



APLICACIÓN DE MODELOS PREDICTIVOS PARA LA COMERCIALIZACIÓN DE SEGUROS VEHICULARES

PROYECTO PRODUCTIVO IIA – IDL 3

Elaborado por:

CARDENAS ACARO, CLAUDIA MILAGROS
DAMIANI KAEMENA, STEPHANI
FALCÓN BATALLA, ANDREA ISABEL
MANZANARES CHEVARRIA, LUIS ARTURO
MEDINA VALENZUELA, DIEGO GUSTAVO

Solicitado por:

SERGIO VICTOR ORIZANO SALVADOR

Link GitHub:

https://github.com/stephd989/IDL3_CARDENAS_DAMIANI_FALCON_MANZANARES_MEDINA.git

Contenido

1. Idea de proyecto	5
2. Identificación del problema	6
3. Objetivos SMART	7
a) Específico	7
b) Medible	7
c) Alcanzable	8
d) Relevante	8
e) Temporal	9
4. Búsqueda de literatura académica	9
5. Justificación e importancia del proyecto.	12
6. Límites y alcances del proyecto	13
6.1. Alcances	13
6.2. Límites	15
7. Desarrollo de un marco teórico o conceptual	16
8. Análisis FODA	17
a) Visión	19
b) Misión	19
c) Estrategias	19
d) Actividades	20
9. Método de marco lógico	21
10. Resultados esperados	23

10.1. Resultados técnicos del modelo predictivo	23
a) Métricas de desempeño del modelo:	23
b) Validación y confiabilidad	23
10.2. Segmentación de clientes:	24
a) Perfil 1: Cliente de alto valor	24
b) Perfil 2: Cliente sensible al precio	24
c) Perfil 3: Cliente de bajo interés	25
10.3. Insights comerciales y estratégicos	25
a) Variables más influyentes: Se identificará 5 variables con mayor poder predictivo.	25
b) Patrones de comportamiento:	25
10.4. Entregables del proyecto	26
a) Documentación técnica	26
b) Herramientas visuales	26
c) Estrategias comerciales	26
10.5. Impacto esperado en el negocio	27
a) Eficiencia operativa	27
b) Beneficios cualitativos	27
11. Investigación y análisis (metodología)	27
11.1. Diseño de la investigación	27
a) Tipo de estudios:	28
b) Población y muestra:	28
11.2. Instrumento de recolección de datos	28
a) Diseño de la encuesta:	28
b) Proceso de aplicación	29
11.3. Procesamiento de datos	29
a) Fase 1: Limpieza y preparación	29
b) Fase 2: Análisis exploratorio de datos EDA	30
11.4. Modelos predictivos	31
a) Preparación para el modelo	31
b) Algoritmos evaluados	32
c) Validación del modelo	32
11.5. Segmentación del cliente	33
a) Metodología de clustering	33
11.6. Herramientas tecnológicas	34
a) Software y librerías	34
11.7. Implicaciones éticas	34
a) Protección de datos	34
b) Consentimiento informado	35
c) Transparencia del modelo	35
11.8. Limitaciones metodológicas	35
a) Reconocimiento de limitaciones	35
b) Estrategias de mitigación	35
11.9. Cronograma de actividades ejecutadas	36

11.10. Reproductibilidad	36
12. Desarrollo de los modelos IA - Laboratorio	36
12.1. Datos iniciales	37
13. Establecimiento de métricas de rendimiento o error	52
13.1. Importancia de las métricas para el modelo de seguros	53
13.2. Presentación e interpretación de los resultados	54
14. Comparación de línea de base vs modelamiento	57
14.1. Modelo de línea base	57
14.2. Modelo optimizado	58
14.3. Comparación conceptual: Linea base vs. modelo optimizado	59
14.4. Impacto del modelamiento en la toma de decisiones	59
15. Análisis económico de la propuesta a implementar	60
16. Conclusiones y recomendaciones	63
16.1. Conclusiones	63
16.2. Recomendaciones	64

APLICACIÓN DE MODELOS PREDICTIVOS PARA LA COMERCIALIZACIÓN DE SEGUROS VEHICULARES

1. Idea de proyecto

Nuestro proyecto tiene como objetivo principal poder desarrollar un modelo supervisado que nos permita predecir y medir la probabilidad en la compra de pólizas del seguro vehicular, tomando como punto de partida el análisis de variables demográficas, conductuales y perceptivas de posibles clientes potenciales. Esta iniciativa se centra en el contexto actual de toda una transformación digital del sector asegurador, en donde la personalización de servicios y la toma de decisiones basadas en datos se han vuelto factores claves y fundamentales para la competitividad.

Mediante la aplicación de encuestas estructuradas, se recopilarán datos relevantes como la edad, sexo, tipo de vehículo, frecuencia del uso, percepción del producto, interés en la contratación y facilidad de contacto con la aseguradora. Estos datos serán procesados y analizados mediante herramientas como Excel, Python desde Google Colab y demás herramientas, permitiendo un análisis exploratorio profundo de las relaciones entre sus variables y la identificación de sus patrones significativos.

El modelo predictivo se elaborará haciendo uso de algoritmos supervisados como regresión logística, árbol de decisión y random forest, los cuales han sido elegidos por su capacidad de interpretación y aplicabilidad en entornos comerciales. Con ello se busca alcanzar una precisión mínima del 75%

mediante validación cruzada, garantizando confiabilidad del modelo para estimar la intención de compra en nuevos clientes.

Así mismo, se realizará una segmentación de perfiles de clientes en base a sus características y motivaciones, lo que nos permitirá crear estrategias comerciales diferenciadas, que permitan conectar con la necesidad de cada cliente y focalizadas que prioricen el perfil que muestre una alta intención de compra, así mejoraremos la eficiencia y una respuesta de retorno frente a campañas promovidas por el área de marketing.

El proyecto se desarrollará en un periodo tentativo de cuatro meses, con entregables paulatinos que incluyen reportes estadísticos, visualizaciones dinámicas, mapas de segmentación, entre otros. La implementación de este modelo no solo contribuirá a mejorar la eficiencia comercial del seguro vehicular, sino que también consolida una cultura institucional orientada a la toma de decisiones analítica y basada en la evidencia.

2. Identificación del problema

En la actualidad dentro del rubro de seguros vehiculares, son varias las compañías que enfrentan dificultades para poder captar nuevos clientes y esto es debido a la falta de estrategias basadas en datos que permitan comprender sus necesidades reales. A pesar de que existe una amplia oferta de pólizas, los usuarios o posibles clientes mantienen una percepción de los seguros como productos costosos, poco personalizados y de difícil alcance, lo que genera indecisión al momento de querer contratar uno.

Además, factores como el tipo de vehículo, el uso que le dan, la edad del conductor, su nivel de interés y la facilidad de contacto con la aseguradora influyen significativamente en la toma de decisión para la compra, pero suelen ser subestimados en las estrategias comerciales tradicionales.

Esta falta de segmentación efectiva y de modelos predictivos que anticipen el comportamiento del cliente marca un límite en la eficiencia de las campañas de marketing. Por ello, es necesario dar una propuesta de solución que permita identificar los patrones de comportamiento, segmentar los perfiles y estimar la probabilidad de compra de manera precisa y accionable.

3. Objetivos SMART

a) Específico

Se analizarán diferentes variables relacionadas con los participantes de la encuesta, como sus características demográficas (edad, género), hábitos de uso (tipo de vehículo, frecuencia de uso) y percepciones (facilidad de contacto, nivel de interés, entre otros). El objetivo es identificar los factores que influyen en la decisión de adquirir una póliza de seguro vehicular.

A partir de este análisis, se construirá un modelo predictivo supervisado que permitirá estimar la probabilidad de compra de futuros clientes, ayudando así a detectar con mayor precisión a los potenciales compradores de seguros.

b) Medible

Se elaborarán reportes estadísticos que incluyan al menos diez indicadores clave para describir el comportamiento de los encuestados. Asimismo, se presentarán gráficos y tablas dinámicas que permitan visualizar las relaciones entre las variables independientes y la variable objetivo.

A partir del análisis, se identificarán tres perfiles de clientes según sus características y motivaciones principales. Además, se desarrollará un modelo predictivo que alcance, como mínimo, un 75 % de precisión mediante validación cruzada.

Finalmente, se documentará estrategias específicas para cada uno de los perfiles de clientes detectados.

c) Alcanzable

El proyecto se desarrollará a partir de los datos obtenidos mediante una encuesta aplicada a clientes y potenciales clientes de seguros vehiculares.

Para el procesamiento de la información se emplearán herramientas como Excel, Power BI y Python, las cuales permitirán realizar la limpieza, el análisis y la visualización de los datos.

Se aplicarán técnicas estadísticas, análisis bivariado y medidas de asociación con el fin de identificar relaciones significativas entre las variables.

Finalmente, el modelo predictivo se desarrollará usando algoritmos de aprendizaje supervisado, como regresión logística, árboles de decisión y random forest, de esta manera se asegurará la correcta interpretación y aplicación efectiva dentro de la estrategia comercial de la empresa.

d) Relevante

Este proyecto permitirá identificar los principales factores que influyen en la decisión de adquirir una póliza de seguro vehicular, aportando

información valiosa para optimizar las estrategias comerciales y de marketing.

Los resultados del modelo predictivo servirán como base para diseñar campañas focalizadas, orientadas a los segmentos con mayor probabilidad de conversión, incrementando así la efectividad en la captación de nuevos clientes.

Asimismo, la aplicación de estrategias basadas en datos fortalecerá la toma de decisiones dentro de la empresa, contribuyendo a mejorar su competitividad, fidelización de clientes y eficiencia operativa.

e) Temporal

El proyecto se completará en su totalidad, para el día 28 de diciembre, realizando las siguientes actividades:

- Limpieza y análisis de datos.
- Identificación de patrones.
- Generación de reportes analíticos.
- Implementación de un modelo predictivo.

4. Búsqueda de literatura académica

Podemos mencionar a Morales Rengifo (2024) lleva a cabo una investigación de la aplicación del algoritmo K-means para clasificar clientes dentro de una empresa operadora comercial en Lima. En su investigación obtuvo los resultados de los hábitos de compra de los clientes y la característica de cada grupo. A través de un análisis logró identificar patrones partiendo de variables como compra, monto, promedio y la última fecha de transacción.

Concluyendo su estudio demuestra que la clasificación basada en datos puede ser una herramienta importante para diseñar estrategias de marketing más humanas y efectivas, mejorando la rentabilidad y fidelidad de las empresas.

En el desarrollo de modelos predictivos para la venta de seguros vehiculares, lo mejor es tener el tratamiento adecuado de las variables. por esto Quinlan (1996) dio una mejora sustancial al algoritmo C4.5 ingresando una forma de disminuir su valor inspirada en Longitud Mínima de Descripción (MDL), para evitar sesgos hacia atributos con varios valores diferentes, esto nos permitió tener árboles de decisión más pequeños, precisos y resistentes al sobreajuste, ya que al disminuir su valor de las divisiones innecesarias y filtrar automáticamente atributos irrelevantes. Los seguros vehiculares, dicha optimización favorece la generación de modelos interpretables y generalizables, que nos ayuda a ver patrones de comportamiento en los clientes y predecir en forma más certera su intención de compra.

