Predicción del riesgo de abandono de un asociado para una cooperativa multiactiva de Santiago de Cali, mediante técnicas de aprendizaje automático

Victor hugo males

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

Director

Jurado

Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en ingeniería

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali 31-05-2023

Pontificia Universidad JAVERIANA Cali

Facultad de Ingeniería y Ciencias

Acta de Correcciones al Documento de Trabajo de Grado

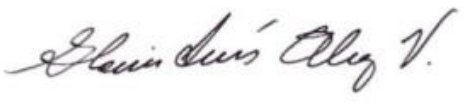
Santiago de Cali, 20 de junio de 2023

Autor: Victor Hugo Males

Título del Trabajo de Grado: "Predicción del riesgo de abandono de un asociado para una cooperativa multiactiva de Santiago de Cali, mediante técnicas de aprendizaje automático"

Director: Gloria Inés Álvarez

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.



Firma del Director del Trabajo de Grado

Datos del estudiante

Nombre completo: Victor hugo males.
Dirección: carrera 121 a # 48-100, unidad aguaclara apto 502 torre f
Correo electrónico: victorhugomales@gmail.com
Teléfono fijo: 3274895
Celular: 3005795627
Profesión: Estadístico
Empresa: Compensar
Cargo: Científico de datos I

RESUMEN

Predecir la tasa de abandono (churn rate) o riesgo de abandono de los clientes, es importante para las empresas y más para la cooperativa multiactiva objeto de estudio, debido a que es importante mantener activa la base social de asociados a la cooperativa. En este marco, el presente trabajo se enfoca en la implementación de técnicas de aprendizaje automático a un conjunto de variables cuantitativas y cualitativas y realizar un despliegue de una visualización con la mejor técnica de aprendizaje automático entrenada, que permita a las áreas comerciales de vinculación y retención tomar mejores decisiones en la implementación de sus estrategias para la reducción del riesgo de abandono. Con esto, la visualización realiza funcionalidades de exploración de análisis de datos, predicción de una lista de usuarios y según las variables que riesgo puede tener un asociado de no continuar en la cooperativa.

Palabras clave: Cooperativa, Riesgo de abandono, aprendizaje automático.

ABSTRACT

Predicting the abandonment rate (churn rate) or risk of customer abandonment is important for companies and more so for the multi-active cooperative under study, because it is important to keep active the social base of members of the cooperative. In this framework, the present work focuses on the implementation of machine learning techniques to a set of quantitative and qualitative variables and to display a visualization with the best trained machine learning technique, which allows the commercial areas of linking and retention make better decisions in the implementation of their strategies to reduce the risk of abandonment. With this, the visualization performs data analysis exploration functionalities, prediction of a list of users and according to the variables that a member may be at risk of not continuing in the cooperative.

Keywords: Cooperative, Dropout risk, machine learning.

Pontificia Universidad JAVERIANA Cali
Res. 2333 del 2012

Vigilada Mineducación
Res. 2333 del 2012

Predicción del riesgo de abandono de un asociado para una cooperativa multiactiva de Santiago de Cali, mediante técnicas de aprendizaje automático

Victor Hugo Males

Proyecto de grado para optar al título de
Magister en Ingeniería

Director
Gloria Inés Álvarez Vargas

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN INGENIERÍA
SANTIAGO DE CALI, OCTUBRE DE 2022

CONTENIDO

Pag.

INTRODUCCIÓN 12
1. DEFINICIÓN DEL PROBLEMA 12
1.1 PLANTEAMIENTO DEL PROBLEMA 12
1.2 FORMULACIÓN DEL PROBLEMA 13
2. OBJETIVOS DEL PROYECTO 14
2.1 OBJETIVO GENERAL 14
2.2 OBJETIVOS ESPECÍFICOS 14
3. RESULTADOS ESPERADOS 14
4. ALCANCE 14
5. JUSTIFICACIÓN 15
6. MARCO TEÓRICO DE REFERENCIA Y ANTECEDENTES (ESTADO DEL ARTE) 15
6.1 TASA DE DESERCIÓN DE CLIENTES (RATE CHURN) 15
6.2 APRENDIZAJE AUTOMÁTICO 16
6.3 APRENDIZAJE SUPERVISADO 17
6.4 TÉCNICAS UTILIZADAS 18
6.5 MÉTRICA DE EVALUACIÓN Y RENDIMIENTO 23
6.6 TÉCNICAS DE PROCESAMIENTO DE DATOS 26
7. TRABAJOS RELACIONADOS 27
8. METODOLOGÍA 28
9. DESARROLLO DEL PROYECTO 29
9.1 ESTUDIO DE LAS CARACTERÍSTICAS RELEVANTES PARA ESTIMAR EL RIESGO DE ABANDONO 30
9.1.1. SELECCIÓN DE LA INFORMACIÓN DE LA BODEGA DE DATOS 30
9.1.2. FICHA TÉCNICA DE LA INFORMACIÓN Y TIEMPO DE ANÁLISIS EN LA BASE DE DATOS 31
9.1.3. DICCIONARIO DE VARIABLES 32
9.1.4. ANÁLISIS DESCRIPTIVO E IDENTIFICACIÓN DE VARIABLES RELEVANTES 37
9.2 SELECCIÓN DE VARIABLES A PARTIR DE ANÁLISIS DE CORRRELACIÓN PARA LA

CONTRUCCIÓN DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO 60
9.3 PREPARACIÓN DE LOS DATOS DE FORMA ADECUADA PARA EL ENTRENAMIENTO DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO 61
9.3.1. CODIFICACIÓN DE VARIABLES CATEGÓRICAS Y NORMALIZACIÓN DE VARIABLES CUANTITATIVAS. 61
9.3.2. SELECCIÓN DE LAS TÉCNICAS DE APRENDIZAJE AUTOMÁTICO 61
9.3.3. DIVISIÓN DE LA BASE DE DATOS EN ENTRENAMIENTO Y PRUEBA 62
9.4 TÉCNICAS DE APRENDIZAJE AUTOMÁTICO QUE PERMITEN PREDECIR EL RIESGO DE ABANDONO 63
9.4.1. DEFINICIÓN DE EXPERIMENTOS PARA LA CONSTRUCCIÓN DE LAS TÉCNICAS DE APRENDIZAJE AUTOMÁTICO 63
9.4.2. RESULTADOS DEL EXPERIMENTO 1 (TÉCNICAS DE APRENDIZAJE AUTOMÁTICO POR DEFECTO) 63
9.4.3. RESULTADOS DEL EXPERIMENTO 2 (OPTIMIZACIÓN DE HIPERPARÁMETROS) 64
9.5 EVALUACIÓN DEL DESEMPEÑO DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO ENTRENADOS Y SELECCIÓN DE LA MEJOR TÉCNICA 67
9.6 APLICATIVO WEB PARA VISUALIZACIÓN DE LAS PREDICCIONES DE LOS ASOCIADOS ACTIVOS A LA COOPERATIVA 68
10. DISCUSIÓN DE RESULTADOS 71
11. CONCLUSIONES 72
12. TRABAJOS FUTUROS 73
13. REFERENCIAS BIBLIOGRÁFICAS 74

