LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

MACHINE LEARNING PARA LA SEGMENTACIÓN Y OPTIMIZACIÓN DE LOS COSTOS DE ADQUISICIÓN DE CLIENTES

ÁLVARO ANTONIO FORERO GONZÁLEZ

ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

FACULTAD DE INGENIERÍA

PROFESOR JOHN GONZÁLEZ VELOZA

4 DE DICIEMBRE DE 2021

LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

DEDICATORIA

A mis padres, seguramente estarían muy orgullosos

A mis hijas, que les sirva como ejemplo y que ellas sean ejemplo

A mi esposa, sin su motivación y apoyo este trabajo no sería una realidad

Gracias!

MACHINE LEARNING PARA LA SEGMENTACIÓN Y OPTIMIZACIÓN DE LOS COSTOS DE ADQUISICIÓN DE CLIENTES

MACHINE LEARNING FOR THE SEGMENTATION AND OPTIMIZATION OF CUSTOMER ACQUISITION COSTS

Álvaro Antonio Forero González, aagorerog01@ libertadores.edu.co. Estudiante de Especialización en Estadística Aplicada.

José John Fredy González Veloza, jjgonzalezv02@libertadores.edu.co, Fundación Universitaria Los Libertadores.

RESUMEN

Este trabajo desarrolla un avance importante en la clasificación, predicción y segmentación de los afiliados a Casur, como parte de sus objetivos estratégicos facilitando la planeación de la entidad.

Para ello, se adoptaron modelos básicos, de reglas, y posteriormente modelos machine learning de clasificación y clustering. De los cuales los primeros se utilizaron como base de comparación y los últimos dos se utilizaron como insumo en el resto del análisis.

No obstante, debido a que los modelos tienen asociados error, se utilizó el modelo de distribución de probabilidad binomial para hallar el número de acercamientos comerciales necesarios para tener, como mínimo, una venta con una probabilidad del 99%; y optimizar, por esta vía, el modelo de adquisición de clientes, con lo que también se pudo comparar el mejor modelo, de acuerdo con la actualidad de la entidad.

Dentro de los resultados más importantes se tiene que el modelo de regresión logística, clasifica como interesados dentro de la data general a 9.797 personas; permite establecer el costo de adquisición de clientes en los $2.608.695, y con

él se define que la probabilidad de venta promedio para cada persona es del 30%.

Por su parte, en el modelo de clustering por k-means se identificó un solo clúster como el más proclive a adquirir servicios de salud, compuesto en la data general por 4.727 personas. Permite establecer el costo de adquisición de clientes en $1.600.000 y gracias a este modelo, la probabilidad promedio de compra para cada persona se fijó en el 45%.

INTRODUCCIÓN

La segmentación y clasificación de los clientes y consumidores, así como las proyecciones de ventas, son fundamentales en la planificación organizacional (Chen & Lu, 2017).

Como ejemplo de los esfuerzos de las empresas privadas, (Rincon et al., 2021) desarrolla un modelo de Machine Learning para la empresa Dell Colombia, con el objetivo de aumentar la efectividad de las ventas, es decir, reducir el número de clientes que visita un asesor para generar una venta, mediante la clasificación y segmentación de los clientes, evaluando sus características particulares.

El anterior estudio consiguió clasificar las cuentas más prometedoras, permitiendo centrar los esfuerzos comerciales sobre las mismas (Rincon et al., 2021).

Adicionalmente, otros estudios apuntan a incrementar la eficiencia y eficacia del proceso comercial, permitiendo “la incorporación de inteligencia artificial en la previsión de ventas B2B y revelar posibles dificultades en el camino” como lo exponen en su estudio (Rohaan et al., 2022), mediante el cual consiguieron mejorar en 250% la tasa de precisión de las ventas, respecto al proceso manual que se llevaba a cabo en una compañía del sector B2B.

Así mismo, trabajos como el de (Mauricio et al., 2016), proponen utilizar herramientas de machine learning, especialmente de segmentación y clasificación para obtener la priorización de zonas de mercado, dando “solución a la incertidumbre que existe en la mayoría de las organizaciones en torno a la prioridad que tiene una zona de mercado”.

Lo anterior fue desarrollado mediante la “búsqueda y evaluación de los criterios más relevantes que las empresas tienen en cuenta para asignar prioridades a ciertos clientes” y cuyo objetivo fue mejorar las posiciones estratégicas de las organizaciones (Mauricio et al., 2016).

En línea con estos trabajos, este proyecto desarrolla un avance importante en la clasificación, predicción (forecasting) y segmentación de los afiliados a Caja de Sueldos de Retiro de la Policía Nacional – Casur.

Casur “es un establecimiento público del orden nacional con autonomía administrativa y patrimonio independiente, adscrito al Ministerio de Defensa Nacional” (Casur, 2021).

La misión de Casur es desarrollar las políticas y planes generales para el reconocimiento y pago de las asignaciones mensuales de retiro, al igual que desarrollar los programas que adopte el Gobierno Nacional en cuanto a seguridad social, en coordinación con la Dirección de Bienestar Social de la Policía Nacional para el personal de Oficiales, Suboficiales, Agentes y demás estamentos de la Policía Nacional, a quienes la Ley otorgue este derecho, así como la institución a sus beneficiarios (Casur, 2021).

