En este grupo se incluyen variables como el área del accidente, la responsabilidad declarada, así como información temporal relativa al momento del siniestro y de la reclamación.

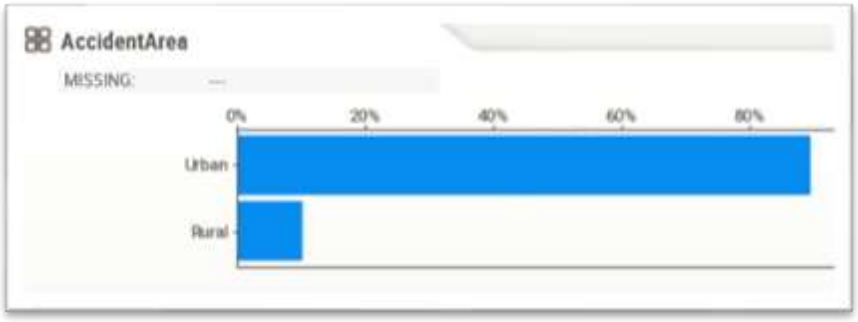

Figura 12: Distribución por área de siniestro

En relación con el área del accidente (*AccidentArea*), se observa una clara predominancia de siniestros ocurridos en zonas urbanas frente a zonas rurales (Figura 12). Esta distribución es coherente con la mayor densidad de tráfico en entornos urbanos, aunque también puede implicar una mayor complejidad en la reconstrucción del siniestro, lo que potencialmente incrementa el riesgo de fraude.

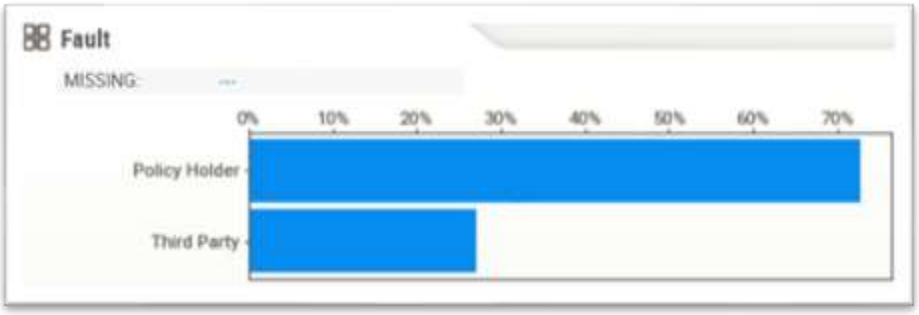

Figura 13: Distribución responsabilidad siniestro

La variable *Fault*, que indica la parte declarada como responsable del siniestro, muestra un patrón especialmente relevante. La proporción de casos en los que el asegurado figura como responsable es significativa (Figura 13), lo cual resulta de interés desde el punto de vista del análisis de fraude, ya que la asunción de responsabilidad por parte del tomador puede estar asociada a intentos de maximizar compensaciones en determinados contextos.

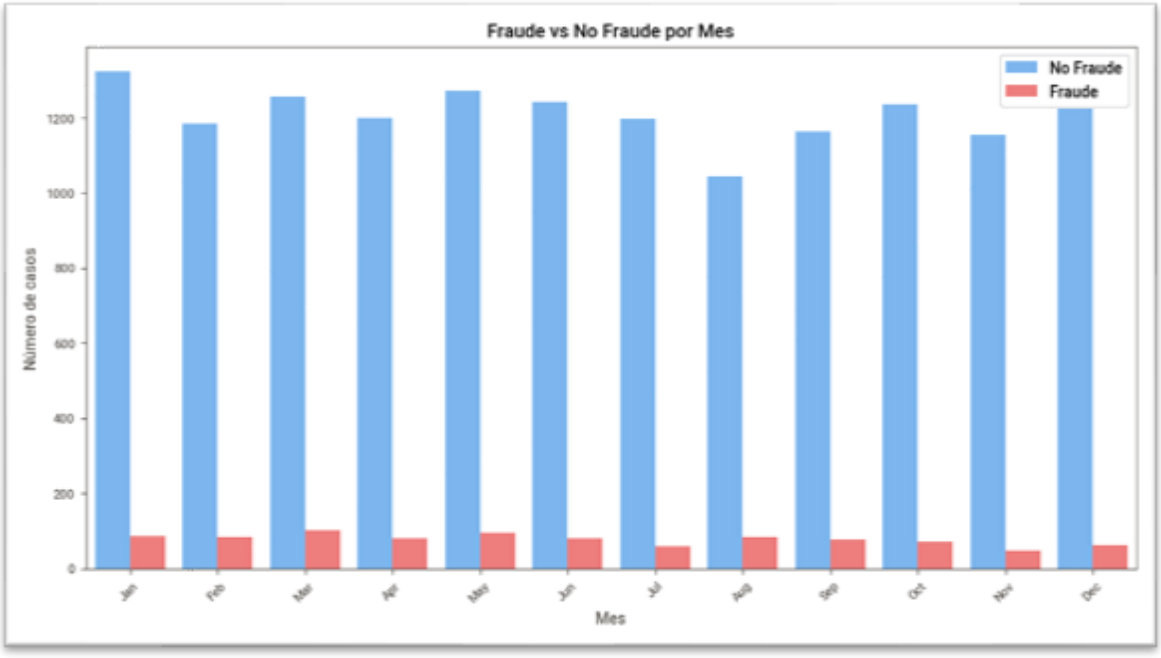

Figura 14: Nºcasos por mes de ocurrencia

En cuanto a las variables temporales, el análisis del mes (*Month*) revela una distribución relativamente homogénea a lo largo del año, con ligeras variaciones en determinados periodos (Figura 14). Este comportamiento sugiere que el

fraude no se concentra exclusivamente en momentos puntuales, sino que puede producirse de forma distribuida en el tiempo.

Las variables del siniestro capturan tanto el entorno físico como el comportamiento temporal asociado a cada reclamación. Su relevancia se verá justificada posteriormente en los análisis de explicabilidad, donde varias de estas variables aparecen entre las más influyentes del modelo final.

2.3.5. Variables administrativas

Las variables administrativas recogen información relacionada con el historial del asegurado y con el proceso de tramitación del siniestro. Aunque no describen directamente las circunstancias del accidente, estas variables resultan especialmente relevantes en la detección de fraude, ya que pueden reflejar patrones de comportamiento reiterados o inconsistencias en la gestión de la reclamación.

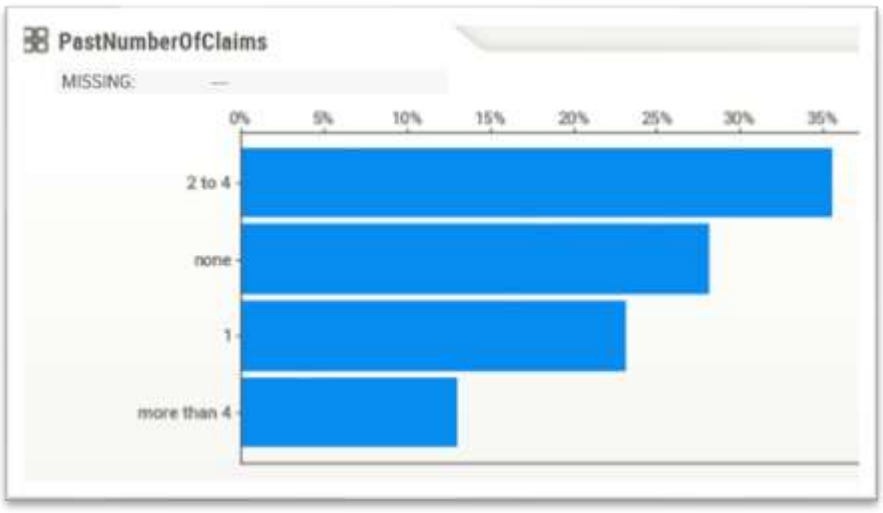

Figura 15: Distribución nºsinistros anteriores

El análisis del historial de reclamaciones previas (*PastNumberOfClaims*) muestra que la mayoría de los asegurados no presenta partes anteriores, mientras que un porcentaje reducido acumula múltiples reclamaciones (Figura 15). Este comportamiento es consistente con la realidad, donde la recurrencia de

siniestros se considera un indicador potencial de riesgo, especialmente cuando se combina con otros factores contextuales.

