



**Universidad
Europea**

UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER OFICIAL EN BIG DATA ANALYTICS

TRABAJO FIN DE MÁSTER

**Impacto del Machine Learning para la
optimización del Pricing Dinámico en la
Industria de Seguros**

Sergio Aranda Valera

Dirigido por:

Jorge Moratalla Collado

CURSO 2026

TÍTULO: Impacto del Machine Learning para la optimización del Pricing Dinámico en la Industria de Seguros

AUTOR: Sergio Aranda Valera

TITULACIÓN: Máster Oficial en Big Data Analytics

DIRECTOR/ES DEL PROYECTO:

Dr. Manuel García Fernandez

FECHA: Abril del 2026

RESUMEN

La industria aseguradora enfrenta el desafío de optimizar la tarificación de sus productos, buscando un equilibrio entre competitividad comercial y rentabilidad técnica. Tradicionalmente, este proceso se ha basado en Modelos Lineales Generalizados (GLM), los cuales ofrecen interpretabilidad y robustez, pero presentan limitaciones para capturar relaciones no lineales complejas en el comportamiento del riesgo.

El presente trabajo propone un enfoque híbrido que integra técnicas de machine learning con modelos actuariales tradicionales, con el objetivo de mejorar la discriminación del riesgo y la precisión en el pricing dinámico. En particular, se emplean redes neuronales y Clusters de aprendizaje no supervisado para la identificación de patrones no lineales en variables vehiculares, cuyos resultados son posteriormente incorporados como variables categóricas dentro de un modelo GLM.

Los resultados evidencian que la incorporación de variables sintéticas derivadas de modelos no lineales mejora significativamente el desempeño predictivo, manteniendo la interpretabilidad del modelo final. Asimismo, se observa una mejor segmentación del riesgo, especialmente en variables que no presentan comportamientos lineales bajo enfoques tradicionales.

En conclusión, el enfoque híbrido propuesto constituye una alternativa viable para la modernización del pricing actuarial, permitiendo aprovechar las capacidades predictivas del machine learning sin comprometer los requerimientos de interpretabilidad y aplicabilidad en entornos regulatorios.

Palabras clave: Aprendizaje automático, modelación actuarial, precios, seguros generales.

ABSTRACT

The insurance industry faces the challenge of optimizing product pricing, seeking a balance between commercial competitiveness and technical profitability. Traditionally, this process has relied on Generalized Linear Models (GLMs), which offer interpretability and robustness but present limitations in capturing complex nonlinear relationships in risk behavior.

This study proposes a hybrid approach that integrates machine learning techniques with traditional actuarial models, with the aim of improving risk discrimination and accuracy in dynamic pricing. In particular, neural networks are used to identify nonlinear patterns in vehicle-related variables, whose outputs are subsequently incorporated as categorical variables within a GLM framework.

The results show that the inclusion of synthetic variables derived from nonlinear models significantly improves predictive performance while maintaining the interpretability of the final model. Additionally, better risk segmentation is observed, especially for variables that do not exhibit linear behavior under traditional approaches.

In conclusion, the proposed hybrid approach represents a viable alternative for modernizing actuarial pricing, enabling the use of machine learning's predictive capabilities without compromising interpretability and applicability in regulatory environments.

Keywords: Machine learning, pricing, actuarial models, general insurance

AGRADECIMIENTOS

Quiero expresar mi más profundo agradecimiento a mi familia, en especial a mis padres, hermanas y mi madrina Soledad. Gracias a ellos estoy aquí hoy, con la alegría de sustentar este proyecto y dar un paso más en mi desarrollo profesional. Su esfuerzo y amor no solo me han brindado las condiciones materiales necesarias, sino que también han sido un impulso constante, mi mayor fuerza y motivación para seguir adelante.

Asimismo, deseo agradecer a la primera persona que confió en mí: mi líder y mentora, hoy también una gran amiga, quien me dio mi primer reto en el mundo de la analítica de datos y fue parte fundamental para que hoy me encuentre aquí. Gracias, Valery L., por tu confianza y por acompañarme desde el inicio de este camino.

TABLA RESUMEN

	DATOS
Nombre y apellidos:	Sergio Aranda Valera
Título del proyecto:	Impacto del ML para la optimización de pricing dinámico en seguros
El proyecto se ha realizado en colaboración de una empresa o a petición de una empresa:	NO
El proyecto ha implementado un producto:	NO
El proyecto ha consistido en el desarrollo de una investigación o innovación:	SI
Objetivo general del proyecto:	Optimizar la precisión y eficiencia de los precios de seguros en Perú

Índice

RESUMEN.....	3
ABSTRACT	4
TABLA RESUMEN	6
Capítulo 1. RESUMEN DEL PROYECTO	10
1.1 Contexto y justificación	10
1.2 Planteamiento del problema	10
1.3 Objetivos del proyecto	10
1.4 Resultados obtenidos	10
1.5 Estructura de la memoria	11
Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE	12
2.1 Estado del arte	12
2.2 Contexto y justificación	12
2.3 Planteamiento del problema	13
Capítulo 3. OBJETIVOS	15
3.1 Objetivos generales.....	15
3.2 Objetivos específicos	15
3.3 Beneficios del proyecto.....	15
Capítulo 4. DESARROLLO DEL PROYECTO	17
4.1 Planificación del proyecto.....	17
4.2 Descripción de la solución, metodologías y herramientas empleadas	18
4.3 Recursos requeridos	22
4.4 Presupuesto	22
4.5 Viabilidad	22
4.6 Resultados del proyecto.....	22
Capítulo 5. DISCUSIÓN.....	28
Capítulo 6. CONCLUSIONES.....	31
6.1 Conclusiones del trabajo	31
6.2 Conclusiones personales	31

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO.....	32
Capítulo 8. REFERENCIAS	33

Índice de Figuras

- Figura 1. Distribución de la frecuencia de siniestros según la edad del asegurado 20
- Figura 2. Distribución de la frecuencia de siniestros según la antigüedad del vehículo 20
- Figura 3. Distribución de la frecuencia de siniestros según tipo de póliza (nueva vs. renovación) 20
- Figura 4. Identificación de variables vehiculares con comportamiento no lineal 21
- Figura 5. Distribución de pólizas según ubicación geográfica a nivel país 22
- Figura 6. Mapa de calor de la concentración de pólizas a nivel nacional 22
- Figura 7. Propagación del riesgo mediante clustering en la capital 24
- Figura 8. Segmentación de grupos de riesgo basada en clustering demográfico 25
- Figura 9. Curvas de aprendizaje del modelo (entrenamiento y validación) 26
- Figura 10. Curva Lorenz de evaluación de la discriminación del riesgo por modelo 26
- Figura 11. Transformación e interpretación de los resultados de predicción 27
- Figura 12. Creación de grupos de riesgo basados en redes neuronales 27
- Figura 13. Discriminación del riesgo mediante grupos tradicionales (GLM) 28
- Figura 14. Estructura y resultados del modelo GLM tradicional 29
- Figura 15. Evaluación del rendimiento del modelo GLM tradicional 30
- Figura 16. Estructura y resultados del modelo híbrido propuesto 31
- Figura 17. Evaluación del rendimiento del modelo híbrido (Lift Curve y MAE) 31

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

La tarificación en seguros generales constituye un proceso clave para la sostenibilidad financiera de las aseguradoras. Este proceso busca estimar adecuadamente el riesgo asociado a cada póliza, de modo que la prima refleje de manera precisa la probabilidad y magnitud del siniestro.

Históricamente, los Modelos Lineales Generalizados (GLM) han sido el estándar en la industria debido a su interpretabilidad, estabilidad y aceptación regulatoria. Sin embargo, el incremento en la disponibilidad de datos y la complejidad del comportamiento del riesgo han evidenciado limitaciones en estos modelos.

1.2 Planteamiento del problema

Los modelos tradicionales presentan dificultades para capturar relaciones no lineales y efectos de interacción complejos entre variables. Por otro lado, los modelos de machine learning ofrecen una mayor capacidad predictiva, pero presentan desafíos en términos de interpretabilidad y aplicabilidad en entornos regulados.

Esta situación genera una brecha entre precisión predictiva e interpretabilidad, lo que limita la adopción de técnicas avanzadas en la industria aseguradora.