LISTA DE TABLAS

Tabla 1. Matriz de Confusión 24
Tabla 2 Dimensiones y tablas de hechos del sistema de base de datos 30
Tabla 3. Ficha técnica de la base de datos 31
Tabla 4. Diccionario de variables 32
Tabla 5. Agrupación de variables 37
Tabla 6. Variables agrupadas y valores vacíos 37
Tabla 7. Variables seleccionadas 42
Tabla 8. Ventajas y desventajas de las técnicas de aprendizaje automático 62
Tabla 9. Resultados por clase en los Modelos Defecto experimento 1 64
Tabla 10 Resultados Modelos Defecto, experimento 1 64
Tabla 11 Descripción y definición de hiperparámetros experimento 2 65
Tabla 12 Hiperparámetros seleccionados experimento 2 66
Tabla 13 Resultados por clase en los Modelos optimizados experimento 2 66
Tabla 14 Resultados generales en los Modelos optimizados experimento 2 67
Tabla 15. Resultados generales comparativos de los experimentos en la evaluación de los modelos 67
Tabla 16 Resultados por clase comparativos de los experimentos en la evaluación de los Modelos 68

LISTA DE FIGURAS

Figura 1: Inteligencia artificial. 16
Figura 2. Ejemplo de un problema de clasificación y regresión. 18
Figura 3. Arquitectura perceptrón multicapa. 20
Figura 4. Visualización de funcionamiento de bagging. 22
Figura 5. Visualización de funcionamiento del boosting. 23
Figura 6. Curva ROC. 26
Figura 7. Metodología General. 29
Figura 8. Histograma del valor de la cuota de la cooperativa y logaritmo del valor de la cuota de la cooperativa. 45
Figura 9. Histograma del valor de protección de la cooperativa y logaritmo del valor de protección. 46
Figura 10. Histograma del valor cuota crédito solidario de la cooperativa y el logaritmo del valor cuota crédito solidario. 46
Figura 11. Histograma del valor de la cuota de fundación educación de la cooperativa y logaritmo del valor de la cuota de fundación educación. 47
Figura 12. Histograma del valor de cuota de turismo de la cooperativa y logaritmo del valor de cuota de turismo. 47
Figura 13. Box plot y puntos de las aprobaciones y negaciones pendiente de la cooperativa. 48
Figura 14. Box plot y puntos del uso de turismo de la cooperativa. 48
Figura 15. Box plot y puntos del uso de eventos de la cooperativa. 49
Figura 16. Box plot y puntos tenencia de la tarjeta débito. 49
Figura 17. Box plot y puntos tenencia de la cuenta deposito. 50
Figura 18. Box plot y puntos tenencia de la tarjeta visa con cupo. 50
Figura 19. Box plot y puntos de producto de banca seguros. 51
Figura 20. Box plot y puntos de producto de otras pólizas. 51
Figura 21. Box plot y puntos de producto de soat. 52
Figura 22. Box plot y puntos créditos de educación terminados. 52
Figura 23. Box plot y puntos créditos de educación cancelado. 53
Figura 24. Box plot y puntos participación en la fundación. 53
Figura 25. Comparativo en porcentaje de participación por cantidad de productos. 54
Figura 26. Histograma de la edad de los asociados. 54
Figura 27. Comparativo en porcentaje de participación por segmento ciclo de vida. 55
Figura 28. Comparativo en porcentaje de participación por nivel académico. 56
Figura 29. Comparativo en porcentaje de participación por área de conocimiento. 56
Figura 30. Comparativo en porcentaje de participación por actividad laboral. 57
Figura 31. Comparativo en porcentaje de participación por estrato socio económico. 57
Figura 32. Comparativo en porcentaje de participación por estado civil. 58
Figura 33. Comparativo en porcentaje de participación por corte de la factura. 58
Figura 34. Comparativo en porcentaje de participación por tipo de vivienda. 59

Figura 35. Correlación de variables numéricas. 61
Figura 36. Implementación de la visualización de resultados. 69
Figura 37. Interfaz de usuario para EDA. 70
Figura 38. Interfaz de usuario para la predicción de lista de usuarios. 70
Figura 39. Interfaz de usuario para la predicción del asociado. 71

INTRODUCCIÓN

El Churn Rate o tasa de cancelación es uno de los indicadores más importantes en las empresas, y se define como el porcentaje de clientes o suscriptores que dejan de utilizar los servicios durante un período de tiempo determinado. Aunque es un término asociado al email marketing, también se utiliza en otros sectores haciendo alusión a la pérdida de clientes.

En la cooperativa multiactiva objeto de estudio, bajo la naturaleza del cooperativismo de ahorro y crédito, es importante mantener activa la base social de asociados a la cooperativa bajo el concepto de la tasa de cancelación o pérdida de clientes. El presente trabajo aportará la construcción de modelos aprendizaje automático, para la predicción del riesgo de abandono de la cooperativa, además de una visualización de datos que ayude a entender el perfil de clientes con riesgo de abandono, para la mejor toma de decisiones en el equipo comercial.

1. DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

La Cooperativa multiactiva ubicada en Santiago de Cali (Colombia), que brinda servicios en diferentes áreas: Protección y aseguramiento, vivienda, educación, desarrollo empresarial, financiera, salud, recreación y turismo, ofrece una gama integrada de productos y servicios, con el objetivo de mejorar la calidad de vida del asociado a lo largo de su ciclo de vida. En este sentido, la cooperativa con el objetivo de mantener su base social de asociados ha trabajado en identificar los clientes con mayor probabilidad de retirarse de la cooperativa. Debido a que el impacto del retiro de asociados en 10 años para el periodo enero 2006 hasta diciembre 2016, se calcula a partir de los 506 Mil asociados, que corresponden a la población del cierre de diciembre 2016 (230.648), más los asociados que se habían vinculado en los últimos 10 años, desde Ene-2006 hasta Dic-2016, y ya se han retirado 276.281. Esto quiere decir que la deserción de asociados en los diez años es del 54%. Por este resultado, la cooperativa ha implementado algunos estudios para identificar asociados con alta probabilidad de retirarse.

