UNIVERSIDAD DE ANTIOQUIA

MODELOS DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE LA PREFERENCIA EN EL USO DE CANALES DE ATENCIÓN PARA UN FONDO DE PENSIONES Y CESANTÍAS

Autor(es)
Carolina Alvarez Florez
Leidy Tatiana Molina Ruiz

Universidad de Antioquia
Facultad de ingeniería
Medellín, Colombia

2021

Modelos de aprendizaje automático para la predicción de la preferencia en el uso de canales de atención para un fondo de pensiones y cesantías

Carolina Alvarez Florez

Leidy Tatiana Molina Ruiz

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:
Especialista en analítica y ciencia de datos

Asesores (a):
M. Sc. Antonio Jesús Tamayo Herrera

Línea de Investigación:
Inteligencia computacional

Grupo de Investigación:
In2Lab

Universidad de Antioquia
Facultad de Ingeniería
Medellín, Colombia
2021

Modelos de aprendizaje automático para la predicción de la preferencia en el uso de canales de atención para un fondo de pensiones y cesantías

Alvarez Florez, Carolina
Carolina.alvarezf@udea.edu.co
Molina Ruiz, Leidy Tatiana
Tatiana.molina1@udea.edu.co

RESUMEN

En este documento se presenta la exploración de una serie de modelos de aprendizaje automático, evaluados con el fin de predecir el canal de servicio que usará un cliente en un próximo contacto, teniendo en cuenta los canales de atención disponibles de un fondo de pensiones y cesantías en Colombia. La base de datos de entrenamiento se obtuvo a partir de los contactos de clientes registrados previamente, los cuales se complementaron con características demográficas disponibles para la compañía. Después de la exploración de varios modelos, el mejor clasificador hallado es un modelo de K vecinos más cercanos al que se le aplicó una selección secuencial de característica, llegando a un resultado de 0.69 en la métrica F1.

1 INTRODUCCIÓN

El análisis de las interacciones con clientes se encuentra en su momento más interesante en cuanto a la facilidad y disposición para la recolección de información a través de los diferentes medios y canales de contacto, facilitando así la implementación de modelos de inteligencia artificial que permitan entender a los clientes de acuerdo con los patrones de comportamiento revelados a través de los datos, para así crear estrategias, partiendo del análisis de estos y de la implementación de diferentes algoritmos, no de la heurística, como se ha hecho tradicionalmente.

Las organizaciones enfrentan cada día más desafíos y presiones con el aumento de los canales de venta digital y el cambio en la preferencia de los usuarios al consumir nuevos productos o servicios, dado que se incrementa la oferta y los medios por los cuales se pueden acceder a estos. Dentro de las organizaciones, la digitalización de los servicios se está convirtiendo en un pilar estratégicos, debido a las diferentes bondades en eficiencia y costos que estos conllevan. En el mundo de los servicios, las entidades financieras también se ven enfrentadas a este tipo de desafíos y han convertido la digitalización en una prioridad. Sin embargo, como lo dictan las teorías de consumo, es importante primero comprender las etapas de un proceso de compra o adquisición de los servicios que tiene los diferentes clientes y cómo es el perfil o la preferencia de los usuarios que navegan por internet [1].

Para llegar a comprender los comportamientos de un cliente a la hora de elegir un canal de consumo, es interesante hacerse la siguiente pregunta: ¿existen diferencias entre los consumidores que prefieren el uso de canales digitales versus los que prefieren el uso de canales tradicionales?, este interrogante es fundamental para las organizaciones a la hora de establecer una estrategia de relacionamiento con sus clientes. Algunos estudios empíricos han analizado los factores que pueden influir en la toma de decisiones de consumo de algún servicio basados en las ofertas digitales y tradicionales [1], no obstante, muchos otros han utilizado técnicas de modelado de datos para lograr un agrupamiento de los clientes que por sus características permitan un perfilamiento de estos en cuanto a sus preferencias.

Fortalecer la relación con los clientes es fundamental y es un trabajo continuo para las empresas, especialmente para las que ofrecen servicios financieros donde no se utiliza almacenamiento de producto dado que el consumo es en tiempo real, de ahí que resulta importante conocer qué factores o características son determinantes para que un cliente prefiera utilizar un canal y repetir la experiencia [2].

A través de este estudio se busca establecer las preferencias en el uso de canales de servicios de los

clientes de una entidad financiera, dado que a pesar de que la organización evaluada ha realizado esfuerzos para promocionar y disponer cada vez más servicios a través de diferentes canales digitales y de autoservicio, los clientes siguen utilizando los canales tradicionales como las oficinas de servicios y *call centers*, que son más costosos y congestionados. De aquí, surge la necesidad de poder predecir a partir de las características de un cliente y el motivo de su contacto, cuál es su canal de preferencia con el fin de brindar herramientas concretas para el diseño de estrategias de promoción y fortalecimiento de canales enfocados para diferentes tipos de clientes o servicios.

Para resolver este problema, se propone un análisis detallado de las diferentes características asociadas a cada cliente y a partir de este conocimiento, se prueban una serie de modelos de *machine learning*, los cuales se van ajustando hasta encontrar el modelo con la mejor configuración de parámetros para resolver la tarea propuesta.

2 ESTADO DEL ARTE

Algunos de los factores importantes para lograr identificar las preferencias de los clientes en el uso de canales digitales o tradicionales se basa en gran medida en las teorías de comportamiento del consumidor y manejo de la relaciones con el cliente CRM (*Customer Relationship Management*), donde se han realizado esfuerzos importantes para lograr predecir el comportamiento del cliente apalancados por diferentes estrategias, por tal razón es importante analizar cómo la analítica de datos ha logrado contribuir a esta labor, qué modelos se han utilizado o se utilizan actualmente y sobre todo qué características relevantes se han encontrado para la implementación de estos.

La herramienta idónea para solucionar un problema como el que se plantea en este trabajo, es la inteligencia artificial, la cual se puede definir como la capacidad de una máquina para realizar actividades que generalmente se asocian con el razonamiento o inteligencia humana [3], y dentro de este contexto se hará uso de modelos de *machine learning (ML)* que de acuerdo con lo descrito en [3], hace referencia a un subcampo de la inteligencia artificial que ofrece a las computadoras la capacidad de aprender sin una programación explícita, es decir, que se basa en algoritmos que le permiten generar un aprendizaje a partir de los datos para predecir resultados futuros.

2.1 SEGMENTACIÓN Y PERFIL DE LOS CLIENTES

Las entidades financieras, entre otras, son organizaciones que gozan de la oportunidad de tener de primera mano diferentes datos de los clientes para fortalecer su estrategia de relacionamiento; estos datos incluyen características sociales, demográficas, laborales y financieras que están disponibles para realizar una explotación y gestión confiable de la información con el fin de identificar cual sería la segmentación o mejor agrupación para los clientes según ciertas características, y perfilarlos en grupos o categorías de acuerdo con sus preferencias por ciertos productos o servicios, todo esto buscando fortalecer la estrategia de CRM y mejorar la experiencia en el contacto con la empresa.

La segmentación y el perfilamiento se ha venido aplicando a través de técnicas de minería y analítica de datos. En la literatura, se ha encontrado que generalmente se utilizan algoritmos de clasificación y agrupamiento en empresas del sector financiero con el fin de predecir el comportamiento del cliente y la toma de decisiones. En materia de segmentación, estas agrupaciones buscan principalmente asociar a los clientes según el nivel de riesgo que implica para la organización y los ingresos que estos pueden aportar, de este modo y de forma estándar, algunos clientes quedan perfilados como clientes de alto valor, valor medio y bajo valor [4].

Dependiendo de la necesidad del negocio también existen otras estrategias de segmentación, las cuales no están directamente relacionadas con el valor y nivel de riesgo de un cliente, sino con otros factores como la segmentación transnacional, que está muy ligada a los hábitos de consumo del cliente, hablando en términos de frecuencia de compra; otro factor podría ser la segmentación basada en el nivel de satisfacción del cliente y, finalmente, una última categoría que se puede aplicar al perfilamiento de los clientes cuando hablamos de navegación o ruta de compra, la cual se enfoca en el ámbito digital; en este escenario se utilizan características como el último canal usado o la frecuencia de consulta de servicios digitales, lo

que finalmente se traduce en definir el nivel o porcentaje de digitalización de los clientes [5].