Teniendo en cuenta lo anterior, la entidad no actúa como una empresa privada que debe competir contra otras empresas para mantener su posición competitiva, más bien actúa como un monopolio natural, en el cual “resulta más conveniente que un bien o servicio sea producido por una sola empresa que por dos o más” (Agencia Nacional de Infraestructura, 2021).

Por lo anterior, la entidad no se encuentra regida por las reglas de la competencia de mercado (Rumelt, 2012) y, sin embargo, sus indicadores de calidad y las normas internas y externas -como entidad pública- requieren que mejore cada día la oferta y calidad de los servicios que ofrece a sus afiliados, es decir, a todos los policías que tienen derecho a las asignaciones de retiro pensional.

Para lograr este macroobjetivo, la entidad ha planteado dentro de su “plan Estratégico cuatrienal 2019-2022” (Casur, 2021) la innovación en la oferta de valor de Casur, el cual incluye el **programa de creación de valor para los afiliados**. Y es en ese sentido que se hace indispensable conocer, mediante su

adecuada caracterización, a la población atendida directa e indirectamente, es decir, el afiliado y su familia.

En consecuencia, con el objetivo de mejorar los resultados de previsión que permitan mejorar la efectividad de la oferta para quienes sí están interesados en servicios de salud, en este trabajo se utilizan modelos de Machine Learning en cuanto a clasificación principalmente, y subsidiariamente en cuanto a segmentación (clústering), como una forma de dar solución a esta necesidad.

De igual forma, es importante resaltar que la Caja cuenta con más de 50.000 afiliados, y tiene una base de datos con información que permite el conocimiento de ciertas características sociodemográficas y de preferencias, gustos y necesidades, para 8.000 afiliados.

Por lo tanto, este trabajo pretende mejorar la clasificación de los afiliados interesados en servicios de salud para ellos y sus grupos familiares, para el conjunto global de las 53.000 personas atendidas por Casur.

En consecuencia, la Caja requiere conocer a sus “clientes”, comprenderlos y poder ofrecer servicios ajustados a sus gustos y necesidades, así como en la industria bancaria, los bancos necesitan de un perfilamiento de clientes, que, entre otras variables, ayuda a las entidades a comprender mejor a sus clientes actuales y potenciales (Dawood et al., 2019).

Con este objetivo presente, se ha decidido que se utilizarán las herramientas para la segmentación (o clústering) y la clasificación, la cual permite hacer proyecciones (forecastings) sobre el grupo total de clientes o consumidores, como se evidencia en (Chen & Lu, 2017; Dawood et al., 2019; Mauricio et al., 2016; Rincon et al., 2021; Rohaan et al., 2022).

1. Metodología

Inicialmente, es este trabajo se enmarca en una investigación descriptiva, donde se parte de un conocimiento básico sobre algunos afiliados, y con base en dicha información se desea conocer el comportamiento más probable de todo el universo poblacional de los afiliados, en cuanto a su interés en servicios adicionales en salud.

Es importante resaltar que la mayoría de los afiliados entrevistados muestran una tasa de interés en servicios de salud muy minoritaria, es decir, solo una pequeña fracción de los afiliados que han participado en las encuestas de conocimiento de la Caja, han expresado su interés en este tipo de servicios. Lo cual implica problemas de desbalance de la variable objetivo que se mencionan más adelante.

De forma complementaria, se precisa que este estudio es un paso inicial para predecir el comportamiento o las preferencias del universo poblacional en cuanto a otros temas, tales como servicios financieros, de empleo, educativos y turísticos.

Para el desarrollo de este trabajo, se desarrolló un ejercicio de clústering, siguiendo, de alguna manera, los desarrollos de (Chen & Lu, 2017), enfocados en establecer cuáles de los usuarios que no han sido sujetos de encuestas, estarán interesados y consumirán servicios adicionales (o complementarios) de salud. Aunque los resultados de este método no son tenidos en cuenta en el resto del trabajo, cabe resaltar que parecen muy prometedores para próximos análisis.

Además, se parte del supuesto que Casur no tiene intenciones de comportarse en el mercado de forma similar a aquellas empresas que participan en los modelos de negocio denominados OGB (Leong et al., 2019)[1].

Respecto a la data, es producto de iniciativas, recientemente desarrolladas, con el objeto de conocer a los afiliados, por lo cual contienen información sociodemográfica, así como información respecto a los gustos y necesidades principales de los afiliados a lo largo del país, e incluso fuera de él. Además, está compuesto por más de 8.000 usuarios entre hombres y mujeres, en un amplio rango de edad, y por ende cierta diversidad de gustos, necesidades y formas de pensar; algunos con familia y otros solteros, etc.

También es importante tener presente que la información no cuenta con datos en blanco o vacíos importantes, salvo para el lugar de residencia, por lo que se eliminó esta variable (que tampoco ofrece una capacidad explicativa importante).

[1] En los cuales se concentra una masa importante de clientes para obtener precios y descuentos importantes en un mismo grupo de productos o servicios. Dentro de este modelo de negocio los principales exponentes son Fave (antes Groupon), Gomaji, LivingSocial, MyDeal, MyMart, StreetDeal, y Huala (Leong et al., 2019); más conocidos quizás en el continente asiático

Enfoque Metodológico:

El problema se ha encarado teniendo en cuenta que las herramientas de Machine Learning son un paso **inicial** que permite identificar y atender el nicho específico de la forma más eficiente posible, desde el punto de vista del costo de adquisición de los clientes de acuerdo con (Ang & Buttle, 2006).