1.3 Objetivos del proyecto

El objetivo principal es demostrar que la integración de modelos GLM con técnicas de machine learning, como redes neuronales y Clustering K-means, mejora la discriminación del riesgo y optimiza el pricing en seguros generales.

De manera específica, se busca comparar el desempeño de GLM frente a métodos tradicionales, identificar relaciones no lineales en variables de riesgo, desarrollar una segmentación mediante redes neuronales e integrarla en un modelo GLM, evaluando su impacto en la precisión, discriminación del riesgo y optimización de la cartera.

1.4 Resultados obtenidos

Los resultados evidencian que los modelos GLM permiten una adecuada generalización en segmentos con baja exposición al riesgo, mejorando la precisión en la tarificación al capturar patrones intuitivos asociados a características del conductor y del vehículo.

Asimismo, el uso de redes neuronales permitió identificar categorías de riesgo basadas en interacciones complejas, cuya incorporación como variables categóricas en el modelo GLM de frecuencia mejoró significativamente el desempeño estadístico frente al uso directo de variables con comportamiento no lineal.

1.5 Estructura de la memoria

La presente memoria se organiza en nueve capítulos.

- El Capítulo 1 presenta el resumen del proyecto, incluyendo el contexto, el planteamiento del problema, los objetivos y los principales resultados.
- El Capítulo 2 desarrolla los antecedentes y el estado del arte.
- El Capítulo 3 detalla los objetivos y beneficios del proyecto.
- El Capítulo 4 describe el desarrollo del proyecto, abarcando la metodología, herramientas, recursos, presupuesto, viabilidad y resultados.
- El Capítulo 5 presenta la discusión de los resultados, mientras que el Capítulo 6 expone las conclusiones del trabajo.
- Finalmente, el Capítulo 7 aborda las futuras líneas de investigación, el Capítulo 8 contiene las referencias bibliográficas.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

2.1 Estado del arte

El pricing actuarial en el sector asegurador se ha basado históricamente en el uso de Modelos Lineales Generalizados (GLM), los cuales constituyen el enfoque estándar para la estimación de primas debido a su base estadística sólida, facilidad de interpretación y aceptación regulatoria [1]. Estos modelos permiten descomponer el riesgo en componentes de frecuencia y severidad, asumiendo relaciones lineales o log lineales entre las variables explicativas y la variable respuesta.

Sin embargo, diversos estudios han señalado que los GLM presentan limitaciones para capturar relaciones no lineales complejas y efectos de interacción de alto orden, especialmente en contextos caracterizados por grandes volúmenes de datos y alta dimensionalidad [2]. Como consecuencia, en los últimos años se ha incrementado el interés por la aplicación de técnicas de Machine Learning en el ámbito actuarial, destacando modelos como árboles de decisión, Random Forests y métodos de Gradient Boosting, los cuales han demostrado mejoras significativas en la capacidad predictiva frente a los enfoques tradicionales [3].

Investigaciones recientes proponen enfoques híbridos que combinan modelos actuariales clásicos con técnicas de Machine Learning, con el objetivo de aprovechar la flexibilidad de estos últimos sin renunciar a la interpretabilidad y estabilidad necesarias en entornos regulados [4]. Estos enfoques buscan un equilibrio entre precisión predictiva y transparencia del modelo, aspecto clave para su adopción en procesos reales de tarificación.

A pesar de los avances logrados, la implementación práctica de modelos basados en Machine Learning en pricing actuarial aún enfrenta desafíos relacionados con la explicabilidad, la robustez y la integración con los marcos tradicionales de tarificación.

En este contexto, el presente proyecto se enmarca en el estado del arte actual, contribuyendo al análisis comparativo y a la complementariedad entre modelos GLM y técnicas de Machine Learning aplicadas al pricing actuarial.

2.2 Contexto y justificación

En la tarificación de seguros generales, los enfoques tradicionales —y, posteriormente, los Modelos Lineales Generalizados (GLM)— han constituido el estándar metodológico debido a su solidez estadística, interpretabilidad y facilidad de implementación en entornos productivos. No obstante, estos modelos presentan limitaciones para capturar relaciones no lineales y estructuras complejas en los datos, particularmente en contextos donde el comportamiento del riesgo es inherentemente estocástico y depende de múltiples factores heterogéneos.

El análisis del riesgo en seguros evidencia que su comportamiento responde a dinámicas complejas y contextuales, que pueden variar significativamente según características propias del mercado. Factores como la cultura de conducción, las condiciones del entorno vial y las características técnicas del vehículo influyen de manera diferenciada en la ocurrencia del siniestro, modificando la relevancia y el efecto de las variables tradicionales utilizadas en la tarificación.

En este contexto, la creciente disponibilidad de datos y el avance de técnicas de analítica avanzada han impulsado la incorporación de modelos de machine learning, los cuales ofrecen una mayor capacidad para capturar patrones no lineales y relaciones de interacción complejas. Sin embargo, su adopción en el ámbito asegurador se ha visto limitada por desafíos asociados a la interpretabilidad, la explicabilidad y el cumplimiento de requerimientos regulatorios.

En respuesta a esta problemática, el presente trabajo propone una metodología híbrida que integra técnicas de machine learning con modelos actuariales tradicionales, con el objetivo de mejorar la precisión en la estimación del riesgo sin comprometer la interpretabilidad del modelo final. Este enfoque busca equilibrar la capacidad predictiva de los modelos no lineales con la transparencia y robustez de los modelos clásicos, facilitando su aplicación en entornos productivos del sector asegurador.

De este modo, la investigación se justifica en la necesidad de desarrollar soluciones metodológicas que permitan optimizar la tarificación y la gestión de cartera, manteniendo coherencia con las exigencias regulatorias y operativas de la industria, al tiempo que se incorporan las ventajas derivadas del uso de técnicas avanzadas de machine learning.

2.3 Planteamiento del problema

A partir del análisis del estado del arte, se observa que las técnicas de machine learning han demostrado mejoras sustanciales en la capacidad predictiva frente a los enfoques actuariales tradicionales, particularmente en contextos caracterizados por alta dimensionalidad y relaciones no lineales entre variables. Diversos estudios evidencian que modelos como redes neuronales, árboles de decisión y métodos de ensamble permiten capturar patrones complejos en los datos, superando en desempeño a los Modelos Lineales Generalizados (GLM) en términos de ajuste y precisión.

No obstante, a pesar de estos avances, su adopción en la tarificación de seguros generales continúa siendo limitada. Esta restricción se explica principalmente por desafíos asociados a la interpretabilidad de los modelos, la dificultad en su validación bajo marcos regulatorios y las barreras de implementación en entornos productivos, donde se requiere estabilidad, trazabilidad y explicabilidad en la toma de decisiones.

En contraste, los GLM se mantienen como el estándar en la industria aseguradora debido a su base estadística sólida, su transparencia en la estimación de efectos y su alineación con los requerimientos normativos. Sin embargo, estos modelos presentan limitaciones estructurales para capturar relaciones no lineales y efectos de interacción complejos, lo que puede derivar en una representación parcial del riesgo, especialmente en contextos donde intervienen múltiples factores heterogéneos.

Esta dualidad evidencia una brecha relevante en la literatura y en la práctica profesional: la necesidad de mejorar la capacidad predictiva de los modelos sin comprometer su interpretabilidad y viabilidad operativa. A pesar de la existencia de propuestas híbridas en investigaciones recientes, no se dispone de un enfoque estandarizado que permita integrar de manera sistemática la capacidad de modelamiento no lineal del machine learning con la estructura explicativa de los modelos actuariales tradicionales.

En este contexto, surge la necesidad de desarrollar metodologías que permitan articular ambos enfoques de forma complementaria. En particular, el presente trabajo propone la utilización de modelos de deep learning, específicamente redes neuronales multicapa, para la identificación y segmentación de grupos de riesgo a partir de características vehiculares. Los resultados de estos modelos son posteriormente transformados en variables categóricas que se incorporan como variables explicativas en un modelo GLM orientado a la tarificación.

De esta manera, se plantea un enfoque híbrido que busca mejorar la discriminación del riesgo y la precisión del pricing, manteniendo simultáneamente la interpretabilidad del modelo final y su aplicabilidad en entornos productivos. Esta propuesta se posiciona como una alternativa metodológica que contribuye a cerrar la brecha identificada en el estado del arte, facilitando la adopción progresiva de técnicas avanzadas de machine learning en el ámbito actuarial.

