UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA INDUSTRIAL Y COMERCIAL
INGENIERÍA DE TECNOLOGÍAS DE INFORMACIÓN Y SISTEMAS

Técnicas de Machine Learning para la clasificación automática de clientes en una empresa de seguros

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los requerimientos para:

Obtener el título profesional de Ingeniero Industrial y Comercial

Obtener el título profesional de Ingeniero de Tecnologías de Información y Sistemas

AUTORES

Luz de los Angeles Manuela Asencio Diaz

Ricardo Hernan Chiang Cornejo

Fernanda Lucía Crisóstomo Fernández

Gisela Vanesa Hernández Quiroz

Almendra Sofia Lajo Aurazo

ASESOR

Junior John Fabian Arteaga

ORCID Nº 0000-0001-9804-7795

Diciembre, 2021

Índice de Contenidos

Tabla de Figuras

Tablas

RESUMEN

Machine Learning y los modelos matemáticos en los que se basa para poder identificar patrones y dar una estimación basada en data histórica son usados cada vez más en diferentes industrias para procesar información que antes se consideraba masiva y por ende difícil de relacionar de manera certera por métodos tradicionales. Con la inclusión de las técnicas de como regresión logística y K-NN, hoy en día es posible formular y proponer un modelo de predicción de aprendizaje supervisado que se ajuste a los requerimientos de clasificación de una empresa. Esta investigación propone la aplicación de las mencionadas técnicas para la elaboración de modelos predictivos de clasificación de tipos de asegurados para una determinada empresa en la industria aseguradora de vehículos automóviles; usando como basede datos los registros históricos recopilados del año 2019.

Palabras clave: machine learning, aprendizaje supervisado, clasificación, seguros vehiculares

ABSTRACT

Machine Learning and the mathematical models on which it is based to identify patterns and give an estimate based on historical data are increasingly used in different industries to process information that was previously considered massive and therefore difficult to relate accurately by traditional methods. With the inclusion of techniques such as logistic regression and K-NN, it is now possible to formulate and propose a supervised learning prediction model that fits the classification requirements of a company. This research proposes the application of the aforementioned techniques for the development of predictive models for the classification of policyholder types for a given company in the motor vehicle insurance industry; using the collected historical records of the year 2019 as a database.

Keywords: machine learning, supervised learning, classification, vehicle insurance

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

Dentro de este capítulo se presenta la descripción detallada de la problemática y un análisis de las principales causas y efectos que se derivan de dicho problema. Asimismo, se realiza una descripción de la tecnología a utilizar dentro de la propuesta de solución.

1.1 Descripción de la Realidad Problemática

En la actualidad, estamos en un entorno empresarial complejo y dinámico; en dónde la tecnología y la innovación tienen un papel importante para el desarrollo de las empresas permitiendo mejorar su desempeño.

¿Qué es lo que una empresa requiere para tener éxito en este entorno tan volátil? necesita poder reconocer los patrones en las necesidades de los clientes, averiguar si su producto y/o servicio tendrá un nicho de mercado, evaluar la creación de nuevos productos, entre otros; para poder verificar esto las empresas recurren al análisis de datos.

Machine Learning es un instrumento innovador que ayuda a los negocios a poder comprender los patrones que hay a partir de los datos recolectados, es así como esta herramienta está siendo utilizada en la mayoría de las empresas que buscan predecir; por ejemplo, la demanda de los clientes. Dentro del Machine Learning existen dos tipos de aprendizajes, el supervisado y el no supervisado, estos tienen la diferencia en que mientras que el supervisado se basa en salidas definitivas a partir de los datos suministrados, el aprendizaje no supervisado no limita las salidas y su objetivo principal es descubrir patrones, relaciones o tendencias presentes en los datos. (Ana González-Marcos, 2017)

Dentro de las empresas que realizan estos estudios en Perú, se encuentra el sector de seguros; considerando que las empresas que venden seguros se basan en datos para ofrecerte sus productos, el análisis que realizan es en base a ciertas características tales como edad, nivel socioeconómico, cantidad de vehículos que posee, cantidad accidentes previos, historial familiar de salud, entre otros; consideramos que es esencial para los seguros que mientras más información tienen de los usuarios, mejores serán los servicios que pueden ofrecerles.

Debido a la coyuntura actual la mayoría de las empresas tuvieron un grave impacto en su crecimiento; esto no ha sido ajeno para el sector seguros en el Perú. Si bien en los últimos 10 años ha tenido un crecimiento de 10,5% según el Informe del Mercado Asegurador Latinoamericano presentado por Manuel Aguilera, Director General MAPFRE Economics en

el marco del Insurance Day 2020, el año 2020 se tuvo una caída de 0.7%; la mejor empresa que supo reaccionar a esta crisis fue MAPFRE debido a su cartera multiproductos; si bien en un tipo de seguros vida se vieron afectados; sin embargo, cuentan con seguros para obras tanto de construcción como en minería y estas obras requieren seguros lo que les permitió seguir creciendo. (Lengua, 2021)

El sector de seguros en el Perú es bastante competitivo, a pesar de que pocas empresas tienen la mayoría de participación de mercado. Usualmente, los clientes optan por el de menor costo y que les brinde más beneficios. Sin embargo, la mayoría de las veces se les ofrece a los clientes paquetes predeterminados y que están parcialmente adecuados a las necesidades de los clientes tomando en cuenta factores como edad, condición de empleado, entre otros.

1.2 Justificación de la Investigación

1.2.1 Teórica

La clasificación de información de índole empresarial, de distintos sectores y áreas, a través del uso de diferentes herramientas de Machine Learning ha sido estudiada en muchas investigaciones a lo largo de los años. Este trabajo busca aportar un nuevo enfoque desarrollando la clasificación de clientes de seguros vehiculares a través del análisis combinado de información del usuario e información del vehículo. Con este propósito se utilizarán herramientas de aprendizaje supervisado (clasificación) de Machine Learning para el procesamiento digital de la información.

1.2.2 Práctica

De acuerdo con Esteban (2020), la personalización en la oferta de servicios es la clave para el éxito de las empresas ya que permite fortalecer la relación entre la empresa y el cliente. Uno de los pasos claves al momento de desarrollar una estrategia de personalización es la identificación y segmentación de los clientes. En este sentido, esta investigación se realiza con el fin de aportar a la empresa estudiada una herramienta automatizada que permita ofrecer a sus clientes un producto más personalizado que se ajuste a su situación y necesidades.

Asimismo, el sector de seguros es altamente dinámico y competitivo, enfatizado el último año por la pandemia de COVID-19, lo que ha generado que las empresas desarrollen proyectos relacionados con la digitalización / robotización con el propósito de generar mayores

beneficios (Willis Towers Watson, 2020). En este sentido la herramienta propuesta permitirá generar mayores eficiencias dentro del proceso de suscripción, permitiendo reducir gastos.

1.2.3 Metodológica

El presente trabajo propone la implementación de un modelo basado en técnicas de Machine Learning para la clasificación de clientes de seguros vehiculares a través de un análisis cuantitativo y científico debido a que está basado en modelos matemáticos de clasificación. Con el propósito de desarrollar este proyecto, se recopilará la base de datos necesaria, luego sehará una limpieza y pre-procesamiento de la información para evitar distorsiones en el resultado,y finalmente se desarrollará el modelo de clasificación utilizando técnicas de Machine Learning.

1.3 Delimitación de la Investigación

1.3.1 Espacial

La investigación se realizará en el Perú y se enfocará en el sector de seguros vehiculares. Dentro de la investigación se utilizará una base de datos proporcionados por una empresa aseguradora (no se mencionará la razón social por temas de confidencialidad), por lo que la presente investigación puede que no sea aplicable en otros sectores y/o países.

1.3.2 Temporal

La presente investigación utilizará información recabada durante el año 2019 para entrenar al algoritmo predictivo. La información fue recolectada por una empresa aseguradora peruana en el tiempo antes mencionados, a través de sus clientes y servicios prestados.

1.3.3 Conceptual

La presente investigación se enfocará en desarrollar un modelo predictivo que pueda identificar la mejor póliza vehicular a ofrecer a un cliente a través del análisis de diferentes variables, tanto del usuario como del vehículo. Se utilizarán diferentes técnicas de Machine Learning para desarrollar el modelo propuesto.

CAPÍTULO II: MARCO TEÓRICO

Dentro de este capítulo se realizará el análisis de estudios anteriores en clasificación de clientes usando técnicas de Machine Learning. Asimismo, se expondrán los aspectos técnicos de la solución propuesta.

2.1 Antecedentes de la Investigación

De acuerdo con Paruchuri (2020), los datos son uno de los principales activos con los que cuenta el sector de seguros; especialmente ahora, cuando estamos viviendo la cuarta revolución industrial y en los últimos 24 meses se han generado más datos que en toda la historia de la humanidad. Este estallido en la generación de datos ha creado la necesidad de tecnologías que puedan procesar y administrar el gran volumen de datos de la industria. En este sentido, el cambio hacia estas nuevas tecnologías ha sido gradual pero firme ya que el entorno del sector exige que las empresas innoven. De acuerdo con el autor, el entorno del sector asegurador se caracteriza por una fuerte competencia, altos niveles de fraude, mercados flexibles, altas perspectivas de los clientes y regulaciones estrictas; todos estos factores obligan a las empresas a buscar nuevas alternativas que puedan crear ventajas competitivas para la empresa. De acuerdo con Vadlamudi (2016), las aseguradoras deben aprovechar los nuevos enfoques tecnológicos, como el Machine Learning, para resolver de manera automática tareas dentro de la cadena de suministro como el proceso de suscripción, la prevención de decomiso, la detección de fraudes o la evaluación de productos y clientes.

Good drivers pay less: A study of usage-based vehicle insurance models.

Bian, et al (2018) redactaron en su artículo que actualmente hay una creciente tendencia dentro de la industria de seguros vehiculares, al igual que de en la industria de transporte en general, a la aplicación del Seguro Basado en Uso (o UBI en sus siglas en inglés), el cual como indica el nombre, basa los precios listados en sus productos de acuerdo al uso que una entidad le dé a sus vehículos. Debido a esto, surgió el interés de poder determinar cómo los datos de comportamiento de los conductores afectan al riesgo de conducción y cómo el comportamiento de los conductores debería afectar a los esquemas de precios del seguro vehicular. En otras palabras, si el usuario se consideraría seguro o riesgoso, en qué grado, y por ende, a qué rango de precios se debería fijar su prima.

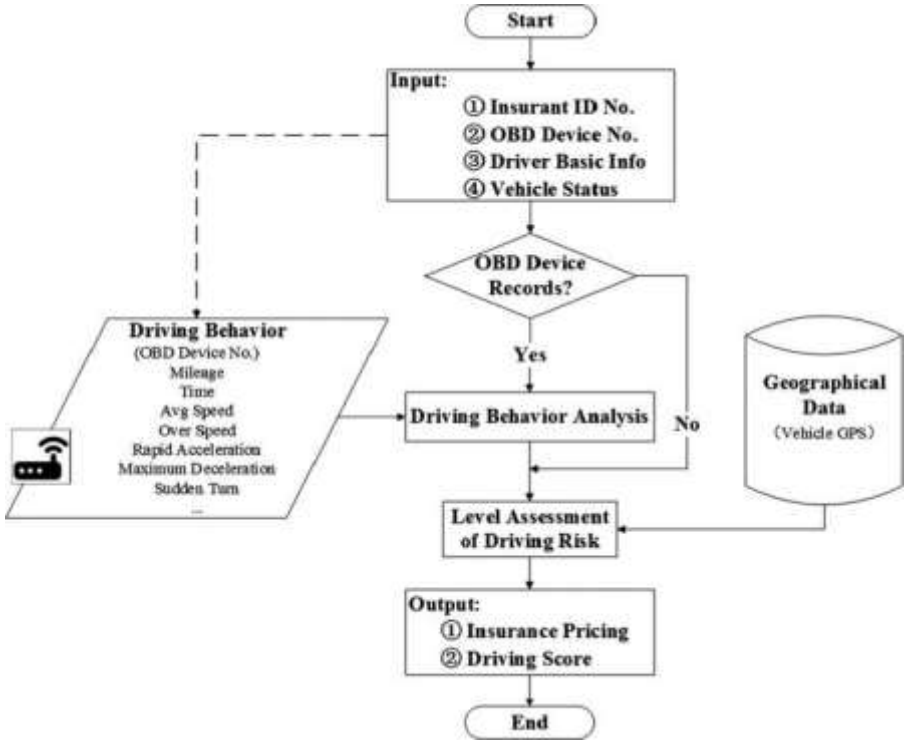
Dado que la investigación busca cómo utilizar los datos de comportamiento masivo para ofrecer asistencia para hacer una estrategia de precios de seguros basados en usos personalizada,

los autores buscan aplicar una combinación de factores usando como bases los modelos de precios y sus variables determinantes, para alcanzar el desarrollo de una estrategia de precios basada en el comportamiento. A esta técnica la definen Bagging o Agregación de BootStrap, la cual permite reducir la varianza de los resultados teniendo en consideración múltiples clasificadores de una misma entidad.