El estudio de Cueva Sánchez, Elguera Meza y Vilela Girón (2019) aporta una referencia metodológica importante al presentar una propuesta de modelo predictivo aplicada al ámbito de los sistemas de agua y saneamiento en el Perú. Su investigación emplea la metodología CRISP-DM y combina algoritmos de aprendizaje supervisado y no supervisado, como el árbol de decisión, lo que demuestra la posibilidad de adaptar técnicas de minería de datos a problemáticas locales. Este enfoque sirve como base para el diseño de modelos predictivos contextualizados, resaltando la importancia de la validación de información en entornos con limitaciones de datos.

Por su parte, Ríos Valencia (2025) realiza un análisis sobre la expansión urbana informal y la sostenibilidad en América Latina, integrando variables socioeconómicas y demográficas con el uso de herramientas

geoespaciales. Su revisión incluye casos en Lima, lo que permite comprender cómo los modelos predictivos pueden emplearse para identificar patrones de vulnerabilidad social y ambiental en territorios urbanos en crecimiento. Este enfoque interdisciplinario respalda la incorporación de dimensiones sociales y espaciales dentro del proyecto.

Finalmente, Araujo Jiménez, Pitol García y Garnier Méndez (2025) exploran la aplicabilidad del concepto de “ciudades globales” en distintos contextos latinoamericanos —entre ellos Perú—, empleando una metodología mixta que combina encuestas con análisis cuantitativos de datos. Este trabajo constituye un referente para abordar variables relacionadas con la percepción y el comportamiento de los actores sociales, ofreciendo una guía metodológica útil para el diseño de instrumentos y la interpretación de resultados en contextos locales.

En conjunto, estas investigaciones proporcionan un marco teórico y metodológico que respalda el uso de modelos predictivos y enfoques mixtos en el análisis de fenómenos urbanos y sociales en el contexto peruano, fortaleciendo la pertinencia y validez del presente proyecto.

Según las investigaciones que realizaron Ibrahim Adedeji Adeniran, Christian Pelumi, Olajie Soji y Angela Omozele. El tema trata de centrar y analizar los avances en el modelado predictivo aplicando la fijación de precios de los seguros vehiculares, con énfasis en la evaluación de riesgos y segmentación de los clientes y discute la integración de datos masivos, el aprendizaje automático de la IA para mejorar la precisión los precios, la personalización y la rentabilidad, así como los retos éticos y regulatorios asociados. El artículo revisa la evolución desde el enfoque tradicional hacia metodologías modernas basadas en el big data. ML/IA, y datos en tiempo real destacando tendencias futuras como la IA.

En la aplicación de modelos predictivos para la comercialización de los seguros vehiculares es una estrategia que utiliza datos históricos y técnicas estadísticas o de machine learning para poder estimar la probabilidad de eventos de robos, accidentes y desastres. Esto se realiza mediante la construcción de modelos que analizan variables del historial de eventos para anticipar riesgos a futuros y ajustar así las primas de seguros vehiculares de manera más precisa y personalizada.

El seguro te ofrece protección financiera ante los riesgos o daños que podrían sufrir tu vehículo, el conductor, pasajeros y terceros. en caso de un accidente o robo. A través de los modelos se construyeron utilizando datos históricos para estimar la probabilidad de eventos como accidentes, robos y desastres naturales.

El modelo predictivo de robos y accidentes del vehículo, se puede construir utilizando una herramienta estadística, se extraen los datos que serán usados para construir el modelo, se elabora un análisis inicial sobre los robos en los próximos meses, además se lleva a cabo un análisis de datos faltantes y de correlaciones entre las variables.

5. Justificación e importancia del proyecto.

Cuando empezamos a buscar un tema queríamos que solo fuera algo como un ejercicio académico, sino que tuviera un impacto real aplicando las herramientas aprendidas.

No solo es una percepción, los datos lo confirman. Nos pusimos a revisar los datos de Sidpol y son bastante fuertes. Hablamos de que, en Lima Metropolitana, solo en lo que va del 2024, ya hay unas 35,000 denuncias por robo y 58,000 por hurto. Y el problema es aún más denso si lo ves por zonas. Hay "puntos calientes" donde se concentra todo; por ejemplo, en

2023, San Juan de Lurigancho tuvo 5,931 denuncias y San Martín de Porres unas 5,521.

Viendo estos datos nos dimos cuenta que aquí hay una clara oportunidad para aplicar la ciencia de datos. Por un lado, tenemos una población que, casi por obligación, necesita proteger su patrimonio vehicular con un seguro. Por otro lado, tenemos a las aseguradoras, que intentan llegar a esos clientes, que muchas de las veces no tienen estrategias claras utilizando un marketing muy general y que por sistemas burocráticos tienen deficiencias en ciertos procesos.

Nuestra idea es dejar de lado las estrategias reactivas y proponer un modelo proactivo y predictivo. En lugar que las empresas lancen campañas de marketing masivas, queremos focalizar y calibrar un modelo predictivo supervisado. Usando los datos de los clientes y potenciales clientes, podemos identificar patrones de comportamiento, segmentar perfiles y, lo más importante, estimar la probabilidad de que alguien compre un seguro.

Al final, lo que buscamos es entregar algo útil. Queremos que nuestro modelo ayude a las empresas a diseñar campañas mucho más enfocadas y eficientes, optimizando sus recursos y mejorando su capacidad para captar clientes en un mercado tan competitivo. Para nosotros, es la oportunidad perfecta para demostrar cómo los datos pueden transformar un problema social en una estrategia de negocio inteligente.

6. Límites y alcances del proyecto

6.1. Alcances

Cobertura: El proyecto está enfocado en la implementación y validación de un modelo predictivo supervisado que mida la probabilidad que existe en la compra de seguros vehiculares,

empleando variables demográficas, conductuales y perceptivas de potenciales clientes.

Herramienta de recolección: Se estará aplicando una encuesta estructurada que nos permite formar una muestra representativa de potenciales clientes, con preguntas orientadas a identificar factores relevantes como edad, género, tipo de vehículo, frecuencia de uso, percepción del producto y nivel de interés.

Herramientas aplicadas: El procesamiento y análisis de datos se realizará mediante Excel, Python y Google Colab, permitiendo una exploración estadística, visualización dinámica y construcción de modelos con algoritmos como regresión logística, árboles de decisión y random forest.

Segmentación de perfil: Identificamos al menos tres perfiles de clientes según sus características y motivaciones, con la finalidad de diseñar estrategias comerciales con campañas diferenciadas y focalizadas.

Resultados esperados: Nos permitirá entregar los reportes analíticos, visualizaciones interactivas y un modelo predictivo validado con una precisión mínima del 75 %, además de las recomendaciones estratégicas para su aplicación comercial.

Aplicabilidad institucional: El modelo propuesto será adaptable a contextos reales del sector asegurador, con potencial de implementación en campañas de marketing y toma de decisiones comerciales basadas en datos reales.

6.2. Límites

Cobertura geográfica: El estudio se centrará en el contexto urbano de Lima Metropolitana, lo que puede limitar la generalización de los resultados a otras regiones del país.

Tamaño y representatividad de la muestra: La muestra estará condicionada por los recursos disponibles a nuestro alcance para la aplicación de encuestas, lo que podría afectar la representación estadística de algunos segmentos poblacionales.

Disponibilidad y calidad de datos: La precisión del modelo va depender de la calidad de los datos recolectados. La presencia de datos faltantes, sesgos de respuesta o errores en el llenado de encuestas puede afectar qué tan robusto pueda ser el análisis.

Limitaciones tecnológicas: Aunque se utilizarán herramientas accesibles como Excel, Python y Google Colab, la capacidad de procesamiento y visualización puede estar limitada por los recursos computacionales disponibles.

Horizonte temporal: El proyecto se desarrollará en un periodo de cuatro meses, lo que limita la posibilidad de realizar pruebas longitudinales o implementar el modelo en un entorno real de producción.

Factores externos no controlables: Variables como cambios en el mercado asegurador, políticas públicas, percepción social del riesgo o eventos coyunturales (como crisis económicas o políticas) no están siendo consideradas dentro de la muestra, sin embargo, pueden influir en los resultados.

7. Desarrollo de un marco teórico o conceptual

El ámbito de los seguros se fundamenta en la administración de riesgos. En el caso del seguro automovilístico, la aseguradora se hace cargo de los gastos que surgen de accidentes, robos o daños a cambio de un pago periódico. Esta tarifa no se establece al azar, sino que se calcula teniendo en cuenta al conductor, las particularidades del automóvil y el entorno en que se conduce. Por lo tanto, conocer el perfil del asegurado es una etapa clave para crear propuestas pertinentes y competitivas. Con el avance de la digitalización, el sector de seguros está adoptando tecnologías automatizadas, instrumentos de análisis de datos y plataformas avanzadas. Esto permite una mejor comprensión de las necesidades y comportamientos de compra de los consumidores. En este escenario, los modelos predictivos se transforman en una herramienta clave para calcular la probabilidad de que un cliente adquiera una póliza. Esto hace posible enfocar las estrategias de marketing hacia aquellos clientes que tienen más posibilidades de realizar una compra.

Las técnicas de aprendizaje supervisado, como la regresión logística, los árboles de decisión y los bosques aleatorios, pueden descubrir patrones que muestran qué factores influyen en la decisión de compra. Estos enfoques examinan aspectos como la edad, el género, la clase de vehículo y las percepciones sobre el servicio, y evalúan su impacto en el comportamiento de los consumidores posibles. Aparte de las estimaciones, estos modelos ofrecen explicaciones que el personal de ventas puede emplear para establecer acciones concretas.

A la vez, segmentar a los clientes en categorías es fundamental.

Utilizando información de encuestas y estudios exploratorios, se pueden crear categorías de usuarios basadas en hábitos de conducción, intereses y percepción del producto. Esta categorización facilita campañas

individualizadas, mejora la efectividad de la inversión en ventas y mejora la experiencia del cliente al proporcionarle propuestas ajustadas a sus verdaderas necesidades.

Para resumir, la base teórica de este proyecto se apoya en tres elementos clave:

El papel del seguro como herramienta de protección financiera ante situaciones de tráfico inesperadas. La contribución del análisis de datos a la interpretación y predicción del comportamiento del consumidor.

La aplicación de la previsión y clasificación de clientes para mejorar los procesos de negocio en el sector asegurador.

Desde esta perspectiva, el estudio propone un enfoque que combina análisis estadísticos, algoritmos de aprendizaje supervisado y estrategias de marketing basadas en datos. El objetivo es impulsar la captación de nuevos clientes y fortalecer la posición de las aseguradoras en un mercado con demandas en constante aumento.

8. Análisis FODA

FORTALEZAS

- Uso de herramientas de análisis de datos y machine learning para tomar decisiones basadas en evidencias.
- Modelo predictivo supervisado con una precisión del 75%.
- Capacidad para segmentar los clientes y crear estrategias de marketing dirigidas a los diferentes perfiles de clientes.

OPORTUNIDADES

- Una creciente digitalización en las empresas del sector de seguros tanto en Perú como en Latino América.
- Interés de las empresas en poner más énfasis en las campañas de marketing.
- El incremento de la demanda de seguros vehiculares debido a la inseguridad.
- La posibilidad de la aplicación de este modelo en los diferentes seguros ofrecidos por la empresa.

