Revista Escuela de Administración de Negocios N° 94- enero a junio de 2023
ISSN: 0120-8160 (Impreso) - ISSN: 2590-521X (En línea)

Medición de *lapses* en seguros de vida mediante modelos de predicción

Yenni Paola Zamora Puentes [1]
Universidad Sergio Arboleda
yennipzamora@gmail.com

Carlos Arturo Peña Rincón [2]
Universidad Sergio Arboleda
carlos.pena@usa.edu.co

Hermes J. Martínez Navas [3]
Values AAA, Banca de Inversión
hermes.martinez@valuesaaa.com

DOI:

Fecha de recepción: 13 de abril de 2023
Fecha de aprobación: 15 de mayo de 2023

Cómo citar este artículo: Zamora Puentes, Y.P.; Peña Rincón, C.A.; Martínez Navas, H.J. (2023). Medición de *lapses* en seguros de vida mediante modelos de predicción. *Revista Escuela de Administración de Negocios*, (94), (páginas). DOI:

Resumen

Las compañías de seguros están en constante medición de sus persistencias, por ello, identificar y analizar las tasas de cancelación denominadas *lapses* o tasas de caducidad, se ha convertido en una actividad de gran importancia debido a su rol determinante para tomar decisiones administrativas y financieras. Entender la dinámica de esta variable facilita la toma de decisiones, y permite identificar las variables que ocasionan cancelaciones de pólizas, es decir: el género, la edad, la ciudad, el tipo de producto, entre otros. Estas variables caracterizan el perfil del asegurado y condicionan una mayor o menor probabilidad de cancelar la póliza. Para analizar los perfiles de asegurados, se consideró utilizar modelos de regresión logística, redes neuronales y máquinas de soporte vectorial, con precisión de 73 %, 81,53 % y 60 %, respectivamente, mediante una base de datos de asegurados del mercado de Colombia con 134 102 registros, con 8 variables, lo que permitió predecir la probabilidad de cancelación y de renovación de una póliza de seguro de vida de acuerdo con las condiciones de las variables que perfilan a un asegurado.

Palabras clave: *lapses*; regresión logística; red neuronal; máquina de soporte vectorial.

****

[1] Magíster en Matemáticas Aplicada. Universidad Sergio Arboleda. Especialista en Actuaria. Universidad Antonio Nariño. Especialista en Matemáticas Aplicadas. Universidad Sergio Arboleda. Licenciada en Matemáticas. Universidad Distrital Francisco José de Caldas. ORCID: https://orcid.org/0009-0006-0106-7022
[2] Doctor en Ingeniería. Universidad Nacional de Colombia. Magíster en Gestión y Evaluación Ambiental. Universidad Sergio Arboleda. Especialista en Matemáticas Aplicada. Universidad Sergio Arboleda. Especialización Gerencia de Medio Ambiente y Prevención de Desastres. Universidad Sergio Arboleda. Físico. Universidad Nacional de Colombia. ORCID: https://orcid.org/0000-0001-9818-3033
[3] Doctor en Matemáticas. Universidad Bonn. Magíster en Matemáticas. Universidad de los Andes. Matemático. Universidad Nacional de Colombia. ORCID: https://orcid.org/0000-0001-7816-1128

Este artículo es un *online first* (versión definitiva del contenido del artículo, con diagramación provisional y asignación de DOI). Aún no cuenta con páginas definitivas, pero es citable utilizando su código DOI.

Measurement of *lapses* in life insurance by means of prediction models

Abstract
Insurance companies are constantly measuring their persistencies, therefore identifying and analyzing cancellation rates, known as lapses or lapse rates, has become an activity of great importance due to its determining role in making administrative and financial decisions. In addition, knowing their dynamics allows direct actions to be taken on business or variables that cause policy cancellations. However, lapses are sensitive to the variables that accompany the insurance contract, that is. gender, age, city, type of product, among others. Because they characterize the profile of the insured and condition a greater or lesser probability of canceling the policy. To analyze the policyholder profiles, logistic regression models, neural networks and support vector machines were used, with 73 %, 81,53 % and 60 % accuracy respectively, using a database of policyholders in the Colombian market with 134 102 records with 8 variables, which allowed to predicting the probability of cancellation and renewal of a life insurance policy according to the conditions of the variables that profile a policyholder.

Keywords: *Lapses*; logistic regression; neural network; support vector machine.

****

1. Introducción

Los seguros de vida tienen como finalidad la protección ante el riesgo de muerte, incapacidad o sobrevivencia. En este mercado, los asegurados realizan un pago periódico a una compañía aseguradora, este pago corresponde a una prima, la cual es asignada según el perfil del asegurado, el riesgo y el tipo de cobertura; en caso de materializarse un riesgo en las condiciones pactadas en el contrato de póliza, la aseguradora deberá pagar o reembolsar los valores cubiertos al usuario hasta el término del contrato, o a terceros beneficiarios en caso de fallecimiento del tomador de la póliza.

Las compañías de seguros se enfrentan a un reto constante: comprender y predecir el comportamiento de renovación y cancelación de pólizas por parte de sus asegurados (Kuo *et al*., 2003). Este comportamiento, conocido como *lapse*, es provocado por los efectos de riesgos propios del mercado, a causa de razones conductuales que son un reto para los modelos actuariales (Bauer *et al*., 2017). Adicionalmente, esta variable tiene una alta incidencia en la provisión o reserva de una compañía, apalancando por sí solo cerca de la mitad del capital de la compañía de seguros, estas provisiones son establecidas según

regulación y corresponden a las reservas mínimas que debe tener la compañía, con el fin de atender los riesgos relativos al mercado de los seguros de vida (Weindorfer, 2012).

Los *lapses* permiten ver la frecuencia en que los asegurados cancelan sus pólizas, esto afecta de manera directa los resultados de las tasas altas de cancelación, y dependiendo de su impacto logran crear perturbaciones en la posición financiera de la compañía, en los esquemas de comisión, gastos de adquisición, presupuesto y el flujo de caja.

La comprensión del *lapse* en los seguros de vida requiere de un enfoque multidimensional, que combine el análisis de datos históricos, la creación de modelos de *marketing* financiero y económicos, para ello, se requiere de la aplicación de modelos de caducidad, la utilización de herramientas analíticas como los modelos máquinas de aprendizajes, basados en agentes y las dinámicas conductuales. Al adoptar este enfoque integral, las compañías de seguros pueden desarrollar estrategias efectivas para reducir el *lapse*, proteger su rentabilidad y fortalecer su posición competitiva en el mercado.

Existen modelos de caducidad que se enfocan en los perfiles de asegurados de inversión racional y neutral al riesgo; que han permitido la implementación de ecuaciones diferenciales parciales, Cox-Ross-Rubinstein y modelos binomiales. No obstante, estos modelos se enfrentan a dificultades a nivel de información y procesos de decisión individuales o colectivos (Eling *et al*., 2013). Por otro lado, los modelos basados en agentes pretenden simular el comportamiento de los asegurados mediante datos históricos, cambios en las condiciones ambientales, la interacción con otros agentes y su entorno para diseñar escenarios presentes y futuros en el sector de los seguros (Bauer *et al*., 2017; Fier y Liebenberg, 2013; Ramchandani *et al*., 2017).

Adicionalmente, la metodología de regresión logística se ha utilizado en el sector empresarial de seguros en Europa para estimar la probabilidad de cancelación de pólizas y contribuir a la medición del riesgo empresarial (Guillén *et al*., 2011).

El uso de técnicas de minería de datos y métodos de aprendizaje de máquina se ha implementado por compañías en la toma de decisiones. Es posible realizar perfilamiento de

clientes utilizando métodos de agrupamiento como *K-means*, un método de segmentación de clientes, la cual consiste en dos fases. La primera, el agrupamiento de los clientes en diferentes segmentos con respecto al modelo de "recencia, frecuencia, monto", en inglés: *recency, frequency and monetary* (RFM), que agrupa clientes con comportamientos similares de compra (Heldt *et al*., 2021). Luego, utilizando datos demográficos en cada grupo, se genera un reagrupamiento para el cliente mediante el valor de por vida (Namvar *et al.,* 2010).

Fier y Liebenberg (2013), desarrollaron un modelo microeconómico para estudiar los *lapses* en seguros de vida, destacando factores clave que afectan la decisión de los asegurados de terminar sus pólizas. Entre estos factores se incluyen los ingresos y la edad, siendo este último más influyente entre los jóvenes y menos relevante en los adultos. El análisis, realizado durante un periodo de 12 años, evidenció cómo las variaciones económicas en los hogares están asociadas con los *lapses*.

En un caso de estudio que abordó características idiosincráticas de los asegurados, como el género, la edad, la duración de la póliza y el hábito de fumar; los resultados indicaron que la población de fumadores presenta una propensión mayor a los *lapses* en comparación con los no fumadores. Como resultado, las compañías en América del Norte optaron por incorporar la pregunta sobre el hábito de fumar a los solicitantes de los seguros de vida, y este factor se consideró en las proyecciones de riesgos según el Canadian Institute of Actuarie (2014).

Otros factores relevantes en los seguros de vida están influenciados por las condiciones macroeconómicas del mercado asegurador, tales como la diferencia entre las tasas de interés del mercado, los tipos de interés contractuales, la tasa de inflación anual, el impacto bursátil y factores demográficos, como el crecimiento poblacional, la esperanza de vida, factores sociales y culturales, el nivel de educación, entre otras; lo anterior, hace parte del conjunto de variables exógenas y las variables idiosincráticas de los asegurados y que tiene alta incidencia sobre los *lapses* (Outreville, 1990; 2013).