Para ello, se parte del supuesto que el interés, y consecuente adquisición, es cierto, sigue el teorema del límite central de acuerdo con (Levin & Rubin, 2004) y que, además, tiene una distribución probabilística binomial (Alvarado Verdin, 2014), así:

$$P(r,n,p) = \frac{n!}{r!\,(n-r)!}\, p^r \cdot q^{(n-r)}$$

Donde,

r = Cantidad de personas interesadas en servicios de salud (compradores). Para el caso, este r se ha definido en igual o mayor que uno (1)

n = Cantidad de personas a las que comercialmente debe contactarse para presentar los servicios de salud.

p = Es la probabilidad de éxito tomada del modelo de Machine Learning. Para el caso del modelo logístico, se obtiene de la matriz de confusión y corresponde a la división de los verdaderos positivos entre la suma de los verdaderos positivos y los falsos positivos. Para el caso del clúster, corresponde al porcentaje de interesados dentro del clúster en el cual hay más interesados en servicios de salud.

q = probabilidad de fracaso (probabilidad de no interés en los servicios de salud, deducida del modelo de Machine Learning mediante: 1 -p).

Es decir, teniendo en cuenta que los modelos de Machine Learning tienen asociados un error se quiere calcular la cantidad de acercamientos comerciales para conseguir una probabilidad de éxito, ya que esto permitirá estimar los costos de adquisición, lo cual es de suma importancia para la institución como para cualquier otra organización en el mercado.

Por lo tanto, si la probabilidad de vender al menos un servicio es cercana al 99% mediante subgrupos pequeños, entonces Casur podrá establecer acuerdos con los proveedores de servicios de salud, que representen mayores beneficios para sus afiliados.

Pero si son necesarios subgrupos más grandes para obtener una probabilidad de éxito del 99%, entonces los beneficios a los que puede aspirar serán menores.

Más adelante se muestra el tamaño de los subgrupos que se requieren para obtener una probabilidad de venta del 99%, y el impacto que ello tiene sobre el costo de adquisición de clientes.

La asignación de probabilidad depende de un modelo de Machine Learning ajustado y confiable, que permita reducir la incertidumbre y establecer los más específicamente posible a quienes se deben “atacar” comercialmente.

1.1. Técnica y diseño

Ahora bien, con el fin de lograr estos resultados, se utilizó Python para el tratamiento y análisis de la información, y dentro de éste, las librerías principales fueron Pandas, Scikit Learn, SweetViz y Pycaret. Las primeras para desarrollar análisis descriptivos de conocimiento de la información, análisis bivariados y correlaciones entre variables; y la última para correr los modelos utilizados.

Finalmente, los pasos utilizados partieron desde el conocimiento intensivo de la información y explotar este conocimiento adquirido, pues no siempre es posible estar familiarizado con la información.

De esta manera, y teniendo en cuenta que la información no tiene valores nulos ni vacíos, se desarrolló un análisis cualitativo de la información, seguidamente se crearon dos modelos de reglas: uno de reglas sencillas, partiendo del supuesto que las afiliadas (género femenino) son más propensas a estar interesadas en servicios de salud.

El segundo, de reglas compuestas, en donde se revisaron todas las variables y manualmente se ajustaron las variables a incluir de acuerdo con el conocimiento adquirido en el análisis cualitativo de la información o análisis descriptivo. Con este modelo se obtuvo un F1 score y se armó una matriz de confusión inicial.

Adicionalmente, al momento de crear un modelo de reglas, tanto el sencillo como el de reglas compuestas, fue necesario hacer un cruce normalizado de las variables independientes respecto a la variable de interés, lo que permitiría conocer mejor las variables que contaban con algún valor con representatividad sobre la variable de interés.

Paso seguido, se transformaron las variables. Las cualitativas con OneHotEncoder, y las numéricas se normalizaron. Así se generó un modelo de

regresión logística que evidenció problemas con el desbalanceo mencionado anteriormente.

Por ende, se creó un modelo que permitiese el balanceo de clases, el método por defecto que utiliza esta herramienta es el método SMOTE (Ackerman et al., 2021). Con ello, se obtuvo un nuevo F1-score y otra matriz de confusión.

Posteriormente, a este modelo se le introdujo un umbral de decisión probabilístico, teniendo en cuenta los resultados de la matriz de confusión mostraban que el modelo era más propenso a predecir falsos negativos mediante la reducción del número de verdaderos positivos, lo cual afectaba la cantidad de personas que podrían atacarse comercialmente.

Para complementar el trabajo, se corrieron modelos K-means de clústering, porque permiten encontrar las características que debe tener un conjunto de personas para ser parte del grupo de interés (aquellos que sí están interesados en servicios de salud).

Finalmente, siguiendo la estructura de (Chen & Lu, 2017), se corrió un modelo de regresión logística entre el clúster obtenido anteriormente y la variable de interés, basados en que este método "combinado" puede mejorar las previsiones sobre la base de datos total, mejorar métricas del modelo y, por consiguiente, las asignaciones de probabilidad de interés en servicios de salud para cada persona.