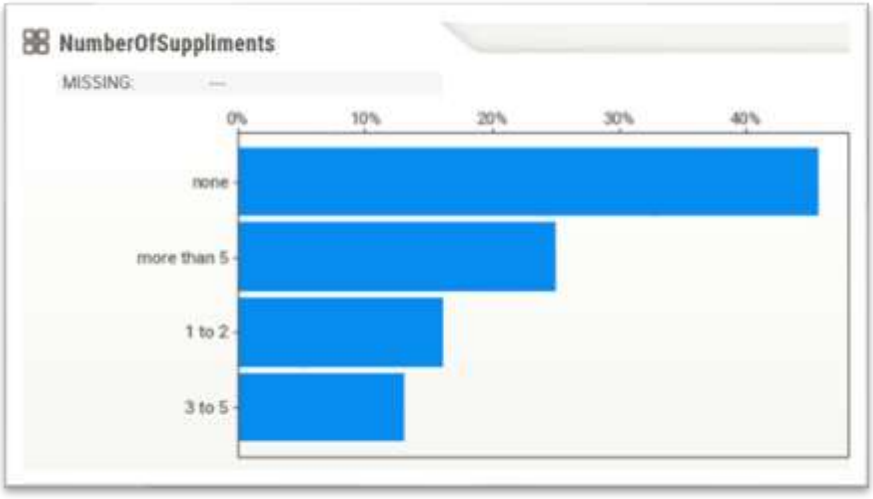

Figura 16: Distribución por nº sumplementos

En relación con el número de suplementos asociados a la póliza (NumberOfSuppliments), se observa que la mayor parte de los casos no presenta modificaciones contractuales, aunque existe un subconjunto de siniestros con varios suplementos añadidos (Figura 16). Este tipo de cambios frecuentes puede introducir complejidad administrativa y ha sido señalado en estudios previos como una posible señal de alerta en procesos de fraude.

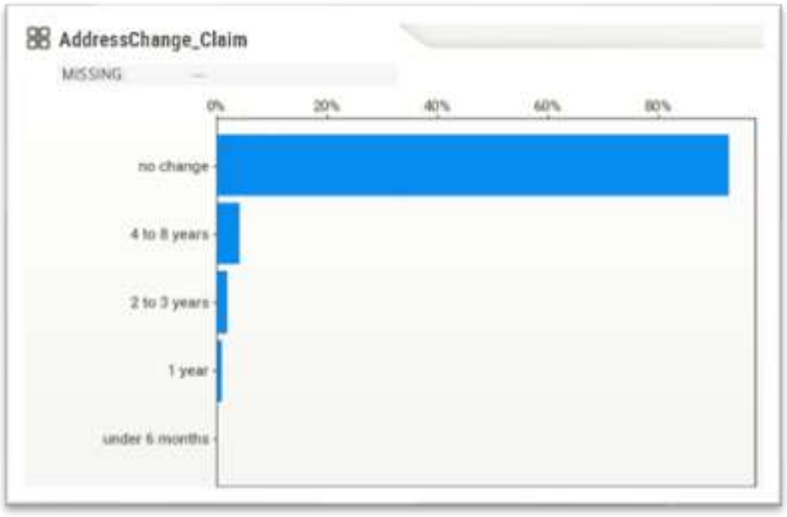

Figura 17: Distribución pólizas con cambios de domicilio

La variable *AddressChange_Claim* indica si se ha producido un cambio de domicilio cercano en el tiempo al siniestro. La mayoría de los asegurados no presenta cambios recientes, pero la presencia de modificaciones en la dirección declarada constituye un factor relevante desde el punto de vista del análisis de riesgo, ya que puede dificultar la trazabilidad del siniestro (Figura 17).

Finalmente, el número de vehículos asegurados (NumberOfCars) y el tipo de agente (AgentType) aportan información adicional sobre el perfil administrativo del asegurado. En conjunto, estas variables complementan el análisis de las características del siniestro y del vehículo, proporcionando una visión más completa del contexto en el que se produce la reclamación.

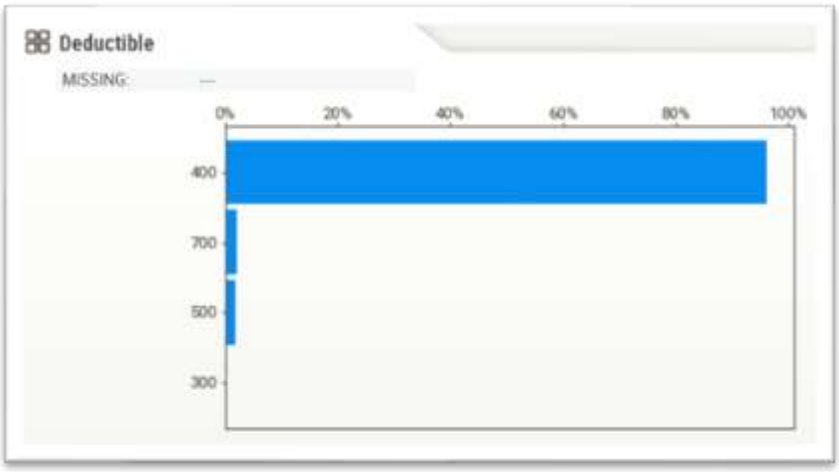

Figura 18: Distribución variable franquicias

La variable *Deductible* (Figura 18) presenta una distribución altamente concentrada, observándose que la gran mayoría de las pólizas tienen una franquicia de 400 dólares, mientras que el resto son residuales. Esta baja variabilidad indica que la franquicia está fuertemente estandarizada en el conjunto de datos, lo que limita su capacidad discriminativa a nivel global. No obstante, desde un punto de vista operativo, la franquicia puede influir en el comportamiento del asegurado, ya que valores más bajos reducen el coste asumido por el cliente y podrían incrementar el incentivo a presentar reclamaciones, lo que justifica su inclusión como variable relevante en el análisis de fraude.

El peso de varias de estas variables se confirma posteriormente en el análisis de explicabilidad del modelo, donde aparecen de forma recurrente entre los factores con mayor contribución a las predicciones de fraude.

2.3.6. Resumen de patrones relevantes y desbalanceos

En conclusión, los gráficos de barras muestran que los primeros meses del año son en los que más siniestros se declararon, aunque con escasa variación en el día de suceso. Existe mayor incidencia de siniestros en coches de gama alta como Pontiac y marcas japonesas (Toyota, Honda, Mazda), en seguros con franquicias de 400$ y vehículos de más de 5 años de antigüedad.

Los vehículos de más siniestralidad, o bien no llevan extras, o llevan más de 5 declarados. La edad de los Tomadores/Conductores de mayor siniestralidad está en la franja de 30 a 45 aproximadamente. Una gran parte de los siniestros son culpa del asegurado (73%) frente a culpa de contrario (27%). La mayor parte de siniestros se hallan en área urbana.

Se comunica el parte principalmente en día laborable(L-V) y durante todo el mes de manera homogénea salvo en la última semana. Hay más de 30 días entre la contratación de la póliza y la declaración/suceso del siniestro en prácticamente todos los partes. No hay denuncia a la policía ni testigos de los siniestros en más de un 90% de los partes. La mayoría son agentes externos a la Compañía y no han cambiado de dirección tras el siniestro. Un 93% solo tiene 1 vehículo implicado en siniestro.

Existe un desbalance en la variable objetivo *FraudFound_P* del 94% vs 6% así como en las variables: *AccidentArea*(90%), *Sex*(84% vs 16%), *Fault*(73% vs 27%), *Deductible*(96%), *DaysPolicyAccident*(99%), *DaysPolicyClaim*(99%), *PoliceReportFiled*(97%), *WitnessPresent*(99%), *AgentType*(98%), *AdressChange*(93%), *NumberofCars*(93%).

2.4. Preprocesado

El proceso de preprocesado del conjunto de datos previo a la modelización consistió en varios pasos. Primeramente, se procedió a realizar una limpieza e imputación de valores faltantes para, después, reasignar correctamente aquellos registros incorrectos de la variable *AgeOfPolicyHolder* para mantener esta información por abarcar la mayor parte de los registros comprendidos en la franja de edad de 16 a 17 años.

Posteriormente, se ha estudiado la correlación entre variables con el objetivo de detectar aquellas variables que presentan mayor correlación y aporten menos variabilidad a los modelos para una posible reducción de dimensionalidad. Se observa en la mayoría de las variables del *dataset* presentan relaciones lineales débiles entre sí, lo cual es característico de datos categóricos y dispersos asociados a siniestros de automóviles.