Como se indicó anteriormente, para aplicar estos algoritmos en la creación de perfiles de clientes se utilizan tanto datos personales como datos transaccionales, los cuales pueden contener atributos como la edad, género, educación, ocupación, estado civil, ingresos, ubicación demográfica, estilo de vida, clase social, nivel de riesgo o historial de transacciones [4]. Otros trabajos incluso sugieren que también se deben incluir atributos psicográficos, los cuales incluyen aspectos éticos como los valores, y otros de comportamiento como los intereses y actitudes de las personas, así como sus tendencias de consumo [6].

Adicionalmente, otro factor importante que está tomando relevancia, aunque no hace parte del objeto de estudio, es la capacidad para procesar grandes volúmenes de datos apalancado de nuevas tecnologías, lo cual ha apoyado otro de los propósitos en el modelado de datos y la analítica en el comportamiento del cliente, el cual es tratar de descubrir elementos ocultos en la información a la cual se tiene acceso [7].

2.2 METODOLOGÍAS UTILIZADAS PARA SEGMENTACIÓN O PERFILAMIENTO DE LOS CLIENTES

La segmentación de los clientes ha estado muy relacionada con diversos estudios, que despiertan interés especialmente en los equipos de servicios y mercadeo; estos estudios se han desarrollado utilizado encuestas a grupos de interés y en algunos casos se han combinado con estadística descriptiva de los datos personales, características demográficas e información transaccional que las empresas van recolectando a través de diferentes plataformas [8]. Para los estudios que se han llevado a cabo únicamente a través de encuestas dirigidas se ha encontrado que estas no pueden ser escalables con el tiempo, o que se vuelven obsoletas dado que se ha evidenciado que la decisión de consumo de algún servicio en las personas puede cambiar con el paso de los años. Al revisar la literatura se ha encontrado incluso que estos estudios han presentado resultados contradictorios comparados con otros similares [9]. Todo esto, ha permitido reconocer en diferentes industrias que ciertas técnicas de recolección de datos como las encuestas, puede inyectar sesgo en los resultados obtenidos y es por este motivo que, a partir de dichos hallazgos, se recomienda buscar técnicas adicionales para fortalecer estos estudios, como el uso de datos transaccionales y características de los clientes que permitan volver los modelos escalables y ajustables con el tiempo [8].

Al revisar la aplicación de modelos analíticos en estudios practicados para diferentes industrias y sectores como el turístico, aéreo y hotelero, se puede encontrar similitud en los tipos de datos o características utilizadas, los cuales son modelados utilizando diferentes técnicas de *machine learning* mediante algoritmos de clasificación supervisados, como es el caso del *Random Forest* y las regresiones lineales, en este caso, si bien se han utilizado modelos clásicos, los resultados han sido relevantes en materia de segmentación y perfilamiento de los clientes [2].

Asimismo, se han realizado estudios que pretenden anticipar el comportamiento de los clientes utilizando técnicas o algoritmos de secuencias de comportamientos, donde se parte de la premisa de que a través del paso del tiempo y de los diferentes canales de contacto se puede "generalizar" la interacción que los clientes tienen con una organización. Para esto, se sugiere utilizar algoritmos de detección de secuencia como el SPMF (*Sequential Pattern Mining Framework*) [6], cuyos resultados se proponen para complementar modelos conceptuales tradicionales.

2.3 REDES NEURONALES

En la década iniciada en 1980, se dió el resurgimiento de las redes neuronales con la exploración de la inteligencia artificial, y su aplicación actual se centra principalmente en procesos industriales y comerciales. En el sector financiero las redes neuronales se han utilizado con frecuencia para la evaluación del riesgo del cliente asociado a operaciones de crédito y para evaluar la rentabilidad en materias de inversiones [10]. Teniendo presente que las redes neuronales en sus aplicaciones han obtenido un gran resultado y muestran una gran capacidad de predicción basándose en las señales de entrada y el conjunto de datos de entrenamiento, y que además estas tienen facilidad para relacionar clases similares, se expone

la gran utilidad que puede llegar a tener en un sistema de relacionamiento con el cliente [10].

Finalmente, es importante considerar que es una generalidad en la literatura, iniciar la segmentación, predicción o generación de hipótesis con el entendimiento de los datos, de cómo están compuestos, cómo se distribuyen, cuál sería un comportamiento normal o anómalo, cómo se generan, cómo se almacena, entre otras características que puedan ser relevantes para la comprensión y aprovechamiento de los resultados obtenidos.

3 DESCRIPCIÓN DEL PROBLEMA

3.1 PROBLEMA DE NEGOCIO

En búsqueda de la eficiencia y calidad en la prestación de servicio al cliente a través de los diferentes canales dispuestos por una compañía de pensiones y cesantías en Colombia, se hace necesario implementar un análisis basado en los datos, que permita a la empresa conocer mejor a sus clientes, sus gustos, necesidades y así llevar a cabo estrategias sustentadas en el conocimiento adquirido a través del proceso analítico, y no de manera intuitiva como se ha hecho tradicionalmente.

3.2 ALCANCE DEL PROYECTO

El estudio realizado se centra en los clientes de la empresa (personas naturales), se excluyen a empleadores ya que desde el negocio se sabe que tienen un comportamiento y necesidades sustancialmente diferentes.

Se considera una muestra de 35.000 contactos de los clientes, realizados entre febrero de 2020 y febrero de 2021 a través de los canales de servicio y registrados en el CRM de la compañía. Esta información se amplía con características relacionadas a la persona que llevó a cabo el contacto. Una vez realizado el preprocesamiento de los datos, el *dataset* resultante fue de 23.596 registros. En el presente trabajo se realizó un análisis descriptivo de cada una de las variables utilizadas para lograr el objetivo del proyecto, sin embargo, teniendo en cuenta que los datos corporativos son de carácter confidencial, aquí sólo se muestra parte de dicho proceso y los resultados obtenidos.

3.3 OBJETIVOS

3.3.1 OBJETIVO GENERAL

Implementar un modelo clasificador para predecir el canal de servicio preferido por un cliente para contactarse con la compañía de pensiones y cesantías.

3.3.2 OBJETIVOS ESPECÍFICOS

- Preprocesar la base de datos.
- Analizar el espacio de características disponibles en la base de datos del proyecto.
- Explorar modelos clasificadores de machine learning con diferentes configuraciones de hiperparámetros.
- Establecer qué características son las más relevantes para lograr predecir el canal preferido del cliente.
- Validar los modelos clasificadores para garantizar su capacidad de generalización.

4 METODOLOGÍA

El desarrollo del proyecto se llevó a cabo en diferentes etapas que incluyeron los pasos detallados en la Figura 1.

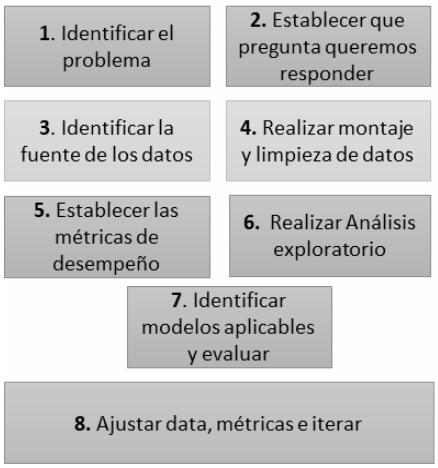

Figura 1 Pasos en la implementación del proyecto

Los dos primeros pasos consistieron en identificar las necesidades del negocio para tener claro los objetivos y alcances del proyecto. Como se mencionó anteriormente el problema abordado se centra en conocer las preferencias en el uso de canales que tiene los clientes de un fondo de pensiones y cesantías para mejorar la calidad en el servicio prestado.

Teniendo esto claro, los pasos siguientes consisten en identificar cuáles son los datos disponibles y su calidad. Inicialmente, se estableció un espacio de características para aproximadamente 10.000 registros con los que se construyó el primer *dataset*, el cual contempló los siguientes datos:

- Información del contacto del cliente como: fecha, canal y tema del contacto.
- Variables del negocio como el estado en cada una de las líneas de negocio y segmento del afiliado.
- Variables demográficas como la edad y ciudad de ubicación.

Después del preprocesamiento de los datos quedaron 8.146 registros, con estos se empezaron a realizar las primeras pruebas con modelos clasificadores paramétricos como: regresión logística y discriminante cuadrático. La evaluación de estos modelos se realizó utilizando el *accuracy* y el *Balance accuracy* como métricas de desempeño, esta última, teniendo en cuenta que se estaba trabajando con un modelo desbalanceado, dado que se encontraron tres canales de contacto predominante como la línea de servicios, oficina de servicio y oficina virtual. Durante esta etapa del proyecto, se exploraron algunos modelos no paramétricos como k vecinos más cercanos y árboles de decisión, pero para ninguno de los modelos se obtuvo un buen resultado, por lo que se realizó una búsqueda de hiperparámetros para este segundo grupo, llegando al mejor resultado obtenido hasta ese momento, con un accuracy de 0.40 para el modelo *Random Forest*.