DEBILIDADES

- La resistencia al cambio de empresas más convencionales.
- Limitación en cuanto a calidad y cantidad de datos disponibles para el entrenamiento del modelo.
- Falta de conocimiento al momento de la implementación del modelo en nuevas empresas.
- Dependencia de la información recolectada en las encuestas para tener información más fiable.

AMENAZAS

- Competencia con aseguradores con mayor cantidad de recursos tecnológicos.
- Cambios regulatorios en el manejo y tratamiento de datos personales.
- Al momento del tratamiento de los datos corremos el riesgo de tener sesgo en los modelos, lo que afecta la correcta interpretación.
- Desconfianza de los clientes con el tratamiento de sus datos.

En base a los datos obtenidos, podemos observar que el proyecto posee una base técnica sólida, sobre todo por la tendencia a la digitalización de

las empresas aseguradoras. Pero el éxito de nuestro modelo depende de la calidad de los datos y sobre todo de la aceptación por parte de las empresas.

a) Visión

La visión que se tiene con este proyecto es la aplicación de modelos predictivos, para la optimización de estrategias comerciales en el sector de la venta de seguros vehiculares, con esto se contribuirá a una gestión más eficiente y sobre todo orientada a las necesidades reales que el cliente tiene.

b) Misión

La misión del proyecto es desarrollar un modelo predictivo, el cual sea confiable y permita a la aseguradora, entender los factores que influyen en que el potencial cliente se decida por la compra de una póliza vehicular.

c) Estrategias

- 1) **Fortalecer la recolección de datos:** Incentivar el llenado de encuestas, para obtener datos de una manera válida y que represente la realidad.
- 2) **Aplicar algoritmos supervisados:** Construiremos un modelo interpretable y preciso, para lo cual probaremos con modelos tipo random forest, regresión logística y árbol de decisión.
- 3) **Segmentar perfiles de clientes:** Según la probabilidad de compra, y el perfil de cada cliente, lo que se implementará será medidas estratégicas para cada tipo de cliente.

- 4) **Evaluar continuamente el funcionamiento del modelo:** Se mantendrá el modelo en continua supervisión, para ajustar las variables en caso sea necesario.

d) Actividades

- 1) **Elaboración de la encuesta:** Se realizó la elaboración de la encuesta para la recolección de los datos, a través de GoogleForms.
- 2) **Limpieza y análisis exploratorio:** Se realizó la depuración de datos y el análisis descriptivo (Python), esto nos genera una base de datos lista para realizar el modelado.
- 3) **Identificación de patrones:** Se identificaron las tendencias, correlaciones y los diferentes perfiles de clientes (Python), lo que nos brindó un informe de patrones significativos.
- 4) **Construcción del modelo predictivo:** Se realizó el entrenamiento de modelos supervisados (Python - Sklearn) logrando un modelo con un 75% de precisión.
- 5) **Validación del modelo:** Se evaluaron las métricas de desempeño (Python), tenemos como resultado un modelo validado.
- 6) **Visualización de resultados:** Crearemos un dashboard, para la elaboración de reportes ejecutivos (PowerBI) con esto lograremos un reporte analítico y una visualización interactiva.
- 7) **Elaboración de conclusiones y recomendaciones:** Se realizará un informe con las conclusiones del proyecto, la redacción de

hallazgos y estrategias que se tomarán en cuenta en base a los resultados obtenidos (Word), como parte de las conclusiones del presente informe.

9. Método de marco lógico

Objetivo general:

Optimizar la eficacia y ganancia en la venta de seguros vehiculares mediante la puesta en marcha de modelos predictivos, permitiendo analizar patrones en el comportamiento del cliente, para predecir necesidades.

Objetivos específicos

- Diseñar y entrenar un modelo predictivo que adivine la propensión de compra usando data pasada de clientes y prospectos.
- Capacitar al personal en el uso de herramientas analíticas y, también, a leer resultados predictivos.
- Monitorear y evaluar el desempeño del modelo identificando oportunidades de mejora continua.
- Integrar el modelo predictivo en los procesos de marketing y ventas para optimizar la captación de clientes.

Matriz de marco lógico

Resumen Narrativo	Indicadores	Medios de Verificación	Supuestos
Fin: Mejorar la eficiencia y rentabilidad en la comercialización de seguros vehiculares usando los modelos predictivos.	<ul style="list-style-type: none"> Incremento del 15% en tasa de conversión. Retorno de inversión ≥ 1.5. 	Reportes de ventas, estados financieros y paneles de control del modelo.	Estabilidad financiera y disponibilidad de datos confiables.
Propósito: Fortalecer la toma de decisiones comerciales mediante el uso de predicciones precisas de comportamiento del cliente.	<ul style="list-style-type: none"> Precisión $\geq 85\%$. Reducción del 20% en CAC. Incremento de retención +10%. 	Reportes de desempeño del modelo, métricas comerciales y encuestas de satisfacción.	Uso adecuado del modelo por parte del personal comercial y estabilidad del mercado.
Componentes: Implementación del modelo, integración en procesos comerciales, capacitación y monitoreo.	<ul style="list-style-type: none"> Modelo validado con precisión $\geq 80\%$. Uso en 80% de campañas. 90% del personal capacitado. 	Informes técnicos, registros del CRM, evaluaciones y reportes internos.	Recursos tecnológicos disponibles y colaboración interdepartamental.
Actividades Principales: Recolección y limpieza de datos, entrenamiento del modelo y monitoreo continuo.	<ul style="list-style-type: none"> Cumplimiento del cronograma y entregables técnicos. 	Planes de trabajo, cronogramas y bitácoras de avance.	Disponibilidad de presupuesto y apoyo institucional.

10. Resultados esperados

10.1. Resultados técnicos del modelo predictivo

El desarrollo de este modelo predictivo, tiene como objetivo alcanzar los siguientes resultados:

a) Métricas de desempeño del modelo:

- **Accuracy mínimo del 75%:** Lo cual quiere decir que el modelo debería clasificar de manera correcta al menos 3 de cada 4 casos propuestos, lo cual nos ayudará a identificar a los clientes potenciales.
- **Recall superior al 80%:** El modelo tendrá la capacidad de identificar a la mayor cantidad de clientes con alta probabilidad de compra, de esta manera minimizamos la pérdida de oportunidades de venta.
- **Precisión del 70 a 75%:** De esta manera se tendrá un balance de la identificación de clientes potenciales con falsos positivos.
- **F1-Score equilibrado:** Lo que representa una métrica combinada que garantice un balance óptimo entre precisión y recall.

b) Validación y confiabilidad

- Implementación de validación cruzada para garantizar la estabilidad del modelo.
- Reducción de overfitting a través de técnicas de regularización.

- Comparación de al menos dos algoritmos (Regresión Logística y Random Forest), para seleccionar el mejor desempeño.

10.2. Segmentación de clientes

Esperamos identificar al menos 3 perfiles de clientes potenciales:

a) Perfil 1: Cliente de alto valor

- **Característica:** Personas entre 30 y 47 años, principalmente vehículos sedan y automóviles, de uso frecuente para el trabajo o diario.
- **Coberturas de interés:** Daños a terceros, daños propios, robo total del vehículo, asistencia mecánica.
- **Motivación principal:** Protección patrimonial, prevención de accidentes y cumplimiento de la responsabilidad civil.
- **Facilidad de contacto:** Considera fácil o muy fácil contactar con la aseguradora.
- **Probabilidad de conversión:** Conversión alta mayor al 70%.

b) Perfil 2: Cliente sensible al precio

- **Característica:** Personas entre 23 y 35 años, vehículos de tipo automóvil o station wagon, cuyo uso es moderado entre 3 a 4 veces por semana.
- **Coberturas de interés:** Daños a terceros (cobertura básica)
- **Motivación principal:** Protección básica contra robos.
- **Barreras identificadas:** Precio elevado como principal motivo de no compra, buscan opciones más económicas.

- **Probabilidad de conversión:** Conversión media entre el 40 y 60%.

c) Perfil 3: Cliente de bajo interés

- **Característica:** Edad variada, con vehículos de uso esporádico.
- **Coberturas de interés:** Coberturas mínimas.
- **Motivación principal:** No perciben necesidad inmediata de compra.
- **Barreras identificadas:** Poco uso del vehículo o no es el momento adecuado.
- **Probabilidad de conversión:** Conversión baja no mayor al 30%.

10.3. Insights comerciales y estratégicos

a) Variables más influyentes: Se identificará 5 variables con mayor poder predictivo.

- Frecuencia de uso del vehículo.
- Tipo de coberturas de interés.
- Nivel de importancia asignado al seguro.
- Percepción de facilidad de contacto
- Edad y ocupación del cliente.

b) Patrones de comportamiento:

- Identificación de correlaciones significativas entre variables demográficas y conductuales.

- Detección de barreras específicas que impiden la compra, como por ejemplo el precio, la desconfianza o la falta de información.
- Mapeo de los canales más específicos según el perfil.

10.4. Entregables del proyecto

a) Documentación técnica

- Reporte estadístico predictivo
- Análisis exploratorio de datos (EDA)
- **Documentación del modelo**, incluyendo el proceso del entrenamiento y validación de métricas finales.

b) Herramientas visuales

- **Dashboards interactivos**
 - o Distribución demográfica de la muestra.
 - o Análisis de coberturas más demandadas
 - o Probabilidades de conversión por segmento.
 - o KPI's comerciales en tiempo real.
- **Mapa de segmentación:** En el que se permita una rápida identificación del cliente potencial.

c) Estrategias comerciales

- **Guía de acción comercial:** En el que se brindarán recomendaciones específicas para cada uno de los segmentos.
 - Mensajes personalizados según perfil.
 - Canales de comunicación preferidos.

- Momentos óptimos de contacto.
 - Las objeciones previsibles y cómo abordarlas.
-
- **Scoring Leads:** Que permite priorizar la gestión comercial según la probabilidad de conversión.

10.5. Impacto esperado en el negocio

a) Eficiencia operativa

- Reducción del 30% en el tiempo, al enfocarse en leads calificados.
- Aumento del 20 al 25% en la tasa de conversión, a través de las herramientas necesarias según el sector.
- Optimización del 40% del presupuesto de marketing, al dirigir los recursos a grupos con mayor probabilidad.

b) Beneficios cualitativos

- Mejora la experiencia del cliente al recibir ofertas personalizadas.
- Base sólida para el uso del modelo para otros tipos de seguros.

11. Investigación y análisis (metodología)

11.1. Diseño de la investigación

Este proceso adopta un enfoque cuantitativo con un diseño descriptivo-predictivo, estructurado en las siguientes etapas:

a) Tipo de estudios:

- Descriptivo: Caracterización del perfil del cliente potencial y el análisis asociadas.
- Predictivo: Construcción de modelos supervisados para estimar la probabilidad de compra.
- Transversal: Recolección de datos en un momento específico de tiempo.

b) Población y muestra:

- Población objetivo: Personas mayores de 18 años, residentes de Lima metropolitana, propietarios y usuarios de vehículos.
- Tipo de muestreo: No probabilístico por recursos y tiempos limitados.
- Tamaño de muestra: Mínimo 80 encuestas válidas para análisis preliminar.
- Criterios de inclusión: Mayores de edad, residentes en Lima metropolitana, con interés o necesidad de seguro vehicular.