2. RESULTADOS

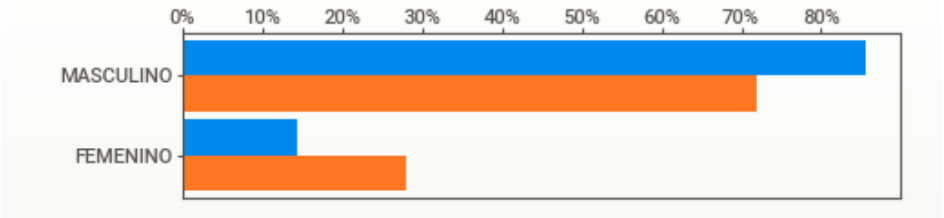
Las variables con las que se cuenta son el género, estado civil, grado de retiro (es decir el grado que tenía la persona en la institución), el municipio y departamento de residencia, el estrato al que pertenece su vivienda, si estaría interesado en recibir información sobre oportunidades laborales, la categoría de retiro (que es una variable más detallada que el grado), el nivel de escolaridad (primaria, bachillerato, universidad, especialización, maestría o doctorado, o alguna de las anteriores sin terminar), la fecha de nacimiento de la que se deduce la edad, el país donde reside, si tiene cónyuge (que es más específico que la variable estado civil, aunque están muy relacionadas), si cuenta con medicina prepagada, cuáles son sus hobbies, si tiene algún tipo de discapacidad, y si está interesado en recreación, vivienda o salud. Lo anterior se resume en la tabla 1.

Tabla 1. Resumen de las variables iniciales

VARIABLES		
GÉNERO	NIVEL DE ESCOLARIDAD	HOBBIES
IDENTIFICADOR ALEATORIO	FECHA DE NACIMIENTO	DISCAPACIDADES2
ESTADO CIVIL	PAÍS DE RESIDENCIA	ESTÁ INTERESADO EN RECREACION
GRADO DE RETIRO	ESPECILIDAD 01	ESTÁ INTERESADO EN VIVIENDA
MUNICIPIO	DISCAPACIDADES	ESTÁ INTERESADO EN SALUD
DEPARTAMENTO	TIENECONYUGE	ESTÁ INTERESADO EN OFERTAS LABORALES
ESTRATO	NUMERODEHIJOS	ESTÁ INTERESADO EN CREDITOS
¿DESEA RECIBIR INFORMACIÓN DE OPORTUNIDADES LABORALES?	OTROSDEPENDEN	ESTÁ INTERESADO EN EDUCACION
CATEGORIA DE RETIRO	MEDICINAPREPAGADA	

Con estas variables, se realizó un análisis bivariado, y se filtraron las variables que podrían tener algún valor para el modelamiento posterior. Por ejemplo, para la variable género, en la figura 1 se encontró que las mujeres muestran una preferencia mayor por los servicios de salud que los hombres, proporcionalmente hablando, ya que tienen un mayor interés (Ver figura 1). Si bien, la cantidad de mujeres es menor que la de hombres, comparativamente hablando, la barra naranja (SI interés en servicios de salud) es mayor que la azul.

Figura 1. Interés en servicios de salud por género. La barra azul significa NO interés en salud. La barra naranja significa SI interés en salud. Del 100% de las personas NO interesadas, más del 85% son hombres. Del 100% de las personas interesadas, el 30% son mujeres.



Lo anterior se replicó para todas las variables, y se desarrolló el modelo de reglas sencillas y el modelo de reglas compuestas. Los modelos de reglas filtraban de algunas variables sus valores, y en caso afirmativo, predecían que sí había interés en salud Puede ver las condiciones de cada regla en el Anexo B). El modelo de reglas sencillas, simplemente predijo interés en salud cuando la variable género era mujer. Por su parte, el modelo de reglas compuestas tuvo en cuenta diferentes variables, y si encontraba alguna condición clasificaba la

predicción como Sí hay interés. De estos modelos, la tabla 2 muestra los resultados de sus métricas.

Tabla 2. Métricas de modelo de reglas y del modelo de regresión logística

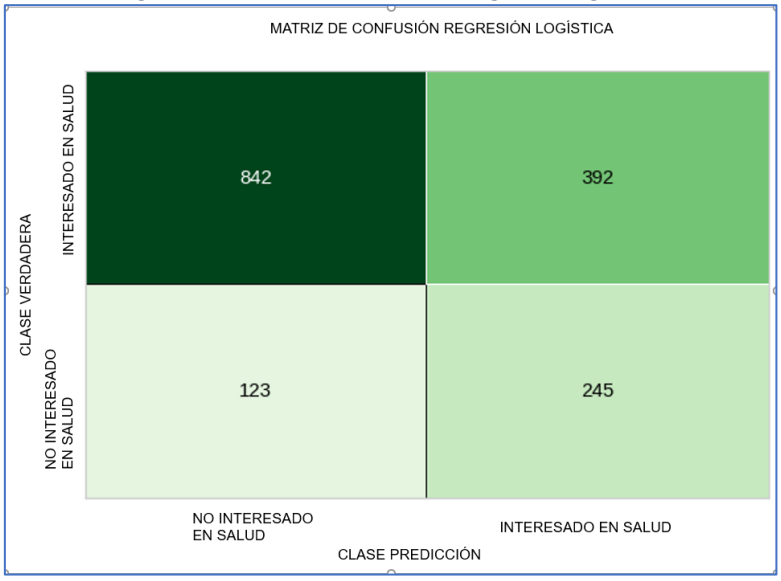
MODELO	F1-SCORE
Regla sencilla	31,75%
Reglas compuestas	43,73%
Regresión Logística	47,92%

Posteriormente, se corre un modelo de clasificación con un arreglo de desbalanceo utilizando la técnica SMOTE, el modelo de **regresión logística** generó un F1-score del 47,92%, como se registra en la tabla 2.

