UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"PREDICCIÓN DEL SINIESTRO DE VEHÍCULOS PARTICULARES EN UNA COMPAÑÍA DE SEGUROS"

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR

TÍTULO DE INGENIERA ESTADÍSTICA INFORMÁTICA

VANESSA JUDITH RAMIREZ NAVARRO

LIMA – PERÚ

2022

Ouriginal by Turnitin

Document Information

Analyzed document	TSP 2022 - VANESSA JUDITH RAMIREZ NAVARRO 02-09-2022.docx (D143633753)
Submitted	9/5/2022 10:10:00 PM
Submitted by	JORGE CHUE GALLARDO
Submitter email	jchue@lamolina.edu.pe
Similarity	0%
Analysis address	jchue.unalm@analysis.urkund.com

Sources included in the report

Entire Document

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN
“PREDICCIÓN DEL SINIESTRO DE VEHÍCULOS PARTICULARES EN UNA COMPAÑÍA DE SEGUROS”
TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO DE INGENIERO ESTADÍSTICO INFORMÁTICO VANESSA JUDITH RAMIREZ NAVARRO
LIMA – PERÚ
2022

_____

La UNALM es titular de los derechos patrimoniales de la presente investigación (Art. 24 - Reglamento de Propiedad Intelectual)

PREDICCIÓN DEL SINIESTRO DE VEHÍCULOS PARTICULARES EN UNA COMPAÑÍA DE SEGUROS

DEDICATORIA

A mi madre y hermana, porque me impulsan a ser la mejor versión cada día. Mis éxitos también son suyos.

AGRADECIMIENTO

A mis padres, abuelos y familia por todo el apoyo incondicional en tiempos buenos y malos. A mi prometido, por el soporte diario y respaldo en este camino profesional. A el profesor Jorge Chue, por los consejos y asesoramientos en el proceso de elaboración de este trabajo profesional.

¡Gracias por todo!

ÍNDICE GENERAL I. INTRODUCCIÓN 1 1.1 Problemática 2 1.2 Objetivo 3 1.2.1 Objetivo General 3 1.2.2 Objetivo Específico 3 II. REVISIÓN DE LITERATURA 4 III. METODOLOGÍA 7 3.1 Tipo de investigación 7 3.2 Población 7 3.3 Muestra 7 3.4 Metodología 7 3.4.1 Modelo Random Forest 7 3.4.2 Modelo Logístico 10 3.4.3 Métricas de evaluación 11 IV. DESARROLLO DEL TRABAJO 14 V. RESULTADOS Y DISCUSIÓN 25 VI. CONCLUSIONES 29 6.1 Conclusiones 29 VII. RECOMENDACIONES 30 7.1 Recomendaciones 30 VIII. REFERENCIAS BIBLIOGRÁFICAS 31

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

“PREDICCIÓN DEL SINIESTRO DE VEHÍCULOS PARTICULARES EN UNA COMPAÑÍA DE SEGUROS”

PRESENTADO POR:

VANESSA JUDITH RAMIREZ NAVARRO

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR

TÍTULO DE INGENIERA ESTADÍSTICA INFORMÁTICA

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

……..….	……..
Dr. Raphael Félix Valencia Chacón	Dr. Jorge Chue Gallardo
PRESIDENTE	ASESOR
.....	…….....
Mg. Sc. Ana Cecilia Vargas Paredes	Mg. Sc. Clodomiro Fernando Miranda Villagómez
MIEMBRO	MIEMBRO

Lima - Perú
2022

DEDICATORIA:

A Dios, por impulsarme a ser mejor cada día.

A mis padres, abuelos y familia por todo el apoyo incondicional en tiempos buenos y malos.

A mi novio, por el soporte diario y respaldo en este camino profesional.

Mis éxitos también son suyos.

AGRADECIMIENTO:

A el profesor,

Dr. Jorge Chue Gallardo.

Por los consejos y asesoramientos en el proceso de elaboración de este trabajo profesional.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Tras el incremento de siniestros vehiculares de los últimos años en los clientes de una empresa aseguradora peruana, se decidió desarrollar diversos modelos estadísticos que permitan identificar aquellos vehículos que generen un siniestro a futuro y cuáles son los factores relevantes asociados a la siniestralidad. Inicialmente se muestra el proceso de tarificación con el cual la empresa comenzó a ofertar el producto de seguro vehicular para luego ser reemplazado con el modelo estadístico escogido y utilizarlo como factor principal en la definición de precios de las primas. En el presente trabajo de suficiencia profesional se muestra las fases de desarrollo de los modelos Logístico y Random Forest, así como el uso de la matriz de confusión para evaluar las métricas de sensibilidad y especificidad, con la finalidad de elegir el modelo que presente la mejor predicción con respecto a la ocurrencia del siniestro. Para la implementación de los modelos se muestra la metodología Cross Industry Standard Process for Data Mining, el cual sirve para asegurar la planificación y cumplimiento de las fases establecidas en los proyectos analíticos. También se muestra los factores relevantes que se incluyeron en los modelos y presentaron una asociación a la variable respuesta, como información demográfica, financiera y de manejo del propietario; y por parte del vehículo, información de sus características como valor comercial, asientos, tipo de vehículo, colore, entre otros. Finalmente se presenta la implementación de los modelos en el negocio y el impacto positivo en los resultados de distintas frentes de la empresa aseguradora. Las herramientas utilizadas para la preparación, construcción y despliegue de los modelos fueron en la plataforma de Google Cloud Platform con el software Python.

Palabras claves: siniestros, seguro vehicular, logístico, random forest.

ABSTRACT

Following the increase in vehicle claims in recent years among the clients of a Peruvian insurance company, it was decided to develop various statistical models to identify those vehicles that generate a claim in the future and which are the relevant factors associated with the claims rate. Initially, the pricing process with which the company began to offer the vehicle insurance product is shown, to be later replaced with the chosen statistical model and used as the main factor in the definition of premium prices. This work of professional sufficiency shows the development phases of the Logistic and Random Forest models, as well as the use of the confusion matrix to evaluate the metrics of sensitivity and specificity, with the purpose of choosing the model that presents the best prediction with respect to the occurrence of the claim. For the implementation of the models, the Cross Industry Standard Process for Data Mining methodology is shown, which is used to ensure planning and compliance with the phases established in the analytical projects. It also shows the relevant factors that were included in the models and presented an association to the response variable, such as demographic, financial and driving information of the owner; and on the vehicle side, information on its characteristics such as commercial value, seats, type of vehicle, color, among others. Finally, the implementation of the models in the business and the positive impact on the results of different fronts of the insurance company are presented. The tools used for the preparation, construction and deployment of the models were in the Google Cloud Platform with Python software.

Keywords: claims, vehicle insurance, logistics, random forest.

I. INTRODUCCIÓN

El presente trabajo de investigación se desarrolló en una empresa aseguradora peruana con reconocido prestigio a nivel nacional en el sector empresarial. Por razones de confidencialidad y sensibilidad de los datos, se mantendrá en reserva el nombre de la empresa. Los productos que ofrece la empresa son: rentas vitalicias, seguros de vida, seguro obligatorio contra accidentes de tránsito (SOAT) y banca-seguros.

En el 2019, la empresa decide incorporar un nuevo producto de gran relevancia para el sector, el seguro vehicular, que está dirigido a clientes con autos de uso particular y con una antigüedad no mayor a 10 años. Las expectativas respecto al producto fueron que se convirtiese en una gran generadora de ingresos debido a la alta penetración que se observaba en otras aseguradoras, sumado al hecho de contar con una gran cartera de clientes provenientes del SOAT, a quienes se les podía ofrecer el producto para cubrir daños personales y materiales.

La empresa no disponía de información histórica de los costos del seguro vehicular en sus clientes. Ante este escenario, la gerencia y equipos involucrados utilizaron como primera aproximación, estudios comerciales como el de (Salazar, 2004) que ayudasen a establecer la tarifa de las primas y planes a ofertar. Como resultado de estas investigaciones se identificaron como principales factores: la marca, el modelo, antigüedad y el nivel de riesgo del vehículo, para definir las tarifas de las primas. Por lo general, dentro del proceso de tarificación, las marcas y modelos son factores estáticos ya que con esta información se clasifica el vehículo por su categoría (económico, clásico, lujoso) y valor asegurado (valor comercial del vehículo). El factor antigüedad es previamente definido según el año de fabricación del auto y el año en el que se realiza la cotización para adquirir un seguro vehicular. En el caso del factor del nivel de riesgo del vehículo, este depende de la evaluación anual de siniestralidad por marca y modelo del vehículo que brinda APESEG (Asociación Peruana de Empresas de Seguros), donde se determinan los modelos de autos que han resultado con mayor y menor riesgo de accidentes por año.