La data usada en el estudio debía reflejar tanto el comportamiento del conductor como su nivel de accidentes. Para el primero se obtuvo la información por parte de una compañía aliada de aseguradoras vehiculares en China la cual ofrece servicios de seguimiento y Diagnóstico a Bordo (OBD en sus siglas en inglés). De este set de datos se obtuvo información como: La descripción del kilometraje total (unidad: kilómetro) por mes (MIL), las horas de conducción nocturna por mes (NDH), las horas de conducción en día laborable por mes (WDH), la velocidad media mensual (AS), los tiempos de sobre velocidad (OST), la aceleración (RA), la desaceleración máxima (MD) y el giro brusco (ST). Para el segundo tipo de datos, la información fue obtenida de una compañía aseguradora aliada a la compañía OBD por un total de 198 usuarios en un registro de tiempo de 6 meses consecutivos de los cuales 125 reportaron accidentes durante ese periodo. Se realizó una referencia cruzada para combinar los datos de acuerdo con cada usuario. De este modo se propuso el modelo de procesamiento de datos mostrado en la Figura 1.

La investigación constó de 2 etapas importantes: demostrar la efectividad del método bagging para el estudio, a comparación de los modelos clásicos de clasificación; y la evaluación de éste método aplicado a un modelo de clasificación enfocado en el comportamiento del usuario. Para la primera etapa se comparó el modelo propuesto con otros modelos como regresiones logísticas, Naïve Bayes, Optimización Media Secuencias y Aprendizaje ponderado localmente. Estos modelos fueron evaluados en aspectos como Porcentaje de casos clasificados correctamente, estadística Kappa, error medio absoluto, error medio cuadrático. Finalmente, se realizó una evaluación T-test la cual determinó que sí hubo una mejora significativa empleando el modelo de Bagging en este estudio.

Figura 1. *Vista del proceso del modelo propuesto para el cálculo de precios*



Fuente: Bian, et al (2018)

Para la segunda etapa de evaluación, se comprobó la efectividad del prototipo del modelo presentado en la imagen anterior. Este proceso se realizó tras comparar el método bagging enfocado en la clasificación basada en comportamiento con otros modelos de clasificación. Para esto se realizó un cuestionario de 10 puntos específicos medido con la escala de Likerd dirigido a usuarios del sistema (ej: asegurados o vendedores de seguros) y a expertos del tema (ej: gerentes de compañías aseguradoras o profesores de universidad) a fin de recopilar sus opiniones acerca del modelo propuesto sobre efectividad y uso.

El estudio llega a la conclusión que el incremento de factores aumenta la facilidad de personalización de precios certeros para cada usuario lo que puede conllevar a una reducción de precios de primas dependiendo de los hábitos de buena conducción del usuario.

Estimating Car Insurance Premia: a Case Study in High-Dimensional Data Inference

Las empresas aseguradoras siempre tienen problemas al momento de estimar el costo de la póliza del contrato. Este costo está condicionado por información disponible del asegurado y del contrato. La dificultad de este cálculo radica en la gran cantidad de casos, gran cantidad

de variables (las cuales la mayoría son discretas y multivariables), la no estacionalidad de las distribuciones y la distribución condicional de la variable dependiente que es muy diferente a las que existen en casos más típicos de Machine Learning.

Se incluyeron en el estudio datos de cinco tipos de pérdidas (se estimó una sub prima por cada tipo de pérdida). Las variables de entrada contenían información de la póliza (deducibles, fechas y opciones), el carro, y el conductor (infracciones pasadas, cobros de seguro anteriores, etc). Para todos los modelos menos para el CHAID se utilizaron codificación one-hot. El número de variables de entrada aleatorias fue de 39, todas discretas menos una. Un dataset promedio contenía alrededor de 8 millones de muestras permutadas de manera aleatoria y divididas en sets de entrenamiento, validación y evaluación, con tamaños de 50%, 25% y 25% respectivamente. Por último, también se utilizó data de benchmarking para realizar la comparación de resultados.

Se utilizaron distintas técnicas de Machine Learning dentro de estas estuvieron:

- Redes neuronales con salidas positivas debido a que el resultado del monto debía salir positivo ya que no se puede cobrar dinero negativo.
- Árboles de decisión: usaron este modelo como comparación con el que realizaron ya que, en la industria de los seguros, esta técnica es bastante utilizada.
- Regresión por Máquinas de vectores de soporte (SVM): los resultados obtenidos no fueron adecuados ya que esta regresión optimiza un criterio cercano a la mediana condicional, mientras que el criterio MSE por la media condicional y como la distribución es altamente asimétrica la mediana condicional está muy lejos de la media condicional.

Para los resultados se comparó un modelo mezclado, que era la mezcla de expertos y en este caso los expertos eran redes neuronales con resultados positivos, con otros modelos como NN, que es una red neuronal con activaciones de salida lineares, Softplus NN con activaciones de salida softplus.

El análisis cualitativo de la prima predicha por el modelo muestra que el modelo mezclado tiene una prima más suave y extendida que el benchmark brindado. El análisis también indica que la diferencia entre la prima mezclada y la prima del benchmark está sesgada

negativamente, con una mediana positiva, resultando en que un cliente típico pagaría menos en la prima mezclada pero los clientes con mayor riesgo pagarían mucho más.

Machine Learning Approaches for Auto Insurance Big Data

Hanafy y Ming (2021) señalan en su artículo que actualmente hay una creciente tendencia al pago justo de los servicios que un cliente usa realmente en el sector de seguros no relacionados a seguros de vida, específicamente en los seguros vehiculares. Para esto se debe anticipar una prima apropiada para los nuevos consumidores; sin embargo, también se debe resaltar el aumento de reclamos falsos para el cobro de las pólizas. Por lo que las compañías de seguros se ven obligadas a evaluar apropiadamente el riesgo que cada cliente representa y establecer una prima adecuada sin perjudicar la oportunidad de realizar una venta debido a altos precios mal establecidos.

El objetivo principal del artículo es crear un algoritmo de ML que prediga con precisión la ocurrencia de siniestros. Para ello, el modelo debe tener en cuenta eficazmente los detalles del consumidor, como el tipo de vehículo o el coste del coche, que difieren entre los clientes. La información de la muestra obtenida fue proporcionada por una aseguradora ubicada en Brasil, consta de un total de 1’488,028 clientes con un total de 35 variables por cada línea de set de data.

Para la evaluación de resultados se repartió los sets de data en una razón de 80-20, siendo el primer grupo la data de entrenamiento y la segunda el test. Se eligieron múltiples modelos de machine Learning los cuales fueron expuestos como usados más frecuentemente en los estudios de antecedentes relacionados al sector de seguros. Estos modelos son: Bosque Aleatorio (RF), Árbol de decisión (C50), XGBoost, KNN y Naïve Bayes. Los modelos fueronevaluados en aspectos como precisión, ratio de error, Estadística Kappa, métrica AUC, sensibilidad, especificidad, exhaustividad y valor-F.

Los resultados de la evaluación determinaron, en primer lugar, que el reclamo para activar la póliza en la compañía de seguros podría predecirse mediante métodos de Machine Learning. En segundo lugar, se evidenció que el modelo que mejor se ajusta a la situación propuesta con múltiples variables es Bosque aleatorio, dado que, bajo los estándares evaluados, era el modelo que se encontraba en el rango aceptable y deseado. Estos resultados comprueban que si es posible e ideal usar una gran cantidad de datos que favorecen la toma de decisiones

respecto al precio de la prima a imponer, pues permite predecir si el cliente es propenso a activar la póliza y con cuánta frecuencia probable.

Computational Intelligence Approach for Estimation of Vehicle Insurance Risk Level

El alto grado de competencia en el mercado de seguros de vehículos obliga a las empresas del sector a mejorar la calidad de sus servicios. En este contexto, las empresas del sector se han enfocado en desarrollar tecnologías digitales, lo que permite la creación de nuevos servicios. Este nuevo enfoque ha permitido que las empresas cuenten con la información suficiente para desarrollar soluciones tecnológicas más avanzadas.

De acuerdo con Vassiljeva et al. (2017), los métodos de minería de datos con aplicaciones de redes neuronales se utilizan ampliamente con fines de modelado, detección y clasificación. Un ejemplo de ello son las redes auto organizadas de Kohonen, las cuales se utilizaron para clasificar los reclamos por lesiones corporales después de accidentes automovilísticos según el grado de sospecha de fraude. Asimismo, se desarrollaron modelos de detección híbridos que integran técnicas de inteligencia con diversos métodos supervisados y no supervisados para mejorar la precisión en la clasificación.

Según el artículo, una de las tareas más importantes dentro del sector es el análisis y predicción de riesgos. Sin embargo, los sistemas de cálculo convencionales se basan principalmente en los registros históricos de otros conductores con características similares. Es por ello que el objetivo de la investigación es encontrar una manera eficaz de evaluar el nivel de riesgo de los clientes de seguros vehiculares utilizando técnicas de Machine Learning. El modelo plantea el análisis de información de entrada relacionada a una persona y su vehículo para obtener un valor de salida en un rango de 0 a 1, lo cual representa la probabilidad de accidente.

El autor utilizó una base de datos con 540,490 pólizas de los períodos 2010 – 2015. Sin embargo, se identificó que la base de datos estaba desequilibrada ya que sólo 11,551 casos hicieron uso del seguro y 1,000 habían resultado en lesiones personales. De acuerdo con el autor, los algoritmos de aprendizaje automático suelen tener dificultades para aprender de problemas de clase desequilibrados, debido a que su objetivo es minimizar la tasa de error general; lo que podría tener como resultado que el modelo clasifique todos los ejemplos de prueba como negativos, lo que resulta en un rendimiento de clasificación deficiente. Para resolver este problema es necesario re-equilibrar la distribución de clases en el conjunto de datos de

entrenamiento. Dentro de la investigación se utilizó un muestreo híbrido en el cual se combinan dos técnicas: submuestreo (reducir la clase en mayoría) y sobremuestreo (incrementar la clase en minoría).

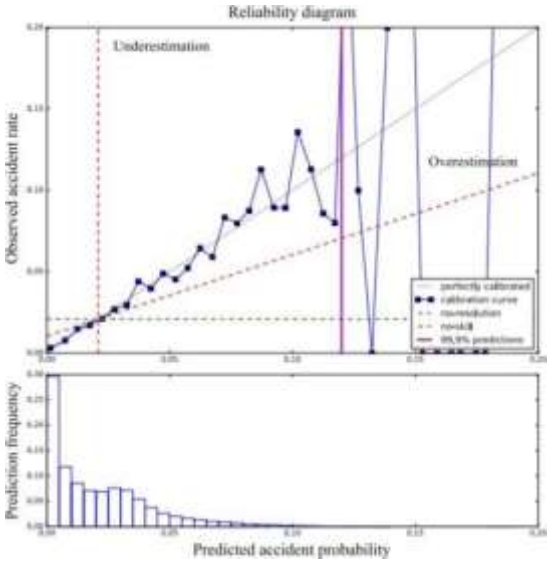
Luego de analizar la data proporcionada, se seleccionaron 17 variables para incluir en el modelo, en el caso de las variables categóricas se les asignó un valor numérico para que puedan ser procesadas.

En primer lugar, se clasificaron los clientes dentro de grupos de riesgo a través de una técnica de discretización, lo que significa que las variables continuas se dividen en intervalos. Sin embargo, se concluyó que este modelo no era aplicable en la vida real ya que los coeficientes de riesgo dados a cada cliente deben ser funciones continuas, no discretas. De lo contrario, los clientes cuyas características iniciales se distinguen ligeramente, según la clasificación se pueden configurar en diferentes grupos cerca del límite de dos clases.

Por lo anterior, se decidió utilizar un modelo de redes neuronales para determinar la probabilidad de riesgo de cada cliente. El conjunto de datos dado fue dividido en tres conjuntos: entrenamiento, validación y prueba. Durante el estado de validación, se seleccionó la muestra para el entrenamiento del modelo, la cual consiste en el 80% de los datos totales. El resto de los datos se divide proporcionalmente para validación y prueba. Cabe mencionar que antes de configurar los conjuntos, todos los datos se barajaron aleatoriamente.

Con el objetivo de encontrar un modelo de redes neuronales con un rendimiento de clasificación óptimo, se llevaron a cabo alrededor de 4,000 sesiones de entrenamiento diferentes. El rendimiento del modelo se verificó a través de un gráfico de confiabilidad. Como se puede observar en la siguiente imagen, si bien el modelo encontrado es bastante confiable para casos con probabilidad de riesgo menor al 12%, otros casos no pueden ser medidos con el modelo seleccionado. Esto se explica principalmente debido a que la muestra de casos con probabilidad mayor al 12% es significativamente menor.

Figura 2. Gráfico de confiabilidad



Fuente: Vassiljeva, et al (2017)

2.2 Bases Teóricas

2.2.1 Inteligencia Artificial

El término de inteligencia artificial (IA) fue introducido por el profesor John McCarthy de la Universidad de Stanford, lo definió como 'la ciencia e ingeniería de hacer inteligentes a las máquinas'. Muchas personas asocian este término con el hecho de usar computadoras para estudiar la inteligencia y la toma de decisiones de los humanos. Sin embargo, esto no es del todo cierto.

La inteligencia artificial es el estudio para simular la inteligencia de los seres humanos utilizando softbots o robots para analizar el entorno, pensar racionalmente y actuar de acuerdo a la situación lo amerite. La IA también se está utilizando para aprender problemas de extrema complejidad que requiere cálculos complicados y encontrar soluciones a través de la experiencia. En otras palabras, solucionar problemas que los seres humanos no podemos hacer con facilidad.