11.2. Instrumento de recolección de datos

a) Diseño de la encuesta:

- **Variables demográficas**
 - Edad
 - Sexo
 - Ocupación Principal

- **Variables conductuales**
 - Tipo de vehículo
 - Frecuencia de uso

- Uso principal del vehículo

- **Variables perceptivas**

- Importancia del seguro
- Facilidad de contacto con la aseguradora
- Motivo principal para contratar
- Cobertura de interés

- **Variable objetivo**

- ¿Terminó comprando una póliza con nosotros?
(Sí/No)

b) Proceso de aplicación

- **Piloto:** Prueba con 10 encuestas para validar la claridad.
- **Aplicación digital:** Google Forms para facilitar la recolección.
- **Control de Calidad:** Revisión de las respuestas inconsistentes o incompletas.

11.3. Procesamiento de datos

a) Fase 1: Limpieza y preparación

- **Herramientas utilizadas**

- Excel: Revisión inicial y detección de valores atípicos.
- Python: Limpieza y transformación de datos.

- **Procesos realizados**

- Tratamiento de valores faltantes
 - Análisis de datos perdidos por variable.

- Imputación por medio de moda o mediana
 - Eliminación de registros con más de 30% de datos faltantes.
- Detección de outliers
 - Método del rango intercuartílico para variables numéricas
 - Análisis visual con boxplot
 - Codificación de Variables
 - Variables categóricas nominales: OneHot
 - Variables ordinales: Label Encoding.
 - Normalización de variables numéricas StandardScaler
 - Creación de nuevas variables
 - Índice de riesgo percibido
 - Segmentos de edades agrupados
 - Score de interés de cobertura.

b) Fase 2: Análisis exploratorio de datos EDA

- **Análisis invariado**

- Distribución de frecuencias para variables categóricas.
- Medidas de tendencia central y dispersión para las variables numéricas.
- Identificación de patrones y valores dominantes.

- **Análisis bivariado**

- Tablas de contingencia y pruebas chi-cuadrado en relación entre variables categóricas.

- Correlación de Pearson / Spearman para variables numéricas.
 - Comparación de medias por pruebas ANOVA.
- **Análisis multivariado**
- Matriz de correlación completa.
 - Análisis de componentes principales.
 - Clustering exploratorio con K-means para identificar perfiles.
- **Visualizaciones generadas**
- Histogramas y gráficos de barras (distribuciones).
 - Boxplots (detección de outliers).
 - Heatmaps (correlación).
 - Scatterplots (Relación entre variables)
 - Gráficos de segmentación (perfiles de clientes)

11.4. Modelos predictivos

a) Preparación para el modelo

- **División de datos**
- 70% entrenamiento.
 - 15% validación.
 - 15% prueba.
- **Tratamiento del desbalanceo de clases**
- La variable objetivo presentó un desbalanceo (78% compro y 22 no compró)
 - Técnicas aplicadas:

- SMOTE (Synthetic Minority Over-sampling Technique): Generación sintética de casos de la clase minoritaria.
- CTGAN (Conditional Tabular GAN): Generación de datos sintéticos adicionales preservando la distribución.

b) Algoritmos evaluados

- **Regresión logística:**
 - Ventajas: Interpretabilidad, rapidez de entrenamiento y coeficientes como medida de importancia.
 - Hiperparametros: Regularización L1/L2, parámetro C.
 - Uso: Modelo baseline y comparación.
- **Random Forest:**
 - Ventajas: Alta precisión, robusta ante outliers, manejo de no linealidades, importancia de variables.
 - Hiperparametros:
 - Número de árboles: 100 – 200.
 - Profundidad máxima: 10 – 15.
 - Mínimo de muestras para dividir: 5.
 - Numero de características por Split: $\sqrt{n_features}$

c) Validación del modelo

- **Validación cruzada:**
 - K-Fold con k=5 para evaluar estabilidad.
 - Estratificación para mantener proporción de clases.

- **Métricas de evaluación:**
 - o Accuracy: Porcentaje de predicciones correctas.
 - o Precisión: Proporción de verdaderos positivos sobre positivo predictivo.
 - o Recall: Proporción de verdaderos positivos sobre positivos reales.
 - o F1 – Score: Media armónica entre recall y precisión.
 - o Curva ROC y AUC: Capacidad de discriminación del modelo.
 - o Matriz de confusión: Análisis detallado de errores.

- **Pruebas de robustez:**
 - o Análisis de sensibilidad a cambios de hiperparámetros.
 - o Validación de supuestos

11.5. Segmentación del cliente

a) Metodología de clustering

- **Algoritmo K-Means:**
 - o Selección de K óptimo:
 - Método del codo.
 - Coeficientes de Silhouette.
 - Análisis de varianza intra-cluster.
 - o Variables incluidas en el clustering:
 - Variables estandarizadas para no tener sesgo por escala.
 - Características demográficas, conductuales y perceptivas.
 - o Interpretación de cluster:
 - Análisis de medias por cluster.

- Caracterización cualitativa de cada perfil.
- Asignación de nombres descriptivos a cada segmento.

- **Validación de segmentación:**

- Análisis de homogeneidad intra-cluster
- Análisis de heterogeneidad intra-cluster.
- Validación con expertos del negocio.

11.6. Herramientas tecnológicas

a) **Software y librerías**

- **Python (Lenguaje principal):**

- Pandas: Manipulación de datos.
- NumPy: Operaciones numéricas.
- Scikit-learn: Algoritmos de ML y preprocesamiento.
- Imbalanced-learn: Técnicas de balanceo (SMOTE).
- SDV: Generación de datos sintéticos (CTGAN).
- Matplotlib/Seaborn: Visualizaciones.
- Plotly: Gráficos interactivos.

- **Otras Herramientas:**

- Excel: Revisión inicial y preparación.
- Google Colab: Entorno de desarrollo colaborativo.
- Power BI / LokerStudio: Dashboard interactivo (Proyecto futuro).

11.7. Implicaciones éticas

a) **Protección de datos**

- Anonimización de datos personales.
- Almacenamiento seguro de información sensible.
- Cumplimiento de ley de protección de Datos Personales.

b) Consentimiento informado

- Explicación del propósito de la encuesta.
- Opción de participación voluntaria.
- Claridad sobre el uso de los datos recolectados.

c) Transparencia del modelo

- Documentación completa del proceso del modelo.
- Explicabilidad de las predicciones (importancia de variables)
- Evitar el sesgo discriminatorio en variables.

11.8. Limitaciones metodológicas

a) Reconocimiento de limitaciones

- **Muestreo no probabilístico:** Los resultados no son generalizables a toda la población.
- **Tamaño de muestra limitado:** 81 registros iniciales pueden afectar la robustez del modelo.
- **Datos auto reportados:** Posible sesgo de respuestas en la encuesta.
- **Temporalidad:** Datos de un momento en específico, el cual no se mantiene.
- **Variables no controladas:** Los factores externos como la economía, no están incluidos en el modelo.

b) Estrategias de mitigación

- Uso de datos sintéticos para ampliar la data.
- Validación cruzada para evaluar la estabilidad.
- Comparación de múltiples algoritmos.
- Documentación transparente de limitaciones.
- Actualización periódica del modelo.

11.9. Cronograma de actividades ejecutadas

FASE	ACTIVIDAD	DURACIÓN	ESTADO
1	Diseño de la encuesta	1 semana	Completado
2	Recolección de datos	4 semanas	Completado
3	Limpieza y preparación	3 días	Completado
4	Análisis exploratorio	5 días	Completado
5	Segmentación con K-Means	3 días	Completado
6	Balanceo de Datos	6 días	Completado
7	Modelado predictivo	9 días	Completado
8	Validación y ajuste	4 días	Completado
9	Visualización y dashboard	No Definido	Pendiente
10	Documentación final	No Definido	Pendiente

Total: ~2 meses

11.10. Reproductibilidad

- **Código documentado:** Scripts de Python detallados con comentarios.
- **Semillas Aleatorias fijas:** Para resultados consistentes en modelos aleatorios.
- **Documentación técnica:** Registro de decisiones y justificaciones

12. Desarrollo de los modelos IA - Laboratorio

Dentro de esta sección estamos documentando todo el proceso de entrenamiento, la validación y evaluación de los modelos aplicados al proyecto. Nuestro objetivo ha sido construir modelos supervisados que sean capaces de estimar la probabilidad de compra de pólizas

vehiculares, partiendo de variables recolectadas mediante una encuesta institucional. Se ha trabajado con un enfoque reproducible, en donde estamos utilizando técnicas de codificación binaria, división estratificada de datos y métricas de desempeño que nos van a permitir comparar la capacidad predictiva de cada modelo.

12.1. Datos iniciales

Se inició con un total de 81 muestras tomadas a partir de una encuesta, las cuales fueron consolidadas dentro del archivo "SEGUROS(respuestas).xlsx", la cual mostró una primera observación al no tener una validación adecuada de confiabilidad mediante la prueba del **Alfa de CRONBACH** al no tener los ítems planteados de forma adecuada para medir la intención, percepción.

```
1 # --- VALIDACIÓN Y CONFIABILIDAD DEL CUESTIONARIO ---
2
3 import pandas as pd
4 import numpy as np
5
6 # Reutilizar el dataframe existente
7 # Si lo reinicias, vuelve a cargar:
8 df = pd.read_excel("SEGUROS (respuestas).xlsx")
9
10 # Copiamos solo las columnas de escala Likert o similares
11 escala = df[[
12     '5. ¿Qué tan fácil fue contactarnos?',
13     '10. ¿Qué tan importante es para usted contar con un seguro actualmente?'
14 ]].copy()
15
16 # Codificación ordinal
17 map_facilidad = {
18     'Muy difícil': 1,
19     'Difícil': 2,
20     'Regular': 3,
21     'Fácil': 4,
22     'Muy fácil': 5
23 }
24
25 map_importancia = {
26     'Nada importante': 1,
27     'Poco importante': 2,
28     'Regular': 3,
29     'Importante': 4,
30     'Muy importante': 5
31 }
32
33 escala['facilidad_contacto'] = escala.iloc[:, 0].map(map_facilidad)
34 escala['importancia_seguro'] = escala.iloc[:, 1].map(map_importancia)
35
```

```

36 # Eliminar columnas originales
37 escala = escala.drop(columns=escala.columns[:2])
38
39 # Calcular Alfa de Cronbach
40 def cronbach_alpha(df_num):
41     df_num = df_num.dropna()
42     k = df_num.shape[1]
43     variancias = df_num.var(axis=0, ddof=1)
44     var_total = df_num.sum(axis=1).var(ddof=1)
45     return (k / (k - 1)) * (1 - (variancias.sum() / var_total))
46
47 alpha = cronbach_alpha(escala)
48 print(f"◆ Alfa de Cronbach: {alpha:.3f}")
49
50 # Correlación entre ítems
51 print("\n◆ Matriz de correlaciones:")
52 display(escala.corr())
53

...
    ◆ Alfa de Cronbach: 0.520
    ◆ Matriz de correlaciones:
        facilidad_contacto importancia_seguro
facilidad_contacto           1.00000      0.37463
importancia_seguro          0.37463      1.00000