Capítulo 3. OBJETIVOS

3.1 Objetivos generales

El objetivo del presente trabajo consiste en demostrar que la integración de modelos GLM con técnicas de machine learning, como redes neuronales aplicadas a la segmentación de riesgo, permite mejorar la discriminación del riesgo en seguros generales, optimizando el pricing dinámico y contribuyendo a una gestión más eficiente y rentable de la cartera.

3.2 Objetivos específicos

- Evidenciar la presencia de relaciones no lineales en las variables asociadas al riesgo vehicular, identificando aquellas características que no son adecuadamente capturadas por la estructura lineal de los modelos GLM.
- Desarrollar un modelo de segmentación de riesgo basado en redes neuronales, orientado a identificar grupos homogéneos de riesgo a partir de características del vehículo.
- Integrar la segmentación de riesgo obtenida mediante redes neuronales como variable categórica dentro de un modelo GLM, evaluando su impacto en la mejora del desempeño predictivo y la discriminación del riesgo.
- Evaluar el impacto del enfoque híbrido propuesto en la optimización del pricing, considerando métricas de precisión, discriminación del riesgo y su contribución potencial a la mejora de la cartera y la rentabilidad.

3.3 Beneficios del proyecto

El enfoque híbrido planteado ofrece una serie de beneficios relevantes tanto a nivel metodológico como operativo dentro del contexto de la tarificación en seguros generales.

En primer lugar, permite mantener la interpretabilidad del modelo, al conservar la estructura de los Modelos Lineales Generalizados (GLM) como componente principal de tarificación. Esto facilita la comprensión del impacto de las variables en la estimación del riesgo, lo cual es fundamental para la validación del modelo y el cumplimiento de requerimientos regulatorios.

Asimismo, el enfoque propuesto presenta una alta factibilidad de implementación en entornos productivos, dado que se apoya en modelos ampliamente utilizados en la industria aseguradora. La incorporación de la segmentación generada por redes neuronales como variable categórica no altera significativamente la arquitectura tradicional del modelo, lo que permite su integración en sistemas existentes con un bajo costo de adopción.

Por otro lado, este esquema facilita el control y ajuste del modelo en producción, ya que la estructura basada en GLM permite realizar calibraciones, monitoreo y actualizaciones de manera transparente y controlada. Esto resulta especialmente relevante en contextos

donde los modelos deben adaptarse a cambios en el comportamiento del riesgo o en las condiciones del mercado.

Finalmente, el enfoque garantiza la compatibilidad con sistemas legacy, al no requerir una transformación completa de la infraestructura existente. Esto permite a las organizaciones incorporar mejoras en la discriminación del riesgo y precisión del pricing sin comprometer la estabilidad operativa ni los procesos establecidos.

En conjunto, estos beneficios posicionan al enfoque híbrido como una alternativa viable y escalable para mejorar la tarificación en seguros generales, equilibrando la capacidad predictiva de técnicas avanzadas de machine learning con los requerimientos prácticos del entorno empresarial.

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto



4.2 Descripción de la solución, metodologías y herramientas empleadas

El proyecto desarrollado corresponde al ámbito de la analítica avanzada, y ha sido estructurado bajo metodologías estándar para proyectos de datos, integrando fundamentos estadísticos, modelos de machine learning y tecnologías modernas orientadas a optimizar la eficiencia en los resultados.

a. Metodologías para proyectos de datos:

El desarrollo del proyecto se basa en la metodología CRISP-DM, la cual permitió establecer un plan de trabajo estructurado y ordenado, asegurando una correcta ejecución y una entrega progresiva de valor a lo largo de todas las fases del proyecto.

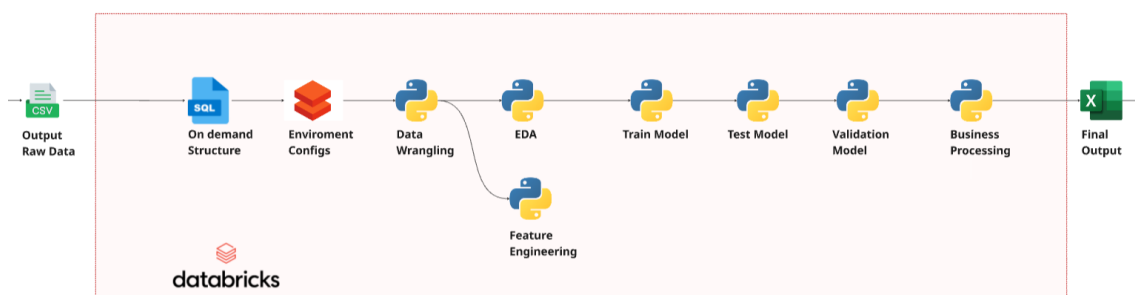
b. Fundamentos estadísticos y Machine Learning:

El enfoque analítico se sustenta en fundamentos de estadística descriptiva y análisis diagnóstico, complementados con la aplicación de diversos modelos de machine learning. Estos modelos han sido optimizados mediante el uso de hiperparámetros adecuados y evaluados a través de métricas estadísticas que garantizan la precisión y confiabilidad de los resultados.

c. Tecnología y herramientas utilizadas:

Para el desarrollo del proyecto se empleó la plataforma Databricks, una infraestructura tecnológica especializada en la gestión de datos y el desarrollo de modelos de machine learning. Asimismo, se utilizó Databricks Free Edition, un entorno de acceso gratuito basado en la nube que permite la carga, procesamiento y análisis avanzado de datos.

d. Diagrama de solución:



e. Análisis Exploratorio (EDA):

Dentro de las variables disponibles, se ha planteado la representación del riesgo a través de la cantidad de accidentes que una póliza de asegurado puede tener durante su vigencia de contrato.

Asimismo, dado que en la realidad no todas las pólizas se están analizando con el mismo tiempo de duración, se creó la variable “Frecuencia” para representar una medida que estandarice la representación del riesgo de acuerdo con el tiempo de vigencia de cada póliza.

A través del análisis exploratorio del dataset, se pudo evidenciar la disponibilidad de variables que presentan una correlación lineal directa con la frecuencia total por segmento.

Esto representa una discriminación del riesgo directa, que sugiere buen performance en el aprendizaje de una regresión lineal. Revisemos las siguientes variables:

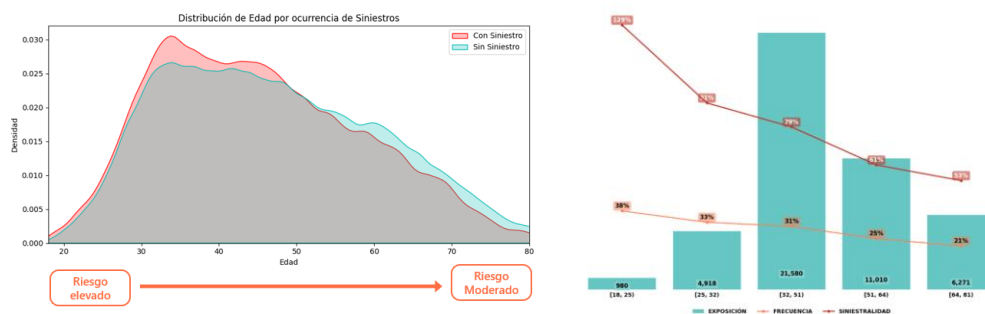


Figura 1. Distribución de la frecuencia de siniestros por edad del asegurado.