Tal como se expuso anteriormente, la cooperativa ha realizado análisis para entender el comportamiento del retiro de asociados, y es a partir del cálculo de la curva de supervivencia construida con el histórico de todos los asociados en los últimos 20 años, que la probabilidad de que un asociado siga en la cooperativa el primer año es de 78 %. La mediana es durar aproximadamente tres años como asociado (35 meses), mientras que la media fue de 5,25 años (63 meses). En otros estudios, se analiza cómo cambia la probabilidad de permanencia de un periodo a otro, y se tiene como resultado que los períodos donde es más importante focalizar políticas de fidelización de clientes son desde su asociación hasta que cumplen casi 2 años.

Adicionalmente, la cooperativa ha realizado algunas técnicas de aprendizaje automático, las cuales predicen si un asociado va a retirarse en un lapso de 6 meses. Estos se han obtenido a partir de información histórica, en una plataforma tecnológica desarrollada bajo un sistema de bases de datos distribuidas, la cual se encuentra en servidores de Google que contiene por cada asociado variables demográficas , las interacciones de compra y cancelaciones históricas en el tiempo de los productos en cada una de las empresas del grupo, agrupadas en sectores como: cooperativo ,financiero, protección(solidaridad y corredor de seguros), salud (EPS, medicina prepagada y CEM) y unidades de educación, fundación y recreación. También las negaciones de productos o servicios, beneficios, atenciones (turnos), peticiones, quejas y reclamos. El sistema de base de datos distribuidas, además almacena datos externos al grupo empresarial que no son por asociado, como los tweets relacionados con la cooperativa y variables macroeconómicas como el IPC, variación Mensual IPC, variación año corrido IPC, variación anual IPC, tasas de interés de los bonos TES a 1 año, tasas de interés de los bonos TES a 5 años, índice de Producción Industrial y Tasa de Desempleo. Con esta fuente de datos las técnicas de aprendizaje automático han obtenido una exactitud, precisión y sensibilidad del 88%, 18.7% y 4.7% respectivamente.

Lo anterior expone un escenario de mejora en métricas como la precisión y sensibilidad para la cooperativa, donde se pueden aplicar técnicas de aprendizaje automático para mejorar la calidad de la predicción y usar la visualización de datos para construir una herramienta de apoyo a las áreas comerciales para la toma de decisiones.

1.2 FORMULACIÓN DEL PROBLEMA

¿Cómo mejorar la calidad de la predicción y visualizar el riesgo de abandono de un asociado de una cooperativa multiactiva usando técnicas de aprendizaje automático?

Esto implica responder los siguientes interrogantes:

- ¿Qué características son relevantes para estimar el riesgo de abandono de un asociado?
- ¿Cómo seleccionar los datos requeridos para la construcción de modelos usando técnicas de aprendizaje automático?
- ¿Cómo preparar los datos para garantizar una adecuada implementación de técnicas de aprendizaje automático?
- ¿Cómo construir modelos usando técnicas de aprendizaje automático que permitan predecir el riesgo de abandono?
- ¿Cómo evaluar el desempeño de los modelos en técnicas de aprendizaje automático?
- ¿Cómo comparar el desempeño de los modelos en técnicas de aprendizaje automático?
- ¿Cómo visualizar el comportamiento del riesgo de abandono?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar una aplicación web que contenga el mejor modelo computacional entrenado que prediga y visualice el riesgo de abandono de un asociado, para una cooperativa multiactiva basado en técnicas de aprendizaje automático.

2.2 OBJETIVOS ESPECÍFICOS

- Analizar las características relevantes para estimar el riesgo de abandono de un asociado.
- Seleccionar los datos requeridos para la construcción de modelos usando técnicas de aprendizaje automático.
- Preparar los datos de forma adecuada para el entrenamiento de técnicas de aprendizaje automático.
- Construir modelos usando técnicas de aprendizaje automático que permitan predecir el riesgo de abandono.
- Evaluar el desempeño de los modelos en técnicas de aprendizaje automático entrenadas, a partir de las métricas de evaluación.
- Comparar las métricas de evaluación de los modelos entrenados en técnicas de aprendizaje automático con los modelos previos.
- Realizar la visualización de las predicciones de los asociados activos a la cooperativa.

3. RESULTADOS ESPERADOS

Los resultados esperados en el desarrollo del proyecto son:

- Conjunto de datos pre-procesados para la construcción de modelos mediante técnicas de aprendizaje automático.

- Documento donde se realiza el análisis de los datos, contenidos en las tablas que reportan los resultados de los experimentos de entrenamiento de los modelos de aprendizaje automático.

- Programa que implementa cada uno de los modelos de aprendizaje automático y construye los gráficos automáticamente, de la predicción del riesgo de abandono.

4. ALCANCE

Creación de una aplicación Web de visualización de datos que incluye el mejor modelo

entrenado de aprendizaje automático, que prediga el riesgo de abandono de cada uno de los asociados activos en la cooperativa. El conjunto de datos históricos para el entrenamiento de los modelos está entre los años 2008 y 2020. La cantidad de modelos de aprendizaje automático se establecerá acorde a las investigaciones previas sobre el tema. Los criterios de selección de los modelos entrenados se trabajarán con la matriz de confusión, especificidad, precisión, sensibilidad y F1-score. Estos indicadores serán usados para comparar con los modelos entrenados por la cooperativa en otros experimentos realizados. Por último, se define la variable a predecir como retirados voluntariamente y activos normales en la cooperativa, para crear la variable de clasificación de dos clases (retiro voluntario y activo normal).

5. JUSTIFICACIÓN

La cooperativa multiactiva, con una base social de 250 mil asociados necesita mantener la mayor cantidad de asociados activos para fortalecer su modelo cooperativo en Colombia. Para ello, en este trabajo se realizará la predicción del riesgo de abandono de la cooperativa usando técnicas de aprendizaje automático, mejorando los modelos previamente construidos en la cooperativa, así como la visualización de datos para una mejor comprensión. De manera que, las técnicas de aprendizaje automático facilitan la obtención de información para comprender, organizar y predecir el riesgo de abandono de asociados con un grado de precisión, mientras que la visualización de datos, ayuda a caracterizar los asociados activos que presentan mayor riesgo de abandono.

La importancia de este proyecto radica en la creación de una herramienta computacional que pueda ser utilizada por la fuerza comercial de la cooperativa, lo que permitirá el diseño de estrategias de retención y la toma de decisiones más informadas para impactar positivamente la tasa de abandono de asociados.