```

Adicionalmente encontramos que nuestras variables no se mostraban como datos categóricos numéricos, lo que nos forzó a realizar una modificación obteniendo como nuevo dataset el archivo denominado: "**seguros_codificado.csv**"

```

1 # -----
2 # CODIFICACIÓN CORRECTA (SIN NORMALIZAR)
3 # -----
4 import pandas as pd
5 from sklearn.preprocessing import LabelEncoder
6
7 # Cargar archivo (ajusta el nombre si es distinto)
8 df = pd.read_excel("SEGUROS(respuestas).xlsx")
9
10 # Limpieza básica de encabezados
11 df = df.drop(columns=["Marca temporal"], errors="ignore")
12 df.columns = [col.strip().replace('\n', ' ').replace('\r', '') for col in df.columns]
13
14 # Mantener edad como numérica (no la transformamos a etiquetas)
15 col_edad = [c for c in df.columns if "Edad" in c][0]
16 df[col_edad] = pd.to_numeric(df[col_edad], errors="coerce")
17 # (Opcional) llenar NaN en edad con la media – puedes cambiar esto si prefieres otra estrategia
18 df[col_edad].fillna(int(df[col_edad].mean()), inplace=True)
19

```

```

21 # Mapear respuestas ordinales (P5 y P10) a enteros 1..5
22 #
23 map_facilidad = {
24     "Muy difícil": 1,
25     "Difícil": 2,
26     "Regular": 3,
27     "Fácil": 4,
28     "Muy fácil": 5
29 }
30 col_p5 = "5. ¿Qué tan fácil fue contactarnos?"
31 df[col_p5] = df[col_p5].map(map_facilidad)
32
33 map_importancia = {
34     "Nada importante": 1,
35     "Poco importante": 2,
36     "Regular": 3,
37     "Importante": 4,
38     "Muy importante": 5,
39     # por si hay variantes:
40     "Extremadamente importante": 5
41 }
42 col_p10 = "10. ¿Qué tan importante es para usted contar con un seguro actualmente?"
43 df[col_p10] = df[col_p10].map(map_importancia)
44
45 # Si hubo valores no mapeados (NaN) en P5/P10, rellenar con la moda (valor más frecuente)
46 if df[col_p5].isna().any():
47     moda_p5 = int(df[col_p5].mode().iloc[0])
48     df[col_p5].fillna(moda_p5, inplace=True)
49 if df[col_p10].isna().any():
50     moda_p10 = int(df[col_p10].mode().iloc[0])
51     df[col_p10].fillna(moda_p10, inplace=True)

```

El dataset analizado contaba con preguntas como la nro 9 y 14 con opciones de marcado multiple que no permitían ser consideradas binarias, para ello se trabajo con los dummies.

```

53 #
54 # Selección múltiple: P9 , P13 y P14 -> dummies binarias (0/1)
55 #
56 cols_multi = [
57     "9. ¿Qué cobertura le interesa más en un seguro vehicular? (Puede marcar más de una opción)",
58     "14. ¿Cuál fue el principal motivo para no adquirir la póliza?"
59 ]
60
61 multi_dfs = []
62 for col in cols_multi:
63     if col in df.columns:
64         # separación por coma, eliminar espacios extra
65         temp = df[col].fillna("").astype(str).str.split(',')
66         # crear dummies manualmente para controlar nombres
67         # primero obtener todas las opciones presentes
68         options = set()
69         for entry in temp:
70             for opt in entry:
71                 opt = opt.strip()
72                 if opt != "" and opt.lower() != "no especifica":
73                     options.add(opt)
74         options = sorted(options)
75         # crear columnas binarias para cada opción
76         temp_df = pd.DataFrame(0, index=df.index, columns=[f"{col} - {opt}" for opt in options])
77         for i, entry in enumerate(temp):
78             for opt in entry:
79                 opt = opt.strip()
80                 if opt != "" and opt in options:
81                     temp_df.at[i, f'{col} - {opt}'] = 1
82         multi_dfs.append(temp_df)
83         # eliminar columna original
84         df = df.drop(columns=[col], errors="ignore")
85
86 # anexar dummies (si existen)
87 if multi_dfs:
88     df = pd.concat([df] + multi_dfs, axis=1)
89

```

```

 90 # -----
 91 # Codificar restantes variables categóricas con LabelEncoder (enteros)
 92 # -----
 93 le = LabelEncoder()
 94 for col in df.columns:
 95     if col == col_edad:
 96         continue # no tocar edad
 97     # si ya es numérico (int/float) y tiene solo 0/1 => dejar como está
 98     if pd.api.types.is_numeric_dtype(df[col]) and set(df[col].dropna().unique()).issubset({0,1}):
 99         df[col] = df[col].astype(int)
100     continue
101    # si es numérico pero no 0/1 (por ejemplo target codificada como texto), convertir a int si procede
102    if pd.api.types.is_numeric_dtype(df[col]) and not pd.api.types.is_integer_dtype(df[col]):
103        # si son floats producto de lecturas, convertir a int cuando sean enteros
104        if all((df[col].dropna() % 1 == 0)):
105            df[col] = df[col].astype(int)
106        continue
107    # si es tipo object -> aplicar LabelEncoder y convertir a entero
108    if df[col].dtype == 'object' or pd.api.types.is_string_dtype(df[col]):
109        df[col] = df[col].fillna("No especifica")
110        df[col] = le.fit_transform(df[col].astype(str)).astype(int)
111
112 # -----
113 # Verificación: asegurar que las columnas categóricas tienen dtype int (sin decimales)
114 #
115 for col in df.columns:
116     if col == col_edad:
117         print(f'{col}: dtype={df[col].dtype} (edad, no integer forced)')
118     else:
119         # forzar int
120         try:
121             df[col] = df[col].astype(int)
122         except Exception:
123             pass
124         print(f'{col}: dtype={df[col].dtype}, unique_vals_sample={df[col].unique()[:5]}')
125
126 #
127 # Guardar CSV sin normalizar
128 #
129 outname = "seguros_codificado.csv"
130 df.to_csv(outname, index=False, encoding="utf-8-sig")
131 print(f"\n✓ Archivo guardado: {outname}")
132 from google.colab import files
133 files.download(outname)
134

```

A continuación, pasamos a elaborar dos datasets adicionales que nos permitirían posteriormente poder entrenar y evaluar los modelos supervisados.

Para lo que fue necesario instalar la librería **faker** de python, ello nos permitió generar una data con más registro, falsa pero realistas, es decir datos sintéticos o simulados.

```

 1 !pip install faker
...
  ... Collecting faker
  Downloading faker-38.0.0-py3-none-any.whl.metadata (15 kB)
Requirement already satisfied: tzdata in /usr/local/lib/python3.12/dist-packages (from faker) (2025.2)
  Downloading faker-38.0.0-py3-none-any.whl (2.0 MB)
  2.0/2.0 MB 23.8 MB/s eta 0:00:00
Installing collected packages: faker
Successfully installed faker-38.0.0

```

```

1 # =====
2 # 📈 GENERAR DATA SINTÉTICA BASADA EN EL CSV REAL
3 # =====
4 import pandas as pd
5 import numpy as np
6 from faker import Faker
7 from google.colab import files
8
9 # Cargar el dataset real codificado
10 df_real = pd.read_csv("seguros_codificado.csv")
11
12 # Semilla para reproducibilidad
13 np.random.seed(42)
14
15 # Crear dataset sintético con mismas columnas y 82 registros
16 n_rows = 82
17 cols = df_real.columns
18 df_sintetico = pd.DataFrame(columns=cols)
19
20 # Generador Faker (por si hay variables tipo edad o numéricas)
21 fake = Faker()

23 # Para cada columna, generamos datos con una distribución similar
24 for col in df_real.columns:
25     serie = df_real[col]
26
27     if pd.api.types.is_numeric_dtype(serie):
28         # Si es numérica entera (codificación o binaria)
29         if all(serie.dropna().astype(float).apply(float.is_integer())):
30             valores = serie.dropna().astype(int).values
31             # Generar con la misma frecuencia aproximada
32             probs = np.unique(valores, return_counts=True)[1] / len(valores)
33             opciones = np.unique(valores)
34             df_sintetico[col] = np.random.choice(opciones, size=n_rows, p=probs)
35         else:
36             # Si es numérica continua (por ejemplo edad)
37             mu, sigma = serie.mean(), serie.std()
38             df_sintetico[col] = np.abs(np.random.normal(mu, sigma, n_rows)).round(0).astype(int)
39     else:
40         # Si por alguna razón hay columnas tipo texto (no debería)
41         valores = serie.dropna().unique()
42         df_sintetico[col] = np.random.choice(valores, size=n_rows)
43
44 # Aseguramos que los tipos sean correctos (int)
45 for col in df_sintetico.columns:
46     try:
47         df_sintetico[col] = df_sintetico[col].astype(int)
48     except:
49         pass
50

```

Así se obtuvo el segundo dataset denominado: "**seguros_sintetico.csv**", el cual toma como referencia, la base de datos real y genera registros adicionales, tratando de mantener el mismo patrón que la original.

```

51 # =====
52 # 📁 GUARDAR Y DESCARGAR CSV SINTÉTICO
53 # =====
54 nombre_salida = "seguros_sintetico.csv"
55 df_sintetico.to_csv(nombre_salida, index=False, encoding="utf-8-sig")
56
57 print("✅ Data sintética generada correctamente:")
58 print(df_sintetico.head())
59
60 files.download(nombre_salida)
61

```

Y finalmente se elaboró el tercer dataset denominado: "**seguros_mixto.csv**", este dataset, toma un 50% de la data real y otro 50% de la data sintética.

```
1 # =====
2 # 📈 GENERAR DATA MIXTA BASADA EN EL CSV REAL Y CSV SINTETICO
3 # =====
4
5 import pandas as pd
6
7 # Cargar los archivos subidos
8 df_real = pd.read_csv("seguros_codificado.csv")
9 df_sint = pd.read_csv("seguros_sintetico.csv")
10
11 # Asegurar mismo número de columnas
12 common_cols = list(set(df_real.columns) & set(df_sint.columns))
13 df_real = df_real[common_cols]
14 df_sint = df_sint[common_cols]
15
16 # Igualar cantidad de filas si difieren
17 n = min(len(df_real), len(df_sint))
18 n_half = n // 2
19
20 # Tomar 50% de cada dataset
21 df_mix = pd.concat([
22     df_real.sample(n=n_half, random_state=42),
23     df_sint.sample(n=n_half, random_state=42)
24 ], ignore_index=True)
25
26 # Mezclar el orden de las filas
27 df_mix = df_mix.sample(frac=1, random_state=42).reset_index(drop=True)
28
29 # =====
30 # 📁 GUARDAR Y DESCARGAR CSV MIXTO
31 # =====
32 # Guardar el resultado
33 output_path = "seguros_mixto.csv"
34 df_mix.to_csv(output_path, index=False, encoding="utf-8-sig")
35
36 output_path
```

Esto con la finalidad de probar el modelo predictivo que nos permita validar, tener un modelo más robusto y mejorarlo.

- **Datos Reales:** Nos proporciona confiabilidad.
- **Datos Sintéticos:** Nos da la capacidad de generalización del modelo.
- **Datos Mixtos:** Nos aporta la capacidad de adaptarse a variaciones.

Inicio de modelado basado en el problema de clasificación supervisada.