Existen distintos tipos de IA según Arend Hintze dentro de ellas están la:

- Reactiva: Es considerado el sistema de IA más básico, ya que no tiene la capacidad de formar recuerdos, ni puede utilizar data pasada para realizar toma de decisiones. Simplemente actúa respondiendo a distintos tipos de estímulos.

- Memoria Limitada: Este tipo de IA puede utilizar data pasada. Es decir, además de la configuración predefinida se les puede añadir algunas experiencias para utilizar por un corto periodo de tiempo, ya que estas piezas de información sobre el pasado no se pueden compilar por mucho tiempo.

- Teoría de la mente: En esta categoría las máquinas son capaces de entender e interactuar con entidades a partir del discernimiento de sus necesidades, emociones, creencias, y procesos pensativos. Lo complejo de este tipo de IA es que para que las máquinas puedan entender las necesidades humanas, estas van a tener que entender lo que piensa un humano.

- Autoconciencia: Este tipo de IA por el momento existe hipotéticamente. En ella las máquinas no solo serán capaces de entender y provocar emociones con los que interactúan, sino que serán capaces de tener sus propias emociones, necesidades, creencias y hasta deseos propios. Algunas personas temen llegar a este nivel, a pesar de que esto ayudaría de manera descomunal a la sociedad humana, también podría significar la decadencia de esta, debido a que las máquinas podrán desarrollar su sentido de autopreservación y afectar de manera negativa la humanidad.

2.2.2 Machine Learning

En la última década, la cantidad de información generada en el mundo ha crecido exponencialmente, esto provoca que sea más difícil el procesamiento de toda esta información sin contar con las herramientas tecnológicas adecuadas. Una de estas herramientas es el Machine Learning, el cual tiene como principal propósito el convertir datos en información útil para el usuario.

Según Harrington (2012), el Machine Learning está basado en un conjunto de ciencias, principalmente informática, ingeniería y estadística. Esta herramienta tiene una amplia gama de aplicaciones en diferentes campos e industrias. Básicamente, cualquier problema que necesite interpretar datos y actuar en base a ello, puede beneficiarse del uso del Machine Learning.

Durante la última mitad del siglo XX, una gran cantidad de la población activa en los países desarrollados ha pasado del trabajo manual a lo que se conoce como trabajo del conocimiento. Como parte de este cambio, las asignaciones de trabajo son cada vez más

ambiguas y objetivos como “maximizar las utilidades”, “reducir los riesgos” y “optimizar la cadena de suministro” son cada vez más comunes. Asimismo, la gran cantidad de información que hay disponible en la red hace aún más complejo el trabajo de los trabajadores del conocimiento. En este sentido, dar sentido a los datos de manera que creen valor en el trabajo se está volviendo, hoy más que nunca, en una habilidad esencial.

De acuerdo con Marcos y Elías (2017), se deben seguir cuatro pasos para aplicar correctamente el Machine Learning:

1. Recopilación de datos: El primer paso es recolectar la información de todas las fuentes disponibles y relevantes para la investigación. Asimismo, existe una gran cantidad en repositorios públicos lo que permite ahorrar tiempo y esfuerzo.

2. Preprocesado: Una vez se tengan los datos identificados, es necesario darles una estructura de base de datos para que pueda ser procesada. Asimismo, se deben limpiar los datos, identificando valores vacíos o fuera de la media que puedan distorsionar el resultado.

3. Aprendizaje: En esta etapa es donde se realiza el aprendizaje automático, se alimenta el modelo con una cantidad considerable de datos para que el sistema pueda entrenar y aprender los patrones requeridos por el usuario. En el caso del aprendizaje no supervisado, no hay ningún paso de capacitación porque no se tiene un valor objetivo.

4. Evaluación: Con lo aprendido por el modelo en el paso anterior, se pueden generar pruebas para determinar el grado de confiabilidad del modelo. En el caso del aprendizaje supervisado, se utilizan datos de validación para obtener el porcentaje de éxito; mientras que, en el aprendizaje no supervisado, es necesario utilizar otras métricas para evaluar el algoritmo.

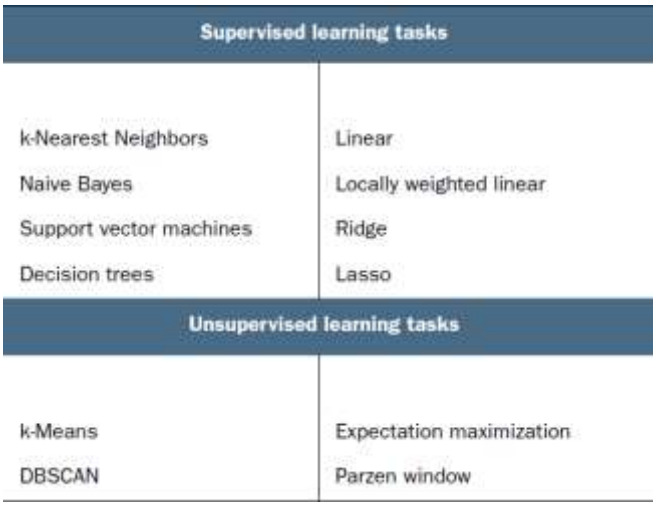
2.2.3 Aprendizaje Supervisado

De acuerdo con Harrignton (2017), dentro del Machine Learning existen dos tipos de aprendizaje: supervisado y no supervisado. En el aprendizaje supervisado no hay una etiqueta ni un valor objetivo para los datos. Un ejemplo de este tipo de aprendizaje es el clustering, el cual tiene como objetivo agrupar los elementos que tienen características similares.

Por otro lado, se encuentra el aprendizaje supervisado se utiliza para realizar tareas de predicción ya que el objetivo es pronosticar / clasificar un resultado específico de interés. El aprendizaje supervisado se ha aplicado a grandes estructuras de datos que incluyen predictores demográficos, clínicos y sociales para desarrollar puntuaciones de riesgo que predicen el inicio y la trayectoria de una variedad de trastornos mentales (por ejemplo, ansiedad, depresión y trastornos relacionados con el trauma) y comportamiento suicida.

En la siguiente figura se describen los principales algoritmos para cada tipo de aprendizaje.

Figura 3. Tipos de aprendizaje y principales algoritmos



Fuente: Harrington (2017)

2.2.4 Algoritmo de K-Nearest Neighbords Algorithm

Dentro de todos los algoritmos existentes dentro del Machine Learning, el KNN es uno de los más fáciles de comprender, pero muy efectivo. Entre sus principales ventajas tenemos que cuenta con una alta precisión, es insensible a valores atípicos y no hace suposiciones sobre los datos. Por otro lado, algunas de sus desventajas son que es computacionalmente costoso y requiere grandes cantidades de memoria.

Este algoritmo funciona de la siguiente manera: inicialmente se debe seleccionar el conjunto de datos de entrenamiento, dentro de esta base se debe especificar la clase (etiqueta) a la que pertenece cada elemento. Luego del entrenamiento, cuando el algoritmo reciba un dato sin etiqueta, el modelo comparará el dato nuevo con todos los datos existentes (base de

entrenamiento) y localizará los más similares. El modelo examinará los "K" elementos más cercanos (vecinos) y la etiqueta que se repita en la mayoría será la etiqueta asignada para el nuevo dato. Al momento de plantear el modelo, el usuario debe determinar "K", que es el número de vecinos cercanos que tomará en cuenta el algoritmo para calcular el resultado. "K" debe ser un número entero y, de preferencia, menor a 20)

2.2.5 Regresión

De acuerdo con Aires (2008), lo que hoy se conoce como regresión fue inventado por el primo de Charles Darwin, Francis Galton. Galton hizo su primera regresión en 1877 para estimar el tamaño de las semillas de guisantes en función del tamaño de las semillas de sus padres. Galton realizó la regresión en una serie de cosas, incluida la altura de los humanos. Observó que, si los padres tenían una estatura superior a la media, sus hijos también tendían a estar por encima de la media, pero no tanto como sus padres. Las alturas de los niños estaban retrocediendo hacia un valor medio. Galton generalizó esta tendencia bajo la Ley de la regresión universal, "cada peculiaridad en un hombre es compartida por sus descendientes, pero en media, en un grado menor". Galton notó este comportamiento en varias cosas que estudió, por lo que la técnica se llama regresión, a pesar de que la palabra en inglés no tiene relación con la predicción de valores numéricos.

El objetivo del modelo de regresión es predecir un valor objetivo numérico. Una de las maneras de lograr esto es a través de una ecuación para calcular el valor objetivo en base a una serie de variables de entrada. Dentro de un análisis de regresión simple se puede encontrar una variable dependiente o de respuesta (Y) y una variable independiente o explicativa (X). El propósito principal de este modelo es obtener una función simple en función de la variable independiente, que sea capaz de describir, de la manera más exacta posible, la variación de la variable dependiente. La variable independiente puede estar formada por un vector de sólo una característica o, también, podría ser un conjunto de diferentes atributos, características o dimensiones; en este caso se le conocería como regresión múltiple.

La ecuación de la regresión lineal está representada en la siguiente ecuación, donde Y representa a la variable dependiente y X es la variable independiente. Por otro lado, β_0 y β_1 son parámetros desconocidos del modelo, los cuales deben estimarse mediante los datos de muestra. Por último, ε representa la variabilidad en Y que no se puede explicar con el modelo lineal, es llamado error.

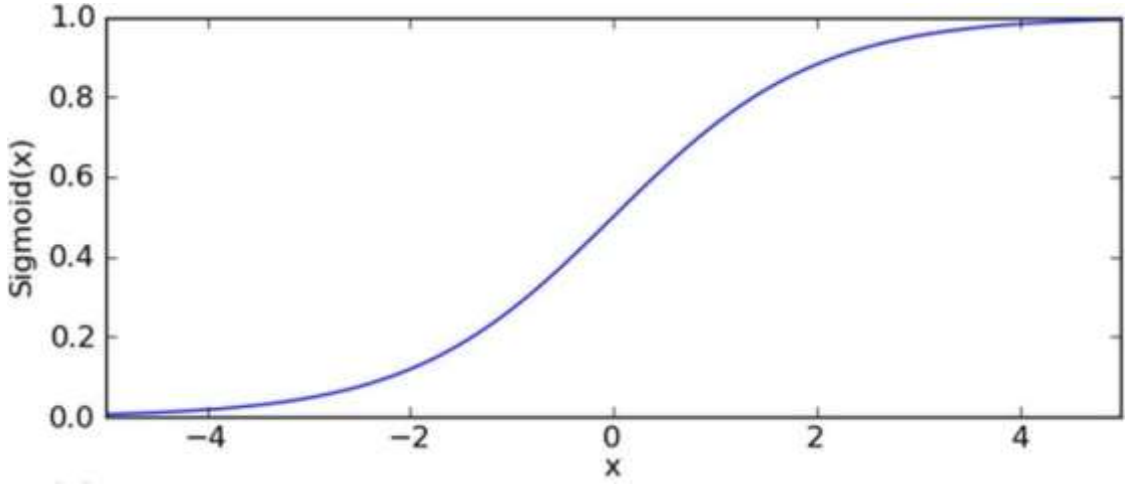
$$Y = \beta_0 + \beta_1 + \varepsilon$$

2.2.5.1 Regresión logística. El objetivo de la ecuación de regresión logística es analizar todas las variables de entrada y predecir la clase de la variable objetivo; por ejemplo, en el caso de tener dos clases definidas, la función arrojará un 0 o un 1. La ecuación utiliza la función sigmoide para determinar la clase del valor objetivo. La función sigmoide responde a la siguiente ecuación:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

En la siguiente figura se muestran dos ejemplos gráficos de la función sigmoide, mientras mayor sea el valor de x, el sigmoide se acercará a 1; mientras que, para valores decrecientes de x, el sigmoide se acercará a 0.

Figura 4. Gráfico de función sigmoide



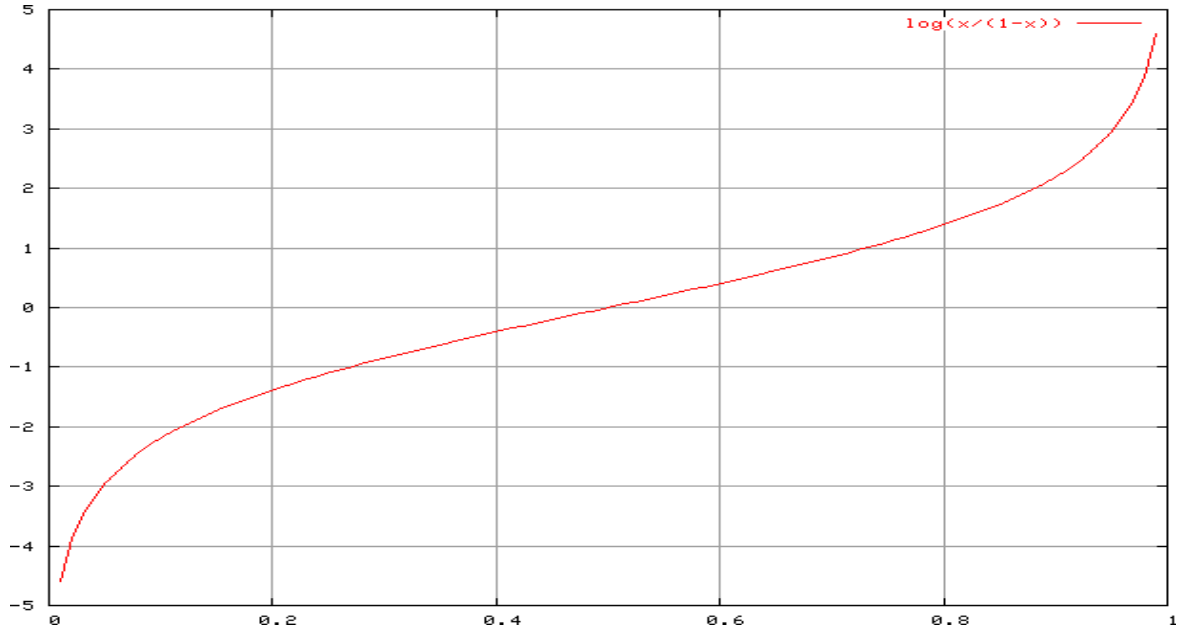
Fuente: Harrington (2017)

Para el clasificador de regresión logística, se multiplican cada una de las variables de entrada por un peso y luego se suman. Este resultado se pondrá en el sigmoide y se obtendrá un número entre 0 y 1. Cualquier valor superior a 0,5 será clasificado como 1, y cualquier valor inferior a 0,5 será clasificado como 0.