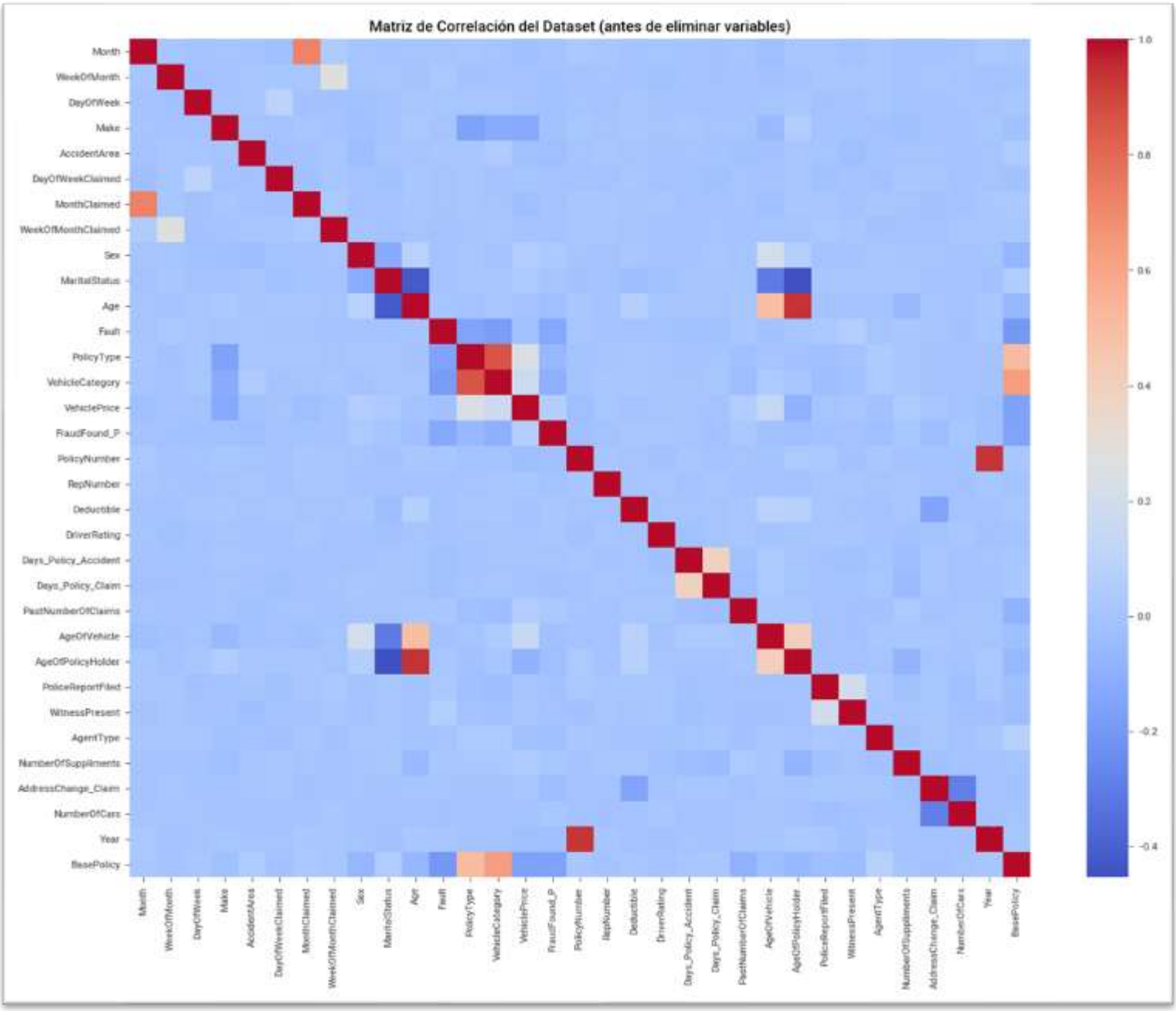

Figura 19: Matriz de correlación

Tras estudiar varias variables relacionadas con el fraude se decide la eliminación de *PolicyType*, *PolicyNumber* y *RepNumber* por ser redundantes o no aportar información añadida al modelo logrando reducir la dimensionalidad y mejorar el rendimiento de los modelos.

El preprocesado del *dataset* para su modelaje se mantuvo común para todas las metodologías probadas con el fin de garantizar comparabilidad y consistió en los siguientes pasos:

- Separación de variable objetivo y predictores.
- Identificación automática de variables categóricas mediante select_dtypes.
- Codificación ordinal de variables categóricas utilizando OrdinalEncoder, aplicando un valor reservado para categorías desconocidas (unknown_value = -1).
- Conversión homogénea de todas las variables a formato numérico, permitiendo su uso en modelos basados en árboles.
- División estratificada *train/test* (80/20) preservando la distribución original de clases.

2.5. Manejo del desbalanceo

Se evaluaron tres técnicas de balanceo: SMOTE (Generación de instancias sintéticas para la clase minoritaria, SMOTENC (Variante de SMOTE diseñada para datasets con variables categóricas codificadas) y *Random OverSampling(ROS)*.

Cada técnica se aplicó exclusivamente al conjunto de entrenamiento, manteniendo intacto el conjunto de test. ROS fue finalmente seleccionada por su mejor rendimiento combinado con *threshold optimization* en el modelo *Random Forest.*

2.6. Modelos y experimentación

REGRESIÓN LOGÍSTICA

La Regresión Logística es uno de los modelos lineales más utilizados en problemas de clasificación binaria. En este modelo, la relación entre las variables predictoras y la probabilidad de fraude se expresa mediante la función sigmoide, que transforma una combinación lineal de las características en un valor comprendido entre 0 y 1.

La probabilidad estimada de fraude viene dada por:

$$p = \frac{1}{1 + e^{-z}}$$

Este modelo es rápido, interpretable y permite analizar la dirección (positiva o negativa) de la influencia de cada variable sobre la probabilidad de fraude. Sin embargo, al asumir una frontera de decisión lineal, puede mostrar limitaciones cuando el patrón de fraude depende de interacciones complejas o relaciones no lineales entre variables.

Este hecho es habitual en la detección de fraude en seguros, donde la interacción entre características del asegurado, del vehículo y de la reclamación suele ser esencial. En este proyecto, la regresión logística se empleó como línea base, proporcionando un punto de referencia para comparar modelos más complejos.

ÁRBOL DE DECISIÓN

Los árboles de decisión dividen el espacio de variables de manera jerárquica, generando reglas de decisión basadas en las características que mejor separan los casos fraudulentos de los no fraudulentos. Para cada división se utiliza típicamente el índice de Gini:

$$G = 1 - \sum_{i=1}^{C} p_i^2$$

Este modelo permite capturar relaciones no lineales, considerar interacciones entre variables y trabajar con datos categóricos y numéricos sin reescalado. Sin embargo, son inestables y tienden al sobreajuste si no se podan adecuadamente y el rendimiento de un único árbol suele ser insuficiente.

RANDOM FOREST

El *Random Forest* constituye una mejora significativa sobre los árboles individuales, al formar un conjunto de múltiples árboles entrenados sobre el conjunto de entrenamiento y subconjuntos aleatorios de variables en cada división. Esto introduce diversidad y reduce el sobreajuste. La predicción final se obtiene por votación mayoritaria entre los árboles.

Ente las ventajas más relevantes de este modelo están que maneja un gran número de variables categóricas codificadas ordinalmente, que es robusto a ruido, valores extremos y correlaciones, además de que su rendimiento es muy estable incluso en presencia de desbalanceo, especialmente cuando se combina con técnicas de sobre muestreo como ROS o SMOTE.

En este trabajo, se entrenaron y compararon varios clasificadores bajo distintos esquemas con: SMOTE, SMOTENC, *class_weight='balanced',* ROS / ROS + optimización de umbral y búsqueda de hiperparámetros mediante *RandomizedSearchCV*. Estos fueron comparados obteniendo como mejor resultado el de *Random Forest* + ROS + optimización de umbral mejorado posteriormente mediante *RandomizedSearchCV*

Este modelo final “best_rf” se utilizó para toda la fase de explicabilidad y para integrarse en el agente LLM.

XGBOOST

XGBoost (Extreme Gradient Boosting) es un algoritmo basado en *boosting*, en el que cada nuevo árbol se entrena para corregir los errores del conjunto anterior. A diferencia de RF, donde los árboles son independientes, XGBoost genera árboles secuenciales mediante la minimización de una función objetivo regularizada.

Aunque es un modelo altamente efectivo en tareas de fraude debido a su capacidad para capturar patrones complejos y a su manejo eficiente del ruido y en contextos con desbalance, en el presente trabajo se descartó como modelo final por dos motivos principales: su mayor sensibilidad a la calibración del

umbral y su menor interpretabilidad cuando se combina con explicadores en comparación *con Random Forest*.