Kiesenbauer Dieter (2012), estudió 5 categorías de productos mediante regresión logística para determinar indicadores económicos en 133 aseguradoras en Alemania, identificando factores de caducidad semejantes en todas sus categorías, donde la dinámica del *lapse*

presentó una relación entre indicadores económicos y características inherentes a las empresas.

El *lapse* en los seguros de vida ha sido un tema de investigación académica desde la década de 1970, donde la mayoría de los estudios se han realizado en países desarrollados, limitando la comprensión del fenómeno en otras regiones del mundo. Además, muchos estudios se han enfocado exclusivamente en el impacto de las tasas de interés en el *lapse,* descuidando otros factores económicos e idiosincráticos relevantes**.**

Desde el Consejo de la Sección de Desarrollo de Productos y el Comité de Investigación de Seguros de Vida de la Sociedad de Actuarios (SOA), la compañía reaseguradora RGA ha reportado sobre su cartera de seguros de vida a largo plazo, destacando el efecto de las tasas de caducidad al término del periodo de la póliza. Los datos, que abarcan 26 compañías de seguros de vida con un producto único, sugieren que estos están centrados en seguros de vida individuales. Además, se ha observado que, en el lapso de las pólizas, las primas muestran variaciones de un periodo a otro, con una correlación significativa que indica que los productos de seguro emitidos más recientemente tienden a experimentar un mayor lapso, debido a variaciones en las tasas de las primas (Rozar *et al*., 2010).

Según Eling y Kochanski (2013), los factores económicos, junto con las bases de datos de asegurados, han brindado la oportunidad de conocer las causas, factores de caducidad de las pólizas de vida y la relación con los atributos de los productos de vida. Es un resultado a partir de una revisión de la literatura con un enfoque teórico y empírico. Además, al analizar la caducidad de los seguros, tanto los modelos teóricos e investigaciones empíricas, muestran dificultades a nivel de información, los procesos de toma de decisiones individuales y colectivos en los resultados para la predicción de tasas de caducidad, técnicas de mitigación de riesgo y la tendencia de *lapses* futuros.

Es importante destacar que, según Gottlieb y Smetters (2021), cuando ocurren cancelaciones de pólizas, el seguro que asume el *lapse* termina subsidiando a aquellos que no cancelan sus pólizas. Además, se observó que la mayoría de las pólizas de seguros de vida caducan antes de que concluya el periodo para el cual fueron contratadas. A su vez, los asegurados no tienen

en su agenda pagar las primas y, por ende, se alcanza el estado de cancelación en forma inmediata, mientras otros resultados indican que los asegurados cancelan la póliza porque consideran que es baja la probabilidad de ocurrencia y prefieren asumir el riesgo.

Es de interés en este trabajo identificar y evaluar con antelación los riesgos de caducidad, esto permite trazar planes de prevención para fortalecer la retención de los aseguradores (Pinquet *et al*., 2011; Tsai *et al*., 2009). Para ello, se han seleccionado, en primera instancia, las técnicas de regresión logística, las cuales permiten estimar la probabilidad de que ocurra el *lapse* de una póliza, basándose en una o más variables independientes, como la edad del asegurado, la duración de la póliza, los cambios en las tasas de primas, entre otros. Luego se especifica la metodología correspondiente a las máquinas de soporte vectorial (MSV), que consiste en un método de clasificación espacial que busca el hiperplano óptimo que separa las pólizas que terminan en *lapse* y las que no. Esto es útil cuando las relaciones entre las variables de la póliza y sus resultados son no lineales.

Posteriormente, las redes neuronales, que mediante modelos computacionales permiten realizar una estimación de *lapses* al modelar las iteraciones entre las variables, lo que ayuda a identificar patrones complejos. Estas técnicas fueron utilizadas en este trabajo para un perfil del asegurado con las siguientes variables: medio de distribución, tipo de seguro, ciudad, asesor, causal de cancelación, género, edad, prima, fuma y valor del asegurado.

Finalmente, se presenta la aplicación de modelos con máquinas de aprendizaje que permite analizar altos volúmenes de información, logrando disminuir tiempos de ejecución en la fase de análisis. Adicionalmente, al combinar el uso de máquinas de aprendizaje, como la regresión logística, máquinas de soporte vectorial y redes neuronales, permite a las compañías de seguros medir la probabilidad de cancelación y renovación de los seguros de vida.

2. Marco teórico

Existen dos líneas de estudio en seguros: la primera enfocada en riesgos propios a la vida y al bienestar de las personas denominados seguros de vida; la segunda, son aquellos riesgos enfocados en bienes materiales, a estos se les denomina seguros generales. En este trabajo se hace mención a los seguros de vida, los cuales, pueden ser tomados como:

2.1. Seguros de vida individual

Los seguros de vida individual hacen referencia a un contrato pactado entre una persona natural y una compañía de seguros denominado póliza, de acuerdo con Moreno (2019), estos contratos/pólizas pueden corresponder a 3 tipos de seguros: dotales, vida entera y temporales.

- **Seguros de vida dotales:** corresponden a seguros que cubren el riesgo por muerte, adicionalmente, funciona como un ahorro que puede ser recibido si la persona natural asegurada sobrevive k cantidad de años. Es decir, este tipo de seguro tiene cobertura en caso de muerte del asegurado y en caso de sobrevivir k años, entonces, el asegurado o beneficiario obtiene la indemnización por el valor pactado de la póliza.
- **Seguros de vida entera**: solo cubren el fallecimiento del asegurado, es decir, la prima se paga hasta determinada edad y la indemnización se hace efectiva sólo en el momento de su muerte.
- **Seguros temporales:** es pactado a un tiempo fijo acordado por las partes contractuales, sin embargo, este puede contener compromisos adicionales con la aseguradora donde se estipule un beneficio económico en caso de sobrevivir al tiempo acordado, pero, este no suele ser inferior al valor asegurado.

2.2. Seguros de vida colectivo

En los seguros de vida colectivo, las tasas y condiciones de la póliza suelen ser iguales para un grupo determinado de individuos que adquieren sus seguros mediante una única figura jurídica, tales como: sociedad limitada, sociedad anónima, sociedad sin ánimo de lucro, entre otras. Esta figura jurídica está en condiciones de adquirir un contrato con una aseguradora, para que sus asegurados tengan el beneficio económico que se pacte si el riesgo se materializa. Este tipo de seguro puede variar según los intereses particulares que marcan la pauta en el colectivo de los asegurados.

2.3. *Lapses*

Estas tasas representan la frecuencia con la que se cancela un contrato de seguro antes de que el riesgo se materialice o alcance su vencimiento. Según Rozar *et al*. (2010), las tasas de caducidad o *lapses* se refieren al momento de una terminación completa de la póliza, o pueden ser definidos como el último día por el que se pagó una prima. Se miden como la proporción de asegurados que cancelan la póliza en relación con el total de asegurados, abarcando a quienes la cancelan. Internamente, estas tasas de caducidad afectan la confiabilidad de la proyección del volumen de primas en las renovaciones y el comportamiento de nuevas primas por ventas. Esto contribuye a la elaboración de presupuestos más precisos y estrategias efectivas para futuros negocios de la compañía.

Un desafío permanente para las compañías de seguros es modelar el comportamiento de los tomadores de seguro de vida, porque la caducidad de las pólizas de vida debilita la rentabilidad y la liquidez, lo cual compromete la capacidad de las aseguradoras para asumir sus obligaciones a corto plazo. Una clasificación dada para los modelos de *lapses* se presenta en lo investigado por Bacinello (2005), la primera, de carácter determinístico, que considera factores externos como lo social, individual y económico; la segunda, dinámica de carácter estocástico, que toma factores como el estado de salud, el producto interno bruto y la tasa de desempleo mediante modelos estadísticos.

Los modelos dinámicos parten de la racionalidad del tomador del seguro y el tipo de producto de seguro de vida. La racionalidad da una clasificación a los modelos tales como: 1) la dinámica óptima para aquellos inversores racionales y neutrales al riesgo; 2) los inversores racionales con aversión al riesgo. Adicionalmente, están los modelos que incorporan la caducidad dinámica y determinista. Finalmente, dependiendo el tipo de producto, los seguros de vida se clasifican en las siguientes categorías: productos tradicionales, productos vinculados a fondos de inversión y rentas vitalicias variables (Eling y Kochanski, 2013).

Otra clasificación es por sus categorías. La primera, considera la tasa de desempleo, la tasa de interés y la tasa de crecimiento económico (Kiesenbauer 2012; Kim 2005). La segunda, contempla el producto y las características del tomador del seguro, como el género, edad, la duración de la póliza y el valor por asegurar (Eling y Kiesenbauer, 2014).

Un enfoque de modelo espacial que considera la estadística descriptiva para investigar patrones espaciales o identificar comportamientos de los consumidores de seguros relacionados con el espacio, el cual consideró la información demográfica y socioeconómica del asegurado, partiendo de la premisa que la riqueza está correlacionado a la zona de residencia, de esta forma, al comprender las características espaciales con técnicas de agrupación y regresión logística, permite identificar la agrupación de los asegurados con atributos similares, con el propósito de reducir las tasas de caducidad en beneficio de las empresas (Hu *et al*., 2021).

3. Metodología

En el comportamiento que caracteriza los *lapses* es imperativo considerar factores económicos y sociales que influyen en la estabilidad o crecimiento del portafolio de servicios. Para ello se consideró, una data de corte transversal de 134012 registros con 12 factores del sector de seguros del 2018 (tabla 1), que caracteriza a un grupo de asegurados del mercado nacional en Colombia, el cual proviene de una empresa de tamaño mediano que opera en 7 ciudades.