Así mismo, la matriz de confusión obtenida muestra un número reducido de falsos negativos (7,67% del total) en favor de una alta tasa de verdaderos negativos (52,55% del total).

Lo anterior afecta el grupo de afiliados a identificar, por cuanto el modelo reduce la cantidad de positivos predichos, siempre, en favor de predecir los negativos (Figura 2).

Figura 2. Resultados del modelo de Regresión Logística.



Es importante tener en cuenta el umbral de decisión probabilístico, debido al costo de adquisición de los clientes, puesto que, si el costo de adquisición es relativamente bajo comparado con la rentabilidad esperada, entonces es mejor tener una base de 1.362 individuos y que haya un interés por parte del 26,13%

(356 verdaderos positivos) que tener una base de potenciales interesados de 637 individuos, y solo venderle a 245.

Ahora bien, si el costo de adquisición es alto, comparado con la rentabilidad del esperada, entonces es mejor acercarse a 637 afiliados y obtener interés de 245 (38,46%).

El análisis del presente trabajo no tiene en cuenta este umbral de decisión porque se incluye una aproximación al modelo de costos de adquisición para la Caja, la cual no con un modelo de gestión de los costos de adquisición, de acuerdo con (Ang & Buttle, 2006).

Lo anterior implica un costo de adquisición comparativamente alto respecto a la rentabilidad esperada, dado que la Caja no tiene ingresos por la venta de servicios de salud (no es su modelo de negocio).

Posteriormente, se utilizaron modelos de clústering, cuyas variables fueron preprocesadas y normalizadas.

Paso seguido, se corrió el modelo, sin la variable dependiente. Los resultados de ambos modelos se detallan en la tabla 3, y muestran que el modelo sin la variable objetivo tiene mejores métricas, con una mejora del 11,31% en el coeficiente de silueta, y del 8,92% en el índice de Davies-Bouldin.

Tabla 3. Métricas de modelos de Clustering por K-means

MODELO	CLUSTERS	COEF. DE SILUETA	ÍNDICE DAVIES-BOULDIN
Clúster inicial con variables de interés	4	0.2166	1.5969
Clúster sin variable objetivo	5	0.2411	1.4544

Mediante este ejercicio, se procedió a revisar cada uno de los clústers, con el ánimo de identificar cuál era el clúster más afín al interés en servicios de salud. Esto permitió evidenciar que el clúster 2 es el más propenso a adquirir servicios de salud, cuyas características principales se resumen en la siguiente tabla:

Tabla 4. Principales características del clúster 2

Características
● Grupo mayoritariamente compuesto por mujeres (98.94%) ● Grupo con categoría de retiro, compuesto en un 81% por Beneficiarios por sustitución. ● El 48.27% del clúster tiene solo primaria ● El 90.15% no tiene cónyuge ● Más del 75% del clúster se concentra en 0 hijos ● El promedio de edad ronda los 70.6 años

Adicionalmente, el clúster 2 llama la atención porque es el único en el cual el 44,97% de sus integrantes está interesado en adquirir servicios de salud, lo que duplica la tasa de interés de la base de datos, en la que solamente el 22,92% de los integrantes tiene interés (El lector puede revisar las características de los demás clústeres en el Anexo C)

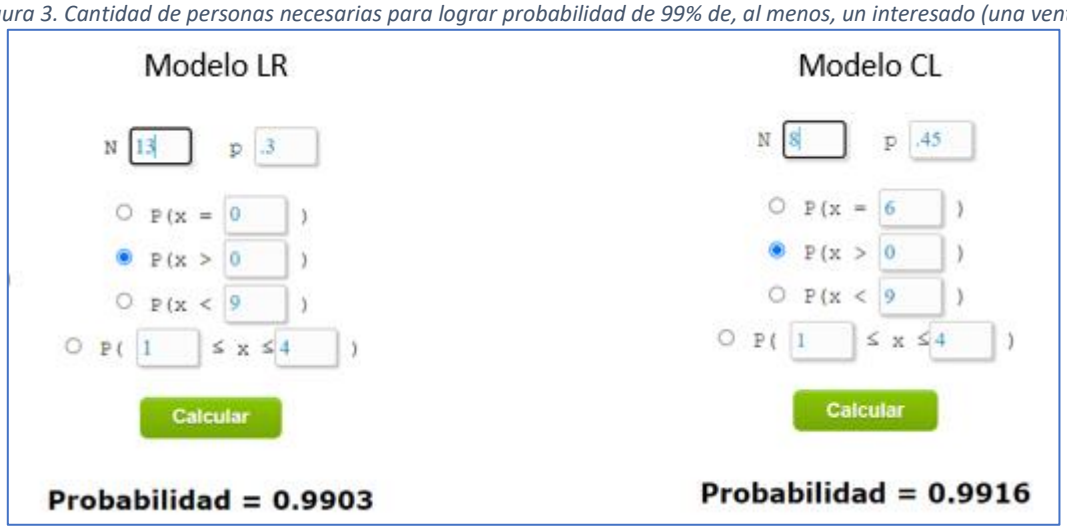
INTEGRACIÓN DE LOS RESULTADOS DE LOS MODELOS DE MACHINE LEARNING, CON EL MODELO DE DISTRIBUCIÓN BINOMIAL, Y LA APROXIMACIÓN AL MODELO DE COSTO DE ADQUISICIÓN DE CLIENTES