Por otro lado, también es posible utilizar la función logit para el análisis con regresión logística, cuya formula viene dada de la siguiente manera:

$$logit(p) = \log\left(\frac{p}{1 - p}\right)$$

Figura 5. Gráfico de función logit



Fuente: Elaboración Propia

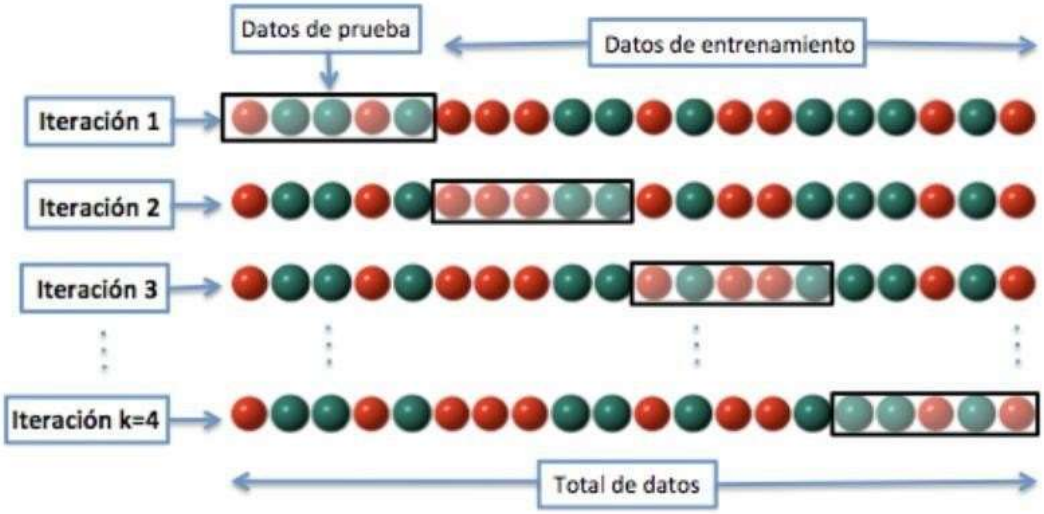
2.2.6 Métricas

2.2.6.1 Cross Validation

Dentro de las investigaciones de Machine Learning existentes, las técnicas de validación cruzada (cross validation) son las más utilizadas por los investigadores para el entrenamiento y validación del modelo. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevará a cabo a la práctica.

Existen varios tipos de Cross Validation, la más utilizada es la de k-iteraciones o K-Folds, la cual consiste en dividir el conjunto de datos en k subconjunto. La técnica consiste en dejar uno de los subconjuntos como datos de validación y el resto (k–1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de validación. Finalmente, se selecciona la que mayor capacidad de generalización posea. (Arlot, 2010).

Figura 6. Técnica de k-iteraciones



Fuente: Caicedo (2014)

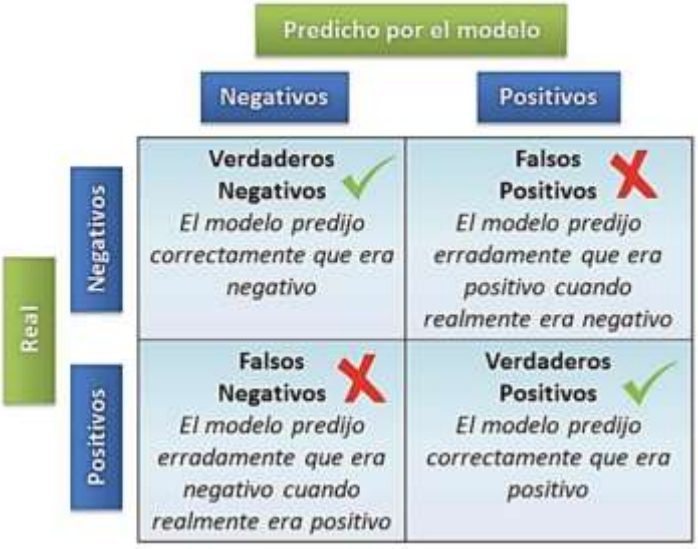
La evaluación de las diferentes validaciones cruzadas normalmente viene dada por el error obtenido en cada iteración. Cabe tener en cuenta que por cada uno de los métodos puede variar el número de iteraciones, según la elección del diseñador en función del número de datos total.

2.2.6.2 Matriz de Confusión

De acuerdo con Swamynathan (2017), la matriz de confusión es una herramienta que explica el desempeño de un modelo de clasificación. Esto se logra analizando los resultados en base a las cuatro posibilidades posibles al comparar el resultado real con el resultado obtenido por medio del modelo.

En la siguiente figura se muestra cada una de estas posibilidades y se explica el detalle de cada una según las definiciones de Swamynathan (2017).

Figura 7. Matriz de confusión



Fuente: Elaboración propia

En base a esta matriz antes mostrada se pueden obtener diferentes métricas que permiten evaluar un modelo de clasificación:

1. Accuracy: Esta métrica busca calcular la proporción de predicciones que el modelo clasificó de manera correcta**. "De todas las clases, cuántas se predijeron correctamente"**

$$Accuracy = \frac{\#\ de\ predicciones\ correctas}{total\ de\ predicciones}$$

$$Accuracy = \frac{Verdaderos\ Positivos + Verdaderos\ Negativos}{Verdaderos\ Positivos + \ Falsos\ Positivos + \\ Verdaderos\ Negativos + \ Falsos\ Negativos}$$

2. Precisión: Esta métrica es conocida como valor predictivo positivo y se calcula en función a la proporción de verdaderos positivos y predicciones positivas. **"De todas las predicciones positivas, cuántas son positivas realmente"**

$$Precisión = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos}$$

3. Recall: También llamada tasa positiva real (TPR) y se calcula en función a la proporción de verdaderos positivos y verdaderos reales. **"De todas las clases positivas, cuántas se predijeron correctamente"**

$$Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos}$$

4. Puntuación F1: Esta es una medida de precisión de prueba, siendo la media armónica entre la precisión y el recall. En conclusión, es una medida de precisión y robustez de su modelo.

$$Puntuación\ F1 = \frac{2\ \times Recall\ \times Precisión}{Recall \times Precisión}$$

CAPÍTULO III: ENTORNO EMPRESARIAL

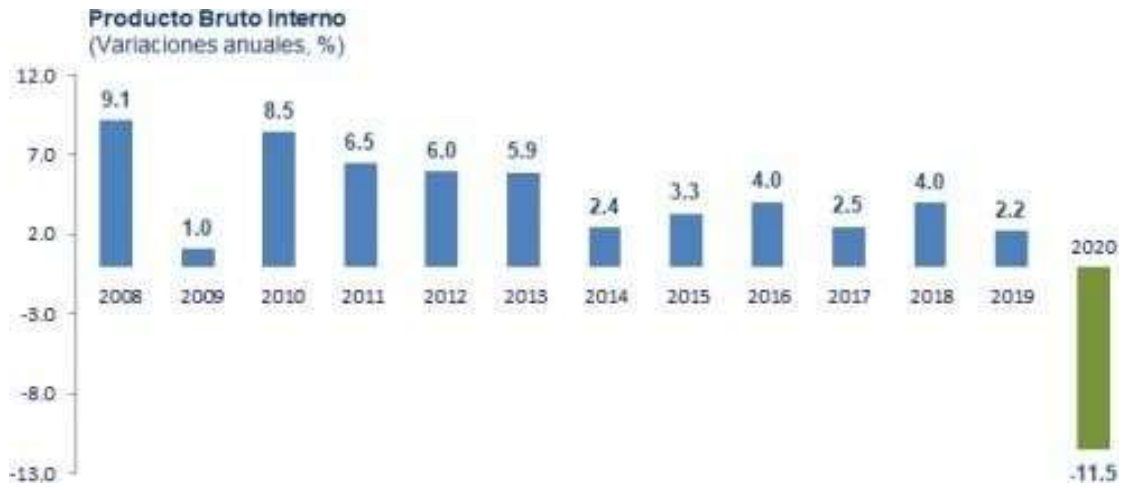
3.1 Descripción de la empresa

3.1.1 Reseña histórica y actividad económica

Debido a la sensibilidad de la información no se mencionará la razón social de la empresa para mantener la confidencialidad. El trabajo se basa en una empresa que forma parte del mercado asegurador que busca brindar una oferta integral de productos y servicios a las familias peruanas.

Para esto nos introduciremos un poco en la información relacionada con el mercado asegurador peruano: En el 2020 y con la nueva realidad enfrentada a nivel mundial. Ha tenido como consecuencia un efecto negativo sobre el PBI de todos los países; en el caso de la economía peruana se ha visto un notable decrecimiento del 11.5% a diferencia de años anteriores según lo reportado por el Banco Central de Reserva del Perú (Figura 2); debido a la nueva normalidad que se vivió especialmente a mediados del mismo año, esto tuvo efectos sobre las siguientes actividades: el consumo de la población, la exportación e importación se vio afectada, al igual que la inversión privada y pública, entre otros que afectaron a muchas trabajadores, empresas y pobladores.

Figura 8. Variación Porcentual del Producto Bruto Interno Peruano en los últimos años



Fuente: Banco Central de Reserva (2021)

Continuando con el mercado de seguros, según lo reportado por la Asociación Peruana de Empresas de Seguros también conocida como APESEG, hasta el segundo trimestre del 2020, existió un total de 20 empresas registradas que operan en el mercado asegurador peruano. Estas han visto un efecto negativo en sus primas anuales teniendo un indicador del 1.83% (ratio de

primas/PBI) de penetración en el mercado versus el 1.87% logrado en el último trimestre del 2019, debido a la disminución de la actividad económica en el Perú y la inestabilidadocasionada por la cuarentena.

En este trabajo nos enfocaremos en el ramo de los seguros generales, para ser más específicos en los seguros vehiculares. En el 2020 este tuvo decrecimiento de 3 puntos porcentuales con respecto al 2019, con un valor del 7.9% (Figura 3) a diferencia del 10.9% (Figura 4) del 2019 que fue reportado por APESEG en base a las primas mostradas en el portal de la SBS.

Figura 9. Composición del Mercado de Primas de Seguros de Principales Riesgos (2019 – 2020)

	2019	2020
Vehículos	7.9%	10.1%
Restos Generales	23.8%	26.3%
Asistencia Médica	8.9%	9.1%
SOAT	3.0%	2.0%
Resto Accidente y Enfermedad	2.0%	1.9%
Seguro de Vida Ind. de Largo Plazo	8.9%	8.1%
Seg. de Desgravamen Hipotecario	9.9%	11.1%
Resto Vida	16.8%	15.2%
Resto SPP	18.8%	15.2%

Fuente: Asociación Peruana de Empresas de Seguros - APESEG

Sin embargo, para nuestra investigación y con el fin de obtener un resultado menos atípico, que no se encuentre afectado por el paro económico que se vivió por el COVID-19 en el 2020, se dará un mayor énfasis en los resultados reportados en el 2019.

Según lo reportado por la SBS al 31 de diciembre del 2019 la empresa con mayor participación en el ramo generales y accidentes y enfermedades, al que pertenece los seguros de autos, fue Rimac con una participación del 35.47%; seguido por Pacífico Seguros con un 23.17% y Mapfre Perú con un 16.68% (Anexo1). Adicionalmente, la Superintendencia de Banca y Seguros reportó que para el 2019 se obtuvo un crecimiento en las primas anualizadas del 9.7% que es un aproximado de S/.14,113 millones.