Tabla 1. Descripción de los factores que componen la data de corte transversal

Factor	Descripción	Factor	Descripción
Medio	Medio de distribución	**Género**	Masculino o femenino
Tipo de seguro	Dotal, vida entera, temporal	**Edad**	Dato del asegurado
Código de la ciudad	Código de la ciudad asegurador	**Rango de edad**	Clasificación en rangos específicos
Asesor	Si el asegurado tiene atención de un asesor	**Prima**	Pago anual por el tipo de seguro
Causal de cancelación	Justificación de la cancelación	**Valor asegurado**	Monto en riesgo
Fecha de nacimiento	Dato del asegurado		Estado de la póliza: cancelado o renovado

Fuente. Elaboración propia.

Se presentan 3 formas distintas de analizar los 134 102 datos mediante modelos de clasificación, como la regresión logística (RL), la red neuronal artificial (RNA) y las máquinas de soporte vectorial (MSV) (Azzone *et al.*, 2022; Xong, 2019).

3.1. Regresión logística

El propósito de toda regresión logística está en obtener el mejor ajuste a sus parámetros, que permita relacionar una variable de salida discreta binaria en función de variables correlacionadas, o al conjunto de variables independientes, que exprese la probabilidad de que ocurra un evento (Fiuza y Rodríguez, 2000). La regresión logística usa variables para representar distintos niveles de estas variables de escala nominal que son identificadas, y no tienen significado numérico. Es decir, se crean tantas variables dicotómicas como números de respuestas, estas suelen ser denominadas *dummy* o variables de diseño.

En general, si la variable de escala nominal tiene k posibles valores, entonces $k-1$ variables de diseño son requeridas. Supongamos que la variable independiente x_j tiene k_j niveles, entonces, k_j-1 variables designadas denotadas como D_{jl}, y los coeficientes de estas variables designadas denotadas por β_{jl} , $l=1,2,\dots,k_j-1$. Así, para el modelo con p variables se considera que g(x) de la siguiente forma:

$$g(x)=\beta_0+\beta_1x_1+\cdots+\sum_{l=1}^{k_j-1}\beta_{jl}\,D_{jl}+\beta_px_p \qquad (1)$$

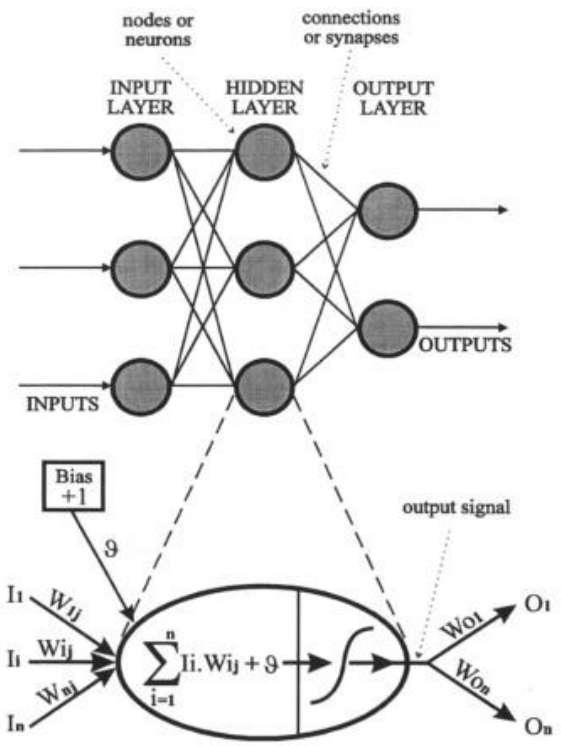
Donde el modelo de regresión logística múltiple $\pi(x)$ es (Harrell, 2015; Hosmer *et al.,* 2013):

$$\pi(x)=\frac{e^{g(x)}}{1+e^{g(x)}} \qquad (2)$$

3.2. Red neuronal artificial

Es una técnica computacional para máquinas de aprendizaje que permite realizar predicciones, clasificación y reconocimiento de imágenes para sistemas complejos con relaciones no lineales en datos multivariados, con un enfoque principal de supervisados; el cual consiste en una calibración multivariada, a partir de un conjunto de datos seleccionados que son transmitidos a través de una conexión entre nodos que contiene las capas de la red, y junto a unos pesos de conexión entre neurona **i** con la neurona **j** w_{ji}, que son modificados mediante una función de activación, establecida en cada nodo (figura 1), donde la máquina realiza un proceso de aprendizaje hasta ajustar los pesos de conexión (Goodacre *et al*., 1996; Tang y Yang, 2021).

Figura 1. Red neuronal de tres capas y el funcionamiento de un nodo



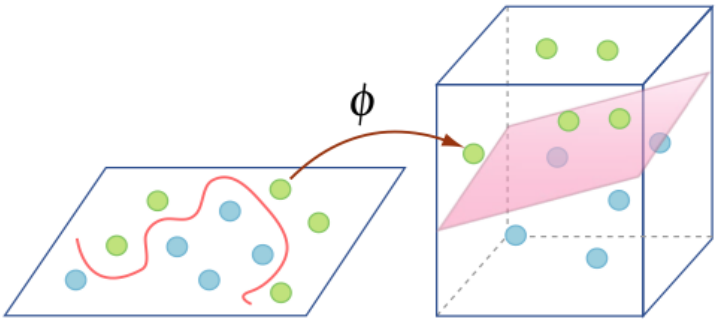
Fuente. Goodacre et al. (1996).

La función de activación es una construcción matemática que define la respuesta de una neurona, entre ellas se considera la función paso, la función sigmoidal, Tanh, Relu, entre otras (He *et al*., 2015; Zhu *et al*., 2021).

3.3. Máquinas de soporte vectorial

Las máquinas de soporte vectorial (MSV), se caracterizan por su gran desempeño en las aplicaciones. Una MSV mapea (ϕ) desde los datos de entrada a un espacio de características de dimensión mayor que el espacio de los datos, y determina un hiperplano que separa y maximiza el margen entre las clases (figura 2) (Betancourt, 2005; Rudin, 2012). Dentro de los problemas de clasificación se consideran los binarios, para casos linealmente separables y no separables.

Figura 2. Mapeo ϕ de características



Fuente. Rudin (2012).

Al considerar un dato $x_i \in R^n$, clasificado en una de las clases mediante $y_i \epsilon \{-1.1\}$ para $i = 1,2$, mediante el criterio de separación: $w \cdot z + b = 0$, siendo $w \epsilon Z$ y $b \epsilon R$ (Cristianini y Shawe, 2000; Hastie *et al.*, 2001; Zaki y Meira, 2014). Se dice que el conjunto de datos es linealmente separable si existe (w, b), con tal que cumpla las siguientes inecuaciones:

$$(w \cdot z_i + b) \geq 1, y_i = 1$$
$$(w \cdot z_i + b) \leq 1, y_i = -1 \qquad (3)$$

En el caso no separable, los datos no pueden clasificarse y pasa a ser un problema de optimización que se propone con la ecuación (4):

$$Función\ objetivo\ :\ \min_{w',b',\xi_i} \left\{ \frac{\| w' \|^2}{2} + C * \sum_{i=1}^{m} (\xi_i)^k \right\},$$
$$Restricciones\ lineales:\ y_i * \left(\vec{\phi}(x_i) \cdot w' - b'\right) \geq 1 - \xi_i, \forall (\phi(x_i), y_i) \in T\ \xi_i \geq 0, i = 1, \dots, m \qquad (4)$$

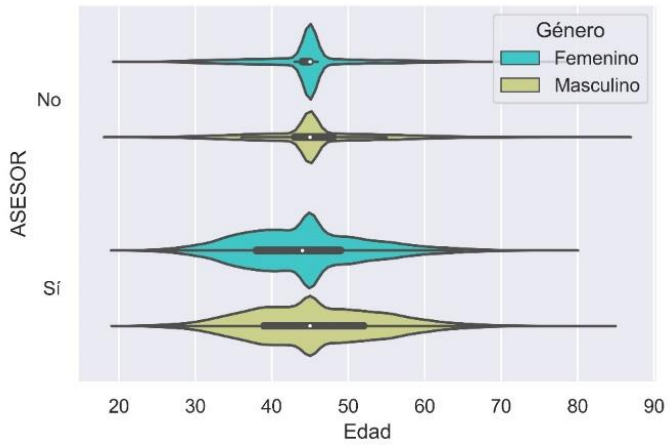
Donde C y k son constantes positivas, lo cual incorpora el costo de la clasificación incorrecta. El término $\sum(\xi_i)^k$ da el error total, el cual se escoge empíricamente, es una constante de regulación que controla el equilibrio entre la maximización del margen (minimizar $\frac{\|w'\|^2}{2}$) y minimizar el error total (Betancourt, 2005; Cortés y Vapinik, 1995; Zaki y Meira, 2014).