Al realizar la predicción sobre la base de datos de los 53.238 afiliados, mediante cada uno de los modelos, se tiene que el modelo de regresión logística (en adelante LR) clasifica como interesados en salud a 9.797 personas, y maneja una probabilidad de éxito del 30% aproximadamente. Es decir, de cada 100 personas que, en el testeo, predice como interesados en servicios de salud, 30 son realmente interesados en salud.

Respecto al modelo de clústering (en adelante CL), se tiene una tasa del 45% de éxito y, al realizar la predicción con la base de datos total, se ubican a 4.727 personas en ese clúster.

Utilizando el modelo de distribución probabilística binomial (Alvarado Verdin, 2014), se tiene que, para lograr una probabilidad igual o superior al 99% de hallar al menos a un interesado (comprador), es necesario acercarse comercialmente a 13 personas para el modelo LR o acercase comercialmente a 8 personas para el modelo CL, según la información que se tuvo de la calculadora de distribución binomial de la figura 3.

Figura 3. Cantidad de personas necesarias para lograr probabilidad de 99% de, al menos, un interesado (una venta)



Referencia: (Calculadoras Online, 2021)

De acuerdo con Casur, el costo de acercarse comercialmente a 30 afiliados, con dedicación a tiempo completa, incluyendo un proceso de capacitación en los servicios ofrecidos (para este caso, los servicios de salud), con un espacio de trabajo adecuado e instalaciones de telecomunicaciones, así como papelería y otros gastos de mercadeo, cuesta alrededor de $6.000.000, en donde la mayor dificultad está en contactar al afiliado y el mayor rubro del costo se concentra en el personal.

Así las cosas, el modelo LR ofrece la posibilidad de atender 2,3 grupos, donde se espera que, por lo menos, se obtenga a 1 interesado por grupo, lo que significa que se obtiene un promedio de 2,3 personas (clientes).

Ahora bien, si el costo de obtener esas 2,3 personas es de $6.000.000, entonces el costo de adquisición esperado es de $2.608.695.

Por su parte, el modelo CL ofrece la posibilidad de atender 3,75 grupos, es decir, se espera obtener como mínimo 3,75 personas, lo que tiene un costo de adquisición de $1.600.000.

Al observar los resultados de los modelos supervisado (LR) y no supervisado (CL), resulta claro que es más eficiente para el modelo de costos de adquisici{on de Casur, el modelo No supervisado o de clustering.

CONCLUSIONES

Las variables probablemente no sean las mejores, pero al correr los modelos se logra reducir la incertidumbre. Queda claro que, si bien los modelos no son perfectos y sus métricas son bastante deficientes, se reduce drásticamente la incertidumbre, con la enorme ganancia de reducir la base de los 53.000 afiliados como potenciales interesados, a 9.797 personas en el modelo de regresión logística, y a poco más de 4.700 personas en el modelo de clustering.

Igualmente, se rescata que realizar un modelo con umbral de decisión probabilístico funciona para otras realidades presupuestarias en cuanto al costo de adquisición de clientes y su rentabilidad esperada, no así para Casur.

De otro lado, realizar un modelo de clusterización, para posteriormente desarrollar un modelo de regresión logística, no fue tenido en cuenta finalmente, pues, aunque sus resultados son mejores que los del modelo de regresión logística, no es claro a qué se deben esos resultados.

Adicionalmente, utilizar modelos de machine learning con análisis de costos y con análisis de distribución de probabilidad, complementa los resultados obtenidos y brinda certeza de analizar el problema desde diferentes ópticas.
Finalmente, es mejor tener cualquiera de los dos modelos que intentar contactar a 53.000 personas para definir cuales sí estarán interesados y quienes no. Evidentemente, el modelo de clustering, evaluado desde la perspectiva del costo de adquisición y del porcentaje de probabilidad que asigna a cada afiliado se muestra como superior para este estudio específico, teniendo en cuenta la realidad de Casur.

REFERENCIAS BIBLIOGRÁFICAS

Ackerman, M., Ben-David, S., Branzei, S., & Loker, D. (2021). Weighted clustering: Towards solving the user•s dilemma | Elsevier Enhanced Reader. *Pattern Recognition*, *120*, 1–13. https://reader.elsevier.com/reader/sd/pii/S0031320321003393?token=D9CB33CF4BCF9D20A4661B59455F3C3A51C2B08C54A28A3EDABD7BB01

DA1BFA34B52CBB531FC0040CBEE5EBFD6809BAF&originRegion=us-east-1&originCreation=20211031232819

Agencia Nacional de Infraestructura. (2021, November). *Monopolio Natural*. Https://Www.Ani.Gov.Co/Glosario/Monopolio-Natural.