6. MARCO TEÓRICO DE REFERENCIA Y ANTECEDENTES (ESTADO DEL ARTE)

En esta sección se presentan los fundamentos teóricos del proyecto, como: concepto de la tasa de deserción de clientes, aprendizaje automático, métricas de desempeño, técnicas de procesamiento de datos y por último la revisión de la literatura.

6.1 TASA DE DESERCIÓN DE CLIENTES (RATE CHURN)

El término Churn Rate hace referencia a la tasa de cancelación de clientes, es decir, se encarga de medir el porcentaje de clientes que se dan de baja de una empresa en un periodo de tiempo determinado. Es una métrica indispensable a la hora de conocer el motivo de la pérdida de clientes y gracias a él, poder elaborar estrategias de marketing que te permitan obtener una mayor fidelización de clientes [15].

Según [16], retener a un cliente resulta aproximadamente diez veces más barato que conseguir uno nuevo, por eso debe ser una prioridad saber aplicar estrategias de retención y fidelización que consigan mantener y desarrollar a los clientes rentables y fieles.

Existen dos tipos de Churn Rate: voluntario e involuntario. El Churn Rate voluntario es cuando por decisión propia el cliente decide cambiar de compañía o servicio, mientras que el Churn Rate involuntario sucede cuando el cliente deja la empresa debido a una causa externa como la reubicación de una zona geográfica, la falta de factibilidad técnica del servicio en la zona, la caída en morosidad, y en caso extremo, la muerte [17].

El problema del abandono de clientes es una constante en sectores en que los clientes se tienen que suscribir o abonar a un determinado servicio como por ejemplo el sector de las telecomunicaciones [15]. En el sector cooperativo el cual es objeto de estudio, el cliente (asociado) es la base central del modelo cooperativo, y mantener la base social de asociados es primordial porque vincular un nuevo asociado, puede ser mucho más complicado que retener a los actuales, es así que un alto retiro de asociados se convierte en un problema importante.

6.2 APRENDIZAJE AUTOMÁTICO

El aprendizaje automático de acuerdo con [1] lo enmarcan como un subconjunto de la inteligencia artificial, el cual permite a un sistema aprender patrones y adaptar un modelo a partir de un conjunto de datos, con un nivel de desempeño mayor que la programación explícita.

En la figura 1, se observa los grupos que componen el aprendizaje automático, dado que es un subconjunto de la inteligencia artificial.

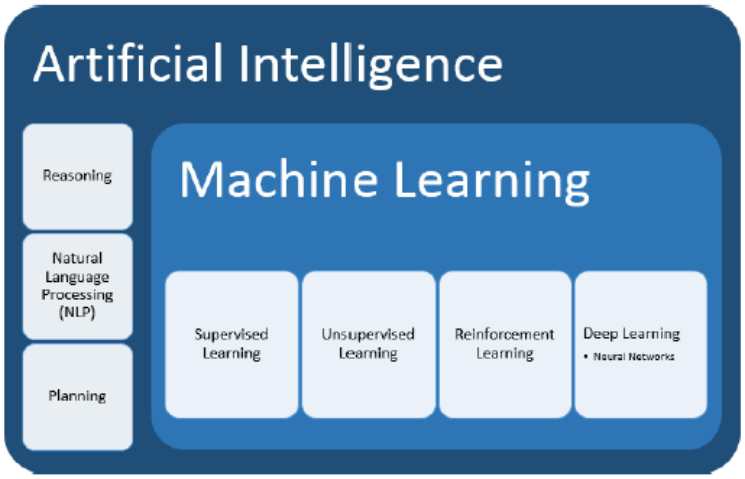


Figura 1: Inteligencia artificial.
Fuente: [1]

En la figura 1, se encuentran los grupos en los cuales se define el aprendizaje automático (machine learning), se definirá cada uno de ellos:

- **Aprendizaje supervisado**: Se le especifica al modelo que debe aprender, por lo cual la variable respuesta puede ser etiquetada, como en aprendizaje supervisado para una clasificación.

- **Aprendizaje no supervisado**: en este caso no se sabe la variable respuesta como etiqueta como en un caso de clasificación, sino que los modelos aplicados descubren patrones, grupos o regularidades de los datos.

- **Aprendizaje por refuerzo**: Es un modelo de aprendizaje conductual a través de ensayo y error y donde lo aprende es un entorno simulado.

- **Deep Learning**: Definido por las redes neuronales profundas, en el cual se pueden agregar muchas capas y neuronas.

Como el proyecto trabajará mediante el aprendizaje supervisado, se profundizará en la siguiente sección.

6.3 APRENDIZAJE SUPERVISADO

De acuerdo con [1], el aprendizaje supervisado generalmente comienza con un conjunto establecido de datos y una cierta comprensión de cómo se clasifican esos datos, por lo cual se habla de datos etiquetados para un caso de clasificación. El aprendizaje tiene en cuenta esa etiqueta para aprender a clasificar según la etiqueta correspondiente.

El aprendizaje supervisado según [8], está destinado a encontrar patrones en los datos que se puedan aplicar a un proceso de análisis. Estos datos tienen características etiquetadas que definen el significado de los datos. [1], además definen el siguiente ejemplo, para el caso de aprendizaje supervisado. Donde al haber millones de imágenes de animales y en cada animal incluye una explicación, que luego se puede crear una aplicación de aprendizaje automático que distinga a un animal de otro, para este caso, etiquetar estos datos sobre tipos de animales, puede tener cientos de categorías de diferentes especies, debido a que se han identificado los atributos y el significado de los datos, así los usuarios que entrenan los datos modelados lo entienden bien para que se ajusten a los detalles de las etiquetas. Cuando la variable que quiere predecir no es una etiqueta sino una variable continua, se aplicará una regresión, que para el aprendizaje supervisado esta le ayuda a comprender la correlación entre las variables. Un ejemplo de aprendizaje supervisado cuando se tiene una variable continúa calificación de una

prueba, que depende de las horas estudiadas, lo cual se puede ajustar la calificación de una prueba dependiendo la relación que tenga con las horas estudiadas.

En la figura 2, se definirá el caso para una clasificación y una regresión.