4. Resultados

4.1. Perspectiva estadística

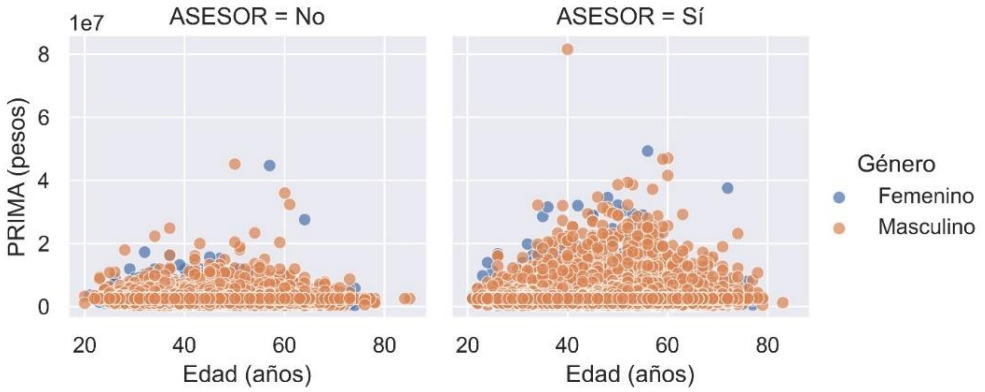
Inicialmente se consideró un análisis estadístico descriptivo con la base de datos, identificando si el asegurado tiene asesor en un rango de edades (figura 3), se muestra que la población de asegurados con asesor (se asigna el valor de 1) se concentra entre 30 y 60 años, sin embargo, hay una mayor concentración de asegurados sin asesor (se asigna el valor de 0), aproximadamente entre 43 y 48 años, adicionalmente, en la figura 4 se genera información por género, en el caso del género masculino (se asigna el valor 1) y femenino (se asigna el valor de 0), se concentra los asegurados con asesor entre 40 y 60 años, así mismo, en el caso femenino hay un acumulado significativo menores a 40 años de no tener asesor.

Figura 3. Edad del asegurado con relación a la presencia del asesor



Fuente. Elaboración propia.

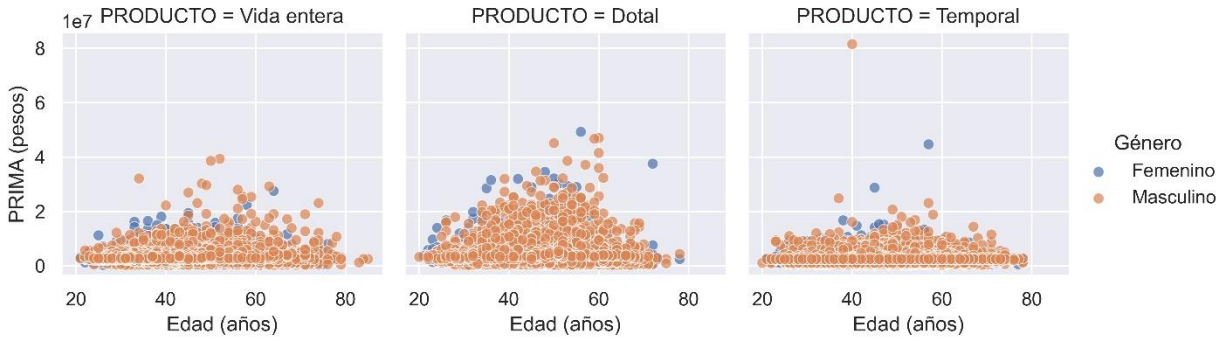
Figura 4. Distribución de la prima en relación con la edad y el género



Fuente. Elaboración propia.

En términos de productos de seguros vida entera (se asigna el valor 1), seguro dotal (se asigna el valor 2), seguro temporal (se asigna el valor 3), la mayor concentración se reporta en el dotal:

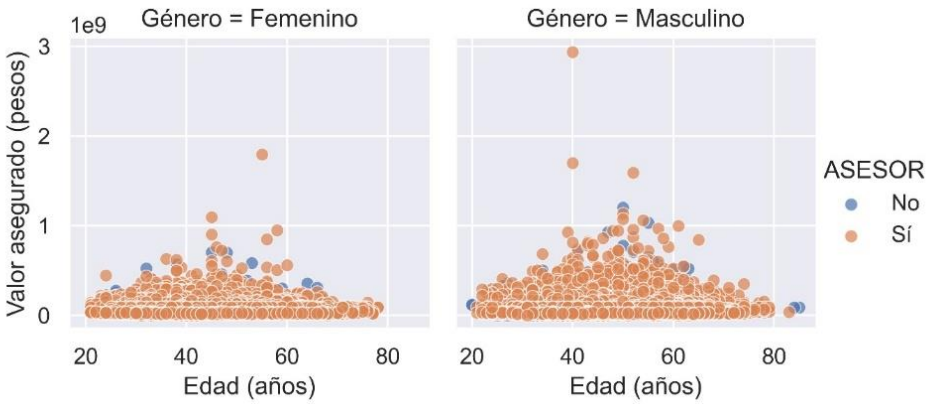
Figura 5. Distribución de los productos de seguros



Fuente. Elaboración propia.

En las figuras 4, 5 y 6, la población de jóvenes no presenta un volumen importante de asegurados, si bien la cultura de seguros de vida se ha caracterizado por tener una mayor frecuencia en personas de edad productiva, entre 30 y 55 años.

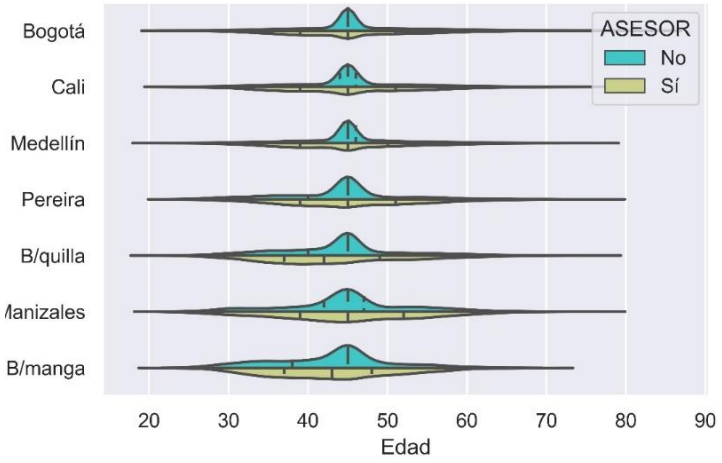
Figura 6. Distribución del valor asegurado según el género y la presencia del asesor



Fuente. Elaboración propia.

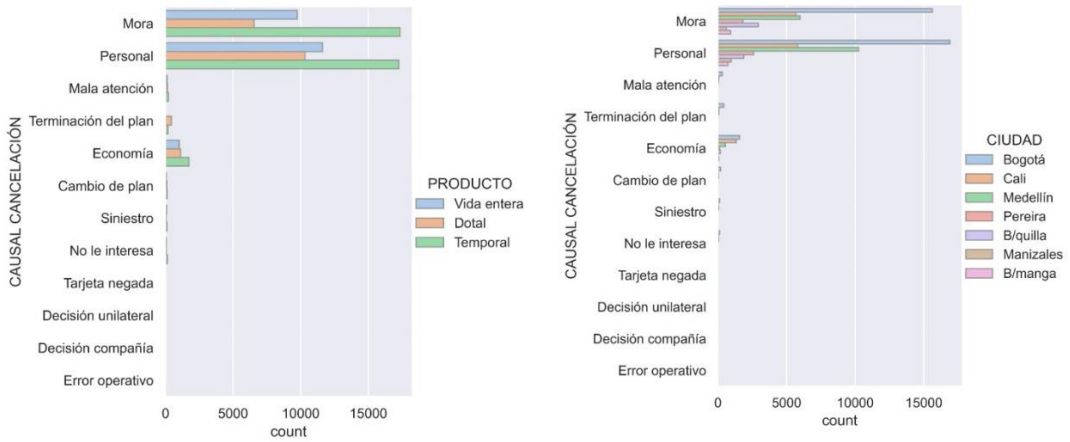
Adicionalmente, en personas de edades mayores los seguros de vida son de un mayor costo por tener una alta probabilidad de fallecimiento. Por otro lado, la tendencia de las personas menores de 30 años es presentar un bajo interés en comprar un seguro de vida, y sus prioridades están enmarcadas en la compra de vivienda, vehículo, estudios, turismo, entre otras (Fier y Liebenberg, 2013), se caracteriza una mayor participación del género masculino en la compra de seguros de vida. En la figura 7, se evidencia que la población entre 40 a 50 años, su tendencia es la no presencia de un asesor en cada una de las ciudades, el cual se asigna los siguientes códigos (Bogotá=1, Cali=2, Medellín=3, Pereira=4, Barranquilla=5, Manizales=6, Bucaramanga=7) esto marca una característica en la población de asegurados.

Figura 7. Distribución de la asesoría por ciudad



Fuente. Elaboración propia.

Figura 8. Causal de cancelación por producto y ciudad



Fuente. Elaboración propia.

Los datos revelan una mayor tasa de cancelación en los siguientes cuatro productos: causas personales, mora, mala atención, factores económicos. El análisis de la distribución por edad dentro de la categoría de "causas personales" revela información interesante, como la cancelación por razones personales, ya que no es confinada a un grupo etario en específico, por otra parte, las personas de mayor edad tienden a buscar orientación profesional antes de tomar decisiones de cancelación (tabla 2).