Alvarado Verdin, V. (2014). *Probabilidad y Estadística* (Vol. 1). Grupo Editorial Patria.

Ang, L., & Buttle, F. (2006). Managing For Successful Customer Acquisition: An Exploration. *Journal of Marketing Management*, *22*(3–4), 295–317. https://doi.org/10.1362/026725706776861217

Calculadoras Online. (2021, December 1). *Calculadora de Distribucion Binomial Online - Probabilidad Binomial*. Https://Calculadorasonline.Com/Distribucion-Binomial-Probabilidad-Binomial/.

Casur. (2021, November). *Casur - La Entidad*. Https://Www.Casur.Gov.Co/La-Entidad. https://www.casur.gov.co/la-entidad

Chen, I.-F., & Lu, C.-J. (2017). Sales forecasting by combining clustering and machine-learning techniques for computer retailing. *Neural Computing and Applications*, *28*(9), 2633–2647. https://doi.org/10.1007/s00521-016-2215-x

Dawood, E. A. E., Elfakhrany, E., & Maghraby, F. A. (2019). Improve Profiling Bank Customer's Behavior Using Machine Learning. *IEEE Access*, *7*, 109320–109327. https://doi.org/10.1109/ACCESS.2019.2934644

Leong, L.-Y., Hew, T.-S., Ooi, K.-B., & Tan, G. W.-H. (2019). Predicting actual spending in online group buying – An artificial neural network approach. *Electronic Commerce Research and Applications*, *38*, 100898. https://doi.org/10.1016/j.elerap.2019.100898

Levin, R., & Rubin, D. (2004). *Estadística para Administración y Economía* (G. Trujano, Ed.; 7th ed.). Prentice hall.

Mauricio, H., Albán, G., Pablo, J., Cabrera, O., Ancízar, Ó., Achipiz, S., José, J., & Bastidas, B. (2016). APLICACIÓN DE MAPAS DE KOHONEN PARA LA PRIORIZACIÓN DE ZONAS DE MERCADO: UNA APROXIMACIÓN PRÁCTICA APPLICATION OF KOHONEN MAPS FOR THE PRIORITIZATION OF MARKET AREAS: A PRACTICAL APPROACH.

Revista EIA, *13*, 157–169. https://doi.org/10.14508/reia.2016.13.25.157-169

Rincon, N. F., Gonzalez, D., & Industrial, I. (2021). *Ejecutar un modelo de machine learning para identificar los clientes potenciales basados en un proceso probabilístico para la empresa dell technologies*. https://repository.unimilitar.edu.co/handle/10654/38562#.YX3nbAqpmGI.mendeley

Rohaan, D., Topan, E., & Groothuis-Oudshoorn, C. G. M. (2022). Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider. *Expert Systems with Applications*, *188*, 115925. https://doi.org/10.1016/j.eswa.2021.115925

Rumelt, R. P. (2012). Good Strategy/Bad Strategy: The Difference and Why It Matters. *Strategic Direction*, *28*(8), sd.2012.05628haa.002. https://doi.org/10.1108/sd.2012.05628haa.002

LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Anexo A. Resumen del clúster 2

	GÉNERO	CATEGORIA DE RETIRO	NIVEL DE ESCOLARIDAD	TieneConyuge	NumeroDeHijos	Edad	salud_modif	Cluster	interes_salud
count	1046	1046	1046	1046	1046.000000	1046.000000	1046.000000	1046	1046
unique	2	6	14	2	NaN	NaN	NaN	1	2
top	FEMENINO	BENEFICIARIO/A POR SUSTITUCION	PRIMARIA	NO	NaN	NaN	NaN	Cluster 2	NO
freq	1035	848	505	943	NaN	NaN	NaN	1046	586
mean	NaN	NaN	NaN	NaN	0.272467	70.615679	0.439771	NaN	NaN
std	NaN	NaN	NaN	NaN	0.613522	10.212081	0.496597	NaN	NaN
min	NaN	NaN	NaN	NaN	0.000000	42.000000	0.000000	NaN	NaN
25%	NaN	NaN	NaN	NaN	0.000000	63.000000	0.000000	NaN	NaN
50%	NaN	NaN	NaN	NaN	0.000000	72.000000	0.000000	NaN	NaN
75%	NaN	NaN	NaN	NaN	0.000000	78.000000	1.000000	NaN	NaN
max	NaN	NaN	NaN	NaN	4.000000	96.000000	1.000000	NaN	NaN

Anexo B. Figura de condiciones de los modelos de reglas

Reglas sencillas	
Variable	**Valor**
Género	Femenino

Reglas compuestas	
Variable	**Valor**
Categoría de retiro	BENEFICIARIO/A POR SUSTITUCION', 'PENSIONADO CIVIL'
Nivel de Escolaridad	DOCTORADO INCOMPLETO', 'PRIMARIA', 'NINGUNA', 'PRIMARIA INCOMPLETA', 'DOCTORADO', 'SECUNDARIA INCOMPLETA'
TieneConyuge	No
NumeroDeHijos	0
Edad	>67

LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Anexo C. Comportamiento de otros clústeres

Cluster 0.