Durante los últimos años se estuvo trabajando con este proceso de tarificación establecido inicialmente. Sin embargo, en el año 2021, al analizar la evolución de la siniestralidad anual, se observó que el porcentaje de siniestros que se generaban en el producto de seguro vehicular de la empresa aseguradora era de 29%, cantidad mucho mayor al porcentaje promedio de

siniestralidad en el mercado de aseguradoras, el cual era de 25%, según (APESEG, 2019). Este fue el problema que motivó a la empresa en la búsqueda de factores que explicasen este incremento significativo de siniestros, ya que los números indicaban que la tasa seguiría en incremento para el siguiente año. Esta necesidad de mejorar el proceso de tarificación de primas utilizando información adicional del vehículo y datos del conductor fueron incorporados para establecer el nivel de riesgo de siniestralidad con mayor precisión.

En este contexto, la Gerencia solicitó al área de analítica de la empresa desarrollar un modelo predictivo que permitiese calcular la probabilidad de que un vehículo particular tuviese un siniestro. Principalmente se buscaba analizar los siniestros que generaron los clientes con seguro vehicular en la empresa, para luego identificar los patrones que conllevan a esta acción y proyectarlos al parque vehicular, el cual contiene un conjunto de vehículos no clientes a quienes se les puede ofrecer un seguro vehicular. Esta solicitud era de alta importancia ya que el modelo reemplazaría la metodología actual de tarificación en este producto. En el proceso de investigación, se encontró que en (Guillen & Pesantez, 2018), se aplicaron 9 métodos de machine learning para predecir la siniestralidad en una aseguradora, teniendo como conclusión que los métodos Naive Bayes, Discrete Adaboost y Random Forest son modelos apropiados para predecir esta ocurrencia, ya que obtuvieron el mejor comportamiento predictivo con la métrica de precisión. En otra investigación realizada por (Herrera, 2021), se utilizó el modelo Random Forest para estimar el nivel de riesgo de los autos en Madrid, obteniendo buenos resultados en las precisiones globales. Con estos resultados de trabajos previos, se decidió aplicar los modelos Random Forest y Logístico para predecir la ocurrencia de un siniestro y seleccionar el modelo que presenta mejor precisión.

1.1 Problemática

El problema identificado en esta aseguradora ocurre al realizar un seguimiento de la evolución de siniestros en los últimos años desde que se creó el producto en la empresa, el cual se identificó que hubo un incremento de siniestros anuales, pasando de 21.5% (2019) a 32.7% (2021) de tasa de siniestralidad sobre el total de pólizas en esos años, además de superar en 8 puntos porcentuales la tasa de siniestralidad base que se tiene como referencia en la industria de seguros vehiculares (25%). En el año 2021, la cantidad de robos parciales y totales tuvo mayor representación, pasando del 15% (2019) a 25% (2021), siendo estos los que generan mayores gastos a cubrir por la empresa. Al tener un incremento de siniestros, y dentro de ello un incremento de robos conllevó a que los gastos totales también se incrementaran y se reduzca

el retorno de la ganancia neta en el último año. Estos resultados motivaron la necesidad de recalcular la tarificación de las primas del seguro vehicular incorporando información que permita perfeccionar la estimación del riesgo de los siniestros.

1.2 Objetivo

1.2.1 Objetivo General

Predecir la ocurrencia de siniestro de vehículos particulares menores a diez años de antigüedad utilizando los modelos Random Forest y Logístico.

1.2.2 Objetivo Específico

- Identificar el mejor modelo para predecir la siniestralidad de vehículos particulares menores a diez años de antigüedad utilizando las métricas de precisión, sensibilidad y especificidad.
- Calcular la probabilidad de ocurrencia de un siniestro en vehículos particulares utilizando información del vehículo y del conductor aplicando el modelo seleccionado.
- Ilustrar el establecimiento de la tarifa de primas de seguros vehiculares utilizando los niveles de predicción de siniestralidad.

II. REVISIÓN DE LITERATURA

Como se mencionó en un inicio, para iniciar el desarrollo de este trabajo se revisaron diversos tipos de investigaciones relacionadas con aplicaciones de diferentes modelos estadísticos que permitieron predecir la ocurrencia de siniestralidad en diversas entidades.

En la investigación de (Guillen & Pesantez, 2018), propone identificar métodos que mejor se relacionan a la predicción de la siniestralidad de vehículos en una aseguradora, aplicando 9 métodos de machine learning, teniendo como conclusión que los métodos Naive Bayes, Discrete Adaboost y Random Forest son modelos apropiados para predecir esta ocurrencia, ya que obtuvieron el mejor comportamiento predictivo con las métricas de Sensibilidad y Especificidad. Adicionalmente observaron que el factor de predicción de siniestro es muy relevante para el cálculo de las primas, ya que mejoraba la precisión del precio estimado a ofertar en sus clientes asegurados. En otra investigación realizada por (Herrera, 2021), aplicó el método de Random Forest para estimar el nivel de riesgo de siniestralidad de los vehículos en Madrid, el cual obtuvo una precisión global óptima para realizar conclusiones relevantes. Además, en la investigación indica que los factores asociados al vehículo de mayor importancia para dicha estimación fueron la cantidad de infracciones cometidas por el conductor y el nivel de magnitud de infracciones cometidas. En la investigación de (Condori, 2020), también buscaron encontrar una relación determinante asociada a las características del vehículo y la ocurrencia de siniestro en una compañía aseguradora peruana. Como resultado, los factores relevantes fueron: el año de fabricación, la marca, el tipo de carrocería y la obtención de GPS. En cuanto al factor humano, (Condori, 2020) buscó encontrar una asociación entre la conducta del individuo y la ocurrencia del siniestro. En la investigación se menciona que las variables determinantes en la ocurrencia del siniestro asociados al conductor fueron la edad, género y estado civil. Asimismo, en (Sanchez, 2021), al aplicar técnicas de optimización de tarifas en seguro vehicular, encontraron que la edad del conductor fue una de las variables más importantes al momento de predecir la ocurrencia de siniestro.

A continuación, se definen los términos que se emplearán en el trabajo a desarrollar:

- **Sistema de seguro**

  Debido a la gran exposición que las personas están frente a una serie de riesgos, como presentar enfermedades, sufrir accidentes o ser víctimas de robo, existe un sistema de seguros que permiten amortiguar el impacto económico que dichas series de riesgos puedan generar. (SBS, 2021)

- **Siniestro**

  Se entiende por siniestro al suceso que queda estipulado en el contrato del seguro, como parte de una prevención a los daños que puedan generar hacia la persona asegurada o a sus bienes. (SBS, 2021)

- **Siniestrado**

  Hace referencia al vehículo que ha sufrido un siniestro durante la permanencia en el seguro vehicular, ya sea daños materiales, personales, robos parcial o totales del vehículo.

- **Seguro**

  Un seguro es un sistema que protege a la persona y sus bienes frente a diversos hechos que la amenazan. Forma parte de una protección y previsión ante un riesgo.
  Según la Asociación Peruana de Empresas de Seguros (APESEG, 2019), el seguro "es una actividad económico-financiera que presta el servicio de transformación de los riesgos de diversa naturaleza a que están sometidos los patrimonios, en un gasto periódico presupuestable, que puede ser soportado fácilmente por cada unidad patrimonial".

- **Seguro Vehicular**

  Este seguro cubre riesgos que provoquen daños personales, que contemplan indemnizaciones por muerte, invalidez, e incapacidad de la(s) víctima(s) de un accidente, así como también el pago de los gastos de atención médica y de recuperación o rehabilitación; y daños materiales, que contemplan la reparación de partes del vehículo o su reemplazo total, así como la indemnización a terceros por los perjuicios ocasionados a su patrimonio, como consecuencia de un accidente. (SBS, 2021)

- **Prima**

  En APESEG se menciona que la prima es una "aportación económica que ha de satisfacer el contratante o asegurado a la aseguradora en concepto de contraprestación por la cobertura de riesgo que esta le ofrece" (APESEG, 2019).

- **Póliza**

  De acuerdo con (APESEG, 2019):

  "Documento que instrumenta el contrato de seguro, en el que se reflejan las normas que de forma general, particular o especial regulan las relaciones contractuales convenidas entre el asegurador y el asegurado. En tanto haya sido emitido y aceptado por ambas partes, se puede decir que han nacido los derechos y obligaciones que del mismo se derivan. Pese al tratamiento unitario que la legislación concede a la póliza de seguro, en la práctica es frecuente distinguir partes diferenciadas de ella, cuya denominación está íntimamente ligada a su contenido. En este sentido, puede hablarse de condiciones generales, condiciones particulares y condiciones especiales."