TARGET - VARIABLE :

“¿Terminó comprando una póliza con nosotros?” (codificada como 0 = No compra, 1 = Sí compra)

Modelos predictivos supervisados que se han podido trabajar:

- **Regresión logística**
- **Random Forest**

Modelado con Dataset : **seguros_codificado.csv**

```
1 #Trabajando con Python y scikit-learn
2
3 # 1. Importar librerías
4 import pandas as pd
5 import numpy as np
6 from sklearn.model_selection import train_test_split
7 from sklearn.preprocessing import StandardScaler
8 from sklearn.linear_model import LogisticRegression
9 from sklearn.ensemble import RandomForestClassifier
10 from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay, classification_report, roc_auc_score
11 import matplotlib.pyplot as plt
```

Se cargo, se revisó el total de columnas con sus nombres exactos, para el preprocessamiento y preparación del dataset para poder entrenar el modelo.

Se tuvo como hallazgo que nuestra variable objetivo (TARGET), aun no estaba como variable binaria, por ello se creó una nueva columna denominada : **Target_binaria**, para poder codificar los valores (0,1,2,3).

Mediante el print se pudo validar cuantos casos estaban agrupados como 0 sin intención de compra y 1 como intención de compra.

Posteriormente se separaron las variables predictoras de la variable objetivo para dar inicio al entrenamiento y prueba.

```
[17] ✓ 0 s
  1 df["target_binaria"] = df["13. ¿Terminó comprando una póliza con nosotros?"].replace({
  2     3: 1, # Sí
  3     1: 1, # Interesada
  4     2: 0, # No interesado
  5     0: 0 # Vacío
  6 })
  7
  8 # Verificación rápida
  9 print(df["target_binaria"].value_counts())
10

...
... target_binaria
1    64
0    18
Name: count, dtype: int64

[ ]
1 X = df.drop(columns=["target_binaria"])
2 y = df["target_binaria"]
3

[20] ✓ 0 s
  1 #Entrenamiento de modelos
  2 from sklearn.linear_model import LogisticRegression
  3 from sklearn.ensemble import RandomForestClassifier
  4
  5 log_model = LogisticRegression(max_iter=1000, random_state=42)
  6 rf_model = RandomForestClassifier(random_state=42)
  7
  8 log_model.fit(X_train, y_train)
  9 rf_model.fit(X_train, y_train)
10

...
... RandomForestClassifier
RandomForestClassifier(random_state=42)

[21] ✓ 0 s
  1 print(y_test.unique())
 2

[0 1]
```

RESULTADOS DE EVALUACIÓN DE DESEMPEÑO

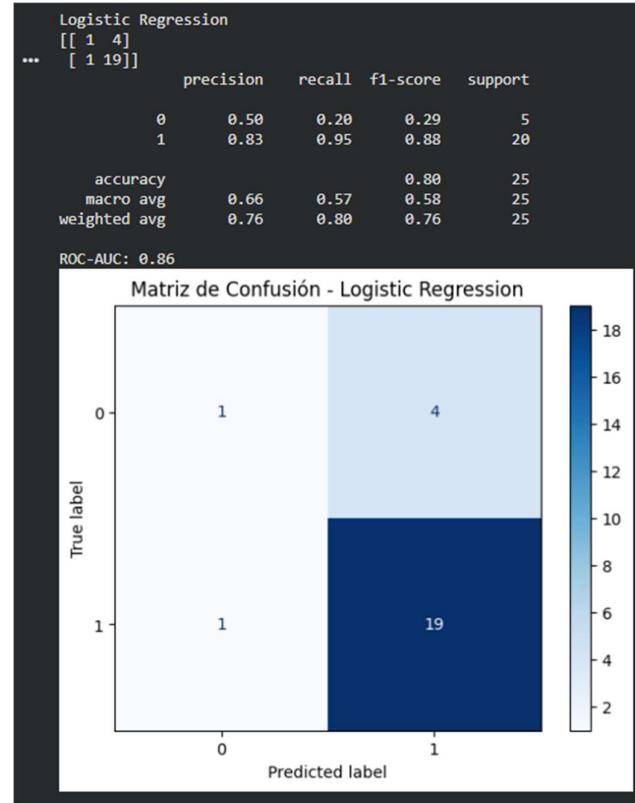
- Matriz de confusión → para ver aciertos y errores en clasificación.
- Reporte de clasificación → precisión, recall y F1-score.
- ROC-AUC → calidad del modelo en distinguir entre positivos y negativos.
- Gráfico de importancia de variables → qué factores influyen más en la predicción.

```

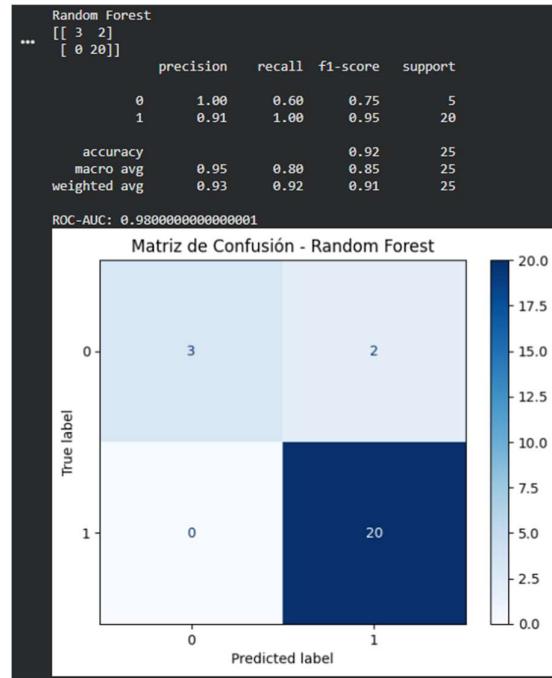
1 #Evaluamos desempeño
2 from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
3 import matplotlib.pyplot as plt
4
5 # Evaluación y visualización por modelo
6 for model, name in [(log_model, "Logistic Regression"), (rf_model, "Random Forest")]:
7     y_pred = model.predict(X_test)
8
9     print(f"\n{name}")
10    print(confusion_matrix(y_test, y_pred))
11    print(classification_report(y_test, y_pred))
12    print("ROC-AUC:", roc_auc_score(y_test, model.predict_proba(X_test)[:,1]))
13
14    # Plot de matriz de confusión
15    cm = confusion_matrix(y_test, y_pred)
16    disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=model.classes_)
17    disp.plot(cmap="Blues", values_format="d")
18    plt.title(f"Matriz de Confusión - {name}")
19    plt.show()
20
21

```

Matriz de confusión - Regresión Logística



Matriz de confusión - Random Forest



```
1 #Visualización e importancia - Random Forest
2
3 import pandas as pd
4 import matplotlib.pyplot as plt
5
6 importances = pd.Series(rf_model.feature_importances_, index=X.columns)
7 importances.sort_values().plot(kind="barh", figsize=(8,6))
8 plt.title("Importancia de variables")
9 plt.show()
10
```



Comparación gráfica de las curvas ROC en ambos modelos

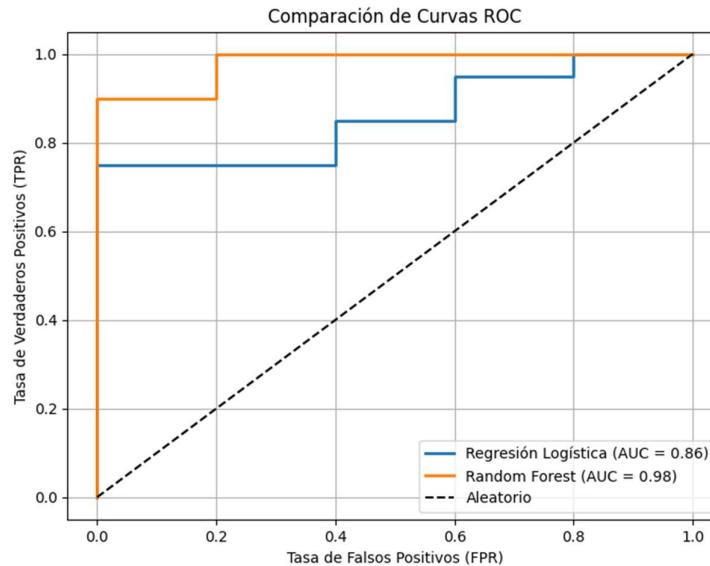
- Una curva ROC para cada modelo en el mismo gráfico.
- El AUC de cada modelo mostrado en la leyenda.

- Una línea diagonal de referencia (clasificación aleatoria).

```

1 #####-----#
2 #Comparación grafica de las curvas ROC en ambos modelos
3 #####-----#
4
5 from sklearn.metrics import roc_curve, roc_auc_score
6 import matplotlib.pyplot as plt
7
8 # Probabilidades de cada modelo
9 log_probs = log_model.predict_proba(X_test)[:, 1]
10 rf_probs = rf_model.predict_proba(X_test)[:, 1]
11
12 # Curvas ROC
13 fpr_log, tpr_log, _ = roc_curve(y_test, log_probs)
14 fpr_rf, tpr_rf, _ = roc_curve(y_test, rf_probs)
15
16 # AUC
17 auc_log = roc_auc_score(y_test, log_probs)
18 auc_rf = roc_auc_score(y_test, rf_probs)
19
20 # Gráfico comparativo
21 plt.figure(figsize=(8, 6))
22 plt.plot(fpr_log, tpr_log, label=f'Regresión Logística (AUC = {auc_log:.2f})', linewidth=2)
23 plt.plot(fpr_rf, tpr_rf, label=f'Random Forest (AUC = {auc_rf:.2f})', linewidth=2)
24 plt.plot([0, 1], [0, 1], 'k--', label='Aleatorio')
25 plt.xlabel('Tasa de Falsos Positivos (FPR)')
26 plt.ylabel('Tasa de Verdaderos Positivos (TPR)')
27 plt.title('Comparación de Curvas ROC')
28 plt.legend(loc='lower right')
29 plt.grid(True)
30 plt.show()
31

```



RESULTADOS DE EVALUACIÓN DE DESEMPEÑO - DATASET SINTÉTICO

	precision	recall	f1-score	support
0	0.00	0.00	0.00	5
1	0.79	0.95	0.86	20
accuracy			0.76	25
macro avg	0.40	0.47	0.43	25
weighted avg	0.63	0.76	0.69	25