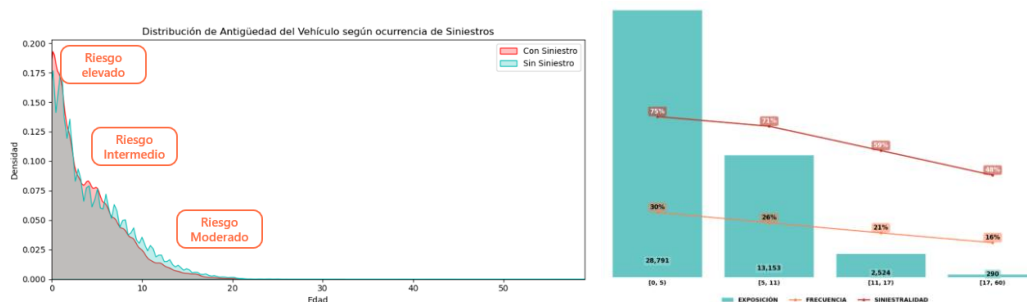


Figura 2. Distribución de la frecuencia de siniestros por antigüedad vehículo

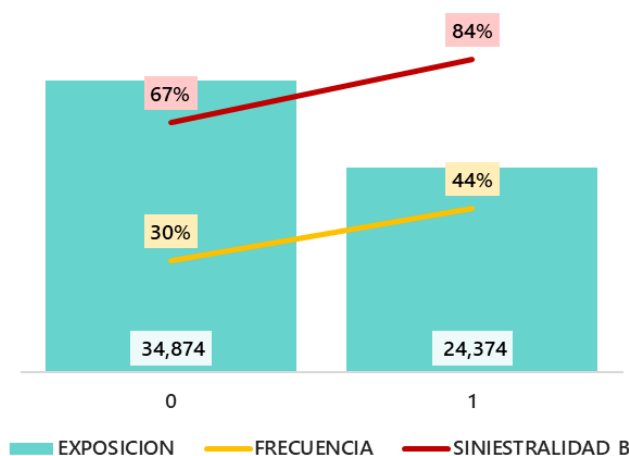


Figura 3. Distribución de la frecuencia de siniestros por póliza nueva

Sin embargo, si bien es cierto que se evidencian variables con una discriminación lineal directa, se evidencian también grupos de variables específicamente de carácter vehicular que no cuentan con un comportamiento lineal del riesgo.

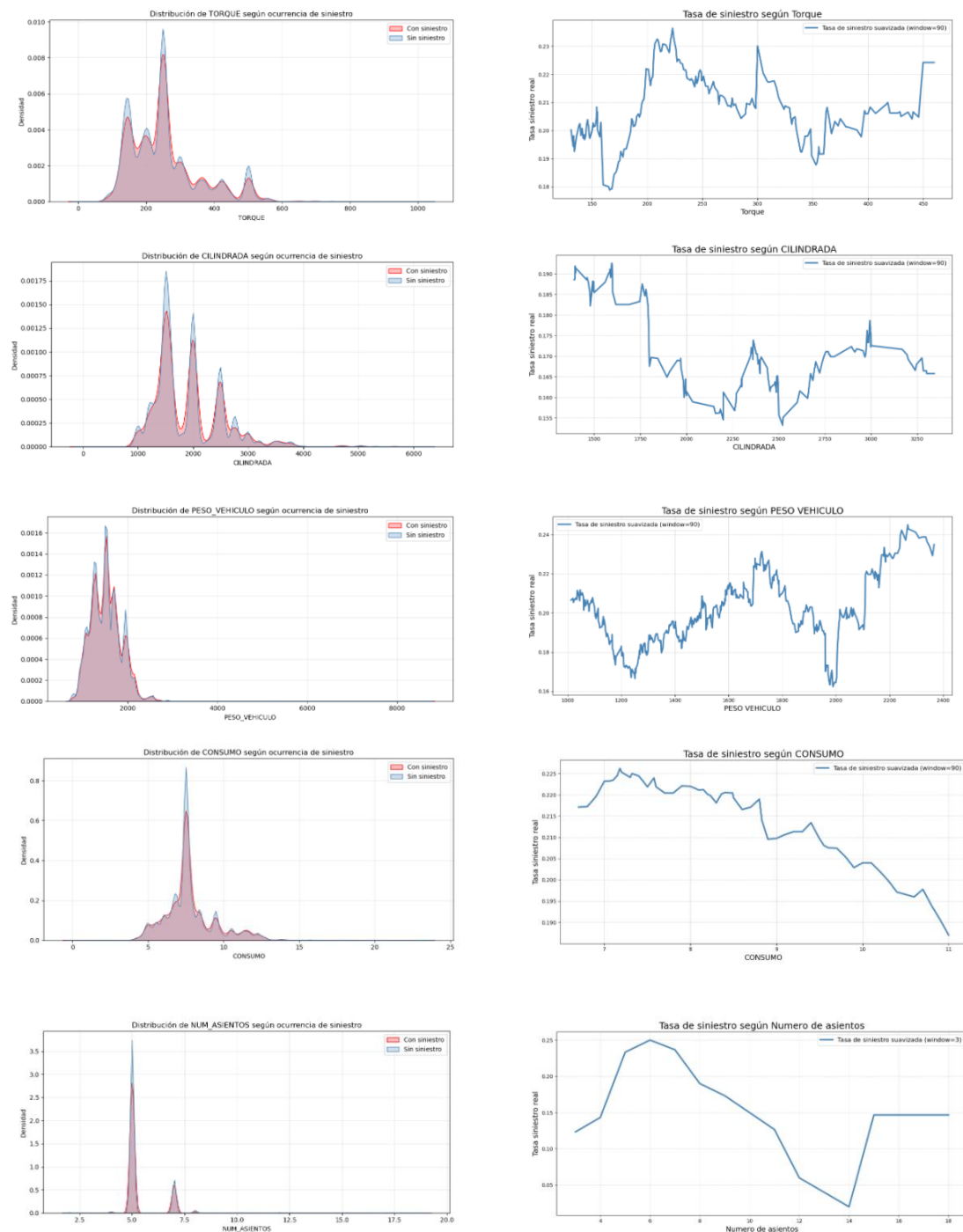


Figura 4. Portfolio de variables no lineales automotor

Adicionalmente, validamos variables de carácter geográfico que si bien es cierto, cuentan con características asociadas al riesgo, no tienen suficiente muestra por distribución categórica para representar correctamente el riesgo. Debido a esto el ajuste se expone al riesgo de no cumplir con los intervalos de confianza ni significancia adecuada.

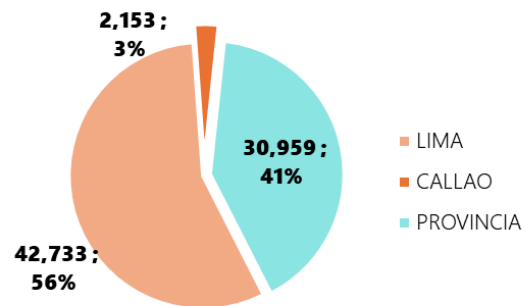


Figura 5. Distribución de pólizas por ubicación país

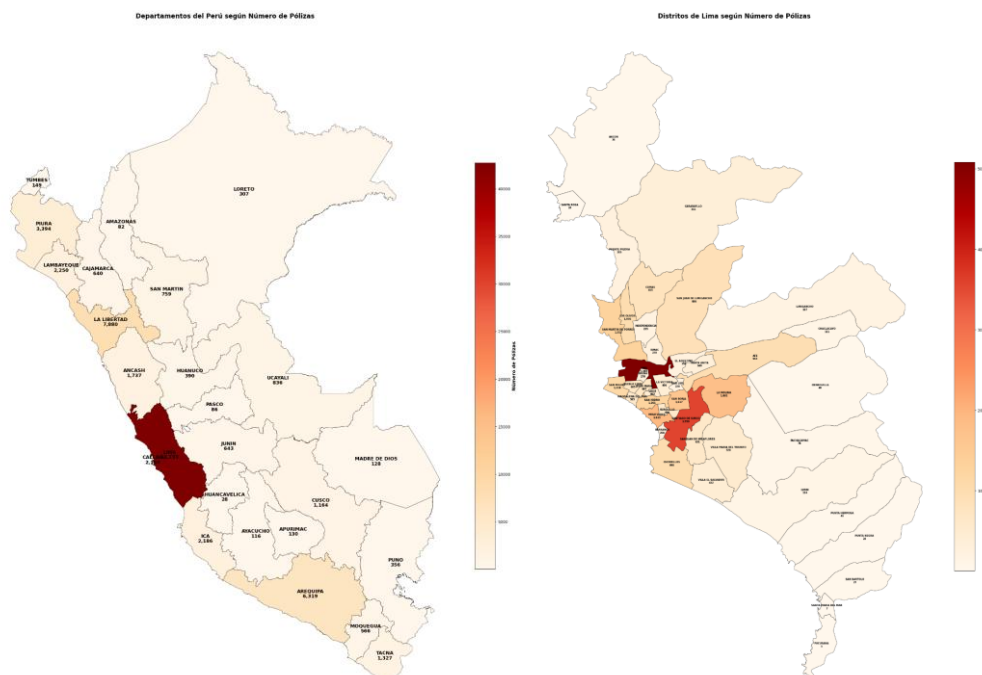


Figura 6. Mapa de calor por concentración de pólizas

4.3 Recursos requeridos

1. Investigador principal – Especialista en Machine Learning y actuarial
2. Equipo computacional - laptop HP 255 G10 con procesador core i7, 16 GB de memoria RAM y 512 GB de almacenamiento SSD, junto con teclado.
3. Tokens GPT4 que corresponden al consumo único de tokens de prompt engineering para la extracción de características vehiculares de dominios públicos.
4. Extensión Databricks Free Edition – Capacidades extra necesarias.