2.7. Optimización de hiperparámetros

La optimización se realizó mediante búsqueda aleatoria y evaluada con *F1-score* y validación cruzada. Una vez obtenido el mejor estimador, se ajustó el *threshold* de decisión usando la curva *precision–recall*, seleccionando el valor que maximizaba el *F1*.

2.8. Técnicas de explicabilidad

En este proyecto se emplearon dos técnicas complementarias de explicabilidad: SHAP y LIME.

SHAP permite explicar cada predicción localmente, obtener la importancia de las características y visualizar mediante gráficos dependencia y summary, complementando al agente del modelo LLM. Además, permitió identificar sistemáticamente las variables con mayor contribución a las predicciones del modelo que fueron Fault, BasePolicy, VehicleCategory, VehiclePrice, AgeOfVehicle, entre otras.

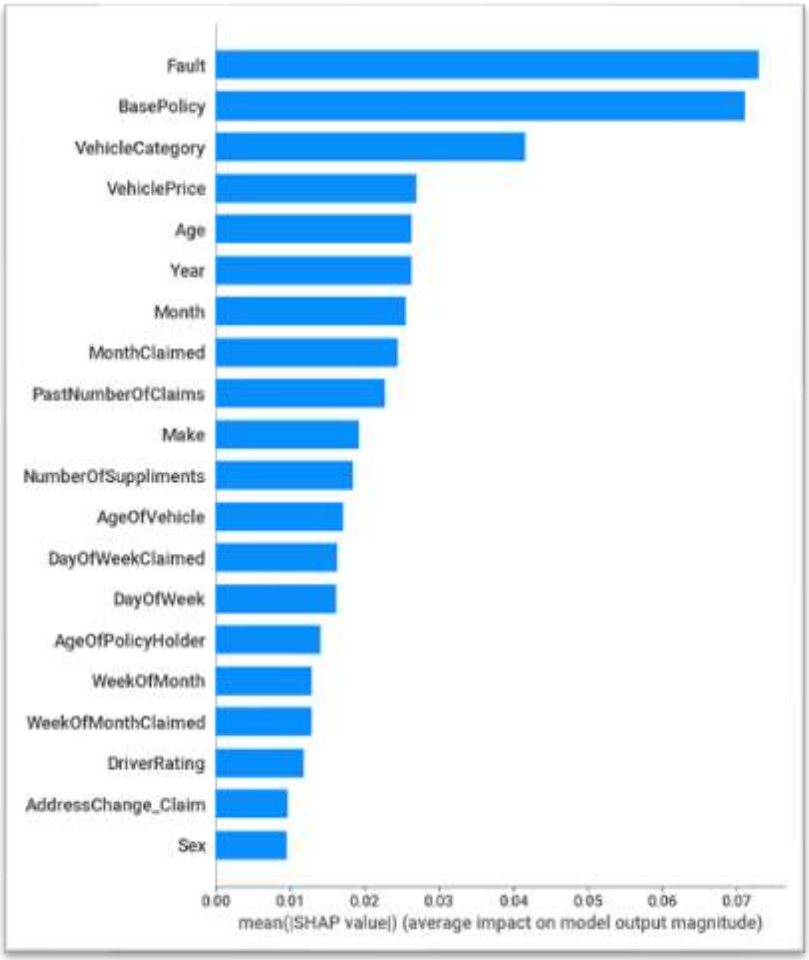

Figura 20: Gráfico importancia de variables SHAP

Se aplicó LIME como técnica complementaria para la explicación local de predicciones puntuales y se utilizó para validar la coherencia de las explicaciones producidas por SHAP, especialmente en casos frontera.

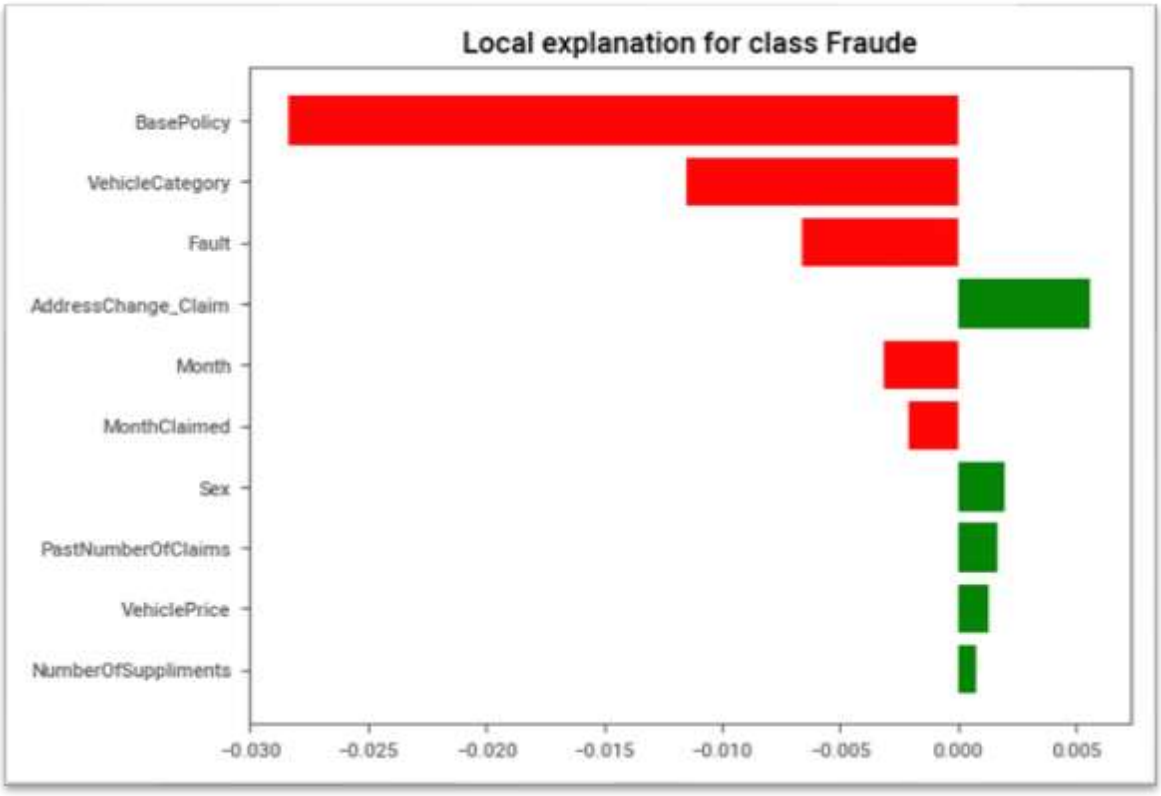

Figura 21: Explicabilidad para clase de fraude LIME

2.9. Agente LLM integrado

En el presente trabajo, se ha desarrollado un agente experto capaz de recibir una reclamación en lenguaje natural y extraer las características estructuradas (parseo) mediante un modelo LLM local (Mistral 7B Instruct) usando un modelo LLM (*llama.cpp)*. También se realiza un mapeo de textos a valores numéricos compatibles con el *Random Forest* mediante un diccionario de categorías para ejecutar el modelo predictivo (best_rf) y generar una explicación técnica utilizando el mismo LLM.

Este diseño permite un flujo de trabajo end-to-end totalmente autónomo:

2.10. Entorno técnico

El desarrollo del proyecto se llevó a cabo utilizando Python 3.12 como lenguaje principal, debido a su amplia adopción en el ámbito del análisis de datos, la ciencia de datos y el aprendizaje automático. Todo el proceso experimental se realizó en Jupyter Notebook permitiendo un flujo de trabajo iterativo y transparente, especialmente útil para analizar el comportamiento de los modelos.

Para la construcción de los modelos predictivos se empleó la biblioteca *scikit-learn,* para los algoritmos de clasificación, las herramientas de preprocesado y los procedimientos de validación utilizados. El manejo del desbalanceo de clases se gestionó mediante *imbalanced-learn.* El procesamiento y manipulación de datos se realizó con *pandas* y *numpy*, mientras que la creación de representaciones visuales de gráficos exploratorios se realizó en *matplotlib*, *seaborn* y *sweetviz*.

La explicabilidad del modelo se ha abordado mediante dos bibliotecas especializadas: SHAP que permite obtener explicaciones tanto globales como locales, basadas en valores de Shapley, y LIME que permite generar explicaciones lineales locales para casos individuales.

Finalmente, para integrar el componente basado en lenguaje natural se utilizó la librería *llama-cpp-python*, que permite ejecutar un modelo de lenguaje de gran tamaño (LLM) de manera completamente local. El modelo empleado fue “Mistral-7B-Instruct-v0.2.Q4_K_M.gguf”, alojado y ejecutado sin conexión a servicios externos. Esta decisión técnica es clave para garantizar la privacidad de los datos, evitar dependencias con APIs comerciales y reducir el coste computacional asociado al uso de modelos en la nube.