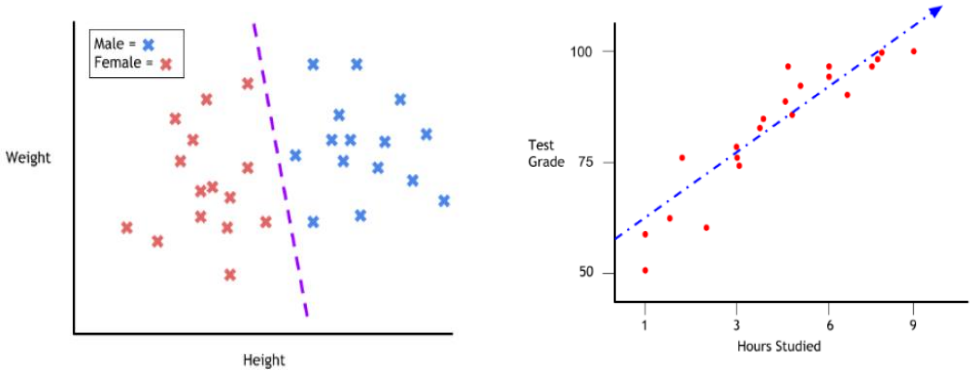


Figura 2. Ejemplo de un problema de clasificación y regresión.

Fuente: https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590

Dentro del aprendizaje supervisado, existen una gran cantidad de métodos para el caso que se requiera clasificar, a continuación, se expone la definición de las metodologías que serán implementadas en el presente trabajo.

6.4 TÉCNICAS UTILIZADAS

ÁRBOLES DE DECISIÓN

En [6], se define a los árboles de decisión como un método que consiste en particionar recursivamente el espacio de las variables independientes, de tal forma que se minimice el índice de impureza entre las regiones conformadas. Si se denota la variable dependiente como **y** y las variables independientes x1, x2, …. xp, la partición recursiva dividirá el espacio dimensional **p** de las variables **x** en conjuntos disyuntos siguiendo el siguiente mecanismo recursivo. En primer lugar, se selecciona una de las variables xi, para esa variable se selecciona un valor si a partir del cual se dividirá el espacio de dimensiones p en dos partes, una parte que contiene todas las observaciones donde xi <= sí y otro donde se encuentre las observaciones xi > sí. Luego, cada una de las partes conformadas se dividirá siguiendo un mecanismo similar. El proceso continúa hasta obtener regiones cada vez más pequeñas. El objetivo, entonces, es dividir, todo el espacio de x en regiones lo más homogéneas posibles (regiones que contengan sólo una clase).

RANDOM FOREST

Esta técnica de aprendizaje automático, según [7], lo define como una extensión del árbol de clasificación que consiste en una colección independiente, idénticamente distribuida y al azar de clasificadores organizados en árboles, en donde cada árbol aporta un único voto a la clase más popular de X. Básicamente, para los agrupamientos Random Forest selecciona al azar un subconjunto de los atributos para luego volver a seleccionar el mejor corte entre estos. Posteriormente el proceso se repite en cada uno de los árboles (muchos árboles crecen de la misma manera) para así construir un bosque. Finalmente, todos los árboles son usados en el resultado final, la etiqueta que obtenga mayor cantidad de incidencias es reportada como la predicción.

MÁQUINAS DE VECTORES DE SOPORTE

Según [8], la característica principal de esta técnica se basa en construir un hiperplano de separación óptimo entre dos clases perfectamente separadas en el cual existe un límite lineal, donde las clases pueden no ser separables. Esto quiere decir que las MVS buscan construir un hiperplano como plano de decisión, el cual separa las clases positivas (+1) y negativas (-1) con el mayor margen de separación, para un problema de clasificación., Cuando las dos clases son linealmente separables en **Rd** se desea encontrar un hiperplano separador que entregue un error de generalización más pequeño entre el número infinito de posibles planos.

Las MVS tienen dos parámetros especificados que se deben ajustar por el usuario:

- c: Parámetro que controla la penalización del error una vez se fija el kernel.
- Parámetros del kernel correspondiente para un kernel polinomial es **p**.

REDES NEURONALES ARTIFICIALES

De acuerdo con [1], en la figura 3 se presenta la arquitectura básica de un perceptrón multicapa.

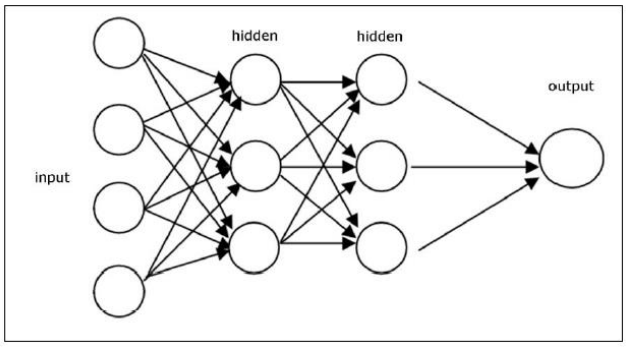


Figura 3. Arquitectura perceptrón multicapa.

Fuente:https://medium.com/@joshua.payne/an-introduction-to-recurrent-neural-networks-8151823daeb7

Los componentes según [9], principales del funcionamiento de una red neuronal son:

- **Entradas**: Las variables de entrada y salida pueden ser binarias (digitales) o continuas (analógicas) dependiendo del modelo de aplicación.
- **Pesos sinápticos**: Representan la intensidad de interacción entre cada neurona presináptica y la neurona postsináptica.
- **Reglas de propagación**: Proporciona el valor del potencial postsináptico de la neurona en función de sus pesos y entradas. La función más habitual es de tipo lineal, y se basa en una suma ponderada de las entradas con los pesos sinápticos (unión sumadora).
- **Función de activación o de transferencia**: Proporciona el estado de activación actual de la neurona en función de su estado anterior.
- **Función de salida**: Proporciona la salida actual de la neurona.

Ventajas

- **Aprendizaje adaptativo**: Las RNA aprenden a realizar tareas a partir de un conjunto de datos dados, en el proceso de aprendizaje, estos datos son representados como las entradas y pesos.
- **Auto-organización**: Pueden crear su propia organización o representación de la información recibida
- **Operación en tiempo real**: Las operaciones realizadas pueden ser llevadas a cabo por computadores paralelos, o dispositivos de hardware especiales que aprovechan esta capacidad.