4.4 Presupuesto

Tipo de coste	Valor	Comentarios
Horas de trabajo en el proyecto	1000 €	El proyecto se desarrollo progresivamente con un único investigador principal.
Equipo técnico utilizado	1000 €	El equipo fue adquirido con las condiciones y características necesarias para el proyecto.
Software utilizado	100 €	Aumento de capacidades Databricks para tareas de aprendizaje específico.
Software utilizado 2	60 €	Coste total por cantidad de tokens OpenAI utilizados en web scraping de características vehiculares.

4.5 Viabilidad

En el marco del proyecto se prioriza la modalidad de publicación sin coste, garantizando la viabilidad económica de la difusión científica sin comprometer el rigor, el impacto ni la visibilidad internacional de los resultados, y manteniendo abierta la posibilidad de optar por acceso abierto en caso de disponibilidad de financiación adicional específica para este fin.

4.6 Resultados del proyecto

De acuerdo con los hallazgos obtenidos en el EDA, se evidencian los resultados de grupos sintéticos creados con técnicas de Machine Learning que permitan capturar características limitadas para los modelos de regresión lineal tradicionales.

a. K-Means para grupos de riesgo por frecuencia y cercanía geográfica

En este apartado revisamos los resultados del método no supervisado de Clusters K Means basados en las variables de frecuencia y coordenadas geográficas.

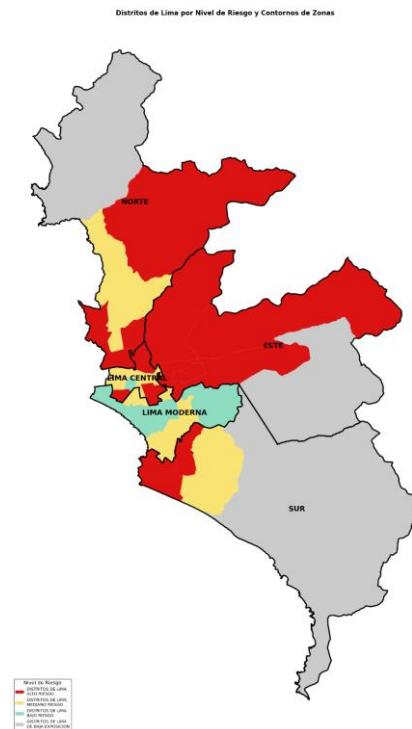
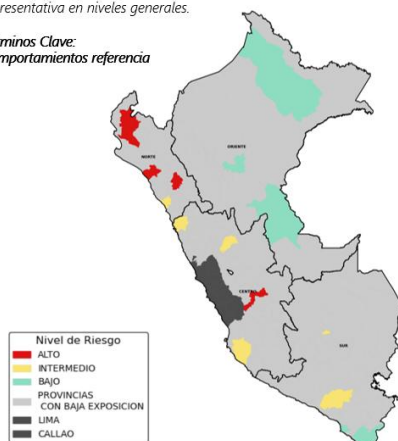


Figura 7. Propagación de riesgo por cluster de capital

A diferencia de Lima, la posición en provincias es dispersa y poco representativa en niveles generales.

Términos Clave:
Comportamientos referencia



Propagación Espacial Del Riesgo vecino

Grupos Representativos



Figura 7. Propagación de riesgo por cluster de provincias

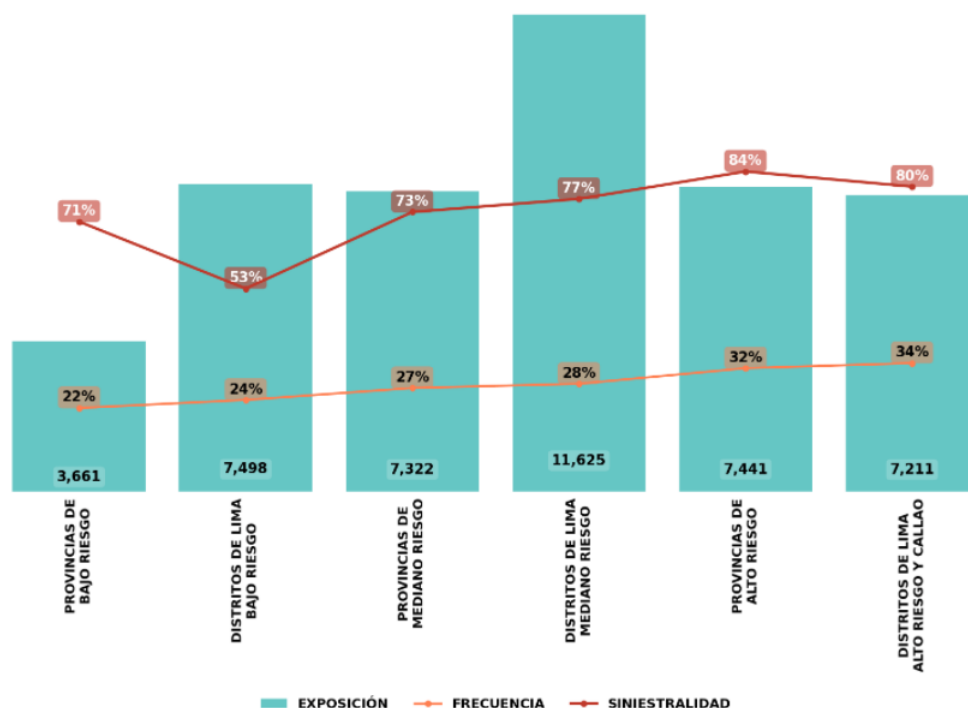


Figura 8. Grupos de riesgo resultado de Cluster demográfico

b. Modelo de frecuencia multivariable con Redes neuronales multicapa

El desarrollo del modelo multivariable modelando la frecuencia de accidentes muestra una convergencia entre el entrenamiento y validación donde se evidencia un aprendizaje no lineal que discrimina parcialmente pero eficientemente el riesgo.

El resultado de esta propuesta desarrolla una función de interacciones complejas que captura las redes neuronales para representar de manera precisa el riesgo parcial que puede ser provocado netamente por el concepto vehicular.