3.1.2 Descripción de la organización

3.1.2.1 Organigrama

Figura 10. Organigrama de la Empresa de Seguros Peruana

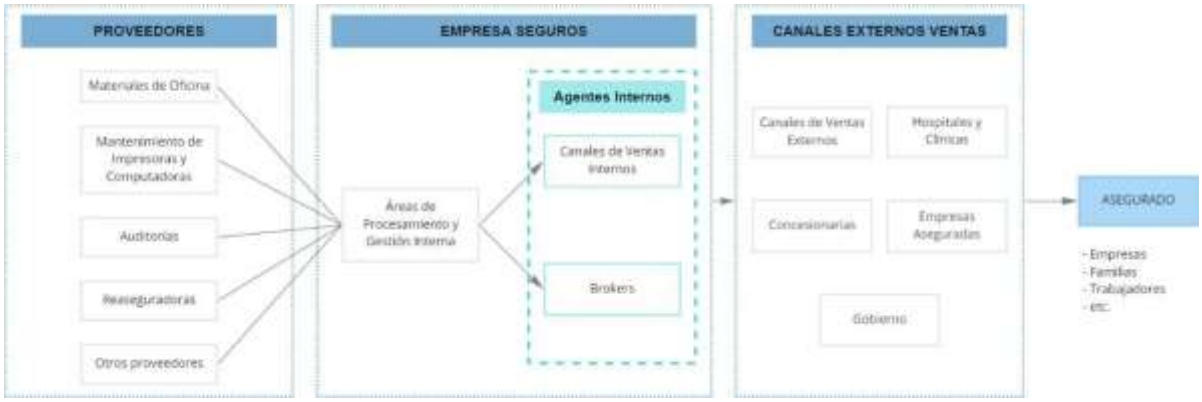
Fuente: Elaboración propia

3.1.2.2 Cadena de suministros

La empresa cuenta con un aproximado de 35 proveedores, que suministran los siguientes bienes y servicios: papel bond, impresiones, servicio y mantenimiento de impresoras en oficina, servicio y mantenimiento de computadoras y laptops, merchandising, útiles de oficina, consultoría, logística, limpieza, servicio de refrigerios y catering, organizadores de eventos, auditoría, evaluaciones, entre otros. Actualmente la aseguradora cuenta con uncontrato con sus proveedores en el que se compromete a cumplir los diez principios del Pacto Mundial de las Naciones Unidas los cuales están relacionados con: los derechos humanos, estándares laborales, medio ambiente y la lucha contra la corrupción. Con los proveedores recurrentes se generan procesos de evaluación anuales que constan de la revisión de su situación financiera y tributaria, gestión comercial, seguridad y salud, prácticas laborales, calidad y operatividad; de fallar las evaluaciones, los proveedores serán informados de la situación dándoles la oportunidad de cambiar y realizar mejoras para continuar trabajando con ellos.

Adicional a lo anterior, cuenta con canales de ventas internos y externos. Los internos están conformados por los brokers y vendedores directos que se encuentran en las oficinas o sucursales de ventas. Los asegurados se acercan a estos centros para ser asesorados por los vendedores y obtener el mejor seguro dependiendo de sus necesidades. Mientras que las Empresas o el Estado cuenta con la intervención de los brokers, los cuales le ofrecen planes de seguros para cubrir a la corporación o institución. Los canales de ventas externos son aquellos brokers externos que se pueden encontrar en los centros de salud como hospitales o clínicas, centros comerciales, bancos, la compra de seguros en concesionarias (al momento de realizar la compra de un carro).

Figura 11. Cadena de Suministros de una Empresa de Seguros



Fuente: Elaboración propia

3.1.3 Datos generales estratégicos de la empresa

3.1.3.1 Visión, misión y valores o principios

Visión: Ser una de las cinco mejores aseguradoras de Latinoamérica: simple, transparente, accesible, rentable y con colaboradores altamente competentes y motivados.

Misión: Salvaguardar la estabilidad económica de sus clientes, ofreciéndoles soluciones que protejan aquello que valoran y aseguren el cumplimiento de sus objetivos.

Valores: Para enfocar y guiar sus esfuerzos, Pacífico se basa en los siguientes principios de gestión:

- Construir relaciones a largo plazo: se creen en las relaciones a largo plazo y se enfocan en desarrollarlas con sus asegurados, corredores y canales de distribución.

- Son especialistas en la gestión de riesgos: trabajan junto con sus clientes para entender sus necesidades y les ofrecen soluciones que les permiten manejar sus riesgos de manera eficiente.

- Cumplen con sus obligaciones de manera justa y oportuna: resuelven los siniestros con un alto criterio de justicia y los pagan de forma justa.

- Buscan la excelencia en el servicio al cliente: asesoran a sus clientes en la gestión de sus riesgos y se enfocan cada día para darles la calidad de servicio que merecen.

- Son una compañía sólida y confiable: su fortaleza financiera, así como una gestión profesional y prudente del negocio de seguros, garantizan la más alta capacidad de pago de sus obligaciones ahora y en el futuro.

3.1.3.2 Objetivos estratégicos

El objetivo principal de la empresa es el compromiso con el bienestar y felicidad de sus clientes a través de los servicios de seguros que brinda y la mejora de los procesos de la empresa, junto con la estabilidad laboral de sus colaboradores. Para lograr esto cuentan con 3 pilares en sus estrategias:

- Crecimiento, esto se lograría a través del crecimiento de la rentabilidad de la empresa, lo que le permitiría una mejorar en la productividad de sus trabajadores y los canales de ventas junto con la innovación en sus procesos internos

- Experiencia del Cliente, como se menciona anteriormente, el salvaguardar la estabilidad emocional y brindar una experiencia satisfactoria a sus asegurados es uno de sus pilares, esto se puede lograr a través de la agilización de la atención al cliente, el fácil acceso a los productos que ofrecen como también a la obtención de información de manera sencilla y entendible para el usuario. Adicional a esto es el reducir los tiempos de atención cuando ocurren siniestros o se quiere hacer uso del seguro. Para esto se generaron estrategias de relacionamiento con los clientes, que permiten la fidelización de los mismos con la empresa.

- Eficiencia, para alcanzar la satisfacción del cliente y mejorar su experiencia con la empresa se hace uso de nuevas técnicas de transformación digital, innovación en los procesos internos y la reducción de los incidentes o fraudes.

3.1.3.3 Evaluación interna y externa

Tabla 1. Matriz de Evaluación de Factores Externos

	FACTORES EXTERNOS CLAVE			
	OPORTUNIDADES	**PONDERACIÓN**	**CALIFICACIÓN**	**PUNTUACIÓN PONDERADA**
1	Economía habría crecido hasta 15% en julio del 2021	0,04	2	0,1
2	La situación causada por el COVID 19 a generado una gran concientización de la importancia de seguros de salud	0,11	3	0,33
3	Persistencia en la inseguridad ciudadana	0,1	3	0,3
4	Alto grado de fidelidad en los clientes	0,05	3	0,15
5	Deficiencia en la salud pública	0,07	1	0,07
6	Crecimiento del poder adquisitivo en las provincias	0,11	4	0,44
7	Baja facilidad de ingreso a competidores a la industria aseguradora	0,1	4	0,4
8	La industria presenta un moderado nivel de utilidades	0,05	3	0,15
	AMENAZAS			
9	Sector con alto grado de competitividad	0,12	4	0,48
10	Creación de Ley de Contrato de Seguros	0,04	4	0,16
11	Reducción de la demanda en líneas mayorista de seguros (empresarial)	0,04	2	0,08
12	Aumento en los precios de las medicinas	0,03	2	0,06
13	El tráfico urbano ralentiza el servicio	0,08	3	0,24
14	Costumbre de automedicarse y evitar atención en centros de salud	0,04	1	0,04
15	Las primas de seguros durante el 2020 experimentaron una caída de 0,7% respecto del año anterior.	0,02	2	0,04
	Total:	**1,00**		**3,04**

Fuente: Elaboración propia

Tabla 2. Matriz de Evaluación de Factores Internos

	FACTORES INTERNOS CLAVE			
	FORTALEZAS	**PONDERACIÓN**	**CALIFICACIÓN**	**PUNTUACIÓN PONDERADA**
16	Cuenta con el respaldo de Holding Financiero más importante del Perú:Credicorp	0,08	4	0,32
17	La moral de los empleados es alta.	0,05	3	0,15
18	Las Primas Netas crecieron en 8% con respecto al año anterior.	0,05	4	0,2
19	Formó un conglomerado con Banmédica Chile.	0,07	4	0,28
20	Interfaz amigable y confiable de la página web.	0,06	3	0,18
21	Alta comunicación interna entre los departamentos y los diferentes niveles jerárquicos	0,05	4	0,2
22	Ofrece una amplia variedad de productos.	0,06	3	0,18
23	Líder como Empresa de Carbono Neutral-	0,05	3	0,15
24	Alto estado del arte en la infraestructura y modernas instalaciones.	0,06	3	0,18
	DEBILIDADES			
25	La publicidad masiva no es tan efectiva.	0,07	1	0,07
26	Falta de personal capacitado para sucursales en provincias.	0,08	2	0,16
27	Precios por encima del mercado.	0,05	2	0,1
28	Caída de la Utilidad Neta.	0,04	1	0,04
29	Las líneas Provisionales cayeron con respecto a las demás líneas.	0,05	2	0,1
30	Aumento de gastos técnicos en 33%.	0,04	1	0,04
31	Decrecimiento de las ventas de Seguros de Vida.	0,07	1	0,07
32	Alta rotación de personal en los niveles jerárquicos más bajos.	0,07	2	0,14
	Total:	**1,00**		**2,56**

Fuente: Elaboración propia

3.2 Modelo de negocio actual (CANVAS)

Tabla 3. Modelo CANVAS

<table>
<tr>
<td rowspan="2">Socios clave

Empresa de servicios financieros
Red de clínicas peruanas.
Una de las compañías de Seguros de salud más grandes del mundo</td>
<td>Actividad Clave
Publicidad, Pre y Post venta, Asesoría y Seguimiento a clientes, Área Legal</td>
<td rowspan="2">Propuesta de valor

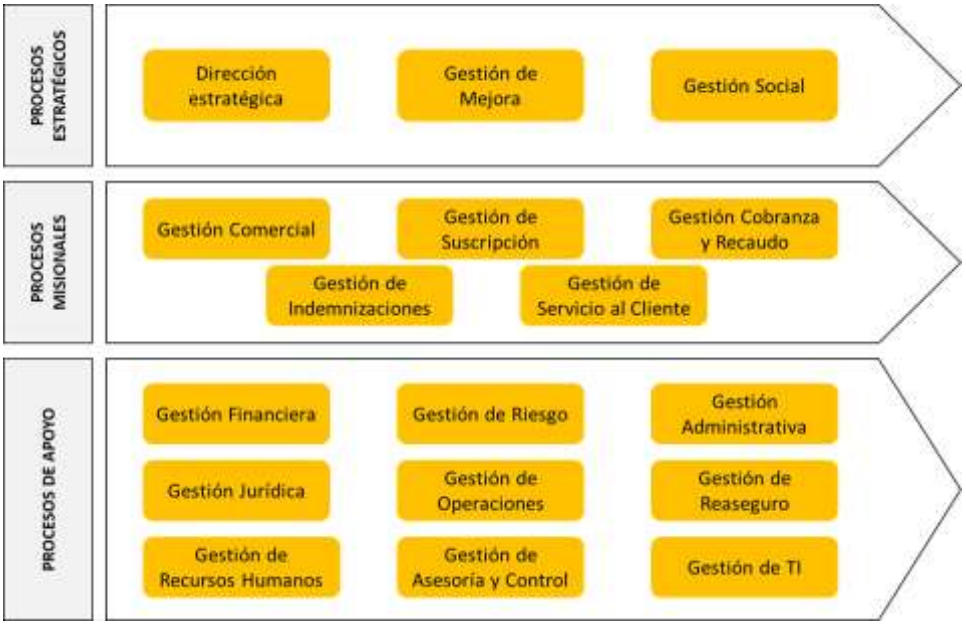
Seguros vehiculares ajustados a la necesidad del conductor/consumidor</td>
<td>Relación con cliente
Asistencia personalizada
Planes base preelaborados</td>
<td rowspan="2">Segmentos de clientes

Personas individuales pertenecientes a sectores socio-económicos A-D con vehículo propio.
Empresas que realizan alguna actividad de transporte (personal u objetos)</td>
</tr>
<tr>
<td>Recurso Clave
Personal + motorizados
Página web</td>
<td>Canales
Teléfono
En línea
Oficinas y agencias</td>
</tr>
<tr>
<td colspan="3">Estructura de Costos
Mercadotecnia y publicidad, salarios y comisiones a personal, gastos de cobertura de pólizas</td>
<td colspan="2">Fuentes de Ingresos
Primas de seguros</td>
</tr>
</table>

Fuente: Elaboración propia

3.3 Mapa de procesos actual

Figura 12. Mapa de procesos de compañía de seguros



Fuente: Elaboración propia

CAPÍTULO IV: METODOLOGÍA DE LA INVESTIGACIÓN

4.1 Diseño de la Investigación.

4.1.1 Enfoque de la investigación

De acuerdo con Hernandez et al. (2014), un planteamiento con enfoque cuantitativo "es útil para evaluar, comparar, interpretar, establecer precedentes y determinar causalidad y sus implicaciones"

En este sentido, el enfoque de la presente investigación es cuantitativo ya que busca diseñar y desarrollar instrumentos que permitan reconocer y medir atributos del objeto de estudio, el cual comprende la clasificación de clientes de seguros vehiculares usando técnicas de Machine Learning. El modelo resultante de esta investigación se podrá evaluar mediante indicadores matemáticamente interpretables.