- **Asegurado**

  El asegurado es aquella persona que está expuesta al riesgo en sí misma o en sus bienes o intereses económicos. (APESEG, 2019)

- **Cliente**

  Se define como cliente aquel vehículo que tuvo un seguro vehicular en la compañía aseguradora.

- **Propietario**

  El propietario es aquella persona que ejerce posesión de un bien en particular, en este caso un vehículo.

III. METODOLOGÍA

3.1 Tipo de investigación

El tipo de investigación aplicado en el trabajo es de carácter explicativo predictivo, ya que se busca explicar la importancia de las variables y encontrar el mejor modelo que prediga con la mínima tasa de error en la ocurrencia de siniestralidad (Hernández, 2021).

3.2 Población

Propietarios de vehículos particulares registrados en la Superintendencia Nacional de los Registros Públicos (SUNARP, 2015) que contrataron un seguro vehicular dentro del periodo 2019 al 2021 en la empresa aseguradora.

3.3 Muestra

Se trabajará con todos los propietarios de vehículos particulares registrados en la Superintendencia Nacional de los Registros Públicos (SUNARP, 2015) que contrataron un seguro vehicular dentro del periodo 2019 al 2021 en la empresa aseguradora.

3.4 Metodología

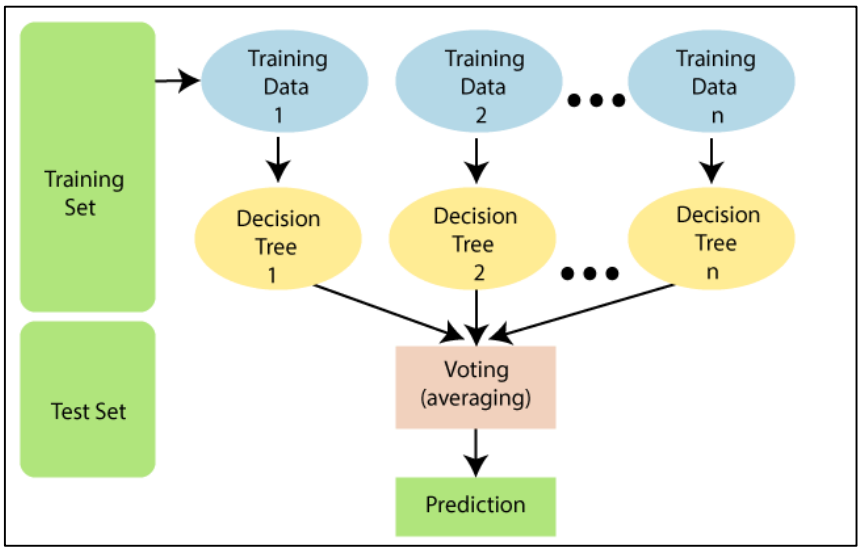
3.4.1 Modelo Random Forest

En (Ruiz, 2018), se detalla que el modelo de machine learning random forest surge a partir de los árboles de decisión, como una evolución de ellos. Los árboles de decisión se basan en clasificar los datos provenientes de un modelo de una serie de particiones binarias, accediendo a realizar predicciones para nuevos casos en base a esta clasificación. Por lo tanto, los random forest estarán formados por un gran número de árboles de decisión, creados a partir de la técnica de baggin o bootstrap aggregating; permitiendo formar bosques que mediante un algoritmo que introduce una aleatoriedad logran reducir la correlación entre árboles. Este algoritmo mejora los árboles de decisión, creando un modelo más fiable y preciso evitando problemas como el overfitting.

Para la construcción del modelo, se divide en dos fases. Por un lado, se crean el conjunto de árboles combinando N árboles de decisión, y por el otro lado, se realizan predicciones para cada árbol creado en la primera fase.

En la Figura 1, se muestra cómo funciona el modelo random forest. Con un conjunto de datos que contiene "N" valores, se dividen y crean "n" subconjuntos de entrenamiento de manera aleatoria para luego ser asignados a "n" árboles de decisión, cada árbol genera un resultado de predicción; y cuando se produce un nuevo registro de datos, el algoritmo predecirá la decisión final en base a la mayoría o promedio de resultados.

Figura 1: *Funcionalidad del modelo Random Forest*



Fuente: (JavaTpoint, 2021)

- **Importancia de variables:** Para llevar un control del aporte significativo que tienen las variables en el modelo, se utilizará la importancia de variables para identificar quienes generan mayor aporte al modelo en comparación a otras variables. Para (Ruiz, 2018), la importancia de variables afecta a la salida del modelo cuando se realizan cambios en las variables de entrada. Las variables de entrada que más variabilidad produzcan en la salida, serían aquella que más influencia tengan y, por tanto, aquellas que mejor explicarán el modelo y serán más importantes. Para la medición de la importancia de variables se utilizará:
    - **Reducción de la impuridad nodal media.** Esta medición consiste en medir la reducción en la impuridad que contribuye la variable elegida en

la partición. Haciendo la media sobre todos los árboles, de todas las reducciones de todas las variables, sacando la media de este valor de impuridad, la cual utiliza la medida del índice de Gini. Finalmente, la variable que más reduzca la impuridad, será la más importante.

- **Supuestos**

  Existen premisas que deben tomarse en consideración al usar Random Forest, según (JavaTpoint, 2021):

  - Debe haber algunos valores reales en la variable de característica del conjunto de datos para que el clasificador pueda predecir resultados precisos en lugar de un resultado adivinado.
  - Las predicciones de cada árbol deben tener correlaciones muy bajas.

- **Ventajas del Random Forest:**
  - Es capaz de realizar modelos de clasificación y regresión.
  - Es apto para manejar grandes volúmenes de datos con alta dimensionalidad.
  - Mejora la precisión del modelo y evita problemas de overfitting.
  - Evita tener que recurrir al uso del proceso de validación cruzada para la optimización de hiperparámetros.
  - Requiere menor tiempo de entrenamiento en comparación con otros modelos.
  - Predice los outputs con alta precisión, incluso para grandes conjuntos de datos realiza el proceso de manera eficiente.
  - Mantiene la precisión incluso cuando se presenta gran proporción de datos ausentes.

- **Desventajas del Random Forest:**
  - Pese a que random forest es apto para modelos de clasificación y regresión, recomiendan no aplicarlo para casos de regresión.

3.4.2 Modelo Logístico

En (Condori, 2020) mencionan que el modelo logístico dicotómicas resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de variables predictoras. Es similar a un modelo de regresión lineal, pero está adaptado para modelos en los que la variable dependiente es dicotómica. Los coeficientes del modelo logístico pueden utilizarse para estimar la razón de las ventajas (odds ratio) de cada variable independiente del modelo. El modelo logístico tiene principalmente dos objetivos: Investigar cómo contribuye en la probabilidad de acontecimiento de un suceso, la existencia o no de diversos factores y el valor o nivel de los mismos; y determinar el modelo más parsimonioso y mejor ajustado que describa el comportamiento entre la variable respuesta y las variables regresoras.

- **Modelo Logístico Binaria**

  El modelo logístico establece la siguiente relación entre la probabilidad de que ocurra el suceso, dado que el individuo presenta los valores ($X = x_1, X = x_2, \dots, X = x_k$).

$$P[Y = 1/x_1, x_2, \dots, x_k] = \frac{1}{1 + e^{(-\beta_0 - \beta_1 x_1 - \cdots - \beta_k x_k)}} \qquad 1$$

  El objetivo es hallar los coeficientes ($\beta_0, \beta_1, \dots, \beta_k$) que mejor se ajusten a la expresión funcional.

  - **Odds Ratio:** Se conoce como odds (ratio del riesgo) al cociente de probabilidades.

$$Odds\ (Ratio\ de\ riesgo) = e^{(\beta_1 - \beta_2 x_2 - \cdots - \beta_k x_k)} \qquad 2$$

    Cuando se hace referencia al incremento unitario en una de las variables explicativas del modelo, aparece el concepto de oods-ratio como el cociente entre los dos odds asociados (el obtenido al realizar el incremento y el anterior al mismo). Suponiendo que ha habido un incremento unitario en la variable X.

$$Odds\ (Ratio\ de\ riesgo) = \frac{odds2}{odds1} = e^{(\beta_i)} \quad (3)$$

De donde se desprende que, un coeficiente βi cercano a cero, es decir, un odds-ratio próximo a 1, indicará que cambios en la variable explicativa xi asociada no tendrán efecto alguno sobre la variable dependiente Y.

- **Bondad de ajuste del modelo:** Se utilizan dos tipos de contrastes: (a) Contrastes que analizan la bondad de ajuste desde un punto de vista global. (b) Contrastes que analizan la bondad de ajuste paso a paso.