- **Tolerancia a fallos parciales**: La destrucción parcial de una red, daña el funcionamiento de la misma, pero no la destruye completamente. Esto es debido a la redundancia de la información contenida.
- En el proceso de aprendizaje, los enlaces ponderados de las neuronas se ajustan de manera que se obtengan ciertos resultados específicos. Una red neuronal no necesita un algoritmo para resolver un problema, ya que ella puede generar su propia distribución de pesos en los enlaces mediante el aprendizaje. También existen redes que continúan aprendiendo a lo largo de su vida, después de completado su período de entrenamiento.
- **Operación en tiempo real**: los cómputos neuronales pueden ser llevados a cabo en paralelo; para esto se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad.
- **Inclusión flexible en la tecnología vigente**: se pueden obtener chips especializados para redes neuronales que mejoran su capacidad en ciertas tareas. Ello permitirá la integración de módulos en los sistemas existentes.
- **Las redes neuronales se autoajustan a los elementos procesales**: Son dinámicas, pues son capaces de estar constantemente cambiando para adaptarse a las nuevas condiciones.
- La función del diseñador es únicamente la obtención de la arquitectura apropiada. No es problema del diseñador el cómo la red aprenderá a discriminar. Sin embargo, sí es necesario que desarrolle un buen algoritmo de aprendizaje que le proporcione a la red la capacidad de discriminar.

Desventajas

- Complejidad de aprendizaje para grandes tareas, cuantas más cosas se necesita que aprenda la red, más complicado será enseñarle.
- Tiempo de aprendizaje elevado. Esto depende de dos factores: primero se incrementa la cantidad de patrones a identificar o clasificar y segundo se requiere mayor flexibilidad o capacidad de adaptación de la red neuronal para reconocer patrones que sean sumamente parecidos, se deberá invertir más tiempo en lograr que la red converja a valores de pesos que representan lo que se quiere enseñar.
- No permite interpretar lo que se ha aprendido, la red por si sola proporciona una salida, un número, que no puede ser interpretado por ella misma, sino que se requiere de la intervención del programador y de la aplicación en si para encontrarle un significado la salida proporcionada [9].

MÉTODOS DE ENSAMBLE

Según [10], los métodos de ensamble son técnicas en el aprendizaje automático que combinan una serie de modelos base con el objetivo de producir una predicción más precisa. Los tipos más comunes de este tipo de técnicas son:

- **Promedio simple**: Las predicciones se combinan a partir de todos los modelos bases y se toma el promedio simple de dichas decisiones. Esta técnica ha tenido una eficiencia del modelo final debido a que reduce la varianza de la variable respuesta.

- **Generalización apilada**: Se utiliza un nuevo modelo base para corregir los errores del modelo base anterior, de manera que se crea un modelo padre que toma como valores de entrada las salidas de las predicciones de cada uno de los modelos base.

- **Bagging**: es una técnica que genera múltiples conjuntos de entrenamiento a partir de la base de entrenamiento inicial, a cada uno de estos se les genera un modelo base individual. A partir de las predicciones de estos modelos base, se genera un modelo más grande el cual agrega los resultados para generar una predicción final, tal y como se muestra en la Figura 4.

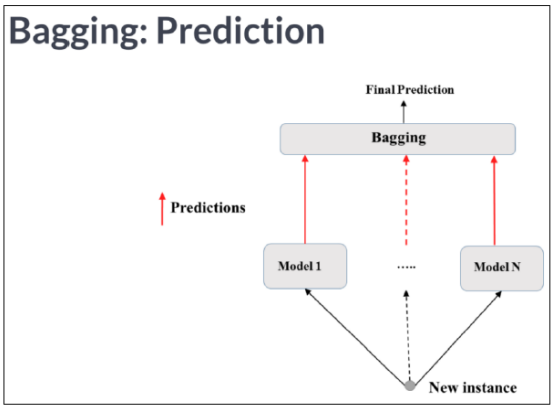


Figura 4. Visualización de funcionamiento de bagging.
Fuente: http://deepnote.me/2019/08/25/datascience-18-machine-learning-with-tree-based-models-in-python/

- **Boosting:** es una técnica donde se generan varios modelos de manera individual con baja capacidad predictiva, y así cada modelo toma una entrada diferente y le entrega al siguiente modelo una versión mejorada de la base de entrenamiento para que este maximice los indicadores de desempeño y al finalizar el entrenamiento, se genera la predicción con base en las predicciones de cada uno

de los modelos individuales. La estructura de este modelo se describe en la Figura 5

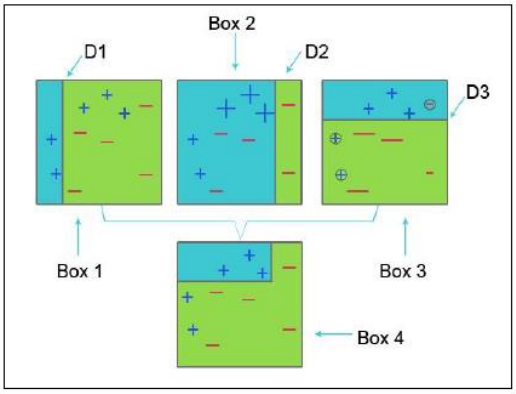


Figura 5. Visualización de funcionamiento del boosting.
Fuente: https://www.educba.com/boosting-algorithm/

Ventajas de estas técnicas de ensamble:

- Disminución de sesgo y varianza debido a la compensación entre los modelos base.
- Poca probabilidad al sobre entrenamiento.

Las principales desventajas de las técnicas de ensamble:

- Estas técnicas al tener que combinar métodos de aprendizaje automático para generar mejoras en la predicción, son difíciles de interpretar por lo que son caja negra y la salida es un modelo complejo, algo que puede ser útil para la interpretación es la importancia de las variables predictoras.
- Son computacionalmente costosos, ya al integrar varias técnicas se pueden consumir rápidamente la memoria del cómputo por el cálculo que debe hacer.

6.5 MÉTRICA DE EVALUACIÓN Y RENDIMIENTO

Cuando se ajustan técnicas de aprendizaje automático, y se quiere seleccionar la mejor de ellas de acuerdo con su mejor predicción dado para un caso de clasificación. Es necesario evaluar el desempeño de la técnica de aprendizaje automático con métricas para seleccionar la que mejor clasifique de acuerdo con el problema en estudio.

MATRIZ DE CONFUSIÓN

En [11] se afirma que la matriz de confusión representa una de las formas más

interpretativas para medir el rendimiento de clasificación de una técnica de aprendizaje automático. El esquema de la matriz de confusión se puede observar en la Tabla 1:

Tabla 1. Matriz de Confusión

		Observación	
		Positivos	**Negativos**
Predicción	**Positivos**	Verdaderos Positivos (VP)	Falsos Positivos (FP)
	Negativos	Falsos Negativos (FN)	Verdaderos Negativos (VN)

Fuente: Propia

Donde VN son los verdaderos negativos, FN son los falsos negativos, FP son los falsos positivos y VP son los verdaderos positivos. Se conocen como negativos las marcaciones iguales a cero y positivos a las observaciones cuya marcación sea uno, para un caso binario de clasificación.