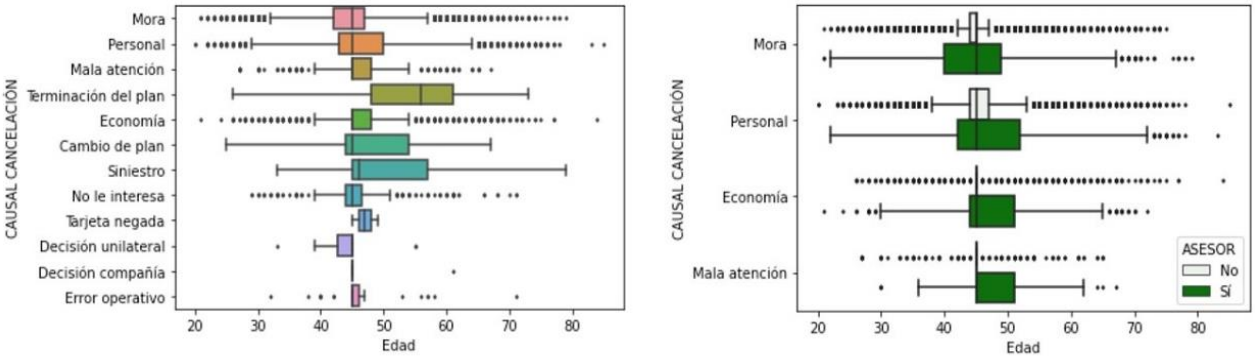
Tabla 2. Descripción estadística

Características	Descripción
Rango	La tasa de cancelación por causas personales se presenta entre los 23 y 70 años.
Mediana	La mediana de edad para aquellos que cancelan sin asesoría es de 45 años, mientras que para aquellos que consultan con un asesor es de 47 años.
Promedio	El promedio de edad para quienes cancelan sin asesoría es de 46 años, mientras que para quienes consultan con un asesor es de 47 años.
Datos del 50 %	El 50 % de los que cancelan sin asesoría tienen una edad cercana a los 45 años, mientras que el 50 % de los que consultan con un asesor tienen una edad cercana a los 47 años.
Asimetría	La distribución de la edad presenta una asimetría positiva en ambos casos, lo que significa que hay más personas que cancelan por causas personales en edades más jóvenes que en edades más avanzadas.

Fuente. Elaboración propia.

Finalmente, en el causal por mora con asesor presenta una distribución normal y en todos los causales de cancelación hay presencia de valores atípicos.

Figura 9. Causal de cancelación según la edad

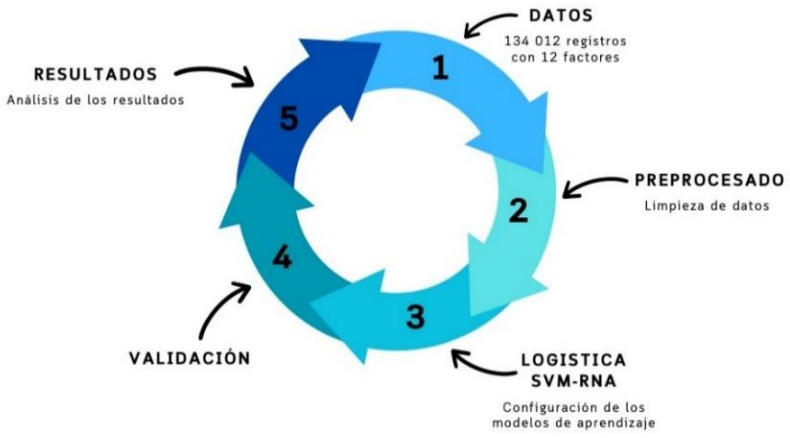


Fuente. Elaboración propia.

4.2. Máquinas de aprendizaje

Ahora en el desarrollo de la aplicación de las máquinas de aprendizaje se consideró el siguiente proceso para para la regresión logística, red neuronal y máquina de soporte vectorial.

Figura 10. Descripción de la metodología



Fuente. Elaboración propia.

4.2.1. Regresión logística

En el modelo de regresión logística se consideró la ecuación 2 y se entrenó con el 80 % de los 134 102 registros, para predecir si hay o no renovación de un perfil de asegurado con características definidas en la tabla 1, mediante variables predictoras como: medio, tipo de seguro, código de ciudad, asesor, género, fuma, edad, valor asegurado y prima, junto con la variable *y* que corresponde a la decisión de renovación o cancelación. En su implementación se utilizó las librerías de Sklearn de Python para crear y definir el modelo, el entrenamiento y su validación (Pölsterl, 2020). La precisión del modelo junto con la matriz de confusión verificó los resultados de salida (anexo 1).

La matriz de confusión presentó un 73 % de acierto, donde de 10 109 asegurados, el modelo predijo la cancelación y, en efecto, se cancelaron, por otro lado, para 9450 asegurados, el modelo predijo que renovaban y, efectivamente, renovaron, a su vez, el modelo predijo que 5645 asegurados cancelaron, pero corresponden a renovación, y 1589 asegurados se predijo que renovaban, pero corresponden a cancelación.

Tabla 3. Reporte de precisiones

	Precission	***Recall***	***F1-score***	***Support***
0	0.86	0,64	0,74	15754
1	0,63	0,86	0,72	11049
Accuracy			0,73	26803
Macroavg	0,75	0,75	0,73	26803
Weightedavg	0,77	0,73	0,73	26803

Fuente. Elaboración propia.

En la columna *support* de la tabla 3 se registran cancelados 15 754 y renovados 11 049, para un total de 26 803 registros. En la columna *recall* la precisión con la que se acierta, es decir, 64 % de precisión en cancelados y 86 % en renovados, con una media de 75 %, y en relación con *el F1-score* tenemos una precisión total del 73 %, es decir, una correcta clasificación, siendo una medida de los datos que aportan información al modelo. Al tomar los pesos de la precisión esta varía a 77 % y en *recall* varía a 73 %, lo cual nos sigue llevando a una precisión total de 73 %.

De cancelados se tiene una precisión del 86 %, es decir que los aciertos correctos dentro de lo que se predijo tienen un buen porcentaje, en cuanto a renovados, la precisión es del 63 %, por otro lado, su *recall* tiene un mejor valor dado que en el ajuste de clasificación esta medida indica que no se están pasando tantos falsos a positivos, es decir que un 86 % de la muestra

no contiene falsos positivos. Adicionalmente, el *F1-score* siendo la media entre la precisión y el *recall* nos muestra un 74 % y 72 % para cada clase, respectivamente, por lo que el modelo con sus 75 % de los datos aportan información correcta. Ahora, se consideró los perfiles de asegurados para establecer su predicción de cancelación.

Con el modelo entrenado se aplicó a un perfil de asegurado 1 (tabla 4), el cual da una probabilidad de cancelar del 61,27 %.

Tabla 4. Perfil de asegurado 1

Características	
Medio	1
Tipo de seguro	1
Asesor	Si
Género	Femenino
Edad	25
Fuma	Si
Código de la ciudad	1
Valor asegurado	Tipo A
Prima	Tipo 1

Fuente. Elaboración propia.

Se incluyen las variables en el modelo de regresión logística:

$$X\_new = pd.DataFrame(\{'Género': [1], 'CODCIUDAD': [1], 'MEDIO': [1], 'ASESOR': [1], 'TIPOSEGURO': [1], 'FUMA': [1], 'Edad': [25], 'PRIMA': [1], 'VA': [1]\})$$

$$model.predict(X\_new)$$

$$lr = LogisticRegression()$$

$$lr\,.\,fit\,(\,x\_train\,,y\_train\,)$$

$$person\;=\;np\,.\,array\,([[\,1\,,1\,,1\,,1\,,1\,,1\,,25\,,1\,,1]])$$

$$lr\,.\,predict\_proba\,(\,person\,)$$

Ahora, para el segundo perfil de asegurado no tiene la variable asesora:

Tabla 5. Perfil de asegurado 2

Características	
Medio	1
Tipo de seguro	1
Asesor	No
Género	Femenino
Edad	25
Fuma	Si
Código de la ciudad	1
Valor asegurado	Tipo A
Prima	Tipo 1

Fuente. Elaboración propia.

En el segundo perfil asegurado al no contar con un asesor tendrá una alta probabilidad de cancelar, con un 96,83 %.

4.2.2. Red neuronal

En la red neuronal, para predecir la probabilidad de cancelación o renovación de un asegurado con un perfil asociado a las características de las tablas 1 y 2, fue necesario separar las variables, dejando en X las características y en Y las clases, de igual forma que en la regresión logística:

Creación de las variables *X, Y* para la red neuronal:

$$X = dataset.iloc[:,:9].values$$
$$y = dataset.iloc[:,9:10].values$$

El conjunto de datos tiene 9 valores de entrada y 2 valores de salida, con 2 capas ocultas, una de 16 y 12 neuronas con funciones de activación *relu* y *softmax*. Adicionalmente, con 100 épocas se realizó el entrenamiento de la red neuronal, obteniendo una precisión del 80,88 % (anexo 2). El modelo obtuvo una precisión del 81,48 %, con la siguiente matriz de confusión (anexo 3).

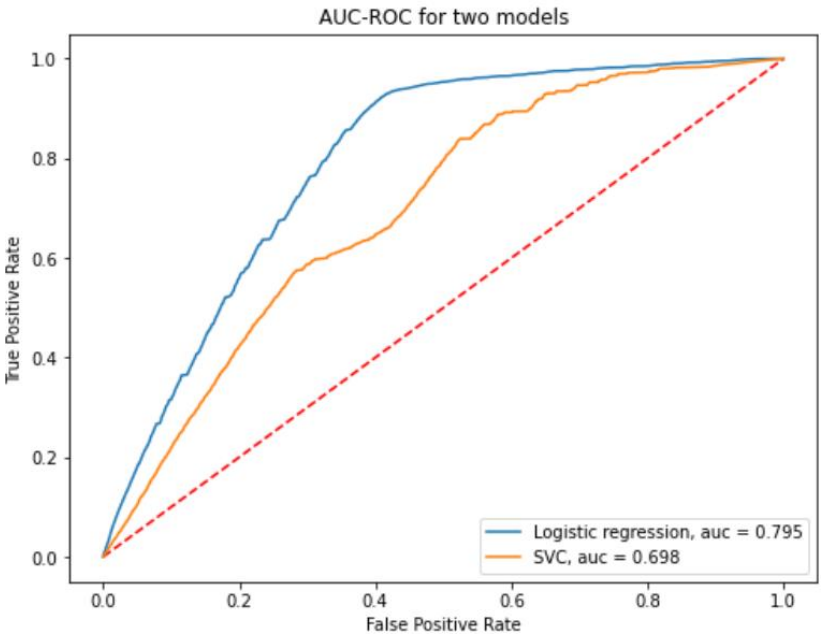
Mediante la matriz de confusión se determinó una exactitud del 81 %, y su precisión de 89 % indica que está cerca del valor verdadero. El 72 % de los casos negativos han sido clasificados correctamente, por lo que la tasa de falsos negativos es del 28 %, y la especificidad muestra que el 90 % de los casos negativos son clasificados correctamente, por lo tanto, la tasa de falsos positivos es solo del 10 %.