  Contraste de bondad de ajuste global de Hosmer-Lemeshow, el índice de bondad de ajuste:

$$z^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)} \quad (4)$$

  Donde $\hat{p}_i = p\left(x_{i1}, x_{i2}, \dots, x_{ik}; \hat{\beta}\right)_{i=1,2,\dots,n}$, $z^2 \approx x^2_{n-k}$ si el modelo ajustado es cierto.

  El estadístico desviación viene dado por la expresión:

$$D = 2\sum_{i=1}^{n} y_i \ln[\frac{y_i}{\hat{p}_i}] + 2\sum_{i=1}^{n-m}(1-y_i)\ln[\frac{1-y_i}{1-\hat{p}_i}] \quad (5)$$

  Donde m=número de observaciones con $y_i = 1$ y $D \approx x^2_{n-k}$ si el modelo ajustado es cierto.

- **Aplicaciones**

  El modelo logístico es aplicado en las siguientes situaciones:

  - Cuando los datos son binarios: 0/1, Verdadero/Falso, Sí/No.
  - Cuando se necesite resultados probabilísticos.
  - Cuando se necesite comprender el impacto de la característica.

3.4.3 Métricas de evaluación

Para la evaluación de los modelos Random Forest y Logístico, se empleará la matriz de confusión para calcular y comparar las métricas de sensibilidad y especificidad en los modelos aplicados.

- **Matriz de confusión:** (Barrios, 2019) menciona que una matriz de confusión es una herramienta el cual permite visibilizar el desempeño de un modelo de aprendizaje supervisado. Esta matriz consiste en una tabla cruzada que contiene información de los valores reales comparados con los valores predichos por el modelo utilizado; cada columna de la matriz hace referencia al número de predicciones de cada clase, por otro lado, cada fila representa la clase real en las observaciones. En total se generan 4 valores que se ejemplifican con los valores utilizados en este trabajo, "siniestrado" y "no siniestrado":

    - **Verdadero positivo:** El valor real es "siniestrado" y el valor predicho también es "siniestrado".
    - **Verdadero negativo:** El valor real es "no siniestrado" y el valor predicho también es "no siniestrado".
    - **Falso negativo:** El valor real es "siniestrado" y el valor predicho es "no siniestrado".
    - **Falso positivo:** El valor real es "no siniestrado" y el valor predicho es "siniestrado".

Figura 2: *Matriz de Confusión Binaria*

Fuente: (Barrios, 2019)

- **Sensibilidad**: También llamada tasa de verdaderos positivos. Hace referencia a la proporción de casos positivos que fueron correctamente identificados por el modelo (Barrios, 2019). Comparando con los valores del trabajo, la sensibilidad detecta correctamente a los que realmente siniestran entre los siniestrados.

$$Sensibilidad = \frac{VP}{VP+FN} \quad (6)$$

- **Especificidad**: También llamada tasa de verdaderos negativos. Hace referencia a la proporción de casos negativos que fueron correctamente identificados por el modelo (Barrios, 2019). Comparando con los valores del trabajo, la especificidad detecta correctamente a los que realmente no siniestran entre los no siniestrados.

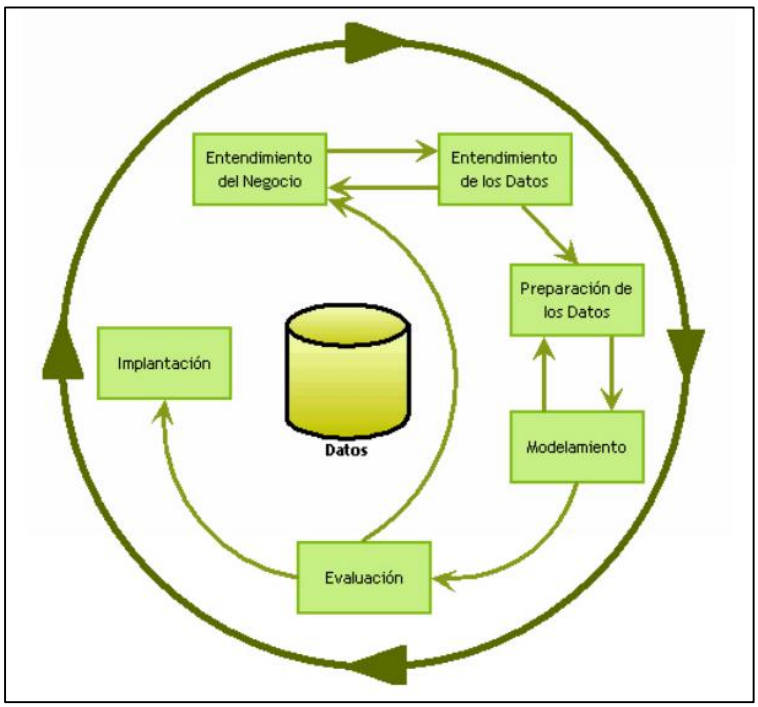
$$Especificidad = \frac{VN}{VN+FP} \quad (7)$$

IV. DESARROLLO DEL TRABAJO

Como integrante del área de analítica y desempeñando el cargo de "Científico de Datos", el autor de este trabajo de suficiencia profesional lideró de inicio a fin el proyecto solicitado por la gerencia de la empresa, el cual comprendía en encontrar un modelo analítico que ayude a predecir la ocurrencia de siniestro de vehículos particulares y optimice el proceso de tarificación mejorando la oferta a usuarios que potencialmente deseen adquirir un seguro vehicular.

Para la realización del proyecto, se aplicó la metodología CRISP - DM (Cross Industry Standard Process for Data Mining), un método dirigido para proyectos de minería de datos que incluye fases que se relacionan entre sí, para la planificación y cumplimiento de este tipo de trabajos. A continuación, se muestra en la Figura 3, la secuencia del proceso CRISP – DM que se utilizó en este trabajo de suficiencia profesional:

Figura 3: *Secuencia del proceso CRISP - DM*



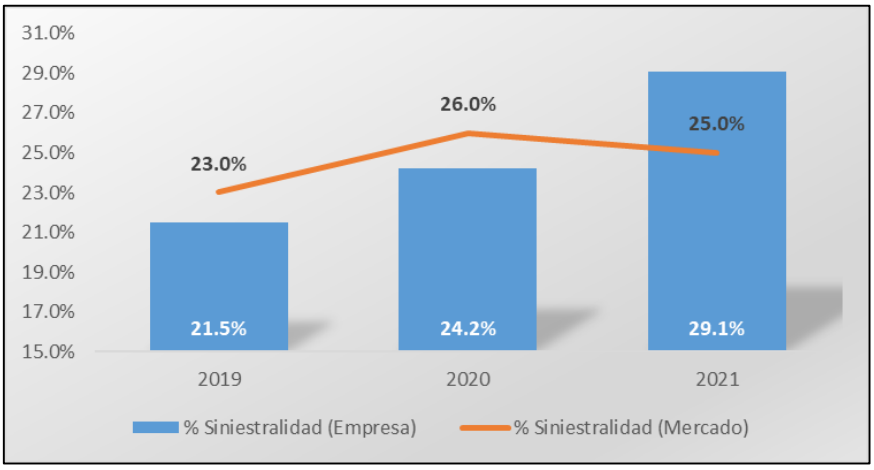
Fuente: (Galán, 2015)

a) Entendimiento del negocio

En primer lugar, se realizó un levantamiento de la situación del producto para identificar los principales puntos de dolor que conllevaron al requerimiento de crear un modelo analítico para

el seguro vehicular. Como se observa en la Figura 4, la tasa de siniestralidad vehicular en el año 2021 incrementó a 29.1%, superando en 4 puntos porcentuales el % de siniestralidad del mercado en ese mismo año el cual fue 25%. El impacto era muy alto en el periodo analizado e incluso si se compara con el año anterior (2020) el incremento superaba los 5 puntos porcentuales, de 24.2% (2020) a 29.1% (2021).

Figura 4: *Evolución de la tasa de siniestralidad anual*



Fuente: Elaboración propia

Con el objetivo de identificar la situación del producto, se menciona la tarificación inicial de las primas del seguro vehicular que la empresa estableció al momento de crear el producto, el cual se basó en 4 factores principales: marca, modelo, antigüedad y nivel de riesgo otorgado por APESEG. En la Figura 5, se observa la categorización de los vehículos para la tarificación de las primas, en base a la marca y modelo se categorizan los vehículos dependiendo al tipo de gamma al que pertenecen estos autos, ya sea por la suma asegurada o por el tipo de gama alta, media o baja; obteniendo 4 clases de auto: A, B, C y D. Por otro lado, según el nivel de riesgo por marca y modelo que se obtuvo en último reporte de siniestralidad de APESEG, se clasifican los vehículos como alto y bajo riesgo.