Teniendo en cuenta los resultados anteriores, para las siguientes etapas, la variación consistió en aumentar el número de muestras a 35.000, con el fin de triplicar los datos iniciales y darle más posibilidades de aprender a los algoritmos. Después del preprocesamiento inicial de los datos, quedaron 23.596 muestras. Se volvieron a evaluar los modelos paramétricos y no paramétricos, y esta vez se optó por la métrica F1, con el fin de contemplar en igual medida tanto la precisión como el *recall* de la matriz de confusión generada. Bajo este escenario, los modelos que mejor se perfilaron fueron el *Random Forest* y el modelo de los k vecinos más cercanos, sin embargo, el resultado continuó estando por debajo del 0.5.

Posteriormente, se incluyeron variables adicionales al *dataset* inicial, con el fin de ampliar el vector de características utilizado para la predicción de la variable de interés y entregar nueva información discriminante a los modelos. Esta búsqueda consistió en revisar el estado del arte para determinar información utilizada en otros proyectos similares y que fueran relevantes para el resultado del modelo; se encontraron, para el sector financiero, estudios en los cuales se utilizó el salario y la ocupación del afiliado y en otros sectores se incluyeron variables más transaccionales como el uso de clave, en caso de tener un portal transaccional. Al validar al interior de la organización, se identificó que se contaba con una fuente confiable y de calidad para agregar estas características al dataset inicial. Más adelante se detallan cada una de las variables definitivas utilizadas en las siguientes etapas del proyecto.

Para lograr un mejor resultado, se realizó una búsqueda más amplia de los hiperparámetros para los modelos ya explorados, como los k vecinos más cercanos, adicionalmente, se evaluaron nuevos modelos como las máquinas de soporte vectorial y las redes neuronales artificiales. Estos experimentos se realizaron en la nube, en un ambiente provisionado por la compañía utilizando la plataforma de *Google Cloud Platform (GCP)*, pero considerando que se incrementó el número de características, el número de muestras, y que alguno de estos modelos tiene un costo computacional más alto, fue necesario utilizar otro ambiente con mayor capacidad de procesamiento en esta misma plataforma. El código fuente y los modelos implementados están publicados en un repositorio público de Github[1]

Otro punto importante a mencionar, es la selección de los datos de modo tal que, para un usuario en específico, únicamente se tuviera el canal de contacto más usado, esto con el fin de evitar ruido producido por el solapamiento de las muestras en el espacio de características, dado que un mismo usuario (representado con exactamente el mismo vector de características) pudo haber realizado

[1] https://github.com/caroalvarezf/Monografia_EACD_UDEA_2021.git

diferentes consultas por varios canales en el periodo de tiempo seleccionado.

4.1 DESCRIPCIÓN DE LOS DATOS

Los contactos de los clientes son registrados en la plataforma de CRM de la compañía y se almacena en *Google Cloud Platform* (GPC). Además de la información del contacto en un canal específico y motivo del contacto realizado, como se describió anteriormente, se incluyen datos del cliente para su caracterización demográfica, económica, ocupacional y geográfica. El *dataset* fue sometido a una limpieza inicial para validar los registros nulos o los datos que podrían homologarse. Después de este preprocesamiento, se analizaron sus distribuciones individuales y con relación a la variable que se pretende predecir, a continuación, se detallan cada una de las variables utilizadas dentro de la base datos final.

4.1.1 DATOS DEL CONTACTO

Canal Radicación: Variable a predecir. Es de tipo categórico, lo que define el problema tratado como de aprendizaje automático supervisado. Hace referencia al canal de atención utilizado por el cliente para la atención de su necesidad. Durante la exploración de los datos se identifican registros atípicos de canales que son de uso exclusivo de clientes empleadores o clientes internos, estos datos fueron eliminados del dataset ya que no hacen parte del alcance del proyecto. El canal de radicación tiene los siguientes 7 valores posibles:

- Oficina de servicio
- Línea de servicio
- Oficina virtual
- Portal web
- Estructura comercial
- Chat asesor
- Gestión documental

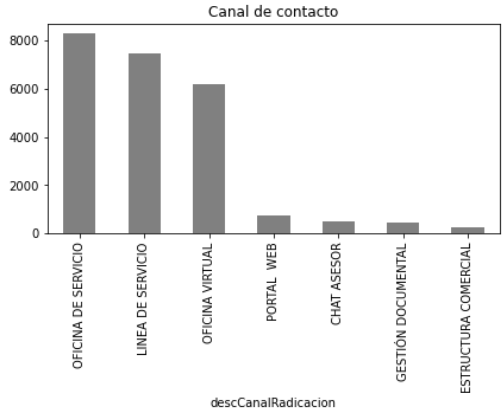

Figura 2 Distribución variable objetivo

Como se ve en la Figura 2, la distribución de la variable a predecir presenta un desbalance importante ya que los canales oficina de servicio, línea de servicio y oficina virtual representan el 90% de los datos.

Tema: Clasificación del motivo de contacto del cliente. Se hace una agrupación de temas similares y de temas que han cambiado de etiqueta, pero hacen referencia al mismo servicio. Es una variable categórica con los siguientes valores posibles:

- Actualización de datos
- Certificados
- Asesoría pensional
- Afiliaciones y traslados
- Aportes y planillas
- Pagos y retiros
- Canales de atención
- Movimientos de cuenta
- Saldos y rentabilidades
- Pensionados
- Información y comunicaciones

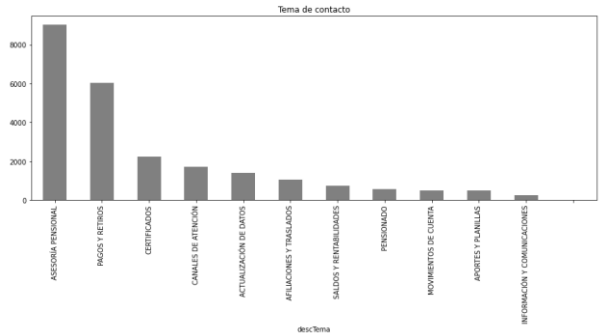

Figura 3 Distribución variable tema

En la Figura 3, se evidencia que la variable tema está concentrada en los valores de asesoría

pensional y pagos y retiros que representan el 63% del total de los datos.

4.1.2 DATOS DEL CLIENTE

Sexo*:* categorizado como:

- Masculino
- Femenino

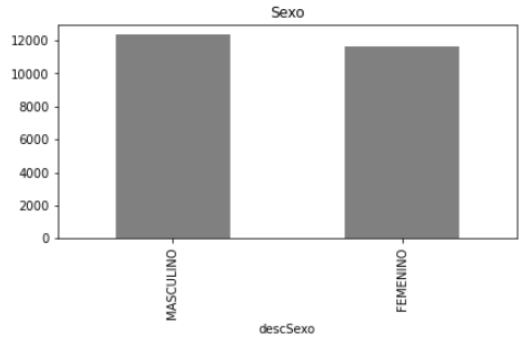

Figura 4 Distribución variable sexo

En cuanto al sexo de los clientes, en la Figura 4, se nota una distribución muy similar en relación al volumen, lo que indica que no hay una diferencia importante en el sexo de los clientes que se han contactado con la empresa durante el período analizado.

Edad Afiliado: Calculada a partir de la fecha de nacimiento hasta la fecha en que se realiza el análisis de los datos. Teniendo en cuenta que la empresa cuenta con clientes menores de edad, pero el contacto para cualquier servicio siempre lo harán sus tutores o representantes legales, se excluyen del análisis los clientes con menos de 18 años.

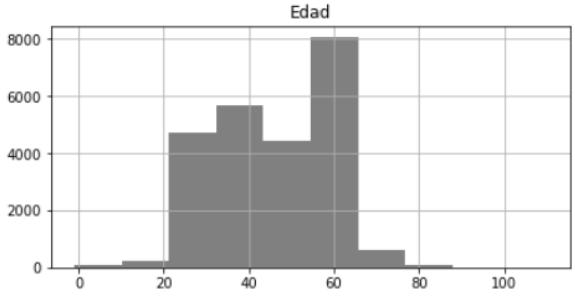

Figura 5 Distribución variable edad

En la Figura 5, se evidencia que se presenta una concentración marcada en los rangos de edades entre los 20 y 70 años. Esto se explica también por la naturaleza del negocio, razón por la cual se toman como datos atípicos los clientes con edades menores a 18 años y mayores a 90, y se eliminan del dataset.