Ahora, retomando el perfil 1, la red neuronal entrega una probabilidad a cancelar de 80,09 %; y con respecto al perfil 2, da una probabilidad de cancelar de 93,95 %. Marcando una diferencia importante con la regresión logística en el perfil 1, que reportó 61,27 %.

4.2.3. Máquina de soporte vectorial

El modelo de máquina de soporte vectorial tiene el objetivo de predecir si el asegurado renueva o no con su probabilidad de ocurrencia. Para ello, la variable X representa todas las variables que identifican las características del asegurado, con *y* dejamos la categoría para predecir, sin embargo, la capacidad de rendimiento fue del 69,8 %, mientras que en la regresión logística fue de 79,5 %:

Figura 11. Modelo máquina de soporte vectorial (SVC) y regresión logística



[915 11]
[49 25]

Fuente. Elaboración propia.

Para la ejecución de la máquina de soporte vectorial se consideraron 5000 registros, permitiendo reducir el tiempo de cálculo computacional y debilitando los resultados esperados, según se visualiza en la curva característica operativa del receptor (ROC), está representa la capacidad del rendimiento del 69,8 % que pueda clasificar la probabilidad de cancelación de un perfil o no. A su vez, la matriz de confusión visualiza el 20 % de los datos. De la matriz de confusión se determina que el modelo tiene una exactitud del 94 %. El 34 % de los casos negativos han sido clasificados correctamente, por el número limitado de datos usados en esta máquina, la tasa de verdaderos positivos es del 34 % y la tasa de verdaderos negativos es del 99 %. Ahora, se determinó la probabilidad de cancelación del perfil 1, con un reporte de probabilidad de cancelar del 64 % y en el perfil 2 de cancelar con el 88 %.

Los resultados presentan una importante diferencia con el primer perfil en términos de porcentaje de cancelación, con respecto a la red neuronal en un valor aproximado de 17 %, el cual puede estar justificado por el rendimiento del 66 % de la máquina de soporte vectorial determinada.

5. Análisis

Conocer el comportamiento de los perfiles de sus asegurados da un espectro de mayores oportunidades para definir líneas estratégicas de negocios en el mercado de seguros, es imperativo considerar algunas de las variables que caractericen al asegurado según las necesidades requeridas, a las que llamaremos categorías, para este trabajo se consideran 8 variables tomadas de la tabla 1, a su vez, cada una de ellas establece subniveles denominados subcategorías, cuya asignación de códigos se estableció en la sección de resultados de perspectivas estadísticas, y se incluyó las probabilidades de cancelación mediante el modelo de regresión logística obtenido.

En la tabla 6 se observan los resultados en cada una de ellas, para el caso de la variable o categoría medio, la cual corresponde al medio de distribución del seguro, obteniendo que es más efectivo el medio 1, puesto que es el que menor probabilidad de cancelación (Pc) presenta; por otra parte, los medios de distribución 2, 3 y 4 tienen una alta probabilidad de tener asegurados que cancelen, por ello, para esta variable seleccionaremos el medio 1. En cuanto al tipo de seguro las 3 opciones tienen probabilidades muy cercanas de ser canceladas, en este caso tomaremos el tipo de seguro 1. De igual manera, con la ciudad donde se distribuye el seguro, las probabilidades de cancelar son muy cercanas, para el ejercicio tomaremos la ciudad de Bogotá (figuras 7 y 8), siendo la ciudad donde hay mayor concentración de asegurados, es decir, la ciudad 1. Para el caso de asesor, aquellos asegurados que tienen este acompañamiento, su probabilidad de cancelar es menor a quienes no lo tienen, en esta variable el valor del asesor es 1. Para la variable género no hay grandes diferencias en las dos categorías (figura 6), por lo tanto, se asume que la de menor probabilidad de cancelación es la categoría 0. Para el caso de los asegurados que fuman hay

una mayor probabilidad de cancelar respecto a los que no fuman, esta última población se definió en la clase 0. En el caso de los valores asegurados y prima, se fijaron rangos para obtener diferentes subcategorías, entre más bajo del valor asegurado mayor es la probabilidad de cancelar, de igual manera con la prima, entre más bajo es el valor de esta, mayor es la probabilidad de cancelar, por esta razón en estas categorías se definió las subcategorías 7 y 9, respectivamente.

Tabla 6. Perfil de asegurado 3

Selección de un perfil				Selección de un perfil			
Categoría	**Subcategoría**	**(Pc)**	**Selección**	**Categoría**	**Subcategoría**	**(Pc)**	**Selección**
Medio	1	0,57457476	1	Fuma	1	0,79930949	0
	2	0,84604704			0	0,57964952	
	3	0,95719346		Valor asegurado	1	0,62000434	7
	4	0,98912874			2	0,40049818	
Tipo de seguro	1	0,57234406	1		3	0,21478058	
	2	0,58423987			4	0,10071524	
	3	0,59603834			5	0,04384501	
Ciudad	1	0,59207338	1		6	0,01842924	
	2	0,58794131			7	0,00762876	
	3	0,58794131		Prima	1	0,64735262	9
	4	0,57964042			2	0,58089232	
	5	0,57547272			3	0,51136117	
	6	0,57129425			4	0,44138759	
	7	0,56710559			5	0,37366692	
Asesor	1	0,40156266	1		6	0,31056001	
	2	0,9295372			7	0,25379235	
Género	1	0,59536824	0		8	0,20432604	
	0	0,57981732			9	0,16240282	

Fuente. Elaboración propia.

Dadas las características seleccionadas en el tercer perfil del asegurado (medio, tipo de seguro, ciudad, asesor, género, fuma, valor asegurado, prima) = (1,1,1,1,0,0,7,9), se utilizaron las tres máquinas de aprendizaje y variando la edad con el objetivo de revisar cual es la incidencia en su resultado mediante las tasas de cancelación (P_c) y la tasa de renovación (P_r).

Tabla 7. Lapses y renovación

Edad	Log (Pc)	Log (Pr)	RNA (Pc)	RNA (Pr)	MSV (Pc)	MSV (Pr)
20	0,001172140000	0,998827860000	0,009944850000	0,990055200000	0,000220183441	0,999779816559
25	0,001297710000	0,998702290000	0,010221530000	0,989778460000	0,000001046209	0,999998953791
30	0,001436710000	0,998563290000	0,011882980000	0,988117100000	0,000000000556	0,999999999444
35	0,001590580000	0,998409420000	0,015400440000	0,984599530000	0,000000000107	0,999999999893
40	0,001760900000	0,998239100000	0,019938100000	0,980061950000	0,000000000001	0,999999999999
45	0,001949420000	0,998050580000	0,025777750000	0,974222240000	0,000000000000	1,000000000000
50	0,002158080000	0,997841920000	0,033269520000	0,966730400000	0,000000000000	1,000000000000
55	0,002389020000	0,997610980000	0,042843270000	0,957156800000	0,000000000000	1,000000000000
60	0,002644610000	0,997355390000	0,055014810000	0,944985200000	0,000000000000	1,000000000000
65	0,002927460000	0,997072540000	0,070390180000	0,929609900000	0,000000000000	1,000000000000
70	0,003240470000	0,996759530000	0,089654760000	0,910345300000	0,000000000004	0,999999999996
75	0,003586820000	0,996413180000	0,102442120000	0,897557850000	0,000000000448	0,999999999552
80	0,003970050000	0,996029950000	0,091829690000	0,908170300000	0,000000068593	0,999999931407
85	0,004394040000	0,995605960000	0,077474440000	0,922525600000	0,000127597427	0,999872402573

Fuente. Elaboración propia.

En la tabla 7 se reportan los resultados de los modelos de regresión logística (log), red neuronal (RNA) y MSV aplicado a P_r y P_c, confirmando la tendencia de este perfil a renovar, las subcategorías permitieron que no se presentaran diferencias probabilísticas entre los modelos. Así mismo, este resultado indica que cualquier asegurado con edades entre 20 y 85 años, con características del perfil 3 (1,1,1,1,0,0,7,9), la probabilidad de renovar es muy alta, por ende, este perfil de asegurado tiene probabilidades muy bajas de cancelar y se acentúa la importancia de tener un asesor para el rango de edades analizadas.

Ahora, en la construcción de un cuarto perfil del asegurado (tabla 6) se consideró probabilidades altas de cancelar en cada categoría (3,3,1,2,1,1,1,1), como se muestra en la tabla 8, y sus resultados están en la tabla 9.