Como se mencionó anteriormente, el objetivo es desarrollar un modelo predictivo que permita predecir la ocurrencia de siniestro en vehículos particulares con menos a 10 años de antigüedad y así encontrar un óptimo proceso de tarificación de primas personalizadas con precios atractivos para los clientes.

Figura 5: *Categorías para la tarificación de seguros vehiculares*

CATEGORÍA	MODELO
 **Bajo Riesgo 1**	Chevrolet Aveo, Sail, Spark, Cruze, Sonic, Optra, Hyundai Elantra, Accent, Eon, i10, Fiat Punto, Kia Cerato, Picanto, Rio, Mitsubishi Lancer, Nissan Almera, Qashqai, Versa, Peugeot 206, 207, 307, Toda marca Renault, Toda marca Seat, Ssangyong Actyon, Toda marca Subaru (excepto Impreza), Suzuki Alto, Swift, Celerio, VW Gol, Golf, Polo, Bora, Voyage, Mazda 2
 **Alto Riesgo 2**	Toyota Yaris, Toyota Corolla, Nissan Sentra, Nissan Tiida, Subaru Impreza, Mazda 3
 **Alto Riesgo 1**	Toyota Land Cruiser, Land Cruiser Prado, FJ Cruiser, Rav4, Nissan Patrol, Pathfinder, Suzuki Grand Nomade, Honda CRV, Hyundai Santa Fe, Tucson, Kia Sportage, Camionetas Rurales / SUV mayores a US$50,000 y Autos y Station Wagon mayores a US$50,000
 **Bajo Riesgo 2**	Otros modelos que no están incluidos en las tres categorías anteriores (incluye Van de uso particular) y que no sea vehículo Chino e Hindú

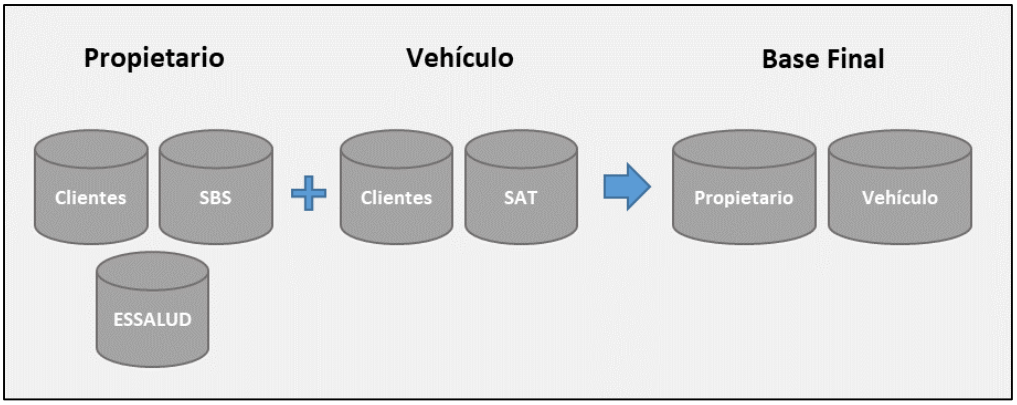
Fuente: Elaboración propia

b) Comprensión de datos

Los datos utilizados para el proyecto provienen de clientes con seguro vehicular que tuvieron una póliza vigente entre el periodo de enero del 2018 a diciembre del 2021, en la compañía aseguradora, obteniendo finalmente 8,900 clientes. Y el periodo de medición para cada cliente fue en el momento en que realizó la compra del seguro.

Por un lado, en la base de clientes se contaba con información netamente del vehículo como: marca, modelo, año de fabricación, tipo de vehículo, color y el valor comercial. Sin embargo, por el lado del propietario, inicialmente sólo se contaba con datos principales, como su edad, género y estado civil. Por ello se buscó fuentes que ayuden a enriquecer el análisis para la construcción del modelo, como es la fuente proveniente de la (SBS, 2021), que contiene información financiera: líneas de crédito, cantidad de bancos, deudas crediticias; también la fuente de (ESSALUD, 2021) que proporciona información salarial relativa del cliente; y finalmente la fuente de (SAT, 2021), que proporciona información de infracciones y montos papeletas. En el Figura 6, se muestra la identificación de fuentes que se utilizaron para la construcción del modelo:

Figura 6: *Identificación de fuentes de datos*



Fuente: Elaboración propia

Para la construcción del modelo, se definió como variable dependiente el campo "Siniestrado" como una variable dicotómica con las siguientes dos clases:

- **Siniestrado (Si):** Haciendo referencia aquel cliente que generó un siniestro durante su vigencia con su seguro vehicular. Se tipifica con el valor de "Si".
- **Siniestrado (No):** Haciendo referencia aquel cliente que no generó un siniestro durante su vigencia con su seguro vehicular. Se tipifica con el valor de "No".

En la Tabla 1, se muestra la distribución de la variable dependiente:

Tabla 1: *Tabla de distribución de la variable dependiente*

Siniestrado	Número de vehículos	% Distribución
Si	2,403	27%
No	6,497	73%
Total	**8,900**	**100%**

Fuente: Elaboración propia

Interpretación: Se cuenta con 8,900 vehículos que tuvieron un seguro vehicular entre el periodo del año 2019 al año 2021. El 27% de clientes generó un siniestro durante su permanencia en el producto con un total de 2,403 vehículos, mientras que el otro 73% de clientes no generó ningún siniestro con un total de 6,497 vehículos.

Además, se definió las variables independientes que caracterizan al vehículo, como marca, modelo, antigüedad, entre otros; y variables que caracterizan al propietario, como

demográficas, financieras, y comportamiento de manejo. A continuación, en la Tabla 2, se muestra la lista de campos que participarán inicialmente en la construcción del modelo:

Tabla 2: *Lista de variables independientes*

Variables Independientes	Descripción
Marca	Nombre de la marca
Modelo	Nombre del modelo
ValComercial	Valor comercial del vehículo
AnioFab	Año de fabricación del vehículo
TipoVehiculo	Nombre del tipo del vehículo
Color	Color del vehículo
NroAsientos	Número de asientos del vehículo
Edad	Edad del cliente
Genero	Género del cliente
EstCivil	Estado de civil del cliente
IngrSalarial	Monto del ingreso salarial del cliente
LinCred	Línea de crédito del cliente
DeudaCred	Deuda crediticia total del cliente
NivCred	Nivel de puntaje crediticio
ZonReside	Zona de residencia del cliente
NivDelin	Nivel de delincuencia en la zona de residencia
NivPob	Nivel de pobreza en la zona de residencia
CantPapeletas	Cantidad de papeletas acumuladas en el vehículo
MontPapeletas	Monto acumulado de papeletas
CatInfractor	Categoría según tipo de infracciones cometidas

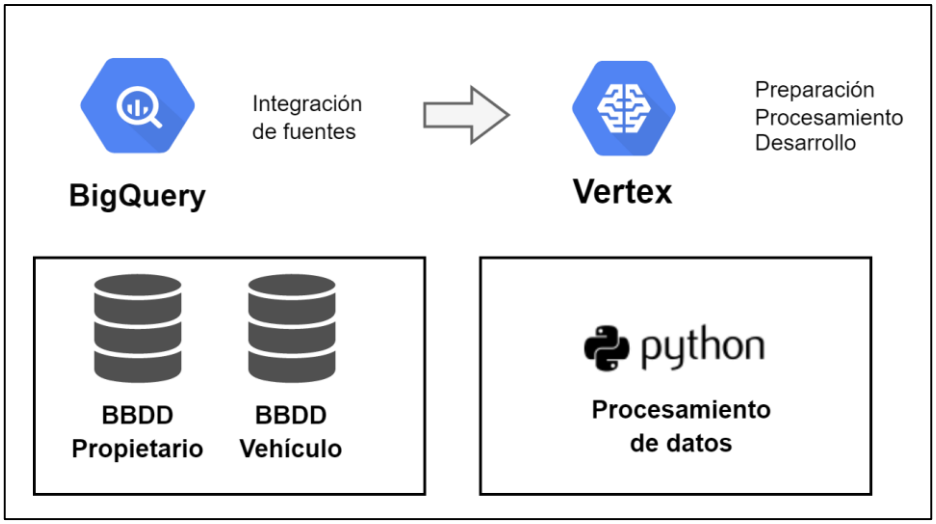
Fuente: Elaboración propia

c) Preparación de datos

Las fuentes mencionadas anteriormente se integraron y almacenaron en la plataforma de Big Query en el ambiente nube de (Google Cloud, 2019). Para la preparación y tratamiento de los datos, se utilizó el software de programación (Python Software Foundation, 2001) que es parte de la plataforma Vertex (Google Cloud, 2019).