3. Resultados

En este apartado se presentan los resultados experimentales obtenidos con los distintos modelos y técnicas aplicadas. Se incluyen comparaciones métricas, análisis del modelo final y evaluación de los métodos de explicabilidad y del agente LLM integrado.

3.1. Comparación global de modelos y estrategias

En este apartado se presenta una comparación global de los distintos modelos y estrategias de balanceo evaluados a lo largo del trabajo, con el objetivo de identificar de forma clara cuál de ellos ofrece el mejor rendimiento para la detección de fraude en siniestros de automóvil. Dado el fuerte desbalanceo existente en la variable objetivo, la evaluación no se ha centrado únicamente en la métrica de accuracy, sino que se ha puesto especial énfasis en métricas más representativas del problema, como el recall de la clase fraudulenta, el F1-score y el área bajo la curva ROC (ROC-AUC). La Tabla X resume los resultados obtenidos para cada combinación de modelo y técnica aplicada, permitiendo una comparación directa y facilitando la selección del modelo final. Esta visión agregada resulta fundamental para comprender el impacto real de cada estrategia y justifica las decisiones tomadas en las fases posteriores de optimización y análisis detallado.

Modelo / Estrategia	Accuracy	Precision	Recall (fraude)	F1-score	ROC-AUC
Random Forest + SMOTE	0.940	0.54	**0.04**	**0.07**	0.830
Random Forest + SMOTENC	0.936	0.26	**0.04**	**0.07**	0.802
Random Forest + class_weight	0.941	1.00	**0.02**	**0.03**	0.840
Random Forest + ROS	0.914	0.20	**0.46**	**0.28**	0.841
Random Forest + ROS + threshold opt.	0.869	0.23	**0.49**	**0.31**	0.844
Random Forest optimizado (best_rf)	0.872	0.24	**0.51**	**0.33**	0.848

Aunque la precision se ha calculado para todos los modelos, no se ha tenido en cuenta como métrica principal en la tabla comparativa debido a que el objetivo

prioritario del sistema es maximizar la detección de fraude (recall), manteniendo un equilibrio razonable mediante el F1-score.

3.2. Optimización del modelo final

Tras la comparación global de modelos y estrategias de balanceo, el modelo Random Forest combinado con Random Oversampling (ROS) fue seleccionado como base para la optimización final, al mostrar el mejor compromiso entre capacidad predictiva y estabilidad en la detección de la clase fraudulenta. No obstante, los resultados iniciales evidenciaron que el rendimiento del modelo podía mejorarse mediante un ajuste más fino tanto del umbral de decisión como de los hiperparámetros internos del clasificador.

En primer lugar, se abordó la optimización del umbral de decisión. Dado que el modelo produce probabilidades continuas de pertenencia a la clase fraudulenta, se analizó la curva precision–recall para identificar el valor de probabilidad que maximizaba el F1-score. Este análisis permitió desplazar el umbral desde el valor estándar de 0,5 hasta un valor aproximado de 0,20, logrando un incremento sustancial del recall sin degradar en exceso la precisión. Este ajuste resulta especialmente relevante en un contexto de detección de fraude, donde el coste de no detectar un fraude suele ser superior al de revisar un falso positivo.

Posteriormente, se llevó a cabo una optimización de hiperparámetros mediante RandomizedSearchCV, utilizando validación cruzada y el F1-score como métrica objetivo. El proceso permitió identificar una configuración óptima caracterizada por un mayor número de árboles y una profundidad superior, lo que indica que el problema requiere modelos con suficiente capacidad para capturar interacciones no lineales entre múltiples variables. Asimismo, el hecho de que la opción bootstrap=False resultara óptima sugiere que, en presencia de una clase minoritaria muy reducida, resulta más eficaz entrenar árboles sobre el conjunto de datos sobremuestreado completo, reduciendo la variabilidad introducida por el muestreo interno.

Como resultado de este proceso, el modelo final (best_rf) alcanza un recall cercano al 50 %, con un F1-score en torno a 0,30 y un ROC-AUC superior a 0,84. Este rendimiento lo posiciona como una herramienta adecuada para tareas de preclasificación y priorización de siniestros sospechosos, sirviendo como apoyo a los analistas humanos en la fase inicial de revisión.

3.3. Análisis mediante Lift y Gain Chart

Con el objetivo de evaluar la capacidad del modelo final para priorizar correctamente los siniestros más sospechosos de fraude, se han utilizado las métricas de *Gain Chart* y *Lift Chart* [23]. Estas herramientas permiten analizar no solo el rendimiento global del clasificador, sino también su utilidad práctica en un entorno real, donde los recursos de revisión son limitados y resulta necesario priorizar los casos con mayor riesgo.

El *Gain Chart* muestra el porcentaje acumulado de fraude capturado en función del porcentaje de siniestros revisados, ordenados de mayor a menor probabilidad estimada por el modelo. En la Figura X se observa que el modelo Random Forest supera ampliamente al comportamiento aleatorio. En particular, al revisar aproximadamente el **20 % de los siniestros con mayor score**, el modelo es capaz de capturar **en torno al 60 % del fraude total**, y al analizar cerca del **40 %**, se alcanza casi el **90 % de los casos fraudulentos**. Este comportamiento evidencia una alta capacidad de priorización, lo que permitiría a una aseguradora concentrar los esfuerzos de investigación en un subconjunto reducido de expedientes con un retorno significativamente mayor.

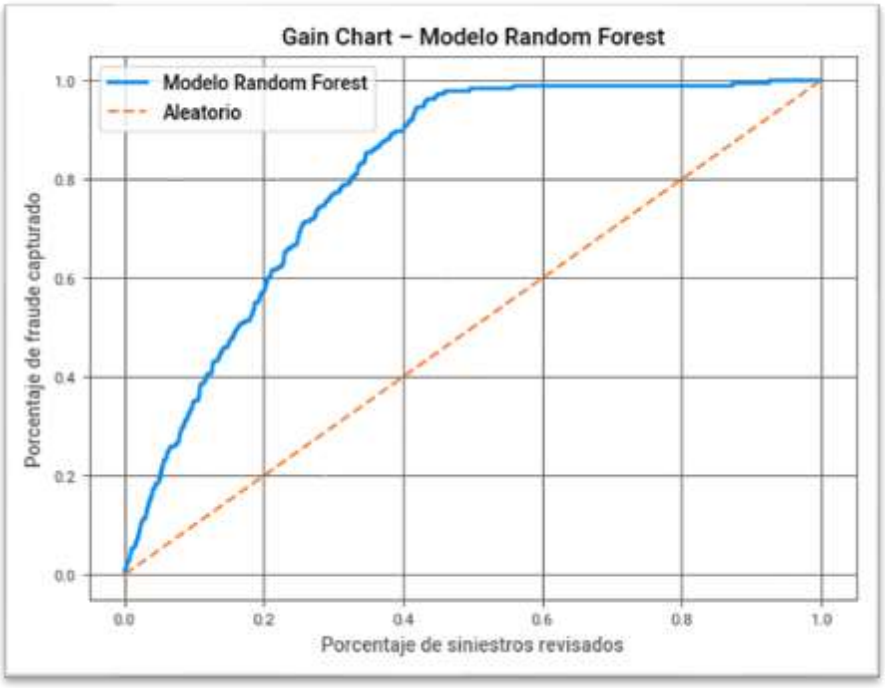

Figura 22: Gain chart

Por su parte, el *Lift Chart* refuerza esta conclusión al mostrar el rendimiento relativo del modelo por deciles de riesgo. En la Figura Y se aprecia que el primer decil, correspondiente al 10 % de siniestros con mayor probabilidad estimada de fraude, presenta un *lift* superior a 3, lo que implica que en ese grupo se detecta más de tres veces el fraude esperado bajo una selección aleatoria. A medida que se avanza hacia deciles de menor riesgo, el *lift* disminuye progresivamente, lo que confirma que el modelo concentra de forma efectiva los casos más sospechosos en los primeros segmentos. Este patrón es coherente con un sistema de detección de fraude bien calibrado y resulta especialmente valioso desde el punto de vista operativo.