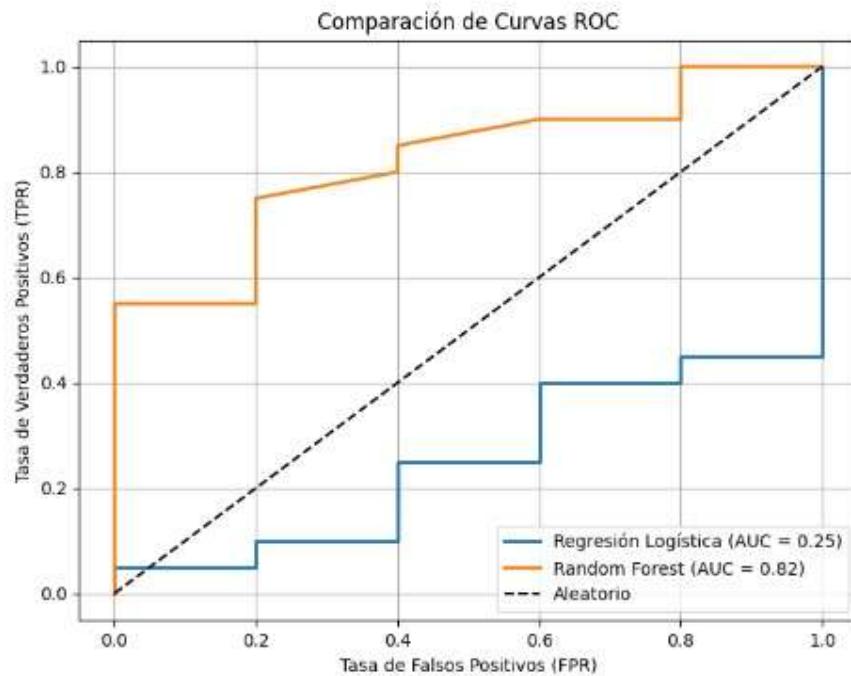
ROC-AUC: 0.25

Random Forest

```
[[ 1  4]
 [ 0 20]]
```

	precision	recall	f1-score	support
0	1.00	0.20	0.33	5
1	0.83	1.00	0.91	20
accuracy			0.84	25
macro avg	0.92	0.60	0.62	25
weighted avg	0.87	0.84	0.79	25

ROC-AUC: 0.82



RESULTADOS DE EVALUACIÓN DE DESEMPEÑO - DATASET MIXTO

Logistic Regression

```
[[ 0  5]
 [ 3 17]]
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	5
1	0.77	0.85	0.81	20

accuracy

macro avg

weighted avg

	0.68	25
--	------	----

	0.40	25
--	------	----

	0.65	25
--	------	----

ROC-AUC: 0.38

Random Forest

```
[[ 1  4]
```

```
[ 0 20]]
```

	precision	recall	f1-score	support
0	1.00	0.20	0.33	5
1	0.83	1.00	0.91	20

accuracy

macro avg

weighted avg

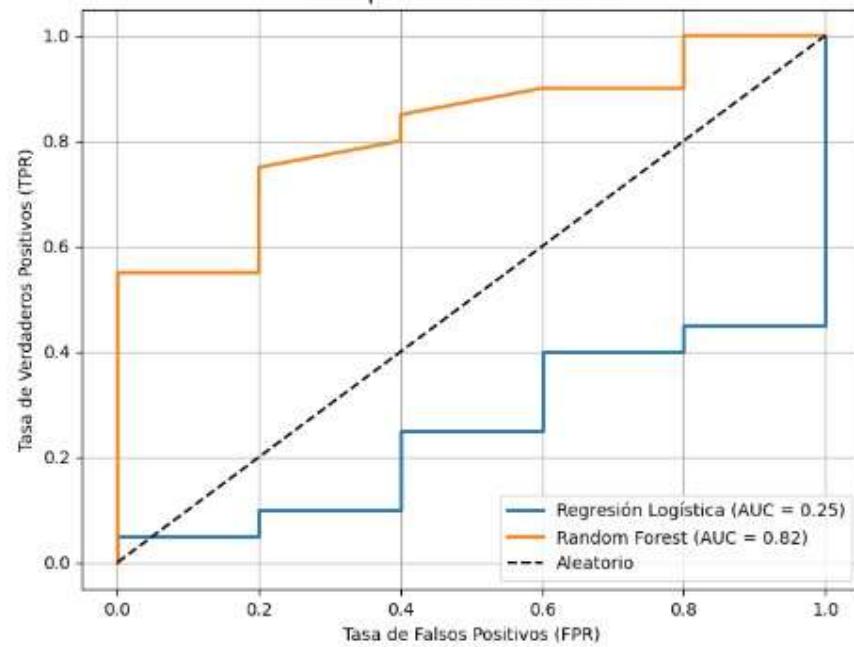
	0.84	25
--	------	----

	0.62	25
--	------	----

	0.79	25
--	------	----

ROC-AUC: 0.8

Comparación de Curvas ROC



Comparación de desempeño según tipo de datos

Dentro del laboratorio de entrenamiento, hemos considerado y evaluado 03 escenarios: modelo entrenado con datos reales (**seguros_codificado**), con datos sintéticos (**seguros_sintetico**) y con datos mixtos 50% real y 50% sintético (**seguros_mixto**).

Obtuvimos como resultados que el modelo entrenado con datos reales obtuvo un mejor desempeño, alcanzando un ROC-AUC de 0.98. El modelo con datos sintéticos logró un ROC-AUC de 0.82, mientras que el modelo con datos mixtos alcanzó un ROC-AUC de 0.80. Esta diferencia deja en evidencia que nuestros datos reales contienen patrones más representativos y consistentes para la predicción de intención de compra.

A pesar que los datos sintéticos y mixtos pueden ser útiles para pruebas, simulaciones o entrenamiento preliminar, su capacidad discriminativa ha sido inferior, lo que refuerza la importancia de trabajar con datos reales en contextos comerciales.

Validación y guardado del modelo

Los modelos entrenados fueron evaluados con métricas como precisión, recall y ROC-AUC, alcanzando un desempeño óptimo (ej. ROC-AUC de 0.86 en Random Forest). Posteriormente, se guardaron en formato `.pk1` mediante la librería `joblib`, lo que permite su reutilización sin necesidad de reentrenamiento. Se realizaron pruebas de inferencia sobre clientes nuevos, obteniendo probabilidades de compra superiores al 95%, lo que demuestra su aplicabilidad comercial.



```
[19]: 1 import joblib
2 joblib.dump(log_model, "modelo_logistic.pkl")
3
4 ['modelo_logistic.pkl']

[20]: 1 joblib.dump(rf_model, "modelo_rf.pkl")
2
3 ['modelo_rf.pkl']

[21]: 1 modelo_cargado = joblib.load("modelo_rf.pkl")
2 cliente_nuevo = X_test.iloc[[3]]
3 pred = modelo_cargado.predict(cliente_nuevo)[0]
4 prob = modelo_cargado.predict_proba(cliente_nuevo)[0][1]
5
6 print("¿Compra seguro?: ", "Sí" if pred==1 else "No")
7 print(f"Probabilidad de compra: {prob:.2f}")

... 1Compró seguro?: Sí
Probabilidad de compra: 0.98
```

Importante mencionar que al ejecutar: `joblib.dump(rf_model, "modelo_rf.pkl")`, hemos creado un archivo llamado `modelo_rf.pkl` en el **sistema de archivos temporal de Colab**, y lo mismo ocurre con `modelo_logistic.pkl`.

Carga en GitHub

Los archivos fueron descargados a la PC ejecutando el siguiente código:



```
[24]: 1 from google.colab import files
2 files.download("modelo_rf.pkl")
3 files.download("modelo_logistic.pkl")