Regional: Partiendo de la ciudad de ubicación del cliente, se hace una agrupación de acuerdo a la región del país en la cual reside. Esta agrupación se hace tomando como base la distribución regional definida por la empresa. Los posibles valores de la variable son:

- Regional Antioquia
- Regional Bogotá
- Regional Centro
- Regional Caribe
- Regional Occidente y Cafetera.

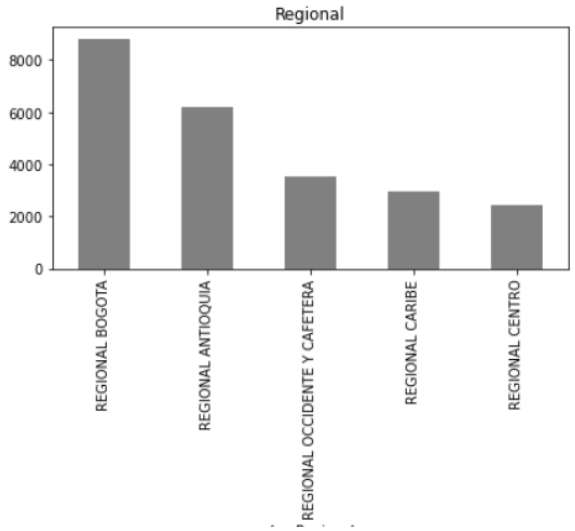

Figura 6 Distribución de la regional del cliente

La Figura 6, representa la distribución regional, cuya principal participación está en Bogotá y Antioquia.

Segmento afiliado: La empresa analizada, realiza una segmentación comercial y de servicio para sus afiliados a partir de una serie de variables internas y propias del cliente, los posibles valores de segmentación son los siguientes y en la Figura 7, es posible observar sus distribución dentro del *dataset*:

- Rentas medias
- Rentas masivas
- Rentas altas
- Sin segmento

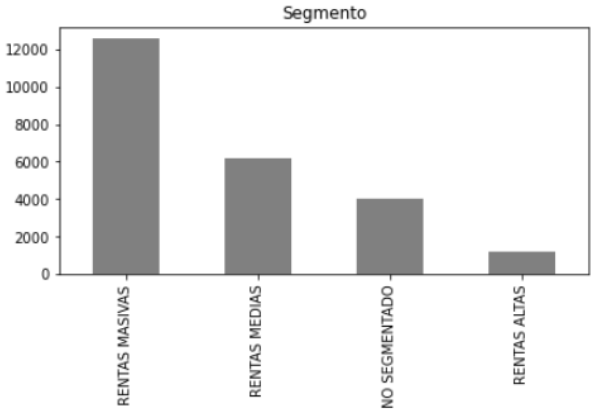

Figura 7 Distribución de la variable segmento

IBC: Estimación del salario registrado para el cliente en la empresa. Es una estimación a partir de la información registrada en la vinculación y complementada con el valor de ingresos reportado en las cotizaciones a pensión obligatoria, su comportamiento en los clientes que tuvieron contando durante el periodo de análisis se muestra en la Figura 8.

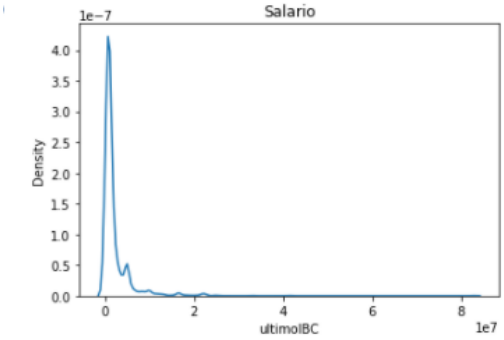

Figura 8 Distribución IBC

Ocupación: Información de ocupación laboral de cliente.

Ciclo de vida: Segmentación interna que se compone de 2 factores relacionados con la edad y el sexo del usuario como se muestra en la Figura 9.

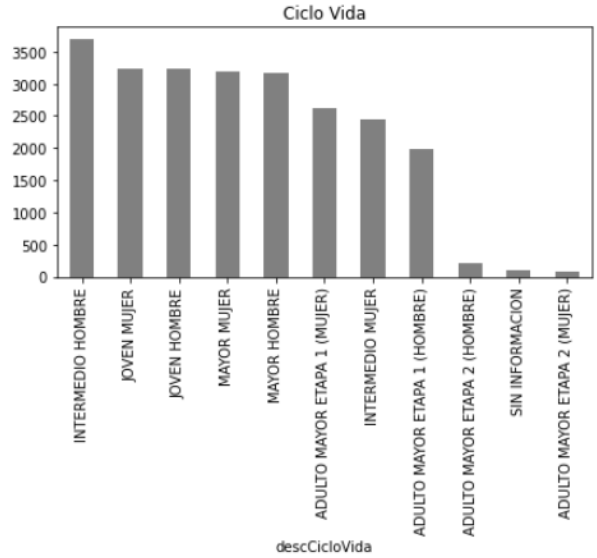

Figura 9 Distribución variable ciclo de vida

4.1.3 DATOS DE PRODUCTOS DEL CLIENTE

Estado PO: Estado general en la línea de producto ahorro obligatorio, cuya principal participación en la base de datos es Activo, en la Figura 10, se pueden observar los valores que puede tomar la variable y su respectiva participación.

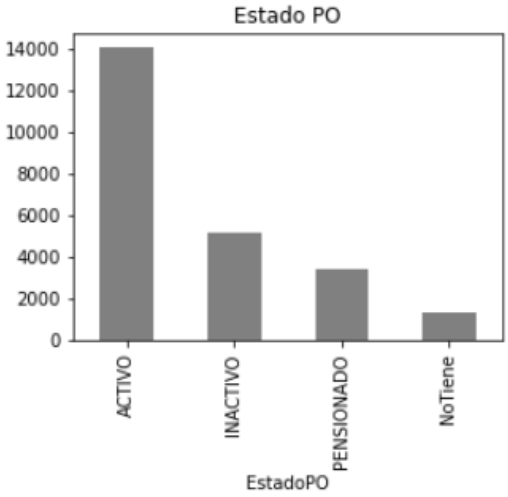

Figura 10 Distribución de la variable estado PO

Estado PV: Estado en los productos de la línea de ahorro voluntario, los posibles valores son activos, inactivos y no tiene, como se muestra en la Figura 11, los clientes analizados, en su mayoría, no tienen el producto.

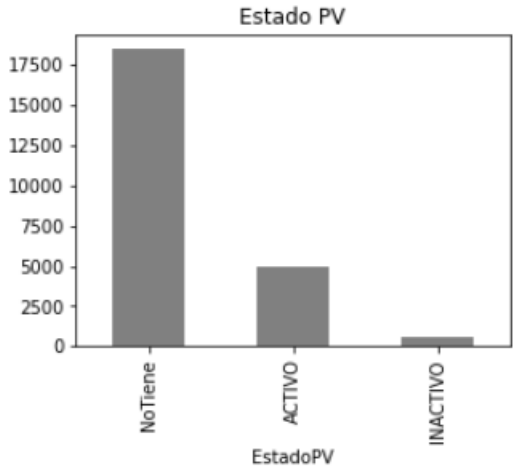

Figura 11 Distribución variable Estado PV

Estado CES: Estado en el producto Cesantías. La distribución de la variable está representada en la Figura 12, al igual que para el caso de la característica estado PO, la mayor proporción de clientes en el estudio están activos.

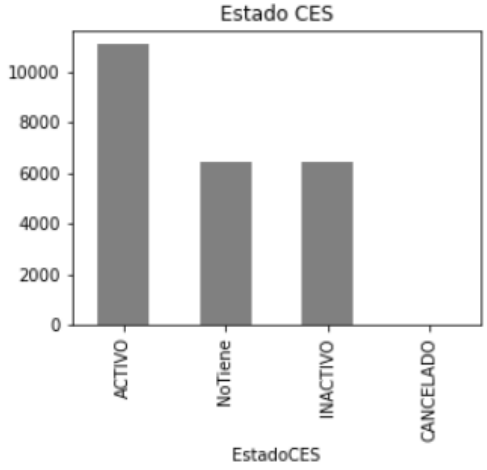

Figura 12 Distribución de la variable Estado CES

Tiene Clave: Variables con los valores sí y no, para indicar que clientes han creado su usuario y clave transaccional para acceder a los servicios digitales que requieren autenticación. En la Figura 13, se puede observar que la composición es muy similar.