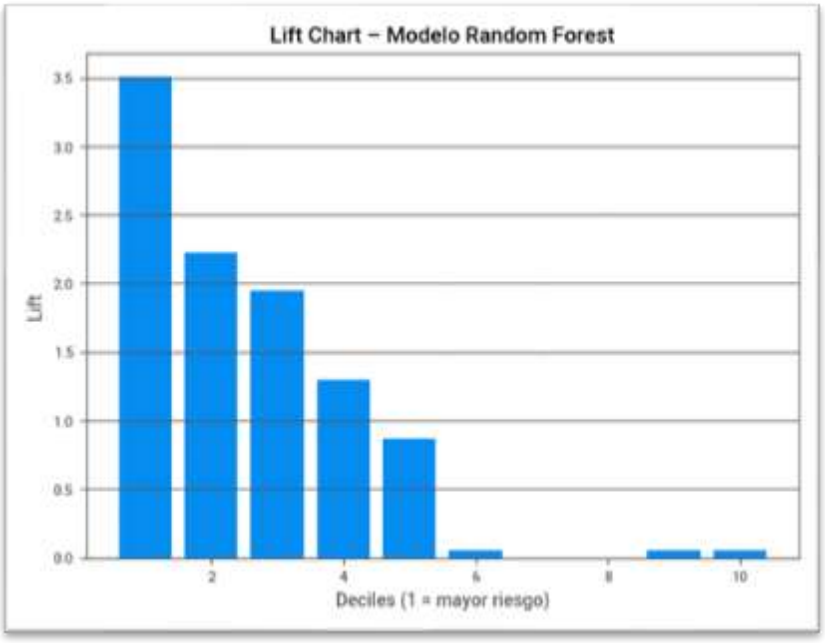

Figura 23: Lift chart

3.4. Evaluación del agente LLM

Las pruebas realizadas en el agente basado en el modelo local "Mistral 7B Instruct" usando *llama.cpp*. han permitido la extracción automática de características relevantes y explicativas sobre casos reales reflejados como texto, aportando descripciones sintéticas de fraude y en otros casos probados, reclamaciones no fraudulentas.

El agente fue capaz de extraer correctamente todas las variables categóricas cuando estaban presentes en el texto, generar predicciones consistentes con best_rf, producir explicaciones con concordancia e identificar correctamente casos altamente fraudulentos al introducir la descripción textual del caso real de fraude.

Con el fin de evaluar de forma cualitativa el comportamiento del agente basado en modelos de lenguaje y su integración con el clasificador *Random Forest*, se presentan a continuación varios ejemplos concretos de entrada y salida del sistema. Estos casos permiten ilustrar cómo el agente interpreta descripciones de siniestros en lenguaje natural, extrae las variables relevantes, ejecuta el

modelo predictivo y genera explicaciones comprensibles sobre la probabilidad estimada de fraude:

※ Ejemplo 1. Caso con alta probabilidad de fraude

La descripción introducida corresponde a un siniestro ocurrido en área urbana, con responsabilidad del asegurado, vehículo de gama media-alta y cambios recientes en la dirección declarada. El modelo asigna una probabilidad de fraude superior al 80 %, clasificando el caso como fraudulento. Las explicaciones generadas mediante SHAP identifican como variables más *influyentes Fault, BasePolicy, VehiclePrice y AddressChange_Claim*. La explicación textual generada por el agente es coherente con estas señales, indicando un patrón típico de riesgo elevado.

※ Ejemplo 2. Caso con probabilidad intermedia

En este caso, la reclamación describe un accidente leve sin testigos, con un vehículo antiguo y póliza estándar. El modelo estima una probabilidad de fraude cercana al umbral de decisión, clasificándolo como caso dudoso. La explicación refleja la coexistencia de factores de riesgo moderado y variables neutras, lo que justifica una recomendación de revisión manual prioritaria, pero no automática.

※ Ejemplo 3. Caso no fraudulento

La descripción corresponde a un siniestro con responsabilidad de un tercero, vehículo de bajo valor y sin cambios administrativos recientes. El modelo asigna una probabilidad de fraude inferior al 10 %, clasificando el caso como no fraudulento. Las variables explicativas muestran contribuciones negativas al fraude, y la explicación generada por el agente resulta consistente con un escenario de bajo riesgo.

Ejemplo	Tipo de caso	Probabilidad de fraude	Clasificación final	Variables más influyentes (SHAP)	Interpretación
1	Alto riesgo	0.81	Fraude	Fault, BasePolicy, VehiclePrice, AddressChange_Claim	El modelo identifica múltiples señales clásicas de fraude, priorizando correctamente el siniestro para revisión.
2	Riesgo medio	0.47	No Fraude (caso dudoso)	VehicleCategory, Fault, AgeOfVehicle, BasePolicy	Caso cercano al umbral, recomendable para revisión manual por la combinación de factores mixtos.
3	Bajo riesgo	0.09	No Fraude	AgeOfVehicle, AgeOfPolicyHolder, VehiclePrice	La ausencia de señales administrativas y económicas reduce significativamente la probabilidad de fraude.

Estos ejemplos evidencian que el agente LLM no solo es capaz de traducir texto libre a variables estructuradas de forma coherente, sino también de generar explicaciones alineadas con el comportamiento del modelo predictivo. Esta capacidad refuerza su utilidad como herramienta de apoyo a la toma de decisiones en entornos reales de evaluación de siniestros.

3.5. Discusión

La detección de fraude en seguros constituye un problema caracterizado por una fuerte descompensación de clases, heterogeneidad en las variables y presencia de ruido en los datos. Por este motivo, los resultados obtenidos en este trabajo deben interpretarse teniendo en cuenta las dificultades inherentes al dominio y las limitaciones prácticas asociadas a la identificación temprana de comportamientos fraudulentos.

En primer lugar, la comparación de técnicas de balanceo pone de manifiesto que los métodos de sobremuestreo sintético, como SMOTE y SMOTENC, no proporcionaron mejoras significativas en este caso concreto. Aunque en la literatura se emplean habitualmente como soluciones estándar para el tratamiento del desbalanceo, su efectividad se vio limitada por la elevada proporción de variables categóricas y por la complejidad semántica del *dataset*. Los modelos entrenados con estas técnicas alcanzaron valores de precisión moderados, pero presentaron un *recall* extremadamente bajo, en torno al 3–4 %, lo que los hace poco adecuados en un contexto donde los falsos negativos suponen un elevado coste económico.

Por el contrario, el uso de *Random Oversampling* (ROS) ofreció un comportamiento claramente superior. A pesar de su simplicidad y del riesgo potencial de sobreajuste asociado a la duplicación de instancias minoritarias, esta técnica permitió mejorar de forma notable la sensibilidad del modelo hacia la clase fraudulenta. Combinado con una optimización del umbral de decisión, el modelo final fue capaz de detectar aproximadamente el 50 % de los casos de fraude, manteniendo al mismo tiempo una precisión razonable para una tarea de preclasificación. Este equilibrio resulta especialmente relevante en escenarios reales, donde el objetivo no es automatizar la decisión final, sino priorizar los casos con mayor riesgo para su revisión manual.

La optimización de hiperparámetros mediante RandomizedSearchCV permitió refinar aún más el modelo. Los mejores resultados se obtuvieron con bosques más profundos y un número elevado de árboles, lo que sugiere que el problema

requiere modelos capaces de capturar interacciones no lineales complejas entre múltiples variables. El hecho de que la configuración óptima incluyera bootstrap=False indica que, en presencia de una clase minoritaria altamente estructurada, resulta beneficioso entrenar árboles más consistentes y menos dependientes del re-muestreo interno.

Más allá de las métricas clásicas de clasificación, el análisis mediante Gain Chart y Lift Chart aporta una perspectiva operativa clave. Los resultados muestran que el modelo es capaz de concentrar una elevada proporción del fraude en los primeros percentiles de riesgo. En particular, revisando aproximadamente el 20 % de los siniestros con mayor score, se captura alrededor del 60 % del fraude total, mientras que el primer decil presenta un lift superior a 3 respecto a una selección aleatoria. Estos resultados confirman que el modelo no solo ofrece un buen rendimiento estadístico, sino que resulta especialmente útil como herramienta de priorización en entornos reales con recursos de investigación limitados.

Desde el punto de vista de la explicabilidad, los análisis realizados con SHAP y LIME revelan patrones coherentes y estables. Variables como Fault, BasePolicy, VehicleCategory, VehiclePrice y AddressChange_Claim aparecen de forma recurrente entre las más influyentes, tanto a nivel global como local. Estos factores coinciden con señales de riesgo tradicionalmente consideradas relevantes en la detección de fraude en seguros. La concordancia entre ambas técnicas de explicabilidad refuerza la confianza en el modelo y facilita su aceptación en contextos regulados, donde las decisiones automatizadas deben poder ser auditadas y justificadas.