A continuación, se muestra en la Figura 7 el proceso de integración de fuentes y el entorno de la preparación y tratamientos de los datos:

Figura 7*: Proceso de integración y procesamiento de datos*



Fuente: Elaboración propia

- Se realizó un análisis exploratorio previo, para identificar la existencia de valores nulos, outliers, teniendo como resultado el 100% de los datos completos en todas las variables y no encontrando presencia de outliers en las variables cuantitativas.
- Para las variables cualitativas, se procedió a validar si los niveles en cada variable tenían una distribución aceptable para su procesamiento, el cual se cumplía en todos los casos.
- Se procedió a realizar un proceso de selección de variables con un análisis de regresión para identificar qué variables cuantitativas son las más relevantes, teniendo como las de mayor importancia: valor comercial, año de fabricación, edad, ingreso salarial, línea de crédito, deuda de crédito, cantidad de papeletas.
- Para el caso de las variables categóricas, se procedió a realizar un análisis de correlación con la variable dependiente para identificar los de mayor asociación. Teniendo como resultado que las variables con una relación relevantes fueron: tipo de vehículo, nivel de puntaje crediticio, zona de residencia, nivel de delincuencia y categoría de infractor.
- Finalmente se trabajó con 12 variables independientes que participaron en la construcción de los modelos aplicados.

d) Modelamiento

Para el modelamiento se procedió a dividir los datos en dos partes; el 70% fue destinado para el entrenamiento, mientras que el otro 30% fue destinado para el testeo y evaluación del

rendimiento del modelo. La distribución en cada conjunto de datos fue seleccionada de manera aleatoria, respetando la misma proporción de la variable dependiente en cada conjunto.

En la Tabla 3, se muestra la distribución de los datos para los conjuntos de entrenamiento y prueba:

Tabla 3: *Distribución de datos de los conjuntos de entrenamiento y prueba*

Conjunto	Total	Siniestrado	Cantidad Vehículos	% Distribución
Entrenamiento (70%)	6,230	Si	1,637	26%
		No	4,593	74%
Prueba (30%)	2,670	Si	701	26%
		No	1,969	74%

Fuente: Elaboración propia

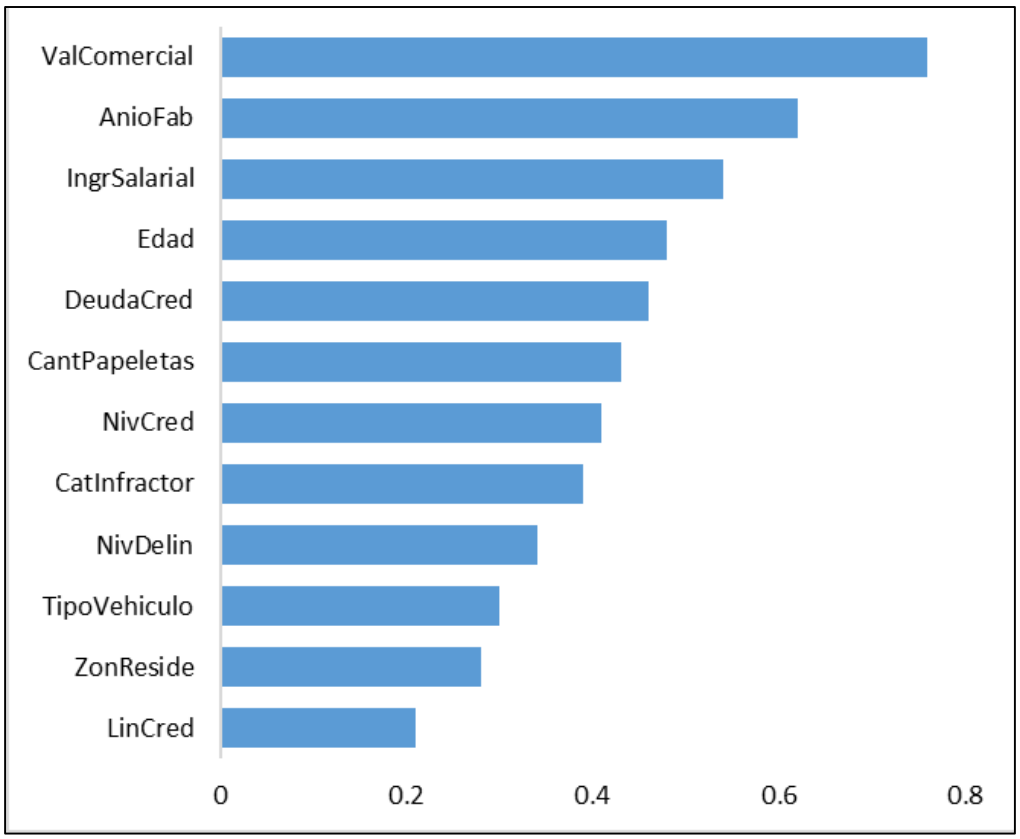
Interpretación: El 70% de clientes fue destinado para el entrenamiento de los modelos con 6,230 vehículos, y manteniendo la proporción de siniestrados con 1,637 (26%) vehículos que generaron un siniestro y 4,593 (74%) vehículos que no generaron siniestro durante su permanencia en el producto de seguro vehicular. El otro 30% de clientes fue destinado para la prueba de los modelos contando con 2,670 vehículos, de los cuales 701 (26%) vehículos generaron un siniestro y 1,969 (74%) vehículos no generaron siniestro.

e) Evaluación de los modelos

Luego del procesamiento de datos y aplicación de los modelos Random Forest y Logítico, se procedió a evaluar e interpretar los resultados que otorga cada modelo.

En la Figura 8, se muestra la importancia de variables que brinda el modelo Random Forest, haciendo referencia al aporte significativo que tienen las variables independientes en la variable dependiente.

Figura 8: *Importancia de variables del modelo Random Forest*



Fuente: Elaboración propia

Interpretación: La variable "valor comercial" hace referencia al valor que tiene el vehículo en el momento de la cotización, y es la característica que mayor influencia tiene en la decisión para predecir la ocurrencia de siniestro; también le sigue el "año de fabricación" del vehículo y por parte del propietario, el "ingreso salarial" y la "edad" son una de las características de mayor relevancia. Mientras que la "línea de crédito" habilitada en la SBS y la zona donde reside el propietario son las variables que generaron menor valor en la importancia, sin embargo, siguen siendo relevantes para la decisión en predecir la ocurrencia de un siniestro en no clientes.

Por parte del modelo Logístico, se generó la tabla de coeficientes y se evaluó la relación significativa de las variables independientes. En la Tabla 4, se muestra los valores de los coeficientes de cada variable independiente que aportan en la decisión de la ocurrencia de siniestro.

Tabla 4: *Coeficientes del modelo Logístico*

Variables	Estimate	P-Value	Sig
LinCred	-0.705	0.1600	
ZonReside_NorteSur	0.530	0.0900	.
TipoVehiculo_Camioneta	0.706	0.0620	.
NivDelin	0.768	0.0436	*
CatInfractor	0.024	0.0521	.
NivCred	-0.637	0.0411	*
CantPapeletas	0.621	0.0165	*
DeudaCred	0.557	0.0261	*
Edad	-0.444	0.0383	*
IngrSalarial	0.348	0.0089	**
AnioFab	-0.870	0.0017	**
ValComercial	-0.647	0.0001	***

Fuente: Elaboración propia

Interpretación: Se obtuvo como resultado que el "valor comercial" del vehículo tiene la mayor significancia para el modelo y tiene una relación negativa con respecto a la ocurrencia de siniestro. Esto quiere decir que, a mayor valor del vehículo al momento de la cotización menor será la probabilidad de que genere un siniestro. Otra variable con alta significancia para el modelo es el "año de fabricación" del vehículo, y presenta una relación positiva, por lo que mientras más antigüedad tenga el vehículo será más propenso a que genere un siniestro. Con respecto a características del propietario, la "edad" e información financiera como "ingreso salarial", "nivel crediticio" y "deuda en créditos" también presentan un alto nivel de significancia con respecto a la siniestralidad. Según el modelo logístico, la "línea de crédito" no es una variable significante, sin embargo, al presentar una relación indirecta se puede utilizar esta información en posteriores análisis, ya que indica que aquella persona que tenga una línea de crédito disponible y el valor sea elevada tendría mayor probabilidad de generar un siniestro en su vehículo.

Para la toma de decisión y elección del modelo, se procedió a evaluar las métricas de sensibilidad y especificidad de los modelos Random Forest y Logístico, los cuales son generadas por la matriz de confusión, reflejadas en las Tablas 5 y 6.

Tabla 5: *Matriz de confusión con el modelo Random Forest*

Siniestrado (Valor Predicho)	**Siniestrado (Valor Real)**		**Total**
	Si	**No**	
Si	499	280	779
No	202	1,689	1,891
Total	701	1,969	2,670

Fuente: Elaboración propia

Interpretación: En la matriz de confusión del modelo Random Forest se puede observar que llegó a predecir correctamente a los "siniestrados" con un 71% (499) sobre el valor real de 701 vehículos siniestrados. Mientras que, para los "no siniestrados", el modelo logró predecir correctamente el 86% (1,689) sobre el valor real de 1,969 vehículos no siniestrados.