Tabla 8. Perfil de asegurado 4

Selección de un perfil			
Categoría	**Subcategoría**	**(Pc)**	**Selección**
Medio	1	0,57457476	3
	2	0,84604704	
	3	0,95719346	
	4	0,98912874	
Tipo de seguro	1	0,57234406	3
	2	0,58423987	
	3	0,59603834	
Ciudad	1	0,59207338	1
	2	0,58794131	
	3	0,58794131	
	4	0,57964042	
	5	0,57547272	
	6	0,57129425	
	7	0,56710559	
Asesor	1	0,40156266	2
	2	0,9295372	
Género	1	0,59536824	1
	0	0,57981732	

Selección de un perfil			
Categoría	**Subcategoría**	**(Pc)**	**Selección**
Fuma	1	0,79930949	1
	0	0,57964952	
Valor asegurado	1	0,62000434	1
	2	0,40049818	
	3	0,21478058	
	4	0,10071524	
	5	0,04384501	
	6	0,01842924	
	7	0,00762876	
Prima	1	0,64735262	1
	2	0,58089232	
	3	0,51136117	
	4	0,44138759	
	5	0,37366692	
	6	0,31056001	
	7	0,25379235	
	8	0,20432604	
	9	0,16240282	

Fuente. Elaboración propia.

Tabla 9. Lapses y renovación

Edad	Log (Pc)	Log (Pr)	RNA (Pc)	RNA (Pr)	MSV (Pc)	MSV (Pr)
20	0,995628090000	0,004371910000	0,999996540000	0,000003460000	0,946036120000	0,053963880000
25	0,996049950000	0,003950050000	0,999968200000	0,000031800000	0,947279790000	0,052720210000
30	0,996431250000	0,003568750000	0,999656560000	0,000343440000	0,948020100000	0,051979900000
35	0,996775870000	0,003224130000	0,996299800000	0,003700240000	0,948532000000	0,051468000000
40	0,997087300000	0,002912700000	0,967158100000	0,032841880000	0,949007500000	0,050992500000
45	0,997368730000	0,002631270000	0,927589060000	0,072410990000	0,949538070000	0,050461930000
50	0,997623030000	0,002376970000	0,832649000000	0,167350980000	0,950122180000	0,049877820000
55	0,997852810000	0,002147190000	0,806583300000	0,193416760000	0,950698810000	0,049301190000
60	0,998060420000	0,001939580000	0,868344660000	0,131655320000	0,951193990000	0,048806010000
65	0,998247990000	0,001752010000	0,849295440000	0,150704600000	0,951562550000	0,048437450000
70	0,998417450000	0,001582550000	0,828035530000	0,171964400000	0,951810400000	0,048189600000
75	0,998570540000	0,001429460000	0,804467000000	0,195532980000	0,951991170000	0,048008830000
80	0,998708850000	0,001291150000	0,821892800000	0,178107140000	0,952178190000	0,047821810000
85	0,998833780000	0,001166220000	0,908500850000	0,091499190000	0,952420900000	0,047579100000

Fuente. Elaboración propia.

En el perfil 4 se obtuvo una alta probabilidad de cancelar, el cual demanda una mayor atención para el rango de edades. Respecto a los modelos utilizados para el análisis de los perfiles, estos tienen una precisión para la red neuronal del 81,53 %, con la regresión logística un 73 % y la SVM 60 %.

6. Discusión

En la actualidad, se emplea un modelo basado en la construcción de triángulos para calcular *lapses*, utilizando la frecuencia y la severidad, una metodología inicialmente diseñada para siniestros y ahora aplicada para prever *lapses* en el conjunto de información de asegurados (Richman, 2018). Aunque este método permite identificar *lapses* en cada periodo, carece de la capacidad de analizar la coherencia y la relación entre las variables que influyen en el crecimiento o decrecimiento de los *lapses* para una póliza o negocio específico.

Este proceso se realiza agrupando información, y no se identifica puntualmente cuál es la variable que ocasiona una alta tasa de cancelación, mientras que, al usar los modelos propuestos en este trabajo, se puede identificar qué tipo de perfiles de asegurados presentan un mayor *lapse*, e incluso se identifican las variables de mayor importancia e impacto en la cancelación.

Es claro que la metodología con triángulos ha funcionado y ha permitido que sus resultados sean usados en presupuestos, pero nos deja con el vacío de cuál es la variable que está ocasionando una mayor renovación en las pólizas, de ahí la importancia de usar las máquinas de aprendizaje propuestas, que es una alternativa para solucionar esta brecha.

Dependiendo del volumen de datos y variables se puede determinar si es óptimo usar máquinas de aprendizaje o realizar un proceso manual mediante el uso de programas comerciales.

Cuando se dispone de información relevante, es decir, con más de tres variables y un amplio conjunto de datos, las técnicas de aprendizaje automático resultan ser herramientas eficaces para analizar y predecir la probabilidad de cancelación (Pc) y renovación (Pr), basándose en variables específicas. En el caso de tener más de 3 variables, es idóneo porque permite entrenar a partir de sus datos la construcción de máquinas de aprendizaje con datos apropiados en las bases de datos para lograr predicciones precisas sobre la probabilidad de cancelación o renovación.

Los 134 102 registros (tabla 1), fueron utilizados para entrenar cada uno de los modelos, que incluyen regresión logística, redes neuronales y máquinas de soporte vectorial con diferentes precisiones: 73 %, 81,53 % y 60 %, respectivamente. Posteriormente, se aplicó el modelo de regresión logística para analizar la probabilidad de cancelación en el perfil 1 (tabla 3), representado por un individuo de 25 años, género femenino, en la ciudad 1 (Bogotá) con asesor. El resultado fue una Pc = 61,27 %. En contraste, para un perfil 2 (tabla 4), donde se prescinde del asesor, su resultado fue una Pc = 96,83 %. Aunque la figura 7 sugiere que la edad de 25 años tiene una información limitada para establecer la importancia de un asesor, pero con este modelo probabilístico, su resultado propone la permanencia del asesor, este modelo probabilístico propone que la presencia del asesor garantiza la continuidad del asegurado.

El segundo modelo, que utiliza una red neuronal, obtuvo una probabilidad de cancelación (Pc) del 80,09 % para el perfil 1, y del 93,95 % para el perfil 2. En cuanto al tercer modelo, la probabilidad de cancelación para el perfil 1 fue del 64 %, y para el perfil 2 fue del 88 %. Ciertamente, existe una disparidad en los resultados, especialmente en el caso del perfil 1, en comparación con el reportado por la red neuronal, marcando una diferencia considerable. Sin embargo, se reconoce que una mejora potencial en futuros trabajos implica ampliar la consideración de registros de 5000 a un 70 % del total, con el fin de considerar un modelo de soporte vectorial más robusto que permita contrastar los resultados de la red neuronal de manera más efectiva.

Con estos resultados los tres modelos indican que el perfil 2, sin asesor, permite al asegurado cancelar el tipo de póliza vigente. Esto se debe a la ausencia de estrategias que faciliten la continuidad activa de la póliza, especialmente ante las causales más prevalentes en la ciudad de Bogotá, como se muestra en la figura 8. Además, se observa una marcada presencia de valores atípicos en la figura 9, lo cual dificulta el análisis desde la perspectiva de la estadística descriptiva.

Adicionalmente, se seleccionaron 8 variables, denominadas categorías, que fueron subdivididas mediante regresión logística. Esta clasificación buscó proporcionar un mayor detalle en el perfil del asegurado 3, como se muestra en la tabla 6. Con el objetivo de realizar

predicciones de renovación para un rango de edades entre 20 y 85 años, se ofrece un análisis más completo de la incidencia del asesor, introduciendo una subcategoría específica para la ciudad de Bogotá. En esta subcategoría, los tres modelos presentan una probabilidad alta de renovación.

Finalmente, se examinó el perfil 4 como se detalla en la tabla 8, y los tres modelos indicaron una probabilidad significativamente alta de cancelación para el mismo rango de edades. Este análisis detallado proporciona una visión más precisa de los factores que afectan las decisiones de renovación y cancelación para un perfil de asegurado en la ciudad de Bogotá, sin embargo, puede extenderse a otras ciudades, lo que permite una comprensión más profunda de los perfiles de asegurados en diferentes productos de seguros, disminuyendo los riesgos de caducidad y fortaleciendo el enfoque estratégico de la empresa.

7. Conclusiones

En el análisis de 134 102 registros con 12 variables, se encontró que los asegurados de entre 40 y 50 años generan cambios significativos en el portafolio de las compañías de seguros, destacando la importancia de los asesores en todas las ciudades. Las principales causas de impacto fueron motivos personales, retrasos en los pagos, deficiencias en la atención y factores económicos. Estos factores pueden ser mitigados mediante estrategias adaptadas por los asesores a cada perfil de asegurado, con el objetivo de prevenir la cancelación de las pólizas.

La aplicación de la regresión logística, redes neuronales y máquinas de soporte vectorial permite calcular las tasas de caducidad con una precisión de 73 %, 81,53 % y 60 %, respectivamente, para un perfil de asegurado con sus características definidas. Adicionalmente, mediante estos modelos permiten identificar variables que influyen en la decisión del asegurador, como fue el caso del ejemplo de la variable asesor, la cual exhorta a estar renovando las alternativas estratégicas comerciales para que los asegurados se interesen en la renovación de sus pólizas, a su vez, promocionar la cultura de seguros de vida.

La aplicación de la regresión logística puede identificar focos donde se presente una mayor renovación y cancelación en las pólizas, de esta manera, puede contribuir en el cálculo de los *lapses*. Ofreciendo información relevante para la proyección de presupuestos en las empresas aseguradoras.

Además de los factores mencionados anteriormente, los perfiles de asegurados también pueden incluir otras variables, como salario, empresa, enfermedades, niveles de estudios, estrato socioeconómico y religión. Estas variables pueden ayudar a mejorar la precisión de las tasas de caducidad y, por lo tanto, a proporcionar información más útil para el análisis de los *lapses*.