4.1.2 Alcance de la investigación.

De acuerdo con Hernandez et al. (2014), un estudio correlacional tiene como principal objetivo "conocer el grado de relación que existe entre dos o más conceptos, categorías o variables en una muestra o contexto en particular"

En este sentido, el alcance del presente estudio es correlacional debido a que se busca conocer el grado de asociación entre dos variables, la clasificación de clientes y el algoritmo de clasificación con Machine Learning.

4.1.3 Tipo de la investigación.

El diseño o tipo de esta investigación es experimental dado que se realiza un experimento en base al análisis relacional entre las variables seleccionadas para clasificar a los clientes de seguros vehiculares mediante técnicas de Machine Learnig. De acuerdo con Hernández et al. (2014) es de tipo pre-experimental debido a que el grado de control es mínimo sobre la muestra.

4.2 Metodología de implementación de la solución

Figura 13. Etapas de desarrollo de la implementación de la solución

Fuente: Elaboración propia

Para la implementación de la solución, se utilizarán dos algoritmos diferentes: KNN y regresión logística. Como visto anteriormente en la base teórica, estos 2 modelos se ajustan al diseño de estudios supervisados cuyo fin es determinar una clasificación específica de una entidad basado en un set de variables definidas. Para la fase de entrenamiento de estos modelos, se procederá a separar el total de la muestra, que ya pasó por una fase de limpieza, y se separará en dos partes, siendo el mayor volumen de sets de datos utilizado para el diseño y entrenamiento de los modelos. El resto de la data se utilizará para la fase de evaluación.

4.2.1 Recolección de base de datos

Los datos que serán utilizados para la presente investigación serán solicitados a la empresa definida en el capítulo anterior. Los datos solicitados deberán ser relevantes para el tema de investigación, es decir, que sea una base de datos con información histórica acerca de la clasificación de clientes y las diferentes variables que se tomaron en cuenta para realizar dicha clasificación. Asimismo, la base de datos debe ser lo suficientemente grande para aplicar técnicas de Machine Learning ya que, de acuerdo a la literatura consultada, se podrán obtener mejores resultados si se utiliza una mayor cantidad de datos.

4.2.2 Limpieza y pre-procesamiento

En esta etapa se realizará una revisión de los datasets obtenidos, comprobando que no haya datos faltantes, dañados y/o incorrectos que puedan afectar el procesamiento de la información y arrojar resultados equivocados. Este paso es sumamente relevante ya que mucha de la información que será proporcionada por la empresa ha sido ingresada a los sistemas de manera manual por lo que hay un cierto grado de error.

4.2.3 Modelado

En esta etapa se realizará el entrenamiento de los algoritmos de clasificación de Machine Learning, experimentando con diferentes parámetros para cada modela de manera que encuentre la combinación que brinde los mejores resultados.

a) K-NN: Dentro del marco teórico se detalla el funcionamiento de este algoritmo, para la implementación se utilizará la librería *sklearn*, el módulo *neighbors* y la función *KNeighborsClassifier*.

Figura 14. Importación de algoritmo K-NN

```
#Para KNN
from sklearn.neighbors import KNeighborsClassifier
```

Fuente: Elaboración propia

b) Regresión Logística: Dentro del marco teórico se detalla el funcionamiento de este algoritmo, para la implementación se utilizará la librería *sklearn*, el módulo *linear model* y la función *LogisticRegression*.

Figura 15. Importación de algoritmo de regresión logística

```
#Para Regresión Logistica
from sklearn.linear_model import LogisticRegression
```

Fuente: Elaboración propia

4.3 Metodología para la medición de resultados de la implementación

La evaluación de los modelos planteados dentro de la investigación será a través del Accuracy y el Cross Validation. A nivel de código, se utilizará la librería *sklearn*, el módulo *metrics* y la función *F1 Score*, ya que no todas las clases están balanceadas y es necesario utilizar diferentes parámetros al *Accuracy* para asegurar la robustez del modelo. Asimismo, se

utilizará la función C*ross_Val_Score* para validar que el modelo sea aplicable a diferentes grupos de datos. Por último, se utiliza la función *Confusion_Matrix* para graficar la matriz de confusión.

Figura 16. Importación de algoritmos de medición

```
#Métricas
from sklearn.metrics import accuracy_score
from sklearn.metrics import f1_score
from sklearn.model_selection import cross_val_score
from sklearn.metrics import confusion_matrix
```

Fuente: Elaboración propia

4.4 Cronograma de actividades y presupuesto

A continuación, se presenta el cronograma de actividades propuesto para llevar a cabo la investigación.

Tabla 4. Cronograma de actividades

Actividades	OCT					NOV				DIC		
	S1	S2	S3	S4	S5	S1	S2	S3	S4	S1	S2	S3
INICIO												
Búsqueda de la problemática												
Búsqueda de investigaciones similares												
PLANIFICACIÓN												
Desarrollo del planteamineto del problema												
Desarrollo del marco de referencia												
Análisis del entorno empresarial												
Desarrollo de la metodología de la investigación												
DESARROLLO												
Propuesta de solución												
Desarrollo de la solución												
Resultados de la solución												
CONTROL												
Comparación de los resultados												
Documentación de la solución												
CIERRE												
Conclusiones y recomendaciones												
Presentación del informe												
Sustentación												

Fuente: Elaboración propia

En la siguiente tabla se presenta el presupuesto estimado para llevar a cabo la investigación. Los costos han sido calculados en la moneda nacional (soles peruanos). Asimismo, es importante resaltar que estos montos no representan un cálculo exacto, sino que son cifras estimadas.

Tabla 5. Presupuesto de la investigación

RECURSO	CONCEPTO	COSTO
Equipo	Laptop con una buena capacidad de procesamiento y memoria RAM	S/ 3.000,00
Software	Software para procesamiento de señales de audio digital (Python, Jupyter Notebook) procesamiento y memoria RAM	S/ -
Otros	Electricidad	S/ 100,00
	Servicio de Internet	S/ 140,00
	TOTAL	**S/ 3.240,00**

Fuente: Elaboración propia

CAPÍTULO V: DESARROLLO DE LA SOLUCIÓN

El presente trabajo tiene como objetivo desarrollar un modelo de clasificación automático que pueda predecir el tipo de póliza vehicular que debe ofrecerse a cada cliente de una aseguradora en función al análisis de diferentes variables, tanto relacionados a la persona como al vehículo. Dentro del alcance contemplado en este trabajo, el total de la solución se desarrollará en lenguaje Python debido a su gran comunidad de soporte en cuanto a paquetes y rapidez de ejecución en modelado de inteligencia artificial.

5.1 Propuesta de Solución

5.1.1 Planteamiento y Descripción de Actividades

5.1.1.1 Obtención de base de datos

Se hará uso de una base de datos de Autos perteneciente a una empresa peruana de seguros. Esta ha sido generada a través de procesos internos realizados en SQL, cuenta con información del año 2019 y 23 variables, algunas de ellas descriptivas y numéricas, como: género del asegurado, edad, tipo de vehículo, tipo de producto o póliza adquirida, entre otros que nos permitirán el desarrollo del modelo final.

Tabla 6. Descripción de variables de base de datos inicial

VARIABLES	DESCRIPCIÓN
ID_CERTIFICADO	Código de identificación interna del asegurado
GENERO_ASEGURADO	Es el generó del asegurado (Masculino, femenino y no aplica)
COD_GENERO	Código para identificar el género del asegurado
EDAD_ASEGURADO_DATA	Edad del asegurado
PROV_ASEGURADO	Provincia del asegurado
DEPARTAMENTO_C	Departamento del asegurado
DISTRITO_C	Distrito del asegurado
COD_UBIGEO	Código de ubicación en Perú del asegurado
TIPO_CLIENTE_ASEGURADO	Tipo de cliente que es el asegurado (Persona o Empresa)
COD_TIPO_CLIENTE	Código del tipo de cliente
COD_PROD	Tipo de producto que adquirió el cliente
COD_TIPO_POL	Código del tipo de producto registrado en la póliza del cliente
ANO_FABRICACION_VEH	Año de fabricación del vehículo registrado
MARCA_VEH_STD	Marca del vehículo
COD_MARCA	Código de la marca del vehículo registrado
TIPO_VEH	Tipo de vehículo registrado
COD_TIPO_VEH	Código del tipo de vehículo registrado
CLASE_AGRUPADA	Clasificación agrupada del tipo de vehículo
COD_CLASE_AGRU	Código de clasificación agrupada del tipo de vehículo
USO_AUTO	Tipo de uso que le dan al vehículo
COD_USO	Código del tipo de uso del vehículo
GRUPO_VEH_ORIGINAL	Clasificación del vehículo según su peso (Liviano o Pesado)
COD_GRUPO_VEH	Código de clasificación del vehículo según su peso

Fuente: Elaboración propia

Tabla 7. Base de datos de seguros de autos (inicial)

ID_CERTIFICA	GENERO_ASE	COD_GENERI	EDAD_ASEG	PROV_ASEGU	DEPARTAME	DISTRITO_C	COD_UBIGEO	TIPO_CLIENTI	COD_TIPO_C	COD_PROD	COD_TIPO_P	ANO_FABRIC	MARCA_VEH	COD_MARCA	TIPO_VEH	COD_TIPO_V	CLASE_AGRU	COD_CLASE_	USO_AUTO	COD_USO	GRUPO_VEH	COD_GRUPO_VEH
321777651	NO APLICA	0	32	LIMA	LIMA	LIMA	1	EMPRESA	1	VM14	3	1947	MACK	29	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
332522906	NO APLICA	0	32	LIMA	LIMA	LIMA	1	EMPRESA	1	VM14	3	1947	MACK	29	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
325755111	NO APLICA	0	32	LIMA	LIMA	LIMA	1	EMPRESA	1	VM14	3	1947	MACK	29	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
328425889	NO APLICA	0	32	LIMA	LIMA	LIMA	1	EMPRESA	1	VM14	3	1947	MACK	29	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
330986761	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1960	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
322883732	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1960	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
326908383	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1960	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
318906475	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1960	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
319396800	NO APLICA	0	32	LIMA	LIMA	SAN JUAN DE	62	EMPRESA	1	VM14	3	1961	VOLKSWAGE	6	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
323377528	NO APLICA	0	32	LIMA	LIMA	SAN JUAN DE	62	EMPRESA	1	VM14	3	1961	VOLKSWAGE	6	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
327412281	NO APLICA	0	32	LIMA	LIMA	SAN JUAN DE	62	EMPRESA	1	VM14	3	1961	VOLKSWAGE	6	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
331481030	NO APLICA	0	32	LIMA	LIMA	SAN JUAN DE	62	EMPRESA	1	VM14	3	1961	VOLKSWAGE	6	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
318906609	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1962	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
322883865	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1962	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
330986888	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1962	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
326908584	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1962	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
320678807	MASCULINO	1	34	LIMA	LIMA	MAGDALENA	42	PERSONA	2	VM12	1	1963	VOLKSWAGE	6	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
324670149	MASCULINO	1	34	LIMA	LIMA	MAGDALENA	42	PERSONA	2	VM12	1	1963	VOLKSWAGE	6	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
330166931	MASCULINO	1	34	LIMA	LIMA	MAGDALENA	42	PERSONA	2	VM12	1	1963	VOLKSWAGE	6	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
334265548	MASCULINO	1	35	LIMA	LIMA	MAGDALENA	42	PERSONA	2	VM12	1	1963	VOLKSWAGE	6	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
318907203	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1963	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
330987490	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1963	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
322884462	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1963	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
326909119	NO APLICA	0	32	CALLAO	CALLAO	VENTANILLA	56	EMPRESA	1	JURI	2	1963	KENWORTH	40	CARROCERIA	7	CAMIONES	5	COMERCIAL	2	PESADOS	2
332800765	MASCULINO	1	63	LIMA	LIMA	SAN ISIDRO	2	PERSONA	2	VM12	1	1964	CHEVROLET	11	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
326063354	MASCULINO	1	63	LIMA	LIMA	SAN ISIDRO	2	PERSONA	2	VM12	1	1964	CHEVROLET	11	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
328701853	MASCULINO	1	63	LIMA	LIMA	SAN ISIDRO	2	PERSONA	2	VM12	1	1964	CHEVROLET	11	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1
322047822	MASCULINO	1	63	LIMA	LIMA	SAN ISIDRO	2	PERSONA	2	VM12	1	1964	CHEVROLET	11	SEDAN	4	AUTOMOVILI	3	PARTICULAR	1	LIVIANOS	1

Fuente: Empresa de Seguros de Autos

Tabla 8. Ejemplo de equivalencias de variables descriptivas a numéricas

GENERO_ASEGURADO	COD_GENERO
NO APLICA	0
MASCULINO	1
FEMENINO	2

TIPO_CLIENTE_ASEGUR	COD_TIPO_CLIENTE
EMPRESA	1
PERSONA	2
GOBIERNO	3

COD_PROD	COD_TIPO_POL
VM12	1
JURI	2
VM14	3
CORE	4
TRAS	5

PROV_ASEGURADO	DEPARTAMENTO_C	DISTRITO_C	COD_UBIGEO
LIMA	LIMA	LIMA	1
LIMA	LIMA	SAN ISIDRO	2
LIMA	LIMA	SAN BORJA	3
LIMA	LIMA	SANTIAGO DE SURCO	4
LIMA	LIMA	ATE VITARTE	5
LIMA	LIMA	SAN MARTIN DE PORRES	6
CHICLAYO	LAMBAYEQUE	CHICLAYO	7
TALARA	PIURA	EL ALTO	8
SULLANA	PIURA	SULLANA	9
CUSCO	CUSCO	SAN SEBASTIAN	10
CALLAO	PIURA	PAITA	11
TALARA	PIURA	PARIÑAS	12
HUANCAYO	LIMA	SAN ISIDRO	13
LIMA	LIMA	MIRAFLORES	14
LIMA	LIMA	VILLA MARIA DEL TRIUNFO	15
AREQUIPA	AREQUIPA	AREQUIPA	16
LIMA	LIMA	BARRANCO	17
LIMA	LIMA	SANTA ANITA	18

Fuente: Elaboración propia

5.1.1.2 Preprocesamiento y Limpieza de datos

Como se mencionó anteriormente la base de datos cuenta con un total de 23 variables entre numéricas y descriptivas, tiene un total de 613,307 filas que pasaron por una etapa de limpieza y adaptación de valores.