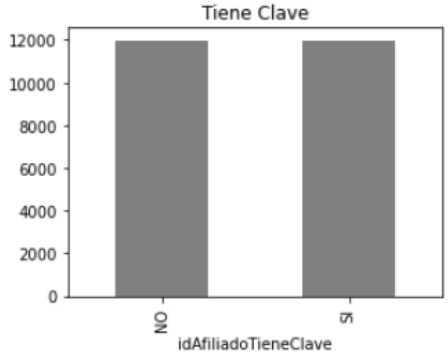

Figura 13 Distribución de la variable tiene clave

Usa Clave: Para clientes virtualizados, es decir, que ya han creado su clave para transacciones y consultas a través de canales digitales y de autoservicio, esta variable indica si la ha usado en algún momento, y como se ve en la Figura 14, el uso es significativamente bajo.

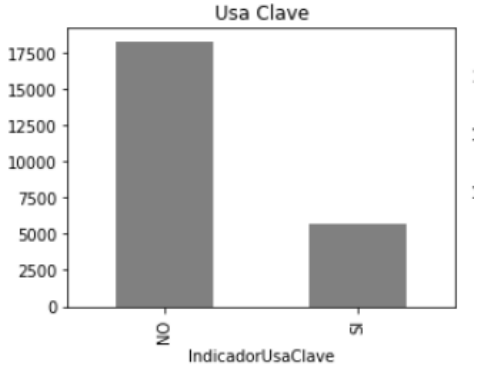

Figura 14 Distribución de la variable usa clave

Tiene email: Para clientes en el proceso de afiliación o de actualización de datos que registraron un correo electrónico, la Figura 15, muestra que los clientes en general si cuentan con este dato, aunque como se mencionó anteriormente, el 50% de los clientes también tienen clave, pero no la usan.

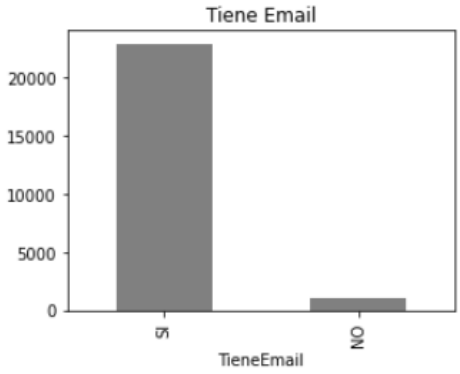

Figura 15 Distribución de la variable tiene email

4.2 SEGURIDAD DE LOS DATOS

Dentro de todo proyecto de analítica es indispensable realizar un análisis responsable y completo del tipo de datos utilizados, autorización por parte del titular para su tratamiento, seguridad de la infraestructura de almacenamiento y procesamiento, finalidad del uso de los mismos dentro de un proyecto en particular y cualquier otro factor que pueda presentar un riesgo para el cliente o la empresa. Teniendo en cuenta el planteamiento anterior, se presenta la siguiente información pertinente a los datos utilizados durante el análisis.

4.2.1 IDENTIFICACIÓN DE DATOS PERSONALES UTILIZADOS

En el apartado anterior se detallan los atributos del cliente, que si bien permiten realizar una caracterización de las personas que utilizaron los canales de servicios, no contiene una identificación directa ni representan un riesgo para el titular de los datos. Adicionalmente, es importante resaltar que dentro del análisis no se presentarán resultados para un cliente en particular, sino de manera general de acuerdo a las características similares de grupos de clientes. Otro factor importante con respecto a los datos utilizados, es que se eliminan por completo de la base de datos cualquier información relacionada con menores de edad.

4.2.2 TRATAMIENTO DE LOS DATOS

Dentro de la autorización de tratamiento de datos de los clientes dispuesta por la compañía para la cual se hizo este trabajo, se indica expresamente que estos pueden ser utilizados para fines analíticos cuando su finalidad sea en pro a mejorar la prestación de los servicios brindados por la empresa, razón por la cual se puede concluir que es posible realizar el presente trabajo.

4.2.3 SEGURIDAD DE LA INFRAESTRUCTURA

Los datos se encuentran almacenados en la plataforma GCP, especializada en almacenamiento de datos de manera segura. Además, se pudo constatar que los datos se guardan de manera anonimizada protegiendo la identidad del cliente y su acceso está restringido a personas de la organización con ciertos perfiles; sumado a esto, se hace una permanente auditoría de las consultas realizadas con el fin de poder identificar conductas sospechosas o no autorizadas. Los modelos de *machine learning* usados durante este proyecto, se implementaron en las instancias de la aplicación Jupyter notebook definidas por la organización, la cual cumple con las mismas políticas de accesibilidad y monitoreo.

4.3 ANÁLISIS DE LOS DATOS

Antes de la implementación de los modelos, se realizó un análisis exploratorio de las características disponibles para la solución del problema, esto con el fin de validar la distribución de las características que serían utilizadas para la predicción de la variable objetivo, canal de contacto del cliente que se puede evidenciar en la Figura 2.

De igual manera se realizó un análisis del espacio de características con respecto a la variable objetivo como se muestra a continuación.

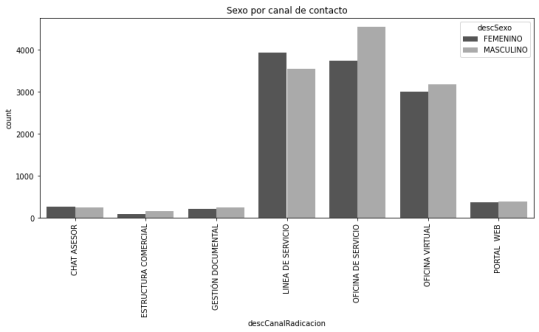

Figura 16 Variable objetivo versos sexo de los clientes

En Figura 16, se puede ver que la composición entre clientes hombres y mujeres es bastante homogénea. Adicionalmente, se logra ver en los canales tradicionales asistidos por un asesor, que las mujeres tienen una ligera tendencia por el canal telefónico y los hombres por el canal presencial.

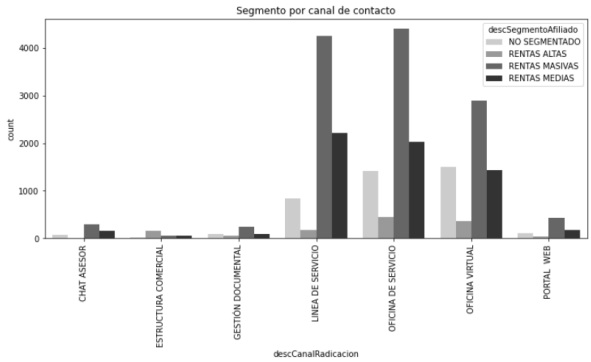

Figura 17 Variable objetivo versos segmento de los clientes

En cuanto a los segmentos que más usan los canales, Figura 17, muestra que existe una diferencia importante en el uso de la línea de servicio y oficinas de servicio por parte los clientes rentas masivas. Esto se debe también a que son los canales definidos para atender a los clientes dentro de estos segmentos. Es de resaltar que los clientes de rentas altas deberían realizar sus contactos principalmente a través del canal estructura comercial, canal especializado y personalizado a su disposición, sin embargo, hay una gran cantidad que siguen utilizando otros canales.