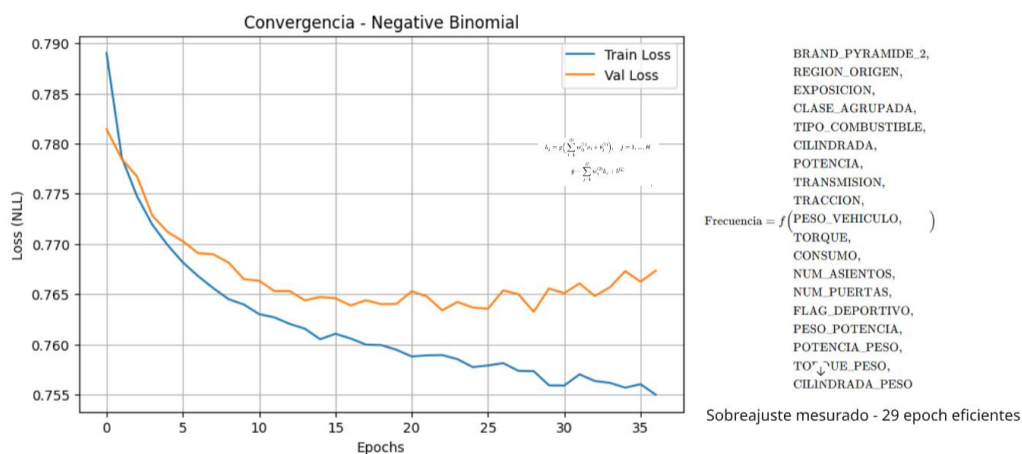


Figura 9. Curva de aprendizaje de entrenamiento y validación

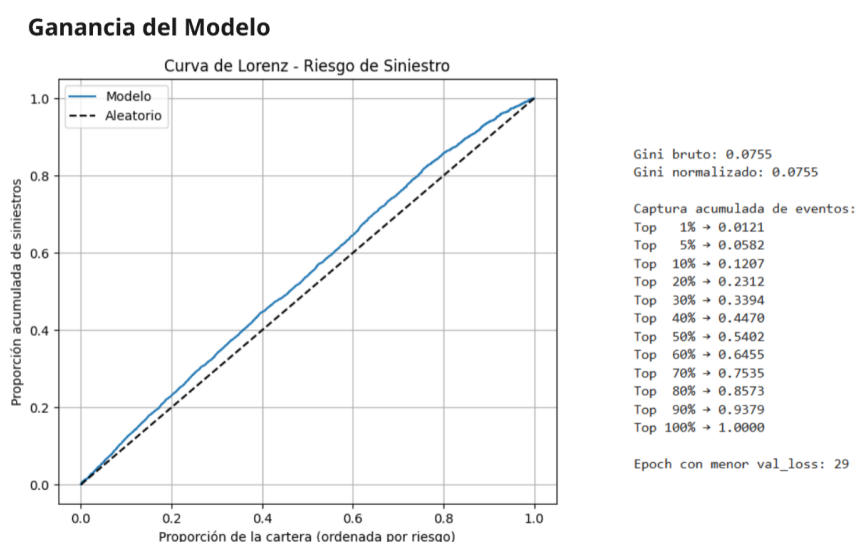


Figura 10. Curva de Lorenz discriminación de riesgo por modelo

Resultados finales e interpretación de la interacción por redes:

El entrenamiento se realiza considerando únicamente la explicación parcial del siniestro, manteniendo constantes (ceteris paribus) las demás variables.

La capacidad predictiva del modelo es parcial, dado que se ha comprobado que existen efectos no lineales que captura el modelo de features vehiculares, pero que deben complementarse con features de carácter persona que ejercen una influencia significativa sobre el siniestro.

Por lo tanto, las variables vehiculares únicamente permiten explicar una fracción del comportamiento del siniestro.

A fin de establecer un marco metodológico de comparación de la discriminación del riesgo, procedimos a realizar la predicción de siniestros por observación y modelamos un histograma que nos permita ver la distribución del riesgo en la cartera.

A partir de la discriminación ordenada que reflejen estos sujetos observados, proponemos un grupo sintético de riesgo para categorizar los vehículos, el cual evidencia una mejora en la discriminación respecto a los grupos tradicionales de negocio.

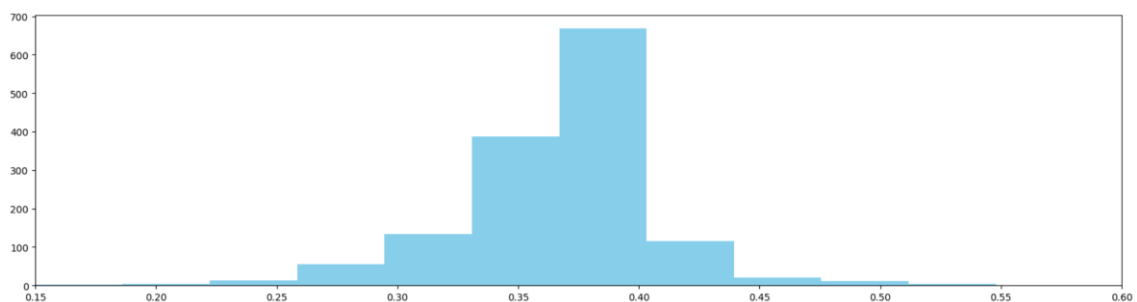


Figura 11. Transformación e interpretación de los resultados de predicción

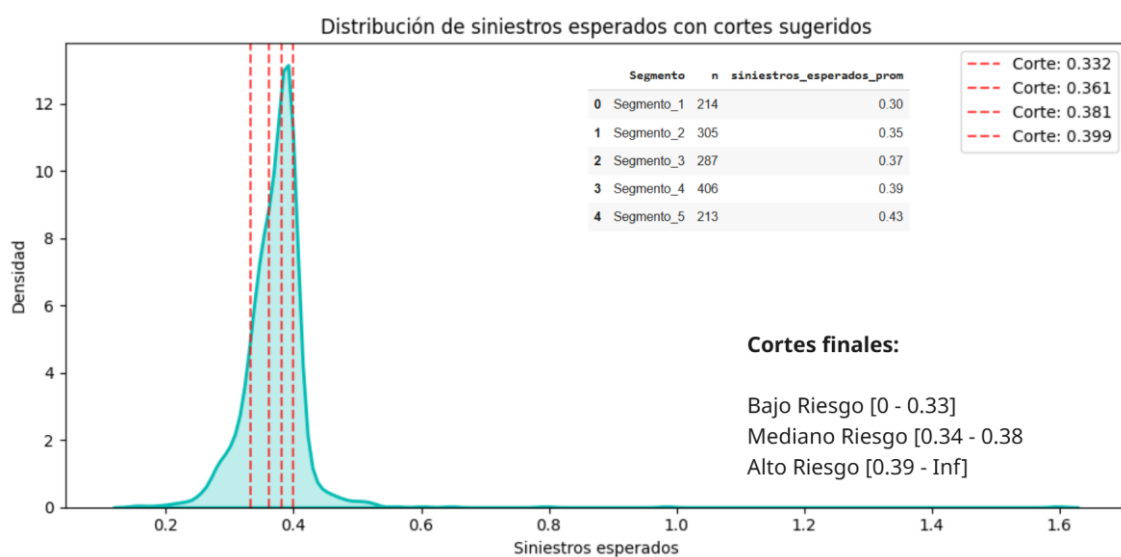


Figura 12. Creación de grupos de riesgo basados en redes neuronales

Comparación de grupos de riesgo basados en GLM y grupos de riesgo de redes neuronales:

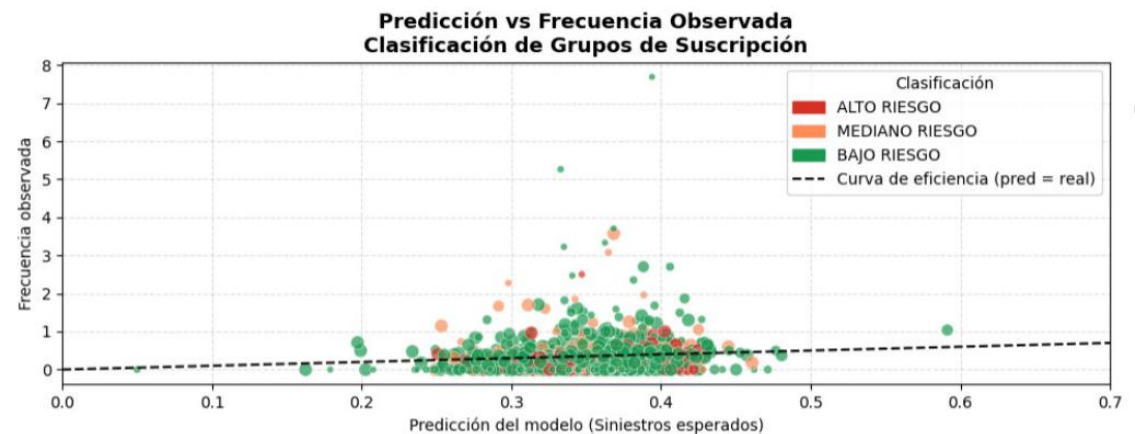


Figura 13. Discriminación del riesgo por grupos de riesgo tradicionales

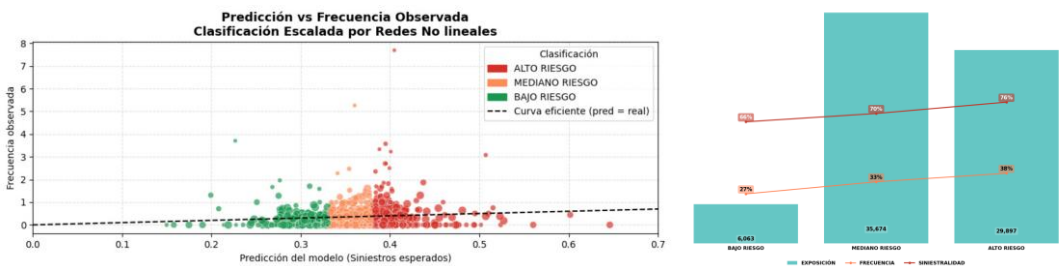


Figura 14. Discriminación del riesgo por grupos sintéticos redes neuronales

Capítulo 5. DISCUSIÓN

Según estas características se resuelve el planteamiento comparativo de un modelo directo con todas las variables y un modelo incorporando variables categóricas sintetizadas a partir de métodos de modelos ML.