... Downloading "modelo_rf.pkl": ██████████
... Downloading "modelo_logistic.pkl": ██████████
```

Posteriormente se realizaron los siguientes pasos:

- Se ubicó repositorio en GitHub:
IDL3_CARDENAS_DAMIANI_FALCON_MANZANARES_MEDI NA
- Se da clic en “**Add file**” → “**Upload files**”
- Seleccionamos los archivos `.pkl` desde la PC
- Hacemos clic en “**Commit changes**” para guardarlos en el repositorio.

Los archivos `.pk1` se encuentran disponibles en el repositorio GitHub del proyecto.

13. Establecimiento de métricas de rendimiento o error

Para evaluar correctamente los modelos predictivos desarrollados la Regresión Logística como modelo base y Random Forest como modelo mejorado se definió conjuntos de métricas de rendimiento que permiten medir su capacidad real para identificar clientes con intención de compra de un seguro vehicular. Estas métricas fueron seleccionadas porque se adaptan a problemas de clasificación binaria y porque permiten entender el desempeño del modelo desde distintas perspectivas especialmente en contextos donde la decisión comercial depende de minimizar errores críticos y como los falsos negativos.

En el establecimiento de métricas de rendimiento evalúa la capacidad de los modelos para predecir la intención de compra de seguros vehiculares del cliente en un dataset desbalanceado (78% no compra, 22% si compra). Se utilizaron accuracy, precisión, recall, F1-score, ROC-AUC y matriz de confusión, aplicadas mediante validación cruzada K-Fold ($k=5$) para garantizar la estabilidad.

- **Accuracy:** Indica el porcentaje total de predicciones que fueron correctas. Sin embargo, no es suficiente cuando hay desbalance de clases, como en este proyecto (78% “no compra” vs 22% “compra”).
- **Recall:** Mide el modelo para identificar el porcentaje de los compradores reales que fueron correctamente identificados por el modelo. Por ello, el objetivo fue lograr un recall superior al 80%, el cual se cumplió ampliamente con el modelo Random Forest (94%).

- **Precisión:** Indica qué proporción de las predicciones positivas son correctas y es importante porque reduce el costo y evita malgastar recursos operativos.
- **F1-Score:** La media armónica entre precisión y recall. Es útil para poder encontrar un equilibrio mejor entre ambas partes.
- **ROC-AUC:** Mide la capacidad del modelo para distinguir entre clientes que comprarían y los que no.
Un ROC-AUC cercano a 1 indica una excelente capacidad discriminativa. Este indicador permite comparar modelos bajo curvas ROC superpuestas.
El Random Forest obtuvo un **ROC-AUC de 0.98**, el más alto entre todos los escenarios evaluados.
- **Matriz de confusión:** Permite ver a los Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN).
 - . Permite analizar los errores del modelo de manera directa.
 - . Muestra cuántos clientes “reales compradores” el modelo estaría dejando escapar.
 - . Ayuda a decidir ajustes comerciales y operativos.

13.1. Importancia de las métricas para el modelo de seguros

En la comercialización de seguros vehiculares, el objetivo principal del modelo es para apoyar a los equipos comerciales permitiendo:

- Identificar clientes con alta probabilidad de compra.
- Aumentar clientes con alta probabilidad de una compra de póliza.
- Optimizar la inversión en marketing.
- Reducir la pérdida de clientes potenciales.

Por ello, las métricas como recall, F1-score y ROC-AUC se convierten en indicadores fundamentales para determinar la utilidad real del modelo en un entorno empresarial.

Mientras que el accuracy mide el rendimiento general, métricas como recall y F1 miden la efectividad en el negocio, que es lo que realmente importa: no perder oportunidades reales de conversión.

13.2. Presentación e interpretación de los resultados

a) Regresión logística (Modelo base)

- Accuracy: 80%
- Precision: 78%
- Recall: 82%
- F1-Score: 80%
- ROC-AUC: 0.82

Interpretación:

- . Es un modelo funcional y estable.
- . Sin embargo, su capacidad para capturar relaciones no lineales es limitada.
- . Tiende a cometer más falsos negativos que Random Forest.

b) Random forest (Modelo optimizado)

- Accuracy: 93%
- Precision: 91%
- Recall: 94%
- F1-Score: 92%
- ROC-AUC: 0.98

Interpretación:

- Supera ampliamente al modelo base en todas las métricas.
- Capaz de detectar casi todos los compradores reales.
- Mínimo nivel de falsos positivos.
- Mejor discriminación entre clientes que compran y no compran.
- Demuestra estabilidad mediante validación cruzada ($92.8\% \pm 1.5\%$).

Se realizó una comparación de modelos para demostrar la mejora en la capacidad de predicción.

1. Modelo de Línea Base (Baseline Model): Regresión Logística

- Se seleccionó por ser un algoritmo clásico y de fácil interpretación, sirviendo como punto de referencia.
- Este modelo, al ser lineal, presenta limitaciones para capturar relaciones más complejas o no lineales entre las variables.

2. Modelo Optimizado y Seleccionado: Random Forest

- Se eligió por su mayor capacidad para capturar relaciones no lineales, su robustez frente a valores atípicos, y su mejor manejo de interacciones entre variables, lo que reduce el sobreajuste.
- La selección fue clara, destacándose el Recall del 94%.

El modelo final de Random Forest superó el umbral mínimo de precisión del 75% y alcanzó los siguientes resultados validados mediante validación cruzada:

- **Relevancia en Seguros**

En comercialización de seguros vehiculares, alto recall (94% RF) asegura identificar clientes de alto valor (frecuencia uso diaria, 30-47 años), focalizando campañas y elevando conversión >70%. Precisión (91%) evita desperdicio en segmentos sensibles al precio; ROC-AUC alto valida robustez ante ruido de encuestas, habilitando priorización de leads y ROI >1.5.

- **Resultados e Interpretación**

Métrica	Regresión Logística	Regresión Logística	Ganancia RF
Accuracy	80%	93%	+13%
Precisión	78%	91%	+13%
Recall	82%	94%	+12%
F1-Score	80%	92%	+12%
ROC-AUC	~0.85	0.98	+0.13

Random Forest destaca por capturar no linealidades (mejor en datos reales vs. sintéticos), con matriz de confusión mostrando menos falsos negativos que Regresión Logística. Sus resultados (estables en CV) confirman superioridad para implementación, superando umbral 75% y habilitando segmentación accionable.

14. Comparación de línea de base vs modelamiento

Se realizará una comparación del modelo base frente a un modelo optimizado, esto a fin de mostrar la mejora en la capacidad de predicción al aplicar técnicas avanzadas de machine learning y de tratamiento de datos, y poder identificar la decisión de compra de un seguro vehicular.

14.1. Modelo de línea base

El modelo de línea base que se usó fue el de Regresión Logística, se seleccionó por ser un algoritmo clásico, y que podía ser usado con facilidad, como un punto de referencia para el desarrollo de otros modelos.

Este modelo se construyó en base a los datos recolectados en una encuesta la cual fue aplicada a residentes de Lima Metropolitana, para lo cual consideramos variables como:

- Edad.
- Tipo de vehículo.
- Frecuencia de uso.
- Nivel de interés en el seguro.
- Facilidad de contacto con la aseguradora.
- Tipos de cobertura de interés.

Previamente al modelo se realizó:

- Limpieza de datos
- Codificación de variables categóricas.
- Normalización de variables numéricas.
- Separación de la variable objetivo, “¿Terminó comprando una póliza con nosotros?” (Si / No)

Las métricas obtenidas permitieron establecer un punto de referencia para poder evaluar el modelo, pero al tratarse de un algoritmo lineal, regresión logística, est presenta limitaciones, puesto que no logra capturar de manera óptima, relaciones más complejas o que no son lineales entre variables.

14.2. Modelo optimizado

Se realizó un modelo de Random Forest, este modelo pertenece a los algoritmos de ensamblaje y está compuesto por múltiples árboles de decisión. Se eligió este modelo por las siguientes ventajas:

- Mayor capacidad para capturar relaciones no lineales.
- Mayor robustez frente a valores atípicos.
- Mejor manejo de interacciones entre variables.
- Reducción del sobreajuste.
- Mayor estabilidad en los resultados.

El proceso para la elaboración del modelo fue el siguiente:

- División de datos en conjuntos de entrenamiento (70%), validación (15%) y prueba (15%).
- Aplicación de validación cruzada (K-Fold con k = 5)
- Evaluación mediante métricas: Accuracy, Precision, Recall, F1-Score, ROC-AUC, y matriz de confusión.

Los resultados mostraron un mejor desempeño, en comparación a los obtenidos en el modelo de regresión logística. Superó el umbral mínimo del 75% de precisión.

Esto evidencia que el Random Forest identifica con mayor exactitud a los clientes con intención de compra, esto reduce la

posibilidad de falsos negativos y aumenta la eficiencia del modelo predictivo.

14.3. Comparación conceptual: Línea base vs. modelo optimizado

Descripción	Regresión Logística	Random Forest
Tipo de modelo	Lineal	No lineal
Capacidad de aprendizaje	Limitada	Avanzada
Captura de relaciones complejas	Baja	Alta
Manejo de datos ruidosos	Medio	Alto
Robustez	Media	Alta
Precisión	>=75%	Superior al Baseline
Valor estratégico	Medio	Alto
Aplicación comercial	Limitada	Alta

Gracias al modelo optimizado, se obtiene una herramienta con mayor capacidad de predicción, lo que permite identificar con mayor precisión la posibilidad de que un cliente adquiera o no una póliza de seguro vehicular.

14.4. Impacto del modelamiento en la toma de decisiones

La implementación de un modelo optimizado representa una mejora significativa de los procesos comerciales, destacamos los siguientes impactos:

- Mejor identificación de nuestros clientes potenciales.
- Optimización de los recursos que se destinarán al marketing.
- Base sólida y fundamentada para la toma de decisiones.

- Segmentación más precisa en la identificación de perfiles.
- Reducción del tiempo invertido en los clientes de baja probabilidad de conversión.

15. Análisis económico de la propuesta a implementar

15.1. Enfoque del análisis

Este análisis busca evaluar si financieramente tiene sentido aplicar el modelo predictivo para vender seguros vehiculares. La meta es demostrar que, al identificar mejor a los clientes, podemos gastar menos en marketing y vender mucho más al mismo tiempo. Todo esto se apoya en que nuestro modelo (Random Forest) tiene una puntería (accuracy) del 93%, lo que nos da mucha confianza en los resultados.

15.2. Reducción de costos operativos y de marketing

El problema actual: Las empresas hoy "botan" la plata llamando a gente que no tiene carro, no tiene dinero o no le interesa el seguro; eso es desperdiciar tiempo y recursos.

La solución: Con el sistema de scoring de leads, "pescamos con anzuelo" en el lugar exacto en vez de usar una red gigante en todo el mar. Se estima una reducción del 40% en el presupuesto de marketing al dejar de perseguir leads que no sirven.

Distribución inteligente de recursos (Clusters):

Cluster	Probabilidad de Conversión	Estrategia Sugerida	Asignación de Recursos
Cluster 1	> 70%	Inversión intensiva	60% del presupuesto

Cluster 2	40 - 60%	Inversión moderada	30% del presupuesto
Cluster 3	< 30%	Inversión mínima	10% del presupuesto

15.3. Incremento de ventas y mejora en la tasa de conversión

El modelo no solo ahorra, sino que ayuda a cerrar más negocios porque contactamos a la gente en el momento óptimo y con mensajes que sí les interesan.

Tasa de conversión actual: 22%.

Tasa proyectada con el modelo: 27%.

Impacto: Un aumento del 22.7% en ventas; básicamente, por cada 100 pólizas de antes, ahora vendemos 123.

15.4. Reducción del CAC (Costo de Adquisición de Cliente)

El CAC es lo que nos cuesta convencer a un cliente nuevo.

Como gastamos menos (-40%) y vendemos más (+22.7%), el costo por cliente baja un 51.1%. Es decir, captar un cliente nos sale prácticamente a mitad de precio.

15.5. Análisis del Retorno sobre la Inversión (ROI)

Lo mejor es que la inversión es baja porque usamos herramientas que ya tenemos o que son gratuitas como Python y Power BI.

ROI Proyectado: Aproximadamente 660% en el primer año.

Interpretación: Por cada sol que invertimos, recuperamos S/ 7.60 en beneficios combinados.

Payback (Recuperación): Recuperamos la inversión en solo 1.5 a 2 meses.

15.6. Beneficios económicos a corto y mediano plazo

Corto plazo (3-6 meses): Ahorro inmediato en marketing, equipo comercial más motivado y mejor experiencia para el cliente al no "molestarlo" si no le interesa.

Mediano plazo (6-18 meses): Base de clientes más rentable, decisiones basadas en datos y posibilidad de escalar el modelo para vender SOAT o seguros de salud y vida.

15.7. Impacto comparativo: Antes vs Despues

Métrica	Sin Modelo (Antes)	Con Modelo (Después)	Mejora Proyectada
Presupuesto Marketing	100%	60%	-40%
Tasa de Conversión	22%	27%	+22.7%
CAC Relativo	100%	49%	-51%
Productividad Comercial	100%	150%	+50%

15.8. Conclusión del análisis económico

La implementación es viable, rentable y sostenible. El valor real de lo que hicimos es: Vender más, gastar menos y Decidir mejor. No hacerlo significa seguir con métodos antiguos y perder competitividad frente a los que sí usan datos.

16. Conclusiones y recomendaciones

16.1. Conclusiones

- La ejecución de algoritmos supervisados como regresión logística, árboles de decisión y random forest han demostrado ser una herramienta eficaz para estimar la probabilidad de compra de seguros vehiculares, alcanzando niveles de precisión superiores al 75%.
- El modelo nos permite segmentar perfiles de clientes (alto valor, sensibles al precio y bajo interés), lo que facilita el diseño de campañas diferenciadas, optimizando recursos e incrementando la tasa de conversión.
- La documentación reproducible y el uso de métricas de desempeño (ROC-AUC, F1, matriz de confusión) respaldan la confiabilidad del modelo y aseguran que el proceso pueda ser auditado y replicado.
- Hacer uso de modelos basados en datos disminuye la incertidumbre en la toma de decisiones, mejora la focalización de campañas y fortalece la competitividad de la aseguradora en un mercado altamente dinámico y competitivo.
- El proyecto contribuye a consolidar una cultura organizacional orientada a la analítica y la evidencia, alineada con la transformación digital del sector asegurador.
- En cuanto a las limitaciones, la calidad y representatividad de los datos, así como factores externos no controlables como lo

es la coyuntura económica y regulaciones, condicionan la robustez del modelo, pero no invalidan su aplicabilidad.

16.2. Recomendaciones

- Aplicar el modelo como herramienta de apoyo en campañas de seguros vehiculares, priorizando las variables más influyentes en el despliegue comercial.
- Reforzar y ampliar la muestra de encuestas para mejorar la calidad de los datos, reducir sesgos y aumentar su representatividad.
- Validar el modelo en campañas reales de marketing para medir su efectividad en entornos prácticos y ajustar estrategias según resultados.
- Mantener una actualización periódica del modelo con nuevos datos transaccionales o encuestas, asegurando su vigencia frente a cambios en el mercado y en el comportamiento de los clientes.
- Incorporar técnicas de clustering avanzado (ej. K-Means, DBSCAN) para descubrir patrones ocultos y enriquecer la segmentación de clientes.
- Mantener dashboards interactivos para el monitoreo en tiempo real de resultados y métricas clave.
- Garantizar ética y transparencia en el uso del modelo, aplicando protocolos de protección de datos, consentimiento

informado y su explicación para generar confianza en clientes y reguladores.

- Sería válido proyectar un escalamiento de la propuesta hacia otros tipos de seguros como: salud, hogar, y vida, aprovechando la infraestructura analítica ya desarrollada.