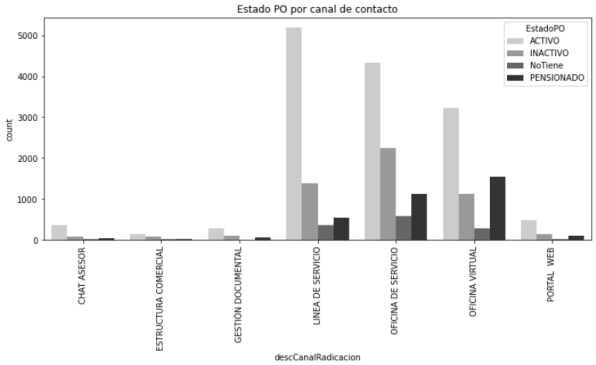

Figura 18 Variable objetivo versos estado de los clientes en producto PO

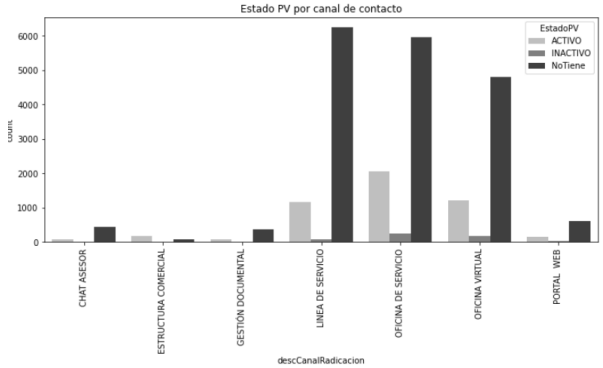

Figura 19 Variable objetivo versos estado de los clientes en producto PV

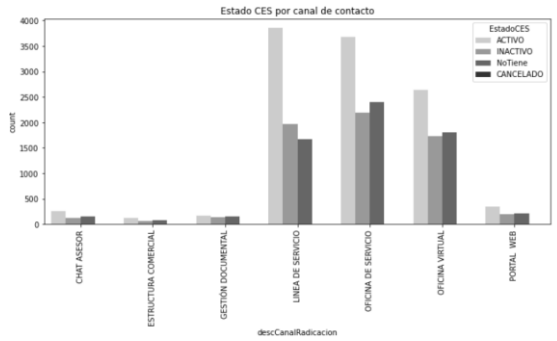

Figura 20 Variable objetivo versos estado de los clientes en producto CES

En las figuras 18, 19 y 20, se observa que hay una tendencia marcada en los clientes dentro la base de datos, ya que la mayoría están activos en PO y CES, pero no tienen el producto PV. Además, estos clientes usan de manera más frecuente los canales línea de servicio y oficina de servicio.

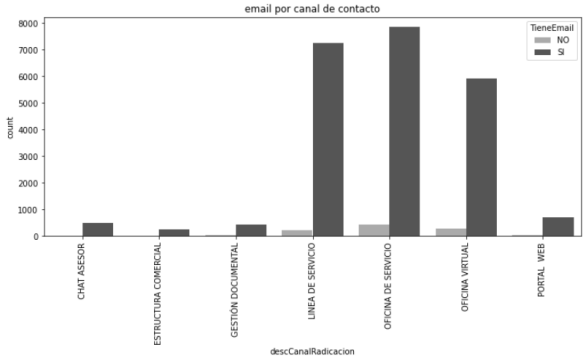

Figura 21 Variable objetivo versos si tienen o no email

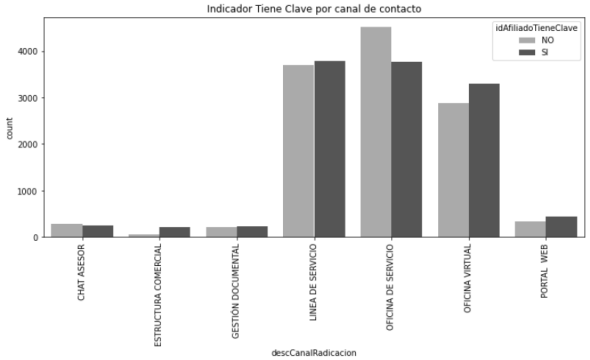

Figura 22 Variable objetivo versus indicador de tiene clave

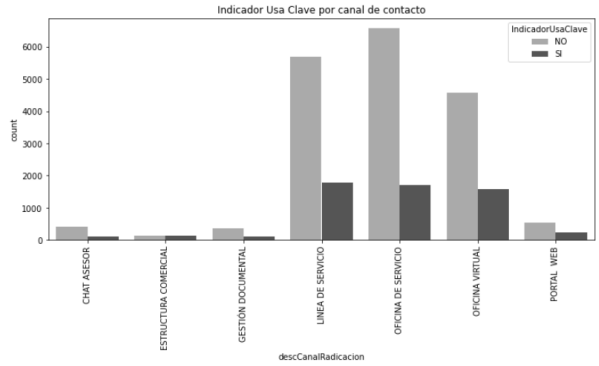

Figura 23 Variable objetivo versus indicador de uso de clave

Al analizar de manera conjunta las figuras 21, 22 y 23, se puede concluir que en los canales tradicionales (línea de servicios y oficina de servicios), los usuarios no suelen utilizar la clave del portal transaccional a pesar de ser clientes con correo registrado y clave creada que de alguna manera se pueden catalogar como clientes virtualizados.

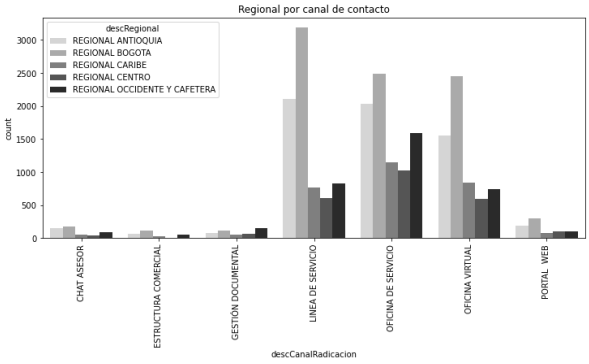

Figura 24 Variable objetivo versus indicador de uso de clave

Al revisar el comportamiento de la regional con respecto a la variable objetivo como se puede observar en la Figura 24 , encontramos que la regional Bogotá es la que presenta mayor participación en cada uno de los canales, esto se puede esperar dado que es la regional con mayor número de usuarios analizados.

4.4 PREPROCESAMIENTO DE LOS DATOS

Antes de iniciar el proceso de entrenamiento con cada uno de los modelos, las variables del *dataset* original fueron revisadas para evaluar la calidad de los datos. Este proceso implicó algunas transformaciones y depuración de campos vacíos o atípicos; se llevaron a cabo estrategias como las descritas a continuación:

- Se homologaron y agruparon algunas etiquetas para el canal de contacto, el tema y el segmento, esto teniendo en cuenta que el conocimiento del negocio determinaba que estas etiquetas hablan de temas similares o se podían unificar.
- Se revisaron los valores nulos o los campos vacíos en el dataset y según la variable se buscaron alternativas para llenarlos, tal es el caso del salario, donde se utilizó información de una variable similar para homologarlo en algunos campos vacíos, también con el campo regional se utilizó la variable ciudad para identificar la que le corresponde al registro evaluado.
- Se homologaron valores registrados en la base de datos con y sin tildes o con nombres que fueron modificados en el proceso de atención del cliente.
- Se eliminaron registros de canales que no son de atención a clientes personas, por ejemplo, canal empresas, ya que no están dentro del alcance del proyecto.
- El segmento del cliente fue agrupado ya que el valor rentas altas estaba distribuido en 3 subcategorías con muy pocos registros que por sí solos no eran significativos.
- Se eliminaron valores para la variable edad que no estaban dentro del rango de 18 a 90 años.

Otro proceso importante llevado a cabo es la codificación de las etiquetas, en el cual se transformaron las variables categóricas en numéricas utilizando para este proceso el *Label Encoder* de la librería *scikit-learn*. Finalmente, para las variables numéricas se hace una normalización utilizando la función *MinMaxScaler* de la librería sklearn [11] para evitar que las diferentes escalas influyan en el resultado de los modelos.

En este punto es importante señalar que, si bien se realizó una limpieza meticulosa de los datos, los modelos evaluados no mejoraron significativamente su desempeño, razón por la cual se planteó una hipótesis sobre si los datos podrían tener un alto grado de solapamiento, dado que un usuario en el periodo evaluado pudo realizar un contacto por canales diferentes inclusive para el

mismo tema. Por tanto, se hizo una selección para cada usuario tomando como referencia el canal más utilizado, de este modo se evita que el modelo reciba un mismo vector de características para dos canales diferentes.

Adicionalmente, y teniendo presente que existen técnicas para trabajar con bases de datos desbalanceadas como la del presente trabajo, en el cual el desbalance es predominantemente fuerte en tres valores de la variable de salida, se optó por trabajar con estos tres canales y agrupar el resto en un canal denominado otros. Los valores principales tomados para los experimentos fueron línea de servicio, oficina de servicio y oficina virtual.

4.5 MODELOS INICIALES

A partir de la siguiente subsección se describen los modelos usados en el presente trabajo. Se implementaron usando la librería scikit-learn de python. Para su entrenamiento se utilizaron los parámetros por defecto. Los datos se dividieron en un 80% para entrenamiento y 20% para validación, indicando al método de separación que la variable objetivo se encontraba desbalanceada. Teniendo presente que el dataset fue evolucionando para mejorar los resultados, y que los modelos iniciales presentan varias iteraciones, se mostrarán los hallazgos principales obtenidos con el fin de evidenciar la evolución de los mismos.