EXACTITUD (ACCURACY EN INGLÉS) (%)

Es el porcentaje de clasificaciones correctas. Esta métrica da una mirada generalizada sobre el desempeño del modelo, pero cuando hay desbalance en la marcación o se desea conocer cómo se distribuye los éxitos y fracasos del modelo por clase no es suficiente.

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

ERROR RATE (%)

Es el porcentaje de Clasificaciones Incorrectas.

$$ERR = \frac{FP + FN}{VP + VN + FP + FN} = 1 - Accuracy \quad (2)$$

ESPECIFICIDAD (%)

Es la capacidad de detectar Verdaderos Negativos, también conocida como fracción de verdaderos negativos.

$$Especificidad = \frac{VN}{VN + FP} \quad (3)$$

PRECISIÓN O VALOR PREDICTIVO POSITIVO (%)

Es el porcentaje de resultados positivos.

$$VPredictP = \frac{VP}{VP + FP} \quad (4)$$

Esta métrica permite conocer que tan acertado es el modelo respecto a su clasificación de positivos.

VALOR PREDICTIVO NEGATIVO (%)

Es el porcentaje de resultados negativos.

$$VPredictN = \frac{VN}{VN + FN} \quad (5)$$

F1SCORE (PUNTAJE) (%)

Es el promedio ponderado de la precisión y el recall, por tanto, es la capacidad de contar tanto los falsos positivos o como falsos negativos.

$$F1Score = 2 * \frac{Recall * Precisión}{Recall + Precisión} \quad (6)$$

CURVA ROC

En [12] se define las curvas ROC como una herramienta para evaluar la habilidad de un clasificador en la discriminación de instancias verdaderas y falsas. En una definición más acertada se puede decir que las Curvas ROC son las que miden la relación de la tasa de verdaderos positivos (predicciones acertadas) versus la tasa de falsos positivos (predicciones erradas). Siendo el positivo el referente a la clase de fuga cuando se trata de un problema de clasificación binario. Estas curvas no tienen una fórmula asociada.

No obstante, sí tienen una métrica, llamada “Area Under the Curve” (AUC), que se define como el área bajo la Curva ROC, tiene la siguiente propiedad estadística: “La AUC de un clasificador es equivalente a la probabilidad que el clasificador posicionará una instancia aleatoria positiva mejor que una instancia aleatoria negativa” [13].

A continuación, se muestran 3 ejemplos de curva ROC. La primera con discriminación perfecta (AUC=1), discriminación adecuada (AUC=0.8) y capacidad de discriminación similar al azar (AUC=0.5).

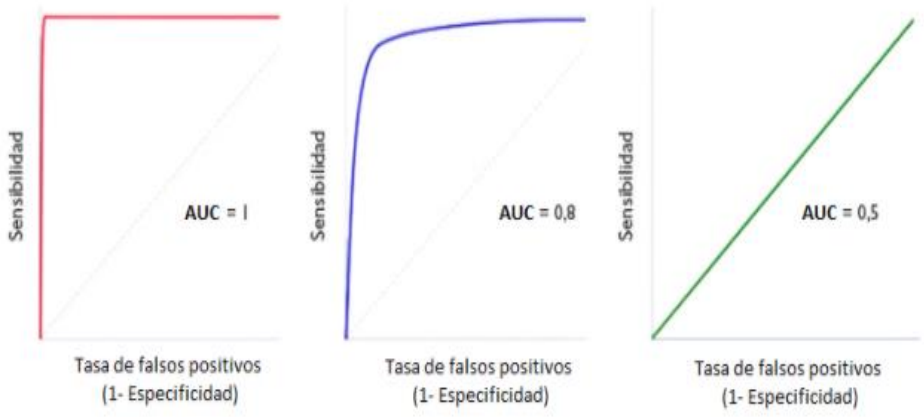


Figura 6. Curva ROC.
Fuente: https://www.cienciasinseso.com/en/privacidad/roc-curve/

6.6 TÉCNICAS DE PROCESAMIENTO DE DATOS

En las técnicas de aprendizaje automático, se debe realizar un procesamiento previo a los datos para poder que funcionen de manera adecuada. Por ejemplo, en el caso de tener datos faltantes, datos atípicos o en el caso de tener duplicidad de información. Esto puede causar un mal ajuste de las técnicas a los datos.

ONE HOT ENCODING

La transformación de variables categóricas con n valores distintos, en un vector de tamaño n (igual al número de posibles valores que puede tomar 38 la variable categórica), se asigna el valor de 1, cuando hay presencia de la variable en la posición correspondiente y 0 para el resto, visto de manera de fila, así existirá presencia o ausencia de la variable.

NORMALIZACIÓN

Esta técnica hace referencia a escalar los dominios de las variables hacia una métrica común, con la finalidad que las diferencias entre las escalas de magnitud de las variables generen un sesgo en el modelo y, por tanto, afecte su desempeño y capacidad de predicción.

7. TRABAJOS RELACIONADOS

Las técnicas de aprendizaje automático expuestas se enmarcan en análisis supervisado y cuando se quiere hacer una clasificación. Así la revisión de la literatura estará enfocada en análisis de predicción de cancelación de un servicio o producto, lo cual genera una variable binaria 0 o 1, cuando se tienen dos clases a predecir.

En [2] se realiza un análisis sobre la deserción de clientes en la industria de las telecomunicaciones, dando importancia al tiempo que tarda el cliente en cancelar el producto. Se considera una variable de respuesta condicionada al tiempo en días de cancelación del producto, argumentando que tener una respuesta de la predicción medida en meses puede no ser suficiente para reaccionar en una industria saturada. La metodología propone cuatro modelos para predecir la deserción diaria. Dos modelos dependen de características extraídas de la serie de tiempo multivariante, como el modelo basado en RFM y el modelo basado en estadísticas, mientras que los otros modelos, aprovechan las técnicas de aprendizaje profundo para la extracción automática de características, como el modelo basado en LSTM y el modelo basado en CNN.

De acuerdo con [3], hoy en día dada la gran cantidad de información, estructurada y no estructurada, es importante para entender mejor a los clientes y mejorar la detección de clientes propensos a la cancelación de un producto o servicio. La huella digital, es información capturada por el uso de aplicaciones e interacción con redes sociales. Esta información es de gran ayuda para capturar patrones externos que no están en las fuentes internas de información de los sistemas transaccionales tradicionales. Estas fuentes no estructuradas como el texto, para el análisis han demostrado que son de gran ayuda para mejorar las métricas de desempeño de las técnicas de aprendizaje automático.