Tabla 6: *Matriz de confusión con el modelo Logístico*

Siniestrado (Valor Predicho)	**Siniestrado (Valor Real)**		**Total**
	Si	**No**	
Si	450	300	750
No	251	1,669	1,920
Total	701	1,969	2,670

Fuente: Elaboración propia

Interpretación: En la matriz de confusión del modelo Logístico se puede observar que llegó a predecir correctamente a los "siniestrados" con un 64% (450) sobre el valor real de 701 vehículos siniestrados. Mientras que, para los "no siniestrados", el modelo logró predecir correctamente el 85% (1,669) sobre el valor real de 1,969 vehículos no siniestrados.

Tabla 7: *Sensibilidad y Especificidad de los modelos Random Forest y Logístico*

Métricas	Random Forest	Logístico
Sensibilidad	71%	64%
Especificidad	86%	85%

Fuente: Elaboración propia

Interpretación: Como se observa en los resultados el modelo Random Forest obtuvo una sensibilidad de 71%, lo cual indica que predice los aciertos en los "siniestrados" de manera óptima, mientras la sensibilidad obtenida en el modelo Logístico fue de 64%. Para el caso de la especificidad, Random Forest obtuve un 86% de aciertos al predecir a los "no siniestrados", mientras que el modelo Logístico obtuvo una especificidad de 85%.

V. RESULTADOS Y DISCUSIÓN

Para el trabajo de suficiencia profesional, se aplicaron dos modelos estadísticos: Random Forest y el modelo Logístico, con la finalidad de encontrar el modelo con mejores resultados de predicción en la ocurrencia de un siniestro.

Si bien existe una mayor importancia en predecir a los siniestrados, pues se puede identificar qué vehículos son más probables a que generan un siniestro en el futuro y realizar diferentes acciones de prevención al momento de ofrecer el producto, también es de suma relevancia el identificar qué vehículos son menos probables en generar un siniestro, ya que se pueden ofrecer precios más atractivos en el producto para este sector.

Finalmente se decide escoger el modelo Random Forest por obtener mejores resultados tanto en la sensibilidad como en la especificidad con 71% y 86% respectivamente. Luego de la elección y para tener un mejor manejo de esta información en la compañía, se decidió agrupar los scores obtenidos por el modelo en intervalos, el cual fue llamado "Grupo de Siniestralidad". En la Tabla 8, se muestra el comportamiento de siniestrados que obtuvo cada "Grupo de Siniestralidad" dentro del conjunto de prueba.

Tabla 8: *Comportamiento de siniestralidad en el conjunto de prueba*

Grupo de Siniestralidad	Cantidad Vehículos	Distribución de vehículos %	Cantidad Siniestrados	Siniestrados %
[00; 10]	15	0.6%	0	0%
(10; 20]	78	2.9%	2	3%
(20; 30]	335	12.5%	23	7%
(30; 40]	759	28.4%	69	9%
(40; 50]	704	26.4%	108	15%
(50; 60]	415	15.5%	223	54%
(60; 70]	217	8.1%	154	71%
(70; 80]	78	2.9%	61	78%
(80; 90]	38	1.4%	32	84%
(90; 100]	31	1.2%	29	95%
Total	**2,670**	**100.0%**	**701**	**26%**

Fuente: Elaboración propia

Interpretación: En la Tabla 8, se observa los scores obtenidos por el modelo Random Forest en el conjunto de prueba, agrupados por rangos. El grupo de siniestralidad de [00; 10], representa el 0.6% (15) de vehículos y tiene el menor porcentaje de siniestrado ya que no presenta algún vehículo con siniestro 0%. Mientras que el grupo de siniestralidad de (90; 100], representa el 1.2% (31) de vehículos y tiene el mayor porcentaje de siniestrados con el 95% (29) de vehículos que generaron un siniestro. Adicionalmente podemos observar que el 29% (779) de vehículos están ubicados en los grupos de siniestralidad de (50; 100] los cuales concentran el 71% (499) de vehículos siniestrados, por lo que da visibilidad a que grupos no se debe recomendar ofrecer un seguro vehicular por tener una alta propensión a generar un siniestro.

Con esta información, se procedió a generar las predicciones al parque automotor, que comprende un conjunto de vehículos y propietarios no clientes, con oportunidad de ofrecer el producto de seguro vehicular. En la Tabla 9, se muestra la distribución del parque automotor por grupo de siniestralidad obtenidos por el modelo.

Tabla 9: *Distribución de Scores de Siniestralidad en el Parque Automotor*

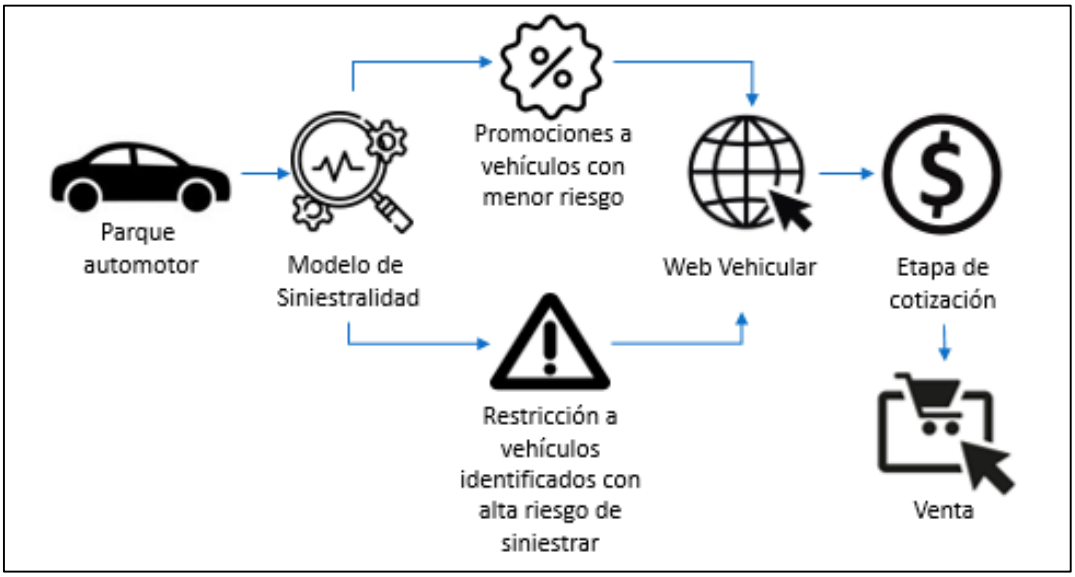
Grupo de Siniestralidad	Distribución de vehículos%	Cantidad Vehículos
[00; 10]	0.6%	17,261
(10; 20]	2.9%	88,032
(20; 30]	12.5%	376,295
(30; 40]	28.4%	852,704
(40; 50]	26.4%	790,564
(50; 60]	15.5%	466,053
(60; 70]	8.1%	243,383
(70; 80]	2.9%	88,032
(80; 90]	1.4%	43,153
(90; 100]	1.2%	34,522
Total	**100.0%**	**3,000,000**

Fuente: Elaboración propia

Interpretación: En la tabla 9, se observa las predicciones para nuevos vehículos que participarán del proceso de ofrecimiento de un seguro vehicular, los cuales están comprendidos en 3 millones de vehículos particulares. Tomando en cuenta el comportamiento de siniestralidad del conjunto de prueba, se decidió ofrecer promociones a los vehículos que se encuentran en los grupos de siniestralidad de [0; 50] que representan el 71% (2,124,856) del parque automotor, ya que al tener una menor probabilidad de que ocurra un siniestro se pueden ofrecer descuentos y promociones por ser más rentables para la compañía. Por otro lado, para los grupos de siniestralidad de (50; 70] que representan el 24% (709,436) de vehículos del parque automotor, se les incrementará la tarifa del seguro vehicular por tener una probabilidad alta de generar un siniestro. Finalmente, para los grupos de siniestralidad de (70; 100], que representa el 6% (165,708) de vehículos, se restringirá la venta ya que cuentan con una muy alta probabilidad de generar un siniestro.

Una vez que se obtuvo la asignación de los grupos de siniestralidad, se procedió a ingresar el nuevo factor de riesgo de siniestralidad en la etapa de cotización del seguro vehicular para generar tarifas y promociones personalizadas acorde al riesgo e identificar a los vehículos con mayor probabilidad de generar un siniestro como se mencionó anteriormente. Este proceso de implementación consiste en identificar al prospecto de cliente que ingresa su placa en la web vehicular y pasa por la etapa de cotización, puede ver reflejado la tarifa personalizada acorde al nivel de riesgo de siniestralidad identificada por el modelo, como se observa en la Figura 9.