De acuerdo con los resultados, procedemos a evaluar la experimentación de las dos alternativas propuestas.

Modelo híbrido de variables directas tradicionales:

	Factor	Nivel	Exposición train	Exposición test	Coefficiente	P> z	[0.025	0.975]	Exponencial coeficiente
0	Intercept	Intercept	35,770.13	8,984.29	-1.548100	0.000	-1.623	-1.474	0.21
1	FLAG NUEVO_CALC	1	15,550.26	3,947.59	0.120000	0.000	0.077	0.163	1.13
2	FLAG NUEVO_CALC	0	nan	nan	0.000000	nan	nan	nan	1.00
3	MVP_REGION_AJ7	Selva	1,106.98	261.15	-0.380300	0.000	-0.520	-0.241	0.68
4	MVP_REGION_AJ7	Sur_Sierra	1,117.47	274.55	-0.101300	0.109	-0.225	0.023	0.90
5	MVP_REGION_AJ7	Lima_Sur	1,245.92	298.26	0.075800	0.183	0.036	0.187	1.08
6	MVP_REGION_AJ7	Lima_Este	1,803.89	431.82	0.209900	0.000	0.119	0.301	1.23
7	MVP_REGION_AJ7	Lima_Norte	2,866.55	708.89	0.134100	0.001	0.055	0.213	1.14
8	MVP_REGION_AJ7	Norte_Costa	2,890.18	717.86	0.150300	0.000	0.072	0.228	1.16
9	MVP_REGION_AJ7	Norte_Costa_II	4,766.17	1,182.92	-0.093900	0.011	-0.166	-0.022	0.91
10	MVP_REGION_AJ7	Sur_Costa_NAs	4,858.83	1,255.97	0.029200	0.408	-0.040	0.098	1.03
11	MVP_REGION_AJ7	Lima_Oeste_II_Callao	7,057.83	1,838.60	0.048100	0.328	0.014	0.110	1.05
12	MVP_REGION_AJ7	Lima_Oeste_I_LaMolina	nan	nan	0.000000	nan	nan	nan	1.00
13	GRUPO_SUSCRIPCION_AS_IS	ALTO RIESGO	6,889.88	1,790.61	0.158900	0.000	0.103	0.215	1.17
14	GRUPO_SUSCRIPCION_AS_IS	MEDIANO RIESGO	9,425.69	2,354.42	0.075100	0.002	0.028	0.122	1.08
15	GRUPO_SUSCRIPCION_AS_IS	BAJO RIESGO	nan	nan	0.000000	nan	nan	nan	1.00
16	MVP_EDAD_ASEGURADO	[18,30]	3,916.76	942.69	0.339600	0.000	0.275	0.404	1.40
17	MVP_EDAD_ASEGURADO	[31,35]	4,520.74	1,117.48	0.305600	0.000	0.244	0.367	1.36
18	MVP_EDAD_ASEGURADO	[36,40]	4,563.52	1,209.75	0.271700	0.000	0.210	0.334	1.31
19	MVP_EDAD_ASEGURADO	[41,50]	8,925.93	2,278.67	0.173800	0.000	0.122	0.226	1.19
20	MVP_EDAD_ASEGURADO	[51,60]	nan	nan	0.000000	nan	nan	nan	1.00
21	MVP_EDAD_VEHICULO	4	2,766.55	673.31	0.105700	0.016	0.020	0.191	1.11
22	MVP_EDAD_VEHICULO	3	2,910.20	758.05	0.070800	0.106	-0.015	0.157	1.07
23	MVP_EDAD_VEHICULO	2	4,213.11	1,006.28	0.012800	0.752	0.068	0.092	1.01
24	MVP_EDAD_VEHICULO	[5,6]	4,862.12	1,218.81	0.070100	0.061	-0.003	0.143	1.07
25	MVP_EDAD_VEHICULO	1	5,861.01	1,467.27	0.029400	0.436	0.045	0.103	1.03
26	MVP_EDAD_VEHICULO	0	7,288.83	1,844.15	0.036900	0.329	-0.037	0.111	1.04
27	MVP_EDAD_VEHICULO	[7,+Inf]	nan	nan	0.000000	nan	nan	nan	1.00
28	MVP_SUMA_ASEGURADA_DP4	e.[60000,Inf]	441.98	123.16	0.239100	0.004	0.076	0.402	1.27
29	MVP_SUMA_ASEGURADA_DP4	a.[0,10500]	4,432.85	1,165.40	-0.348900	0.000	-0.431	-0.267	0.71
30	MVP_SUMA_ASEGURADA_DP4	b.[10500,15000]	8,787.94	2,246.23	-0.201500	0.000	-0.256	-0.147	0.82
31	MVP_SUMA_ASEGURADA_DP4	d.[22500,60000]	9,623.42	2,360.30	0.125800	0.000	0.076	0.175	1.13
32	MVP_SUMA_ASEGURADA_DP4	c.[15000,22500]	nan	nan	0.000000	nan	nan	nan	1.00

Figura 14. Estructura y resultados del modelo tradicional directo

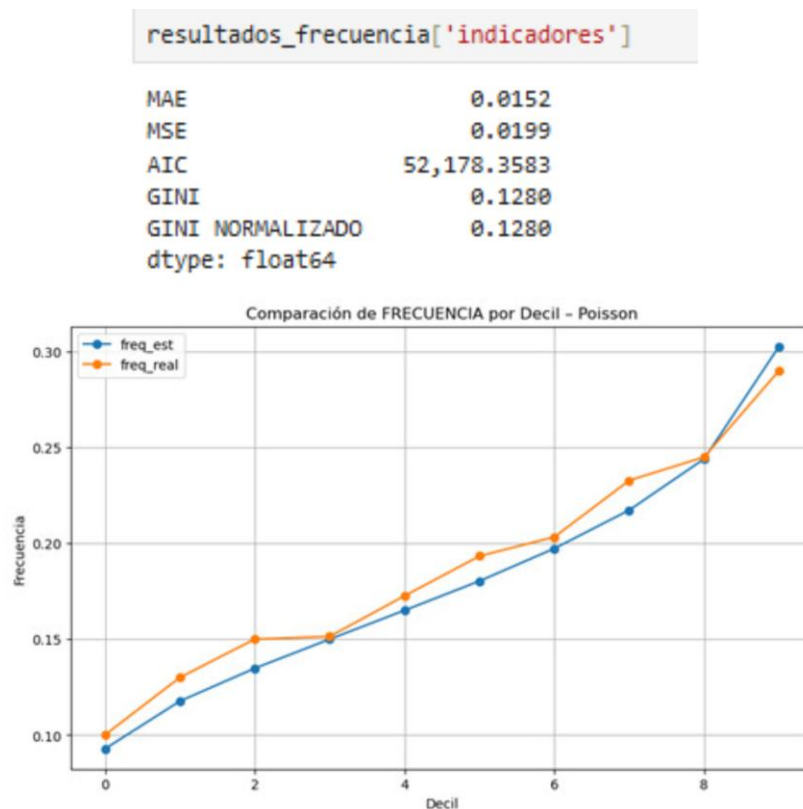


Figura 15. Rendimiento general de modelo GLM tradicional directo

Los resultados del modelamiento tradicional evidencian que la incorporación de determinadas variables tales como ubicación geográfica y grupos de riesgo tradicionales no reflejan aporte significativo al modelo, sino generan por el contrario ruido y volatilidad debido a la baja carga muestral por categorías que representan. Esto nos lleva directamente a la discusión de trabajar técnicas de modelamiento ML a fin de generar variables sintéticas que concentren mejor el riesgo para el modelo general.