En esta etapa se realizaron las siguientes actividades:

- Eliminación o corrección de campos nulos, incompletos o vacíos.
- Corrección de datos erróneos.
- Conversión de variables descriptivas en numéricas, por ejemplo para la variable género que cuenta con 3 valores estandarizados como: Masculino se le colocó un valor1, Femenino se le colocó un valor 2 y No Aplica que los clientes o asegurados del tipoEmpresa se le colocó un valor de 0.
- Revisión de las 23 variables, se evaluó cuáles eran necesarias o importantes para el modelo.

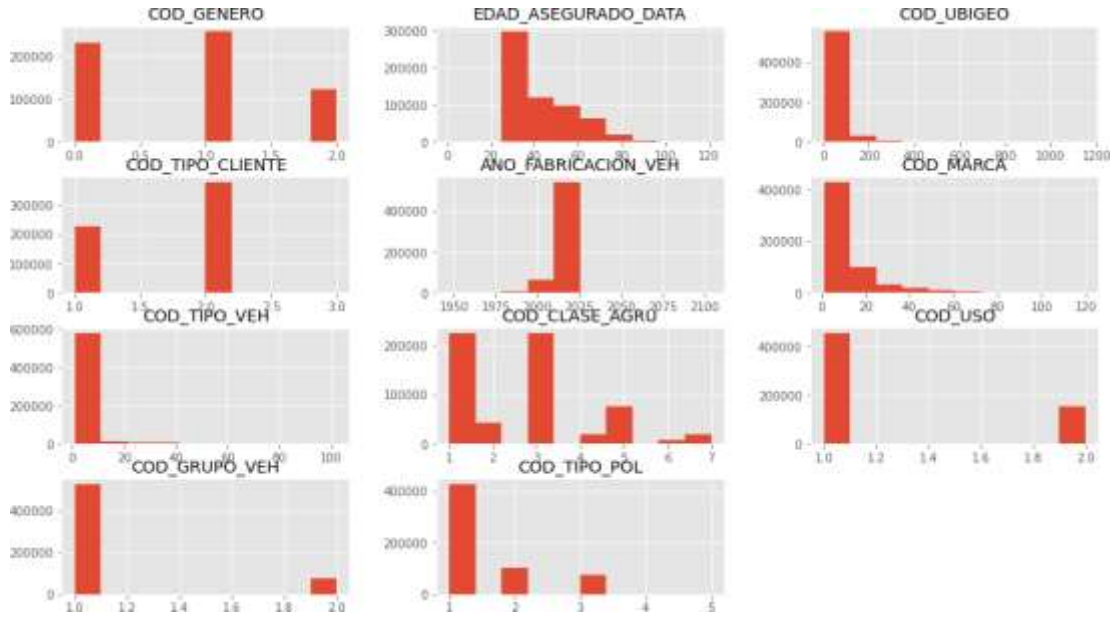
Como resultado final se obtuvo un total de 11 variables y 607,301 filas para elaborar la predicción del modelo de acuerdo con los criterios que afecten de manera directa el cálculo de la póliza a ofrecer a los clientes (marca de automóvil, edad del asegurado, si es empresa o persona natural, entre otros).

Tabla 9. Base de datos de seguros de autos (luego de limpieza y pre-procesamiento)

COD_GENERO	EDAD_ASEGURADO_DATA	COD_UBIGEO	COD_TIPO_CLIENTE	ANO_FABRICACION_VEH	COD_MARCA	COD_TIPO_VEH	COD_CLASE_AGRU	COD_USO	COD_GRUPO_VEH	COD_TIPO_POL
0	32	1	1	1947	29	7	5	2	2	3
0	32	1	1	1947	29	7	5	2	2	3
0	32	1	1	1947	29	7	5	2	2	3
0	32	1	1	1947	29	7	5	2	2	3
0	32	56	1	1960	40	7	5	2	2	2
0	32	56	1	1960	40	7	5	2	2	2
0	32	56	1	1960	40	7	5	2	2	2
0	32	56	1	1960	40	7	5	2	2	2
0	32	62	1	1961	6	4	3	1	1	3
0	32	62	1	1961	6	4	3	1	1	3
0	32	62	1	1961	6	4	3	1	1	3
0	32	62	1	1961	6	4	3	1	1	3
0	32	56	1	1962	40	7	5	2	2	2
0	32	56	1	1962	40	7	5	2	2	2
0	32	56	1	1962	40	7	5	2	2	2
0	32	56	1	1962	40	7	5	2	2	2
1	34	42	2	1963	6	4	3	1	1	1
1	34	42	2	1963	6	4	3	1	1	1
1	34	42	2	1963	6	4	3	1	1	1
1	35	42	2	1963	6	4	3	1	1	1
0	32	56	1	1963	40	7	5	2	2	2
0	32	56	1	1963	40	7	5	2	2	2
0	32	56	1	1963	40	7	5	2	2	2
0	32	56	1	1963	40	7	5	2	2	2

Fuente: Elaboración propia

Figura 17. Distribución de variables



Fuente: Elaboración propia

5.1.1.3 Modelamiento

Para la etapa de modelamiento en el presente trabajo, se hace uso de dos algoritmos, que nos permitirá resolver el problema de clasificación de clientes planteado para la empresa de seguros de autos.

Aplicación del algoritmo K-NN

Como se mencionó anteriormente, para aplicar el algoritmo K-NN se hizo uso de la función *KNeighborsClassifier*, la cual cuenta con los parámetros descritos en la siguiente figura. Dentro del modelamiento se experimentó con diferentes valores del parámetro *n_neighbors* para determinar el modelo con mayor grado de precisión.

Figura 18. Parámetros de KNeighborsClassifier

class sklearn.neighbors.KNeighborsClassifier(*n_neighbors=5, *, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None*) [source]

Fuente: Elaboración propia

Inicialmente, se utilizó la parametrización configurada por defecto en la herramienta, donde k = 5.

Figura 19. Código de prueba inicial (k = 5)

Fuente: Elaboración propia

Se hicieron un total de 30 experimentos con distintos valores de k, en la siguiente figura se muestra el código de algunas de las pruebas realizadas y en la Tabla 10 se muestran los resultados obtenidos en cada una de las pruebas:

Figura 20. Código de pruebas para diferentes valores de k

Fuente: Elaboración Propia

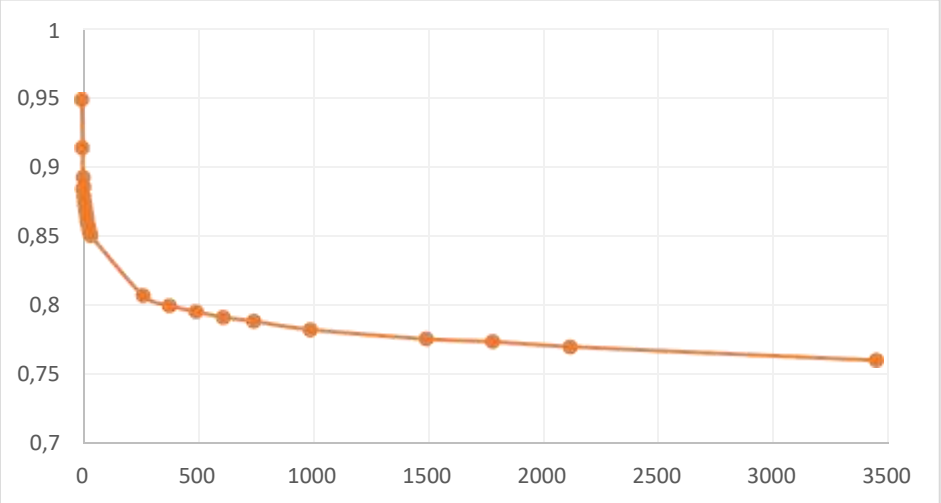
Tabla 10. Resultados de pruebas con KNeighborsClassifier

K	ACCURACY
3	0,9486
5	0,9136
7	0,8838
9	0,8925
11	0,8856
13	0,8786
15	0,8751
17	0,8727
19	0,8696
21	0,8672
23	0,865
25	0,8631
27	0,8606
29	0,8592
31	0,8577
33	0,8561
35	0,854
37	0,8528
39	0,8519
41	0,8503
269	0,807
383	0,7998
499	0,7953
619	0,7913
751	0,7884
997	0,7824
1501	0,7756
1789	0,7737
2125	0,7701
3457	0,7602

Fuente: Elaboración propia

De manera preliminar se puede concluir que el modelo de K-NN con parámetro k=3 es el que tiene un mayor nivel de accuracy.

Figura 21. Nivel de accuracy en cada prueba



Fuente: Elaboración propia

Aplicación de Regresión Logística

Como se mencionó anteriormente, para aplicar la regresión logística se hizo uso de la función *Logistic Regression*, la cual cuenta con los parámetros descritos en la siguiente figura. Debido a que el problema de clasificación planteado es multiclase, se definió el parámetro multi_class = 'multinomial'. Asimismo, se experimentaron con diferentes valores de C para determinar el modelo con mayor grado de precisión

Figura 22. Parámetros de Logistic Regression

class sklearn.linear_model.LogisticRegression(*penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None*) [source]

Elaboración propia

Figura 23. Código de prueba inicial (C = 1)

```
In [56]:
1 modeloRLog5 = LogisticRegression(C = 1, multi_class = 'multinomial')
2
3 modeloRLog5.fit(X_train,Y_train)
4
5 resultadoRLog5 = modeloRLog5.predict(X_test)
6
7 accuracy_score(Y_test, resultadoRLog5)

Out[56]: 0.7137435061460057
```

Fuente: Elaboración propia

Tabla 11. Resultados de pruebas con LogisticRegression

K	ACCURACY
0.0005	0.7133
0.005	**0.7146**
0.05	0.7115
1	0.7137
10	0.7130
100	0.7114

Fuente: Elaboración propia

De manera preliminar se puede concluir que el modelo de regresión logística con parámetro C=0.005 es el que tiene un mayor nivel de accuracy.

5.2 Medición de la solución

5.2.1 Análisis de Indicadores cuantitativo y/o cualitativo

En esta etapa de la investigación se comparan los resultados obtenidos en los diferentes realizados

5.2.1.1 K-NN

Tabla 12. Resultados de pruebas con KNeighborsClassifier

K	ACCURACY	K	ACCURACY	K	ACCURACY
3	0,9486	23	0,865	269	0,807
5	0,9136	25	0,8631	383	0,7998
7	0,8838	27	0,8606	499	0,7953
9	0,8925	29	0,8592	619	0,7913
11	0,8856	31	0,8577	751	0,7884
13	0,8786	33	0,8561	997	0,7824
15	0,8751	35	0,854	1501	0,7756
17	0,8727	37	0,8528	1789	0,7737
19	0,8696	39	0,8519	2125	0,7701
21	0,8672	41	0,8503	3457	0,7602

Fuente: Elaboración Propia

Se hizo uso del algoritmo K-NN con un total de 30 valores, que ha permitido optimizar el valor del resultado. El parámetro n_neighbors en la estructura del algoritmo ha permitido encontrar los elementos más cercanos, que han coincidido con el modelo de entrenamiento. Por lo que se encontró que los mejores valores para k según accuracy son de 3 y 5 con una precisión de 94.86% y 91.36% respectivamente para los resultados de prueba.

No obstante, debido a que todos los entrenamientos se han realizado con el mismo grupo de datos, se puede incurrir en sobreajuste. Esto puede provocar que el modelo pueda tener un alto nivel de acierto con el grupo de entrenamiento actual, pero reducir considerablemente dicho nivel si se utilizan otros datos. En este sentido, la validación cruzada es una forma de evitar el riesgo de sobreajuste ya que los datos de prueba van cambiando con cada iteración realizada.

Asimismo, debido a que los datos no están balanceados, el accuracy no es 100% fiable en este caso. Por ese motivo es importante utilizar otro indicador como el puntaje F1 para asegurar la robustez del modelo. A continuación, se muestran los resultados obtenidos.