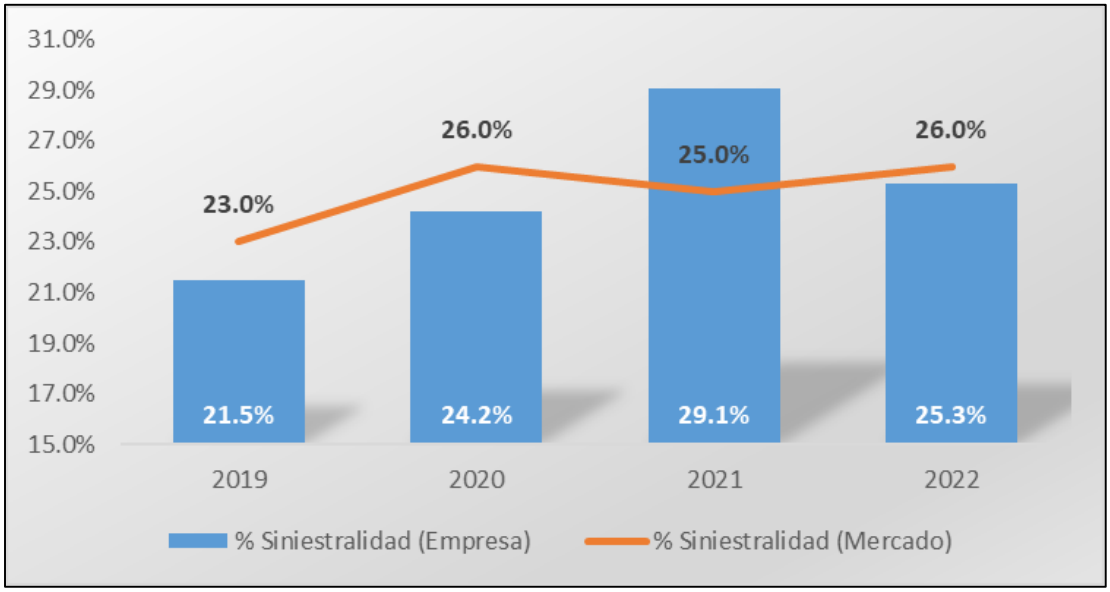
Figura 9: *Proceso de implementación del modelo de siniestralidad*



Fuente: Elaboración propia

Luego de la implementación del modelo y las acciones estratégicas, se tuvo que esperar alrededor de 6 meses de aplicado el modelo para realizar el seguimiento de los indicadores de siniestralidad, esto se debe a que la ocurrencia de siniestro no es un evento que se presente de manera inmediata. El periodo definido para el seguimiento fueron los meses de enero a junio del año 2022, y para uniformizar el seguimiento anual se procedió a proyectar la siniestralidad del segundo semestre del año 2022.

Figura 10: *Evolución de la tasa de siniestralidad después de la implementación*



Fuente: Elaboración propia

Interpretación: En la Figura 10, se observa la evolución de la siniestralidad anual. El año 2022 tiene un 25.3% de siniestralidad el cual es 4 puntos porcentuales menos que el año 2021, y 1 punto porcentual por debajo del % de siniestralidad del mercado.

Se confirma que el uso del modelo analítico para predecir la ocurrencia de siniestros en vehículos particulares, el cual fue generado por el autor de este trabajo de suficiencia profesional, obtuvo resultados óptimos, logrando disminuir la tasa de siniestralidad anual para el año 2022 y manteniendo el % por debajo del mercado.

VI. CONCLUSIONES

6.1 Conclusiones

- Al ser comparados los modelos Random Forest y Logístico, se obtuvieron mejores resultados para el modelo Random Forest, tanto en el indicador de sensibilidad con 71% como en especificad con 86%. Por lo que fue el modelo elegido para predecir la ocurrencia de siniestro en vehículos particulares dentro de la compañía.

- Las variables utilizadas con respecto al propietario, como la información financiera, los niveles de deuda crediticia, ingreso salarial y la información del vehículo como el valor comercial, año de fabricación; sirvieron para aportar en la precisión de las predicciones del modelo.

- El modelo Random Forest permitió identificar de manera óptima a los vehículos con mayor propensión a generar un siniestro, por lo que ayudó a la compañía a tomar decisiones con mayor precisión.

- Como consecuencia de la aplicación del modelo Random Forest en la predicción de la siniestralidad, se logró disminuir la tasa de siniestralidad esperado por la compañía. Obteniendo tasas menores con respecto a los años anteriores y logrando mantener la brecha de siniestralidad por debajo del mercado asegurador.

- La utilización de los scores del modelo de ocurrencia de siniestralidad en el parque vehicular facilitó a la compañía a tener un mejor manejo de la información para desarrollar un proceso de tarificación óptimo, generando acciones comerciales que permitieron personalizar las tarifas con descuentos y promociones más atractivos para los clientes y no clientes.

VII. RECOMENDACIONES

7.1 Recomendaciones

- A futuro se pueden considerar otras variables que mejoren el desempeño del modelo, como información de mayor concurrencia por el conductor, por ejemplo, supermercados, clínicas/hospitales, colegios, restaurantes, etc. Estos datos ayudarían a identificar el trayecto del conductor y en qué lugares es más probable que ocurra un siniestro.

- Se pueden aplicar otros modelos acordes a este tipo de casos y comprobar si presentan mejores indicadores de precisión en la ocurrencia de siniestralidad con respecto al modelo Random Forest.

VIII. REFERENCIAS BIBLIOGRÁFICAS

Amat, J. (Octubre de 2020). *Árboles de decisión, random forest, gradient boosting y C5.0*. Obtenido de https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting

APESEG. (2019). *Asociación Peruana de Empresas de Seguros*. Obtenido de https://www.apeseg.org.pe/

Barrios, J. (26 de Julio de 2019). *Health Big Data*. Obtenido de https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Cichosz, P. (2015). *Data mining algorithms: explained using R.* Wiley.

Condori, N. (2020). *Determinantes de la siniestralidad de autos en una empresa aseguradora peruana: Un enfoque bayesiano.* [Tesis de grado], Lima.

ESSALUD. (2021). *Seguro Social de Salud del Perú*. Obtenido de http://www.essalud.gob.pe/

Galán, V. (2015). *Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario.* Madrid.

Google Cloud. (2019). *Optimiza el almacenamiento en BigQuery*. Obtenido de https://cloud.google.com/bigquery/docs/best-practices-storage

Guillen, M., & Pesantez, J. (2018). Machine learning y modelización predictiva para la tarificación en el seguro de automóviles. *Anales del Instituto de Actuarios Españoles*, 123-147.

Guillermo, M. (2015). *Metodología de minería de datos para el estudio de tablas de siniestralidad vial.* [Tesis de Maestría], Madrid.

Hernández, F. (18 de Marzo de 2021). *Modelos Predictivos.* Obtenido de https://fhernanb.github.io/libro_mod_pred/index.html

Herrera, J. (2021). *Análisis y predicción de la lesividad en accidentes de tránsito mediante la aplicación de Random Forest.* [Tesis de grado], Madrid.

IPython Project. (2014). Jupyter Lab. EE.UU. Obtenido de https://jupyter.org/

JavaTpoint. (2021). *Java T Point*. Obtenido de https://www.javatpoint.com/machine-learning-random-forest-algorithm

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Connecticut: Springer.

Python Software Foundation. (2001). Python. Obtenido de https://www.python.org/

Ruiz, M. (2018). *Análisis de sensibilidad mediante Random Forest.* [Tesis de grado], Madrid.

Salazar, D. (2004). *Valoración actuarial de primas de seguros de vehículos en la provincia de Guayas.* [Tesis de Grado], Guayaquil.

Sanchez, J. (2021). *Optimización matemática a partir de algoritmos híbridos. Una aplicación en la tarificación del seguro de automóviles.* [Tesis de Maestría], Madrid. Obtenido de https://e-archivo.uc3m.es/bitstream/handle/10016/33147/TFM_CCAAFF_Juan_Sanchez_Campillo_2021.pdf?sequence=3&isAllowed=y

SAT. (2021). *Servicio de Administración Tributaria de Lima*. Obtenido de https://www.sat.gob.pe/

SBS. (2021). *Superintendencia de Banca, Seguros y AFP*. Obtenido de https://www.sbs.gob.pe/usuarios/seguros/otros-seguros/seguro-vehicular

SUNARP. (2015). *Superintendencia Nacional de los Registros Públicos*. Obtenido de https://enlinea.sunarp.gob.pe/