Finalmente, la integración de un agente basado en modelos de lenguaje añade una capa adicional de accesibilidad y funcionalidad. La posibilidad de introducir descripciones de siniestros en lenguaje natural y obtener una predicción acompañada de una explicación técnica comprensible demuestra que los LLM pueden actuar como intermediarios eficaces entre modelos tabulares complejos y usuarios no técnicos. El uso de un modelo local como Mistral 7B Instruct valida

además la viabilidad de soluciones híbridas IA+NLP en escenarios donde existen restricciones de privacidad, coste o dependencia de servicios externos.

No obstante, el sistema presenta limitaciones relevantes ya que, a pesar de los avances logrados, un *recall* cercano al 50 % implica que una parte significativa del fraude sigue sin ser detectada. Así mismo, el proceso de extracción de características desde texto mediante LLM puede verse afectado por descripciones ambiguas o incompletas, lo que introduce una fuente adicional de incertidumbre. Estas limitaciones señalan la necesidad de seguir explorando estrategias de mejora en trabajos futuros.

En conjunto, los resultados obtenidos muestran que la solución desarrollada constituye un sistema sólido de preclasificación de fraude en seguros, capaz de equilibrar rendimiento predictivo, interpretabilidad y utilidad operativa, alineándose con los requisitos reales del sector asegurador.

4. Conclusiones y trabajos futuros

4.1 Conclusiones generales

El objetivo principal de este trabajo era desarrollar un sistema capaz de detectar posibles casos de fraude en siniestros de seguros combinando técnicas de *machine learning*, métodos de explicabilidad y modelos de lenguaje (LLM). Los resultados obtenidos permiten afirmar que el enfoque híbrido empleado es viable y proporciona un rendimiento adecuado para una fase inicial de cribado, donde el objetivo no es sustituir a los analistas humanos sino priorizar expedientes sospechosos.

En relación con los modelos de clasificación, una de las conclusiones más relevantes es que las técnicas avanzadas de sobre muestreo sintético no ofrecieron mejoras significativas en este *dataset*. Por el contrario, el *Random Oversampling* simple, junto con una optimización del umbral de decisión, proporcionó un equilibrio más sólido entre *precision* y *recall*. Esto confirma que, en contextos con una alta proporción de variables categóricas y patrones

minoritarios bien definidos, los métodos simples pueden resultar más eficaces que los métodos sintéticos.

Tras la optimización mediante RandomizedSearchCV, el modelo Random Forest alcanzó una capacidad de detección razonable, llegando aproximadamente al 50 % de recuperación de fraude (recall) con un umbral ajustado, lo cual supone una mejora notable respecto a los modelos base. Este comportamiento se ve reforzado por los análisis de Gain y Lift Chart, que muestran que el modelo es capaz de concentrar una proporción elevada de los casos fraudulentos en los primeros segmentos de riesgo, lo que lo hace especialmente útil como herramienta de priorización operativa.

La integración de SHAP permitió comprender por qué el modelo toma sus decisiones, identificando como variables más influyentes *Fault*, *BasePolicy*, *VehiclePrice*, *AddressChange_Claim* y *VehicleCategory*. Este tipo de información resulta especialmente relevante en sectores regulados, donde la trazabilidad y justificabilidad son obligatorias.

Por último, la construcción de un agente basado en LLM, capaz de transformar texto libre en características estructuradas y generar explicaciones comprensibles, añade un valor diferencial al sistema, facilitando su potencial integración en flujos reales de evaluación de siniestros.

4.2 Reflexión crítica sobre la consecución de los objetivos

De los objetivos planteados inicialmente, la mayoría han sido alcanzados:

- Desarrollar y comparar varios modelos supervisados
- Aplicar técnicas de preprocesado y balanceo incluyendo pruebas comparativas.
- Optimizar el mejor modelo mediante *RandomizedSearchCV*.
- Incorporar técnicas explicables (SHAP/LIME con integración práctica y análisis.
- Diseñar un agente IA capaz de interpretar descripciones en lenguaje natural con un LLM local.

- Detectar patrones y justificar predicciones gracias a SHAP y al reporte del agente.

El único objetivo parcialmente alcanzado es lograr una sensibilidad más alta. Aunque el modelo detecta un porcentaje razonable de casos de fraude, todavía se pierden eventos. Dado el desequilibrio extremo y la naturaleza del problema lo cual era previsible, aunque sigue siendo un área de mejora.

4.3 Análisis crítico de la metodología y planificación

En términos generales, la metodología planteada al inicio ha sido adecuada y ha guiado el proyecto de manera coherente. El flujo propuesto de preprocesado, posterior balanceo, modelado, evaluación, explicabilidad y agente experto, demostró ser válido.

Sin embargo, sí fue necesario introducir cambios metodológicos importantes durante el desarrollo. La planificación inicial contemplaba SMOTE/SMOTENC como estrategias principales de balanceo, pero su escaso rendimiento obligó a reconsiderar el enfoque y optar por ROS.

También el comportamiento inesperado de XGBoost, debido a fuertes incompatibilidades con datos categóricos, motivó centrar el análisis en *Random Forest* y la integración del agente LLM requirió replantear la forma de extraer y codificar variables categóricas desde lenguaje natural, incorporando un mapeo explícito de categorías para evitar errores.

En cuanto al seguimiento temporal, aunque la mayor parte de las tareas se completaron según lo previsto, la integración del LLM y la validación de SHAP consumieron más tiempo del esperado debido a incompatibilidades técnicas y al ajuste fino del *parser* de texto.

4.4 Impactos éticos, sociales y de sostenibilidad

Durante el desarrollo del trabajo se han tenido en cuenta los principales impactos identificados en la fase inicial, especialmente aquellos relacionados con la ética, la sostenibilidad y la dimensión social del uso de modelos de inteligencia artificial en la detección de fraude.

Desde el punto de vista ético, el proyecto presenta un impacto positivo al incorporar técnicas de interpretabilidad, que permiten comprender las razones de cada predicción y reducen el riesgo de decisiones opacas. Esto resulta fundamental en un contexto donde un error puede afectar directamente a un asegurado.

No obstante, existe también un potencial impacto negativo derivado de posibles sesgos en el *dataset* original. Para mitigar este riesgo se llevó a cabo un análisis cuidadoso de las variables más influyentes y se revisaron sus efectos mediante explicabilidad local y global, evitando así que el modelo reprodujera patrones discriminatorios no deseados.

En cuanto a los impactos sociales, el sistema desarrollado puede contribuir a mejorar la eficiencia en la evaluación de siniestros, reduciendo tiempos de tramitación y facilitando la detección temprana de casos sospechosos. Sin embargo, también existe el riesgo de generar falsos positivos que puedan perjudicar a clientes legítimos. Para minimizar este efecto, el modelo no se utiliza de manera autónoma y se ajustó el umbral de decisión para evitar decisiones excesivamente agresivas, integrando además explicaciones claras que permiten una revisión humana informada.

Finalmente, en relación con la diversidad, la inclusión de técnicas de explicabilidad facilita la detección de posibles sesgos hacia colectivos concretos, promoviendo una inteligencia artificial más equitativa.

4.5 Líneas de trabajo futuro

De cara a mejorar el rendimiento del sistema, especialmente en la detección de la clase fraudulenta, una línea prioritaria consiste en explorar técnicas más

avanzadas basadas en *embeddings* que podrían mejorar la representación de variables categóricas complejas y aumentar la capacidad del modelo para capturar relaciones no lineales.

Otra línea de trabajo relevante es el perfeccionamiento del módulo de procesamiento de lenguaje natural. El *parser* actual, basado en un modelo de lenguaje generalista, podría complementarse con enfoques más estructurados, lo que permitiría una extracción más precisa y consistente de atributos desde descripciones reales de siniestros.

Finalmente, una posible línea de mejora posible es ampliar el conjunto de variables incorporando información adicional sobre el vehículo y el propio siniestro, sin recurrir a historiales del asegurado para preservar la seguridad en la privacidad. En este sentido, variables como el uso del vehículo (alquiler frente a propiedad, uso profesional o particular), los importes económicos asociados a la reclamación (indemnización solicitada, coste peritado o relación entre daño y valor del vehículo) o detalles más precisos sobre la tipología y coherencia del daño podrían aportar una señal relevante para la detección de fraude.

Así mismo, la incorporación de información contextual agregada, como condiciones meteorológicas o características de la vía, permitiría enriquecer el modelo y mejorar su capacidad de generalización manteniendo un enfoque ético y respetuoso con la protección de datos.