Modelo híbrido con variables sintéticas ML:

a. Estructura y resultados por variable:

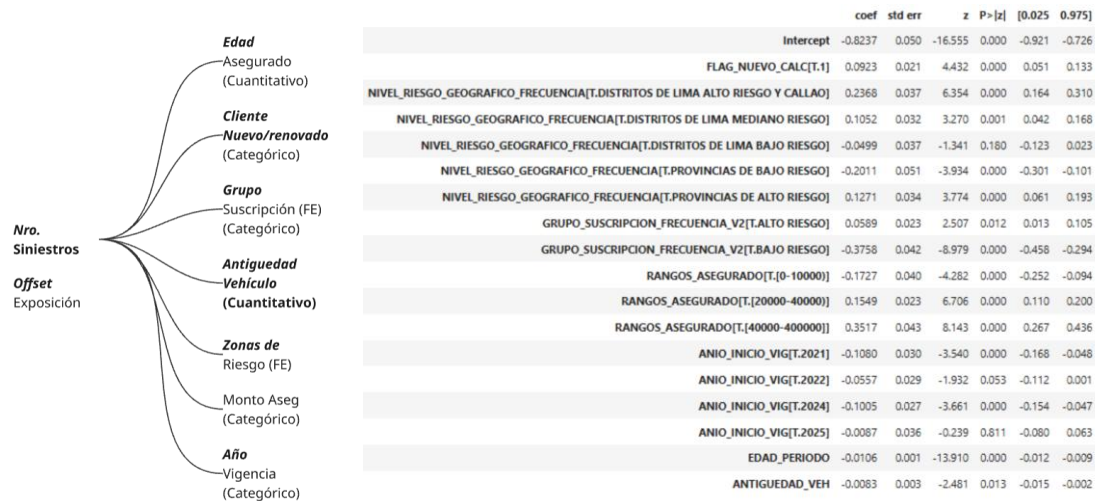


Figura 16. Estructura y resultados de modelo híbrido

b. Performance generalizado del modelo para predicción:

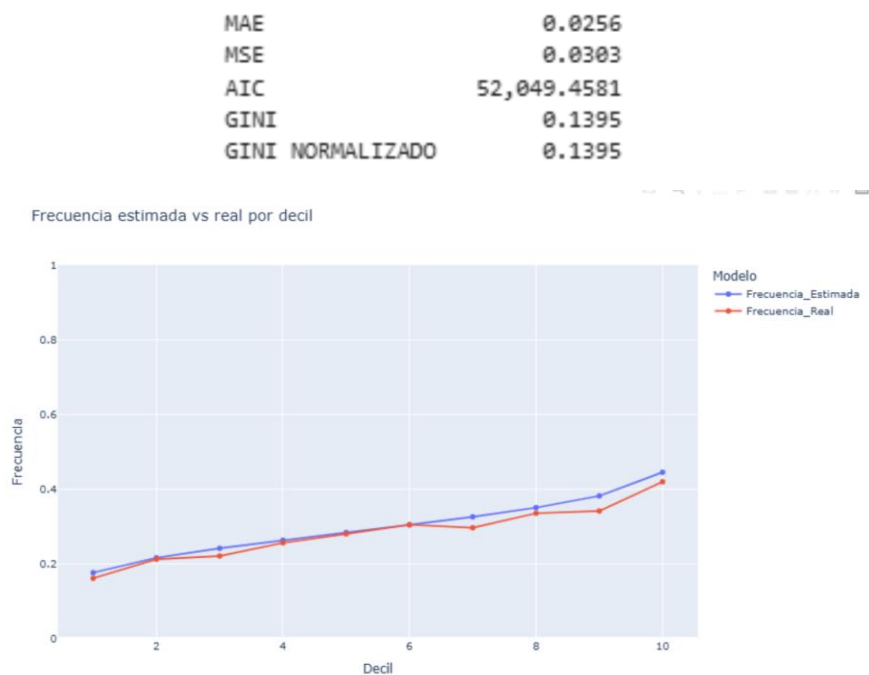


Figura 17. Rendimiento general Lift Curve MAE de modelo GLM híbrido

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

De acuerdo con los resultados obtenidos y realizando la comparativa entre el rendimiento de un modelo lineal generalizado incorporado todas las variables directas, respecto al modelo lineal híbrido con variables sintéticas creadas a partir de técnicas de machine learning y aprendizaje no lineal, podemos evidenciar que, el rendimiento del segundo modelo mejora respecto al primero a nivel de significancia e representatividad por categoría de variable y al nivel de rendimiento general evidenciado con la evaluación del error absoluto y de la curva de deciles de predicción respecto a realidad.

Esto nos permite concluir directamente que las técnicas avanzadas de machine learning permiten catalizar mejor el aprendizaje de riesgo, así como nos permiten generar grupos sintéticos en base a características no supervisadas. Esto conduce a una mejora en el rendimiento y precisión de predicción de siniestros conservando la interpretabilidad a través del modelo generalizado, donde podemos disponer de una explicación e interpretación conceptual y transferible a ambientes de negocio.

6.2 Conclusiones personales

A nivel personal, este proyecto me resulta satisfactorio en un ambiente de comparación de métodos tradicionales y técnicas modernas que permiten, no reemplazar, sino mejorar combinadamente el rendimiento y precisión de las predicciones y sobre todo la discriminación del riesgo.

Esta apertura de enfoques tradicionales a modelos modernos permite conectar y traer los beneficios del contexto actual y modernizar procesos actuariales sin sacrificar ámbitos importantes como lineamientos regulatorios, necesidades explicativas e interpretables y sobre todo, acceso intuitivo a perfiles múltiples que participan de estos objetivos actuariales.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

Estos resultados son un gran avance que promete gran proyección en futuras líneas de investigación, dada los grandes pasos de modernización que hoy vivimos en el contexto de inteligencia artificial.

Se recomienda a los futuros investigadores, profundizar sobre la factibilidad de aplicación de experimentos completamente desarrollados en ambientes de machine learning y el reto que conlleva migrar no solo el contexto técnico sino la modernización del enfoque tradicional a uno nuevo.

Capítulo 8. REFERENCIAS

- [1] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. London, U.K.: Chapman & Hall, 1989.
- [2] A. Noll, R. Salzmann, and M. V. Wüthrich, “Case Study: French Motor Third- Party Liability Claims,” *ASTIN Bulletin*, vol. 48, no. 3, pp. 817–861, 2018, doi: 10.1017/asb.2018.18.
- [3] G. B. Kuo, A. C. Chuang, and C. C. Hsu, “Application of Machine Learning Techniques in Insurance Pricing,” *Expert Systems with Applications*, vol. 168, 2021, Art. no. 114349, doi: 10.1016/j.eswa.2020.114349.
- [4] M. V. Wüthrich, “From Generalized Linear Models to Neural Networks, and Back,” *Swiss Finance Institute Research Paper*, no. 21-08, 2021.
- [5] C. Molnar, *Interpretable Machine Learning*, 2nd ed. Munich, Germany: Leanpub, 2022.
- [6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.
- [7] J. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [8] A. Henckaerts, R. C. C. Breugelmans, and J. Dhaene, “Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods,” *North American Actuarial Journal*, vol. 24, no. 3, pp. 255–285, 2020, doi: 10.1080/10920277.2019.1683151.
- [9] E. W. Frees, G. Meyers, and A. Cummings, “Predictive Modeling Applications in Actuarial Science,” *North American Actuarial Journal*, vol. 18, no. 1, pp. 1–18, 2014. Convocatoria de financiación interna proyectos de investigación 2025.
- [10] S. Chakraborty and S. Joseph, “Explainable AI for Insurance: Challenges and Opportunities,” *IEEE Access*, vol. 9, pp. 63613–63626, 2021, doi: 10.1109/ACCESS.2021.3074568.
- [11] M. Lindholm, “Machine Learning in Non-Life Insurance Pricing: A Review,” *Risks*, vol. 8, no. 1, 2020, Art. no. 17, doi: 10.3390/risks8010017.
- [12] S. Biecek and D. Burzykowski, *Explanatory Model Analysis*, Boca Raton, FL, USA: CRC Press, 2021.

[PÁGINA INTENCIONADAMENTE EN BLANCO]