4.5.1 REGRESIÓN LOGÍSTICA

Método de clasificación básico, en el cual se trata de explicar y predecir una variable categórica, la cual puede tomar dos valores o una serie finita de categorías mutuamente excluyentes o que presentan entre ellas algún orden. Este es un modelo lineal de clasificación en el cual se utiliza la función logística asumiendo que el grupo de variables evaluadas son relevantes, influyentes e independientes entre sí, buscando para cada caso individual o vector evaluado (xi), una probabilidad p [12].

Este modelo fue utilizado con el *dataset* inicial y evaluado con las métricas de *accuracy* y *balance accuracy*, posteriormente fue evaluado con el *dataset* extendido en variables y número de muestras. Adicionalmente, se ajustó la métrica para tomar como base el valor F1, en este caso se entrenó el modelo con los parámetros por defecto que ofrece la librería de sklearn, entre los cuales se encuentra la función de penalización que para este caso fue la L2, pesos asociados a las clases y el número máximo de iteraciones.

4.5.2 BAYES GAUSSIANO INGENUO

El clasificador bayesiano ingenuo está fundamentado en el teorema de bayes. El calificativo de ingenuo se otorga dado que en este modelo se parte del supuesto de que las características son independientes. Se centra en calcular la probabilidad de que ocurra un evento determinado teniendo en cuenta las probabilidades de los eventos anteriores [13]. Al igual que en la regresión logística este modelo fue entrenado con el *dataset* inicial, y posteriormente fue evaluado con el *dataset* extendido; para su entrenamiento se usan los parámetros por defecto establecidos en la librería utilizada.

4.5.3 DISCRIMINANTE CUADRÁTICO

El análisis discriminante cuadrático es un método de clasificación multiclase clásico en el cual se aplica el teorema de bayes. Es un modelo en el cual se asume que cada observación presenta una distribución normal multivariable y que cada clase presenta su propia matriz de covarianza; los parámetros se estiman a partir de una función cuadrática lo que permite generar límites de decisión curvos, una funcionalidad muy útil cuando los límites de decisión no son necesariamente lineales; esto le permite al modelo contar con buena aplicación en la práctica [14]. En su implementación se puede definir la proporción de las clases, pero el modelo las puede inferir de los datos de entrenamiento, y adicionalmente permite definir regularizadores para las matrices de covarianza y permite además almacenar la matriz de covarianza de cada clase.

4.6 OTROS MODELOS EVALUADOS

4.6.1 RANDOM FOREST CLASSIFIER

En estos modelos de clasificación, los algoritmos se basan en la teoría de los árboles de decisión, sin embargo este método utiliza el trabajo de varios árboles mejorando el rendimiento que puede realizar uno solo. En la implementación de

este modelo se realizó una búsqueda de los mejores hiperparámetros utilizando el método de GridSearchCV, en el cual se evalúan diferentes valores, variando la cantidad de árboles y la profundidad máxima del árbol para optimizar los resultados del modelo [15].

Para el entrenamiento del modelo se hallan como mejores parámetros una cantidad de árboles de 70 y profundidad máxima igual a 8.

4.6.2 K VECINOS MÁS CERCANOS

Este es un algoritmo de clasificación no paramétrico, es decir, no hace suposiciones sobre las distribuciones de los datos. Este método no aprende un modelo particular, sino que utiliza las observaciones o el conocimiento de la fase de entrenamiento para las predicciones. Para esto, analiza una cantidad de vecinos (k) más cercanos al momento de la predicción, por lo que este parámetro se convierte en un estimador clave para el desempeño del modelo. Aunque este es un modelo de clasificación simple, en la práctica ha tenido muchas aplicaciones en el campo económico y financiero. Para encontrar el mejor valor de k y la función óptima para la medición de distancia, se utilizó el método de GridSearchCV.

En este caso los mejores hiperparámetros obtenidos fueron un número de vecinos k igual a 50, para la función de pesos utilizada *(weights)* igual a '*distance',* la cual implica que los vecinos más cercanos tendrán mayor importancia a la hora de asignar las clases, y finalmente la función para estimación de distancias que usa el modelo es la euclidiana.

4.6.3 MÁQUINAS DE SOPORTE VECTORIAL

En este método de clasificación, se construyen hiperplanos buscando separar dos clases de forma óptima, es decir, busca márgenes máximos de distancia entre dos puntos de las clases más cercanas, utilizando en su método de separación funciones kernel, permitiendo proyectar los datos en un espacio de mayor dimensión, eliminando los limitantes que puede llegar a tener las funciones lineales. Al igual que en los dos modelos anteriores, se hizo una búsqueda de los mejores hiperparámetros utilizando la metodología de validación cruzada para el parámetro de regularización C y el coeficiente de la función kernel gamma.

Para el entrenamiento del modelo se utilizaron los siguientes parámetros como mejores estimadores, función kernel gaussiano, gamma de 0.1 y C de 10.

4.6.4 RED NEURONAL

El término red neuronal se aplica para un tipo de modelos que se caracterizan por el uso de un amplio espacio de parámetros, una estructura altamente flexible y que proviene de los estudios sobre el funcionamiento del cerebro. Los modelos de redes neuronales tienen la capacidad de utilizar sus conexiones para aproximar funciones y dinámicas aprendiendo a partir de ejemplos; se han convertido en una alternativa muy popular para resolver una gran variedad de problemas [16].

Para la implementación de este modelo se hizo previamente una codificación de la variable de salida con el método *One Hot Encoder*, que consiste en codificar una variable categórica como una matriz numérica, en la cual cada categoría se representa en una columna binaria. Posteriormente se hizo la definición del modelo de redes neuronales secuencial utilizando la librería Keras, con una capa densa y una capa de salida *softmax* [17].

Para el entrenamiento del modelo se implementa un ciclo *for* para validar los datos de pruebas en 3 *subsets* seleccionados a través del mecanismo de kfold de la librería *sklearn.*

4.7 MÉTRICA UTILIZADA

La métrica que se utilizó para la evaluación final de los modelos fue F1, esta métrica es una combinación entre las medidas de *precision* y *recall* y se puede interpretar como un promedio ponderado de la de las dos métricas. Está en un rango entre 0 y 1. Se opta por esta métrica ya que en este proyecto tanto el *recall* como la *precision* son importantes a la hora de realizar una medición de los resultados obtenidos.

Cuando se habla de *recall* se hace referencia a la exhaustividad del modelo, es decir, la cantidad de elementos correspondientes a una clase que el

modelo es capaz de identificar, mientras que la *precision* se centra en la precisión de los resultados, es decir, de los registros clasificados en una clase, que porcentaje realmente pertenece a dicha categoría.

$$F1 = 2\,\frac{precision.recall}{precision + recall}$$

$$precision = \frac{True\ positives}{True\ positives + false\ positives}$$

$$recall = \frac{True\ positives}{True\ positives + false\ negatives}$$

Los valores de *true positives*, *false positives* y *false negatives* utilizados para el cálculo de las métricas, se obtienen a partir de la matriz de confusión resultante en la validación de los modelos y se representa gráficamente como se muestra a continuación:

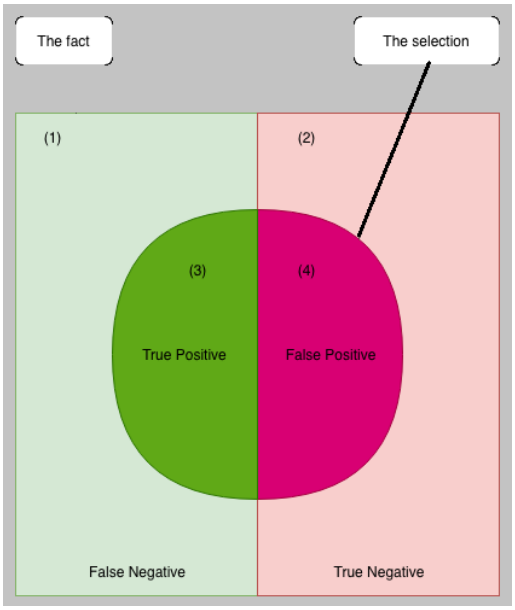

Figura 24. Matriz de confusión

En los modelos explorados, se puede evidenciar que el mejor resultado tomando como métrica de referencia el valor F1 es un K vecinos más cercanos, con 50 vecinos y una puntuación de 0.66.