Glosario

- Accuracy (Exactitud): métrica que indica el porcentaje total de predicciones correctas del modelo sobre el total de muestras evaluadas.
- AddressChange_Claim: variable categórica que indica si ha habido un cambio reciente de domicilio declarado en el proceso de reclamación.
- Agente LLM (Large Language Model Agent): sistema basado en modelos de lenguaje que puede interpretar texto, ejecutar herramientas, generar explicaciones y tomar decisiones asistidas.
- AUC (Area Under the ROC Curve): área bajo la curva ROC; mide la capacidad del modelo para separar clases. Valores cercanos a 1 indican un mejor rendimiento.
- Balanced Dataset (Dataset balanceado): conjunto de datos donde el número de muestras de cada clase es similar, reduciendo el sesgo del modelo hacia la clase mayoritaria.
- BasePolicy: tipo de póliza contratada, categorizada en grupos como *Liability*, *Collision*, *All Perils*, etc.
- Best Threshold (Umbral óptimo): valor de probabilidad a partir del cual el modelo clasifica un caso como fraude. Se selecciona para maximizar métricas como F1-score.
- Categorical Features (Variables categóricas): variables que representan categorías y no valores numéricos —por ejemplo: marca del coche, tipo de póliza o área del accidente.
- Codificación ordinal (Ordinal Encoding): método de transformación que convierte valores categóricos en números enteros manteniendo un orden arbitrario.
- Confusion Matrix (Matriz de confusión): tabla que recoge verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos del modelo.
- DayOfWeekClaimed: día de la semana en que se presentó la reclamación asociada al siniestro.
- Deciles: división del conjunto de siniestros en diez grupos iguales según el score del modelo, empleada para analizar la concentración del fraude en los segmentos de mayor riesgo.
- Explainability (Explicabilidad): conjunto de técnicas que permiten interpretar las decisiones de un modelo de inteligencia artificial.
- False Negative (Falso Negativo – FN): caso de fraude real clasificado incorrectamente como no fraude.
- False Positive (Falso Positivo – FP): caso no fraudulento clasificado erróneamente como fraude.
- Feature Importance (Importancia de características): medida que indica cuánto contribuye cada variable a la predicción del modelo.
- Fraud Detection (Detección de fraude): proceso de identificar comportamientos sospechosos o potencialmente fraudulentos en datos financieros o de seguros.
- F1-score: media armónica entre precisión y *recall*; métrica útil en problemas con clases desbalanceadas.
- Gain Chart (Curva de ganancias): gráfico que representa el porcentaje acumulado de casos de fraude detectados en función del porcentaje de

siniestros revisados, ordenados de mayor a menor probabilidad estimada por el modelo.
- Imbalanced Dataset (Dataset desbalanceado): conjunto de datos donde una clase (por ejemplo, fraude) está significativamente menos representada que la otra.
- Label (Etiqueta): variable objetivo que indica si el caso corresponde o no a un fraude.
- Lift: medida cuantitativa que indica cuántas veces mejora el modelo la detección de fraude en un determinado segmento respecto al comportamiento aleatorio.
- Lift Chart (Curva de elevación): herramienta de evaluación que muestra el rendimiento relativo del modelo por segmentos de riesgo, comparando la tasa de fraude detectada frente a una selección aleatoria.
- LIME (Local Interpretable Model-agnostic Explanations): método de explicabilidad que genera interpretaciones locales, aproximando el comportamiento del modelo cerca de una predicción concreta.
- LLM (Large Language Model): modelo de lenguaje de gran tamaño entrenado para procesar y generar texto humano.
- Machine Learning (Aprendizaje automático): disciplina que desarrolla algoritmos capaces de aprender patrones a partir de datos sin programación explícita.
- Make: marca del vehículo implicado en el siniestro (por ejemplo: Honda, Toyota, Ford…).
- Oversampling: técnica para aumentar el número de muestras de la clase minoritaria, con el fin de corregir el desbalanceo.
- Precision (Precisión): proporción de predicciones positivas que realmente son casos de fraude.
- Recall (Exhaustividad): proporción de casos de fraude correctamente detectados por el modelo.
- Random Forest: modelo basado en múltiples árboles de decisión para producir una predicción robusta.
- Random Oversampling: estrategia que duplica aleatoriamente casos de la clase minoritaria para equilibrar el dataset.
- ROC Curve (Receiver Operating Characteristic): gráfica que muestra la relación entre la tasa de verdaderos positivos y la de falsos positivos para distintos umbrales.
- SHAP (SHapley Additive exPlanations): marco teórico y práctico de explicabilidad basado en los valores de Shapley de la teoría de juegos, que evalúa la contribución de cada variable a una predicción.
- SMOTE (Synthetic Minority Over-Sampling Technique): técnica de oversampling que genera muestras sintéticas de la clase minoritaria mediante interpolación.
- Threshold (Umbral): valor límite que determina a partir de qué probabilidad un caso es clasificado como fraude.
- VehicleCategory: categoría del vehículo que aparece en la póliza (sedán, deportivo, utilitario, etc.).
- VehiclePrice: rango o clasificación del precio del vehículo, variable determinante en varios modelos de riesgo.

Bibliografía

[1] UNESPA. (2022). *Las reclamaciones fraudulentas al seguro son más frecuentes en automóviles y responsabilidad civil.* https://www.unespa.es/notasdeprensa/fraude-seguro-2021/

[2] Imbalanced-learn. (2025). *Documentation: SMOTE, RandomOverSampler*. https://imbalanced-learn.org/stable/

[3] Scikit-learn. (2025). *Logistic Regression — scikit-learn documentation*. https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

[4] Scikit-learn. (2025). Decision Trees — scikit-learn documentation. https://scikit-learn.org/stable/modules/tree.html

[5] Scikit-learn. (2025). *Random Forests — scikit-learn documentation*. https://scikit-learn.org/stable/modules/ensemble.html#random-forests

[6] XGBoost Developers. (2025). *XGBoost Documentation*. https://xgboost.readthedocs.io

[7] Scikit-learn. (2025). *Precision score — scikit-learn documentation*. https://scikit-learn.org/stable/modules/model_evaluation.html#precision

[8] Scikit-learn. (2025). *Recall score — scikit-learn documentation*. https://scikit-learn.org/stable/modules/model_evaluation.html#recall

[9] Scikit-learn. (2025). *F1 score — scikit-learn documentation*. https://scikit-learn.org/stable/modules/model_evaluation.html#f1-score

[10] Scikit-learn. (2025). *ROC and AUC — scikit-learn documentation*. https://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics

[11] Scikit-learn. (2025). *Confusion matrix — scikit-learn documentation*. https://scikit-learn.org/stable/modules/model_evaluation.html#confusion-matrix

[12] SHAP. (2025). *SHAP documentation*. https://shap.readthedocs.io

[13] LIME Documentation. (2025). *Local Interpretable Model-Agnostic Explanations.* GitHub. https://github.com/marcotcr/lime

[14] Llama.cpp. (2025). *Project page*. https://github.com/ggerganov/llama.cpp

[15] United Nations. (n.d.). *Goal 9: Industry, Innovation and Infrastructure*. Sustainable Development. https://www.un.org/sustainabledevelopment/infrastructure/

[16] United Nations. (n.d.). *Goal 11: Sustainable Cities and Communities*. Sustainable Development. https://www.un.org/sustainabledevelopment/cities/

[17] Parlamento Europeo y Consejo de la Unión Europea. (2016). *Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (Reglamento General de Protección de Datos). Diario Oficial de la Unión Europea,* L 119, 1–88. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32016R0679

[18] Gobierno de España. (2018). *Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. Boletín Oficial del Estado,* núm. 294, 6 de diciembre de 2018, 119788–119857. https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673

[19] Hugging Face. (2025). *GGUF models — Mistral-7B-Instruct*. https://huggingface.co

[20] Yankol-Schalck, S. (2022). *Fraud detection in automobile insurance using machine learning and textual claim descriptions*. Expert Systems with Applications, 198, 116789. https://doi.org/10.1016/j.eswa.2022.116789

[21] AXA. (s.f.). *La inteligencia artificial, gran aliada en nuestra lucha contra el fraude.* AXA Seguros. Recuperado el 10 de mayo de 2025, de https://www.axa.es/ca/-/la-inteligencia-artificial-gran-aliada-en-nuestra-lucha-contra-el-fraude

[22] Ayuso, M., Guillén, M., & Artís, M. (s.f.). *Técnicas cuantitativas para la detección del fraude en el seguro del automóvil.* Fundación MAPFRE. Recuperado el 10 de mayo de 2025, de https://documentacion.fundacionmapfre.org/documentacion/en/media/group/1054576.do

[23] Lo, V. S. Y. (2002). The true lift model: A novel data mining approach to response modeling in database marketing. ACM SIGKDD Explorations Newsletter, 4(2), 78–86. https://doi.org/10.1145/772862.772873