Tabla 13. Métricas de evaluación adicionales (K-NN)

K	Accuracy	Cross Validation	F1	K	Accuracy	Cross Validation	F1
3	0,9486	0,7349	0,9158	33	0,8561	0,7923	0,8453
5	0,9136	0,7362	0,9499	35	0,8540	0,7912	0,8439
7	0,8838	0,7380	0,8819	37	0,8528	0,7917	0,8424
9	0,8925	0,7734	0,8891	39	0,8519	0,7925	0,8404
11	0,8856	0,7740	0,8818	41	0,8503	0,7930	0,8393
13	0,8786	0,7747	0,8728	269	0,8070	0,7815	0,7835
15	0,8751	0,7754	0,8675	383	0,7998	0,7775	0,7741
17	0,8727	0,7847	0,8662	499	0,7953	0,7748	0,7688
19	0,8696	0,7849	0,8621	619	0,7913	0,7732	0,7618
21	0,8672	0,7861	0,8589	751	0,7884	0,7707	0,7564
23	0,8650	0,7870	0,8567	997	0,7824	0,7686	0,7479
25	0,8631	0,7900	0,8538	1501	0,7756	0,7644	0,7381
27	0,8606	0,7902	0,8512	1789	0,7737	0,7612	0,7339
29	0,8592	0,7905	0,8493	2125	0,7701	0,7589	0,7288
31	0,8577	0,7907	0,8471	3457	0,7602	0,7518	0,7097

Fuente: Elaboración Propia

En base a los datos obtenidos, se concluye que el modelo de K-NN con parámetro k=41 es el que tiene un mayor nivel de accuracy tomando un enfoque de cross validation; asimismo, presenta un puntaje F1 aceptable lo que demuestra la precisión del modelo.

5.2.1.2 Regresión logística

De igual manera que con el algoritmo anterior, se validan los resultados obtenidos con métricas adicionales como el puntaje F1 y la herramienta de cross validation.

Tabla 14. Métricas de evaluación adicionales (Regresión Logística)

K	Accuracy	Cross Validation	F1
0.0005	0,7133	0,6531	0,6367
0.005	0,7146	0,6622	0,6375
0.05	0,7115	0,6535	0,6326
1	0,7137	0,6494	0,6355
10	0,7130	0,5920	0,6349
100	0,7114	0,5645	0,6321

Fuente: Elaboración Propia

En este caso, se concluye que el modelo de regresión logística con parámetro C=0.005 es el que tiene un mayor nivel de accuracy, tomando un enfoque de cross validation. Sin embargo, el puntaje F1 es de 0.6367 por lo que el modelo no es lo suficientemente robusto en comparación de los modelos antes presentados.

Luego de comparar los resultados, se concluye que el modelo con mayor nivel de robustez y precisión es el modelo K-NN que utiliza un parámetro de k = 41.

5.2.2 Simulación de solución. Aplicación de Software.

5.2.2.1 Importación de información

Figura 24. Paso 1: Importación de librerías y algoritmos a utilizar

```
In [ ]:
import pandas as pd
import numpy as np

#Supervisado - Clasificación
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

#Para KNN
from sklearn.neighbors import KNeighborsClassifier

##Para Regresión Linea y Logistica
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LogisticRegression, LinearRegression
```

Fuente: Elaboración Propia

Figura 25. Paso 2: Carga de base de datos

```
In [2]:  1 dataAutos = pd.read_excel('PlantillaAutosCarga.xlsx')

In [3]:  1 dataAutos
```

Out[3]:

	COD_GENERO	EDAD_ASEGURADO_DATA	COD_UBIGEO	COD_TIPO_CLIENTE	ANO_FABRICACION_VEH	COD_MA
0	0	32	1	1	1947	
1	0	32	1	1	1947	
2	0	32	1	1	1947	
3	0	32	1	1	1947	
4	0	32	56	1	1960	
...	...	...	...	...	...	
607296	0	32	5	1	2104	
607297	0	32	5	1	2104	
607298	0	32	5	1	2104	
607299	0	32	5	1	2104	
607300	0	32	5	1	2104	

607301 rows × 11 columns

Fuente: Elaboración Propia

5.2.2.2 Pre-procesamiento de información

Figura 26. Paso 3: Definición de variables independientes y variable dependiente

```
In [4]:  1 #Variables Independientes:
         2 x = dataAutos.values[:,:-1]
         3 x

Out[4]: array([[ 0, 32,  1, ...,  5,  2,  2],
               [ 0, 32,  1, ...,  5,  2,  2],
               [ 0, 32,  1, ...,  5,  2,  2],
               ...,
               [ 0, 32,  5, ...,  5,  2,  2],
               [ 0, 32,  5, ...,  5,  2,  2],
               [ 0, 32,  5, ...,  5,  2,  2]], dtype=int64)

In [5]:  1 #Variables Dependiente:
         2 y = dataAutos['COD_TIPO_POL'].values
         3 y

Out[5]: array([3, 3, 3, ..., 2, 2, 2], dtype=int64)
```

Fuente: Elaboración Propia

Figura 27. Paso 4: Definición de set de entrenamiento y set de prueba

```
In [6]:  1 X_train, X_test, Y_train, Y_test = train_test_split(x,y,test_size=0.2)

In [7]:  1 X_train.shape

Out[7]: (485840, 10)
```

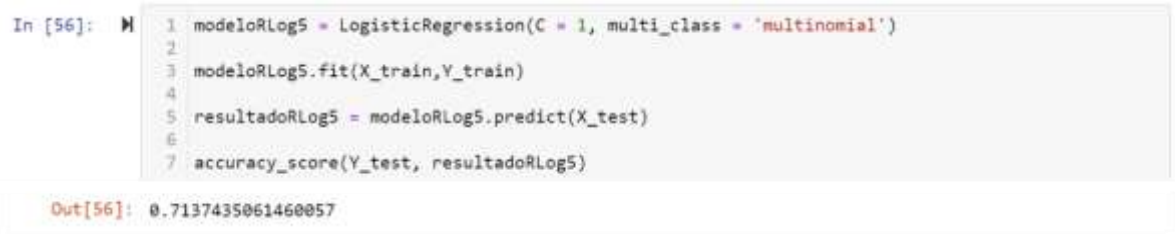
Fuente: Elaboración Propia

5.2.2.3 Aplicación de algoritmos

Figura 28. Paso 5: Aplicación de K-NN

Fuente: Elaboración Propia

Figura 29. Paso 6: Aplicación de Regresión Logística



Fuente: Elaboración Propia

5.2.2.3 Evaluación de modelos

Figura 30. Paso 7: Calcular puntaje F1 para cada modelo

Fuente: Elaboración Propia

Figura 31. Paso 8: Calcular accuracy por medio de Cross Validation

```
In [113]:
1 resultado21=cross_val_score(modeloKNN21, x, y, cv=10)
2 print("Avg accuracy: {}".format(resultado21.mean()))

Avg accuracy: 0.7815119927420946

In [114]:
1 resultado22=cross_val_score(modeloKNN22, x, y, cv=10)
2 print("Avg accuracy: {}".format(resultado22.mean()))

Avg accuracy: 0.7775485572323382

In [115]:
1 resultado23=cross_val_score(modeloKNN23, x, y, cv=10)
2 print("Avg accuracy: {}".format(resultado23.mean()))

Avg accuracy: 0.7748118584544055

In [116]:
1 resultado24=cross_val_score(modeloKNN24, x, y, cv=10)
2 print("Avg accuracy: {}".format(resultado24.mean()))

Avg accuracy: 0.7732475662347479
```

Fuente: Elaboración Propia

CAPÍTULO VI: CONCLUSIONES Y RECOMENDACIONES

6.1 Conclusiones

Los modelos que utilizan Machine Learning tienen la capacidad de "aprender" conocimiento a partir de entrenamientos con información existente. Estos modelos son altamente flexibles lo que ha permitido que puedan desarrollar diferentes aplicaciones para diferentes entornos y sectores.

El mercado de seguros peruano se desenvuelve en un entorno altamente competitivo e impredecible por lo que es importante desarrollar herramientas que permitan capitalizar las fortalezas y oportunidades, y mitigar las debilidades y riesgos. En este sentido, las herramientas de Machine Learning tienen el potencial de generar eficiencias dentro de la empresa y fomentar el desarrollo de ventajas competitivas que impacten positivamente en el desempeño general de la compañía.

En este trabajo se ha propuesto utilizar las técnicas de Machine Learning para categorizar de manera automática los clientes de seguros vehiculares de una aseguradora. Con este propósito, se recolectó información de una compañía aseguradora peruana, obteniendo una base de datos de 179,053 entradas y 23 variables. Luego de la limpieza y pre – procesamiento de la información, se redujo la base a 607,301 entradas y 11 variables; los cuales fueron utilizados para entrenar el algoritmo. Debido a las características del problema, se optó por desarrollar un modelo de clasificación (aprendizaje supervisado), para lo cual se utilizaron 2 algoritmo: K-NN y Regresión Logística.

Finalmente, luego de evaluar ambos modelos, se concluyó que el mejor modelo para este problema es el modelo desarrollado con el algoritmo K-NN, utilizando el parámetro k = 41. Para llegar a esta conclusión se utilizaron diferentes métricas: en primer lugar, se utilizó el *Accuracy* para evaluar el desempeño del modelo; sin embargo, para asegurar que no se genere un sobreajuste, se utilizó la función *Cross Validation* para calcular un nuevo *Accuracy* que mida el nivel de precisión en diferentes grupos de prueba. Por último, se utilizó la función *F1 Score*, la cual evalúa el nivel de robustez del modelo ya que cuando las clases en la base de datos no están balanceadas, el indicador de *Accuracy* no es completamente fiable.

6.2 Recomendaciones

Debido al dinamismo del mercado de seguros, es importante hacer un estudio más extenso que involucre más datos y variables pertinentes para mejorar el nivel de precisión y robustez del modelo. Asimismo, sería relevante evaluar distintas técnicas de limpieza de datos a través de algoritmos de Machine Learning para asegurar la validez de los datos ya que cada empresa de seguros tiene una distinta manera de trabajar su data y utilizan distintos factores para la clasificación resultante.

Para los investigadores que quisieran ahondar más en esta investigación, se recomienda evaluar otras técnicas de Machine Learning como el Support Vector Machines (SVM) o técnicas aún más complejas como la estructura de Redes Neuronales Recurrentes (Deep Learning). Bajo este escenario, es necesario recolectar más datos de la empresa en investigación.

La empresa puede desarrollar una estrategia de marketing personalizado utilizando los resultados del modelo propuesto. De esta manera, se puede lograr un mayor nivel de ventas y fidelización por parte de los clientes.

BIBLIOGRAFÍA

Ana González-Marcos, F. A.-E. (2017). MACHINE LEARNING EN LA INDUSTRIA: EL CASO DE LA SIDERURGIA. La Rioja: Universidad de La Rioja.

Asociación Peruana de Empresas de Seguros - APESEG (2019). Informe Trimestral del Sistema Asegurador - Cuarto Trimestre 2019. https://www.apeseg.org.pe/wp-content/uploads/2020/02/Resultados_Sistema_Asegurador_4T19.pdf

Asociación Peruana de Empresas de Seguros - APESEG (2020). Informe Trimestral del Sistema Asegurador - Segundo Trimestre 2020. https://www.apeseg.org.pe/wp-content/uploads/2020/08/Resultados_Sistema_Asegurador_2T20.pdf

Bian, Y., Yang, C., Zhao, J. L., & Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models. In Transportation Research Part A: Policy and Practice (Vol. 107, pp. 20–34). Elsevier BV. https://doi.org/10.1016/j.tra.2017.10.018

Chapados, N., Bengio, Y., Vincent, P., Ghosn, J., Dugas, C., Takeuchi, I., Meng, L. (2001). Estimating Car Insurance Premia: a Case Study in High-Dimensional Data Inference. Advances in Neural Information Processing Systems.

Hanafy, M., & Ming, R. (2021). Machine Learning Approaches for Auto Insurance Big Data. En Risks (Vol. 9, Issue 2, p. 42). MDPI AG. https://doi.org/10.3390/risks9020042

Harrington, P. (2012). Machine learning in action. Simon and Schuster.

Lengua, C. (16 de Febrero de 2021). Financiero y Seguros fue el sector que más creció el 2020: ¿cómo les fue a las empresas aseguradoras? El Comercio, págs. https://elcomercio.pe/economia/peru/financiero-y-seguros-fue-el-sector-que- mas-crecio-el-2020-como-le-fue-a-las-empresas-aseguradoras-ncze- noticia/?ref=ecr.

Marcos, A. G., & Elías, F. A. (2017). Machine Learning en la industria: el caso de la siderurgia. Economía industrial, (405), 55-63.

Paruchuri, H. (2020). The Impact of Machine Learning on the Future of the Insurance Industry. American Journal of Trade and Policy, 7(3), 85-90. https://doi.org/10.18034/ajtp.v7i3.537

Superintendencia de Banca, Seguros y AFP (2019, Dic). Boletín de Seguros Mensual – Reporte de Diciembre 2019 - P005. https://www.sbs.gob.pe/app/stats/estadisticaboletinestadistico.asp?p=25#

Vadlamudi, S. (2019).How Artificial Intelligence Improves Agricultural Productivity and Sustainability: A Global Thematic Analysis. Asia Pacific Journal of Energy and Environment,6(2), 91-100. https://doi.org/10.18034/apjee.v6i2.542

Vassiljeva, K., Tepljakov, A., Petlenkov, E., & Netšajev, E. (2017, May). Computational intelligence approach for estimation of vehicle insurance risk level. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 4073-4078). IEEE.