	GÉNERO	CATEGORIA DE RETIRO	NIVEL DE ESCOLARIDAD	TieneConyuge	NumeroDeHijos	Edad	salud_modif	Cluster	interes_salud
count	2282	2282	2282	2282	2282.000000	2282.000000	2282.000000	2282	2282
unique	2	6	13	2	NaN	NaN	NaN	1	2
top	MASCULINO	AGENTE	SECUNDARIA	SI	NaN	NaN	NaN	Cluster 0	NO
freq	2207	2021	1884	2038	NaN	NaN	NaN	2282	1825
mean	NaN	NaN	NaN	NaN	0.968449	58.539001	0.200263	NaN	NaN
std	NaN	NaN	NaN	NaN	0.790043	5.365674	0.400285	NaN	NaN
min	NaN	NaN	NaN	NaN	0.000000	43.000000	0.000000	NaN	NaN
25%	NaN	NaN	NaN	NaN	0.000000	55.000000	0.000000	NaN	NaN
50%	NaN	NaN	NaN	NaN	1.000000	59.000000	0.000000	NaN	NaN
75%	NaN	NaN	NaN	NaN	2.000000	62.000000	0.000000	NaN	NaN
max	NaN	NaN	NaN	NaN	3.000000	76.000000	1.000000	NaN	NaN

Cluster 1.

	GÉNERO	CATEGORIA DE RETIRO	NIVEL DE ESCOLARIDAD	TieneConyuge	NumeroDeHijos	Edad	salud_modif	Cluster	interes_salud
count	1264	1264	1264	1264	1264.000000	1264.000000	1264.000000	1264	1264
unique	2	5	14	2	NaN	NaN	NaN	1	2
top	MASCULINO	NIVEL EJECUTIVO	SECUNDARIA	SI	NaN	NaN	NaN	Cluster 1	NO
freq	1212	740	559	1160	NaN	NaN	NaN	1264	1096
mean	NaN	NaN	NaN	NaN	3.513449	51.104430	0.132911	NaN	NaN
std	NaN	NaN	NaN	NaN	0.888820	8.159738	0.339613	NaN	NaN
min	NaN	NaN	NaN	NaN	3.000000	33.000000	0.000000	NaN	NaN
25%	NaN	NaN	NaN	NaN	3.000000	45.000000	0.000000	NaN	NaN
50%	NaN	NaN	NaN	NaN	3.000000	49.000000	0.000000	NaN	NaN
75%	NaN	NaN	NaN	NaN	4.000000	55.000000	0.000000	NaN	NaN
max	NaN	NaN	NaN	NaN	10.000000	90.000000	1.000000	NaN	NaN

Cluster 3.

	GÉNERO	CATEGORIA DE RETIRO	NIVEL DE ESCOLARIDAD	TieneConyuge	NumeroDeHijos	Edad	salud_modif	Cluster	interes_salud
count	1048	1048	1048	1048	1048.000000	1048.000000	1048.000000	1048	1048
unique	2	6	14	2	NaN	NaN	NaN	1	2
top	FEMENINO	BENEFICIARIO/A POR SUSTITUCION	PRIMARIA	NO	NaN	NaN	NaN	Cluster 2	NO
freq	1037	850	505	943	NaN	NaN	NaN	1048	587
mean	NaN	NaN	NaN	NaN	0.273855	70.630725	0.439885	NaN	NaN
std	NaN	NaN	NaN	NaN	0.613758	10.221150	0.496610	NaN	NaN
min	NaN	NaN	NaN	NaN	0.000000	42.000000	0.000000	NaN	NaN
25%	NaN	NaN	NaN	NaN	0.000000	63.000000	0.000000	NaN	NaN
50%	NaN	NaN	NaN	NaN	0.000000	72.000000	0.000000	NaN	NaN
75%	NaN	NaN	NaN	NaN	0.000000	78.000000	1.000000	NaN	NaN
max	NaN	NaN	NaN	NaN	4.000000	96.000000	1.000000	NaN	NaN

Cluster 4.

	GÉNERO	CATEGORIA DE RETIRO	NIVEL DE ESCOLARIDAD	TieneConyuge	NumeroDeHijos	Edad	salud_modif	Cluster	interes_salud
count	2036	2036	2036	2036	2036.000000	2036.000000	2036.000000	2036	2036
unique	2	5	14	2	NaN	NaN	NaN	1	2
top	MASCULINO	NIVEL EJECUTIVO	TECNICO	SI	NaN	NaN	NaN	Cluster 4	NO
freq	1822	1833	928	1659	NaN	NaN	NaN	2036	1826
mean	NaN	NaN	NaN	NaN	1.389489	47.205305	0.103143	NaN	NaN
std	NaN	NaN	NaN	NaN	0.732582	4.994748	0.304221	NaN	NaN
min	NaN	NaN	NaN	NaN	0.000000	7.000000	0.000000	NaN	NaN
25%	NaN	NaN	NaN	NaN	1.000000	44.000000	0.000000	NaN	NaN
50%	NaN	NaN	NaN	NaN	2.000000	47.000000	0.000000	NaN	NaN
75%	NaN	NaN	NaN	NaN	2.000000	50.000000	0.000000	NaN	NaN
max	NaN	NaN	NaN	NaN	2.000000	65.000000	1.000000	NaN	NaN