4.8 ITERACIONES Y EVOLUCIÓN DEL MODELO

4.8.1 SELECCIÓN DE CARACTERÍSTICAS SECUENCIALES

El modelo fue evaluado con 14 variables que, según la investigación de la bibliografía y el conocimiento de la información disponible, eran características que podrían aportar al modelo predictivo. Para validar que tanto contribuía cada una de las variables al modelo, se utilizó la selección de características secuenciales de la librería mlxtend [17], la cual utiliza algoritmos de secuenciales para reducir el número de características que fueron utilizadas inicialmente en el modelo. Con este método es posible determinar cuáles son más relevantes, lo que permite reducir la dimensionalidad inicial del vector sin afectar los resultados, generando un ahorro en el costo computacional y eliminando variables que podrían estar inyectado ruido en el modelo. Uno de los parámetros de entrada en este modelo es el número de características al que se quiere llegar, por tanto se requiere el uso de técnicas que permitan evaluar el mejor rendimiento del modelo para diferentes números de variables [18].

En este caso el modelo realizó una selección de 11 características, presentando como las más relevantes para el modelo las siguientes: sexo, segmento del afiliado, edad, estado PO, estado PV, estado CES, salario, Indicador de si usa clave, tiene email, ocupación y regional. De este modo las características que el modelo encontró como poco relevantes fueron: tema de contacto, ciclo de vida e indicador de si tiene clave.

5 RESULTADOS Y ANÁLISIS

En la Tabla 1, se presenta el resumen general de los resultados obtenidos en las métricas F1, *recall* y *precision* para cada uno de los modelos experimentados con la base de datos extendida y sus respectivas configuraciones de parámetros.

Modelo	Parámetros	F1	Recall	Precision
Regresión logística	Valores por defecto sklearn	0.49	0.51	0.51
Bayes Gausiano Ingenuo	Valores por defecto sklearn	0.45	0.48	0.48
Discriminante cuadrático	Valores por defecto sklearn	0.47	0.49	0.50
Random forest	n_estimators=70, max_depth=8, random_state=0, class_weight='balanced_subsample'	0.59	0.60	0.61
K vecinos más cercanos	Valores por defecto sklearn	0.60	0.60	0.61
K vecinos más cercanos	n_neighbors=30, weights ='distance'	0.65	0.65	0.65
K vecinos más cercanos con GridSearch	'n_neighbors':[20, 25,30, 35, 40,50,60], 'weights':['uniform','distance']	0.66	0.65	0.66
Máquinas de soporte vectorial	kernel='rbf', C = 0.1, gamma = 0.005	0.35	0.44	0.33
Máquinas de soporte vectorial	kernel='rbf', C = 10, gamma = 0.1	0.38	0.46	0.58
Red neuronal	Dense: 5, Input: 14, Activation: relu; Dense: 3, activation: softmax	0.47	0.50	0.49
KNN con selección de características	n_neighbors=50, weights='distance', k_features=11, cv=5	0.69	0.69	0.70

Tabla 1 Métricas de modelos evaluados

Los resultados de los modelos evaluados y la cantidad de exploraciones que se realizaron en primera instancia nos permite resaltar que el problema planteado no tiene una solución sencilla, sin embargo, a través de los diferentes experimentos aplicados es posible llegar a una mejora considerable en los resultados. Como se puede ver en la Tabla 1, se exploraron modelos potentes como las máquinas de soporte vectorial y las redes neuronales, pero éstos no lo logran generar mejores predicciones que las logradas con un modelo más sencillo como lo es el K vecinos más cercanos, que resulta ser potente para el tipo de problema tratado en este trabajo. De igual forma, hay que anotar que, KNN, a pesar de su simplicidad, genera un costo computacional importante para la compañía durante los procesos de entrenamiento y reentrenamiento, esto debido a la cantidad de cálculos que debe realizar y al paso adicional de selección de características agregado para mejorar los resultados con dicho modelo.

6 CONCLUSIONES

Con la tendencia actual de migrar las decisiones y diseño de estrategias con base a la evidencia generada a través del análisis de los datos como sustento lógico y válido para el entendimiento del cliente, las capacidades de la analítica se convierten en un factor sumamente valioso para las organizaciones. En concordancia con lo anterior, durante la implementación de este proyecto se muestra la exploración y conocimiento del espacio de características del cliente, el cual, sumado al entendimiento del negocio, permitió garantizar una depuración adecuada de los datos que mejoró la calidad de los mismos.

Adicionalmente, fue posible realizar la experimentación de una variedad de modelos clásicos de *machine learning*, con diferentes configuraciones de parámetros en búsqueda de un resultado óptimo y aceptable para la empresa, sin embargo, se concluye que el problema propuesto es bastante complejo, así que se utilizan métodos proporcionados por librerías especializadas para ajustar sus parámetros de acuerdo con las mejores configuraciones posibles, lo cual tampoco mejora los resultados de manera significativa.

No obstante, durante el desarrollo del proyecto, se pudo apreciar como una correcta selección de las variables incluidas en el *dataset* y una adecuada validación y exploración de diferentes tipos de modelos, sí generan mejoras significativas en los resultados de la predicción. Es por esto que, una base de datos con una cantidad representativa de registros, un espacio de características construido a partir del conocimiento del negocio y de la revisión del estado del arte para proyecto similares, un modelo con una buena exploración de hiperparámetros y una selección de las variables más relevantes, lograron pasar de una métrica F1 de 0.40 en la primera iteración a un resultado de 0.70 en el mejor modelo encontrado.

Finalmente, se puede concluir que se llega a un resultado aceptable que representa un avance importante en la solución del problema, dejando en claro una serie de aprendizajes que se podrían tomar como base para proyectos futuros.

REFERENCIAS

[1] A. Gupta, B.-c. Su y Z. Walter, «An empirical study of consumer switching from traditional to electronic channels: A purchase-decision process perspective.,» *International Journal of Electronic Commerce*, vol. 8, pp. 131-161, 2004.

[2] M. Muñoz, «IT Travel services,» 13 04 2021. [En línea]. Available: https://ittravelservices.com/se-puede-predecir-la-repeticion-de-un-cliente/.

[3] S. Anastasia, M. Madonna y L. Monica, «Implications of embedded artificial intelligence - machine learning on safety of machinery,» *Procedia Computer Science,* vol. 180, pp. 338-343, 2021.

[4] M. M. Hassan y M. Tabasum, «Customer profiling and segmentation in retail banks using data mining techniques.,» *International journal of advanced research in computer science,* pp. 24-29, 2018.

[5] N.-M. Casariego Sarasquete, «Metodología de análisis y segmentación de clientes usando secuencias de comportamiento.,» Tesis de Maestría., Madrid, 2019.

[6] M. Frasquet, A. Mollá y E. Ruiz, «Identifying patterns in channel usage across the search, purchase and post-sales stages of shopping.,» *Electronic Commerce Research and Applications,* pp. 654-665, 2015.

[7] S. Erevelles, N. Fukama y L. Swayne, «Big Data consumer analytics and the transformation of marketing.,» *Journal of business research,* vol. 69, pp. 897-904., 2016.

[8] S. Nakano y F. N. Kondo, «Customer segmentation with purchase channels and media touchpoints using single source panel data,» *Journal of Retailing and consumer services,* vol. 41, pp. 142-152, 2018.

[9] Ø. GRØNFLATEN, «Predicting travelers' choice of information sources and information channels.,» *Journal of Travel Research,* vol. 48, nº 2, pp. 230-244, 2009.

[10] ,. A. Bojanowska y M. Milosz, «Application of neural networks in CRM systems.,» *En ITM Web of Conferences. EDP Sciences,* p. 04001, 2017.

[11] scikit learn, «sklearn.preprocessing.MinMaxScaler,» 2021. [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[12] scikit learn, «Linear Models,» Mayo 2021. [En línea]. Available: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

[13] scikit learn, «Naive Bayes,» Mayo 2021. [En línea]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes.

[14] scikit learn, «Linear and Quadratic Discriminant Analysis,» Mayo 2021. [En línea]. Available: https://scikit-learn.org/stable/modules/lda_qda.html#lda-qda.

[15] scikit learn, «sklearn.ensemble.RandomForestClassifier,» Mayo 2021. [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[16] N. Kriegeskorte y T. Golan, «Neural network models and deep learning» *Current Biology,* vol. 29, pp. R231-R236, 2019.

[17] Keras, «The Sequential model,» 2021. [En línea]. Available: https://keras.io/guides/sequential_model/. [Último acceso: 2021].

[18] S. Raschka, «Sequential Feature Selector,» abril 2021. [En línea]. Available: http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/#sequential-feature-selector.