Finalmente, la combinación de los análisis entre la estadística descriptiva junto con los modelos de máquinas de aprendizaje, permite identificar variables que contribuyen al pronóstico del portafolio de seguros en cada ciudad, y logra realizar escenarios en los cuales ofrecen insumos para la toma de nuevas decisiones en beneficio del asegurado y de las aseguradoras. Este cambio hacia enfoques más avanzados destaca la necesidad de evolucionar constantemente en la evaluación y predicción de *lapses,* para adaptarse a la complejidad creciente de los mercados aseguradores.

8. Referencias

Azzone, M., Barucci, E., Moncayo, G. G. & Marazzina, D. (2022). A machine learning model for lapse prediction in life insurance contracts. *Expert Systems with Applications*, *191*, 116261. https://doi.org/10.1016/j.eswa.2021.116261

Bacinello, A. R. (2005). Endogenous model of surrender conditions in equity-linked life insurance. *Insurance: Mathematics and Economics*, *37*(2), 270-296. https://doi.org/10.1016/j.insmatheco.2005.02.002

Bauer, D., Gao, J., Moenig, T., Ulm, E. R. & Zhu, N. (2017). Policyholder exercise behavior in life insurance: the state of affairs. *North American Actuarial Journal, 21*(4), 485-501. https://doi.org/10.1080/10920277.2017.1314816

Betancourt, G. A. (2005). Las máquinas de soporte vectorial (SVMs). *Scientia et Technica, 11*(27), 67-72. https://www.redalyc.org/pdf/849/84911698014.pdf

Canadian Institute of Actuarie. (2014). *Lapse experience study for 10-year term insurance - report individual life experience subcommittee*.

Cortés, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273-297. https://doi.org/10.1007/BF00994018

Cristianini, N. & Shawe-Taylor, J. (2000). *An introduction to support vector machine and other kernel-based learning methods.* Cambridge University Press https://doi.org/10.1017/CBO9780511801389

Eling, M. & Kochanski, M. (2013). Research on lapse in life insurance: what has been done and what needs to be done? *Journal of Risk Finance*, *14*(4), 392-413. https://doi.org/10.1108/JRF-12-2012-0088

Eling, M. & Kiesenbauer, D. (2014). What policy features determine life insurance lapse? An analysis of the German market. *Journal of Risk and Insurance, 81*(2), 241-269. https://doi.org/10.1111/j.1539-6975.2012.01504.x

Fier, S. G. & Liebenberg, A. P. (2013). Life insurance lapse behavior. *North American Actuarial Journal, 17*(2), 153-167. https://doi.org/10.1080/10920277.2013.803438

Fiuza Pérez, M. y Rodríguez Pérez, J. C. (2000). La regresión logística: una herramienta versátil. *Nefrología, 20*(6), 477-565. https://www.revistanefrologia.com/es-la-regresion-logistica-una-herramienta-articulo-X0211699500035664

Goodacre, R., Neal, M. J. & Kell, D. B. (1996). Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectra. *Zentralblatt für Bakteriologie, 284*(4), 516-539. https://doi.org/10.1016/S0934-8840(96)80004-1

Gottlieb, D. & Smetters, K. (2021). Lapse-based insurance. *American Economic Review*, *111*(8), 2377-2416. https://doi.org/10.1257/aer.20160868

Guillén, M., Pérez, A. M. & Alcañiz, M, (2011). *A logistic regression approach to estimating customer profit loss due to lapses in insurance.* Document de Treball No. XREAP 2011-13. http://dx.doi.org/10.2139/ssrn.1942278

Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis* (2nd ed.). Springer.

Hastie, T., Tibshirani, R. & Freidman, J. (2001*). The elements of statistical learning: data mining, inference and prediction*. Springer. http://dx.doi.org/10.1007/978-0-387-21606-5

He, K., Zhang, X., Ren, S. & Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on imagenet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 1026-1034. https://doi.org/10.1109/ICCV.2015.123

Heldt, R., Silveira, C. S. & Luce, F. B. (2021). Predicting customer value per product: from RFM to RFM/P. *Journal of Business Research*, *127*, 444-453. https://doi.org/10.1016/j.jbusres.2019.05.001

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. https://onlinelibrary.wiley.com/doi/chapter-epub/10.1002/9781118548387.fmatter

Hu, S., O'Hagan, A., Sweeney, J. & Ghahramani, M. (2021). A spatial machine learning model for analysing customers' lapse behaviour in life insurance. *Annals of Actuarial Science, 15*(2), 367-393. https://doi.org/10.1017/S1748499520000329

Kiesenbauer, D. (2012). Main determinants of lapse in the German life insurance industry. *North American Actuarial Journal*, *16*(1), 52-73. https://doi.org/10.1080/10920277.2012.10590632

Kim, C. (2005). Modeling surrender and lapse rates with economic variables. *North American Actuarial Journal, 9*(4), 56-70. https://doi.org/10.1080/10920277.2005.10596225

Kuo, W., Tsai, C. & Chen, W. K. (2003). An empirical study on the lapse rate: the cointegration approach. *Journal of Risk and Insurance, 70*(3), 489-508. https://doi.org/10.1111/1539-6975.t01-1-00061

Moreno, L. G. (2019). *Notas de clase: teoría del riesgo y contingencias*.

Namvar, M., Gholamian, M. R. & KhakAbi, S. (2010). A two phase clustering method for intelligent customer segmentation. *2010 International Conference on Intelligent Systems, Modelling and Simulation, Liverpool, UK*, pp. 215-219. https://doi.org/10.1109/ISMS.2010.48

Outreville, J. F. (1990). Whole-life insurance lapse rates and the emergency fund hypothesis. *Insurance: Mathematics and Economics*, *9*(4), 249-255. https://doi.org/10.1016/0167-6687(90)90002-U

Outreville, J. F. (2013). The relationship between insurance and economic development: 85 empirical papers for a review of the literature. *Risk Management and Insurance Review*, *16*(1), 71-122. https://doi.org/10.1111/j.1540-6296.2012.01219.x

Pinquet, J., Guillén, M. & Ayuso, M. (2011). Commitment and lapse behavior in long-term insurance: a case study. *Journal of Risk and Insurance, 78*(4), 983-1002. https://doi.org/10.1111/j.1539-6975.2011.01420.x

Pölsterl, S. (2020). Scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *The Journal of Machine Learning Research, 21*, 1-6. https://jmlr.org/papers/volume21/20-729/20-729.pdf

Ramchandani, P., Paich, M. & Rao, A. (2017). Incorporating learning into decision making in agent based models. In E. Oliveira, J. Gama, Z. Vale, H. Lopes (Eds.), *Progress in Artificial Intelligence*: 18th EPIA Conference on Artificial Intelligence, EPIA 2017, Porto, Portugal, (pp. 789-800). https://doi.org/10.1007/978-3-319-65340-2_64

Richman, R. (2018). *AI in actuarial science.* SSRN. http://dx.doi.org/10.2139/ssrn.3218082

Rozar, T., Scott, R. & Susan, W. (2010). *Report on the lapse and mortality experience of post-level premium period term plans*.

Rudin, C. (2012). *15.097 Lecture 13: Kernels*. MIT Open Course Ware. http://ocw.mit.edu/courses/15-097-prediction-machine-learning-and-statistics-spring-2012/resources/mit15_097s12_lec13/

Tang, S. & Yang, Y. (2021). Why neural networks apply to scientific computing? *Theoretical and Applied Mechanics Letters, 11*(3), 100242. https://doi.org/10.1016/j.taml.2021.100242

Tsai, C., Kuo, W. & Chiang, D. M. (2009). The distributions of policy reserves considering the policy-year structures of surrender rates and expense ratios. *Journal of Risk and Insurance*, *76*(4), 909-931. https://doi.org/10.1111/j.1539-6975.2009.01324.x

Weindorfer, B. (2012). *QIS5: a review of the results for EEA Member States, Austria and Germany*. https://www.fh-vie.ac.at/uploads/WP-070_2012.pdf

Xong, L. J. & Kang, H. M. (2019). A comparison of classification models for life insurance lapse risk. *International Journal of Recent Technology and Engineering*, *7*(5S), 245-250.

Zaki, M. & Meira, W. (2014). *Data mining and Machine Learning: fundamental concepts and algorithms.* (2nd ed.). Cambridge University Press. https://www.cambridge.org/co/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/data-mining-and-machine-learning-fundamental-concepts-and-algorithms-2nd-edition?format=HB&isbn=9781108473989

Zhu, H., Zeng, H., Liu, J. & Zhang, X. (2021). Logish: A new nonlinear nonmonotonic activation function for convolutional neural network. *Neurocomputing*, *458*, 490-499. https://doi.org/10.1016/j.neucom.2021.06.067

9. Anexos

Anexo 1. Matriz de confusión

$$print\ (\ confusion\_matrix\ (\ Y\_validation\ , predictions\ )\ )$$

[10109 5645]

[1589 9450]

Fuente. Elaboración propia.

Anexo 2. Construcción del modelo de la red neuronal

```
import keras
from keras.models import Sequential
from keras.layers import Dense
model = Sequential()
model.add(Dense(16, input_dim=9, activation='relu'))
model.add(Dense(12, activation='relu'))
model.add(Dense(2, activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
history = model.fit(X_train, y_train, epochs=100, batch_size=64)
```

Fuente. Elaboración propia.

Anexo 3. Matriz de confusión

[11852 1261]

[3858 9832]

Fuente. Elaboración propia.