El éxito de las campañas de retención en mercados rápidos y saturados como la industria de telecomunicaciones, depende de predecir con precisión los posibles clientes que van a cancelar el servicio. En un estudio como en [4], proponen una forma de incorporar el tiempo para mejorar el comportamiento del cliente y presentar un enfoque dinámico de extremo a extremo para la tarea de predicción de abandono en servicios de telecomunicaciones, que incluso conduce a un mayor rendimiento del modelo. En este caso, se enfoca en construir redes de llamadas semanalmente durante un período de seis meses y extraer características de cada cliente para capturar la dinámica del comportamiento de abandono del cliente, utilizando tres conjuntos de datos distintos. Así se puede predecir la pérdida de clientes, luego se realiza una clasificación con la serie de tiempo multivariante resultante del comportamiento de las fuentes de datos, que contiene variables como la duración de la llamada, fecha y número de celular.

En estudios realizados en [5], usan modelos híbridos, en el cual explican cómo combinar diferentes técnicas supervisadas como no supervisadas, con el objetivo de mejorar las métricas de predicción y determinar focos estratégicos a perfiles encontrados.

8. METODOLOGÍA

En este capítulo, se relaciona cada uno de los pasos propuestos para dar cumplimiento tanto a los objetivos específicos como al objetivo general del presente trabajo.

A. ESTUDIO DE LAS CARACTERÍSTICAS RELEVANTES PARA ESTIMAR EL RIESGO DE ABANDONO

- Seleccionar la información de bodega de datos.
- Realizar una ficha técnica de la información y definir tiempo de análisis en la base de datos.
- Construir un diccionario de variables.
- Realizar un análisis descriptivo e identificar variables relevantes

B. SELECCIÓN DE LOS DATOS PARA LA CONSTRUCCIÓN DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO.

- Analizar correlaciones y seleccionar las variables relevantes.

C. PREPARACIÓN DE LOS DATOS DE FORMA ADECUADA PARA EL ENTRENAMIENTO DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO.

- Aplicar codificación de variables categóricas y normalización de variables cuantitativas.
- Seleccionar las técnicas de aprendizaje automático.
- Dividir la base de datos en entrenamiento y prueba.

D. CONSTRUCCIÓN DE LAS TÉCNICAS DE APRENDIZAJE AUTOMÁTICO QUE PERMITAN PREDECIR EL RIESGO DE ABANDONO.

- Definir los experimentos para la construcción de las técnicas de aprendizaje automático.
- Reporte de los resultados experimento 1(técnicas de aprendizaje automático por defecto)
- Reporte de los resultados experimento 2(optimización de hiperparámetros)

E. EVALUACIÓN DEL DESEMPEÑO DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO ENTRENADOS, A PARTIR DE LAS MÉTRICAS DE EVALUACIÓN.

- Generar tablas comparativas de los experimentos en cada una de las métricas y selección de la mejor técnica.

F. CONSTRUCCIÓN DE LA VISUALIZACIÓN DE LAS PREDICCIONES DE LOS ASOCIADOS ACTIVOS A LA COOPERATIVA.

- Construir la visualización de los datos en un aplicativo web

9. DESARROLLO DEL PROYECTO

El proyecto se trabajó con base en la revisión de la literatura del capítulo 6 y siguiendo el esquema descrito de las actividades del capítulo 8. En la figura 7, se muestra el esquema general de la metodología usada en el proyecto.

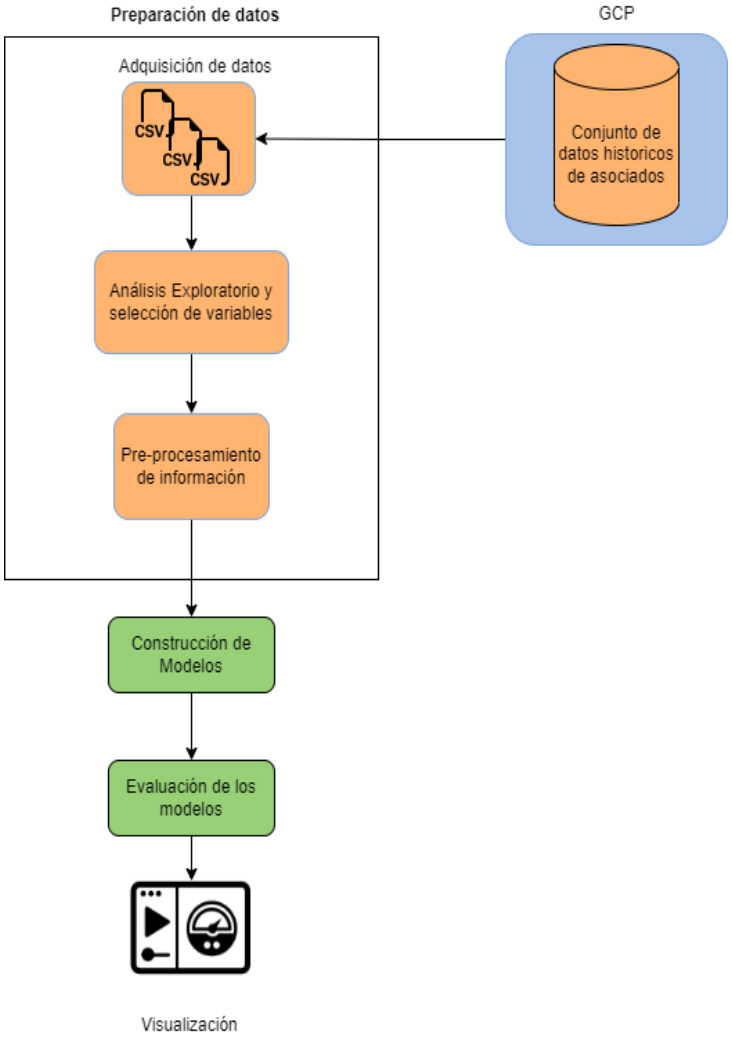


Figura 7. Metodología General.
Fuente: Elaboración propia.

9.1 ESTUDIO DE LAS CARACTERÍSTICAS RELEVANTES PARA ESTIMAR EL RIESGO DE ABANDONO

9.1.1. SELECCIÓN DE LA INFORMACIÓN DE LA BODEGA DE DATOS

La bodega de datos de la cooperativa contiene las tablas descritas en la tabla 2, y son las que están disponibles para estructurar el conjunto de datos, con las variables para la predicción del riesgo de abandono.

Tabla 2 Dimensiones y tablas de hechos del sistema de base de datos

Tablas	Instancia	Descripción
dimCliente	Bodega	Información del cliente
fact asociatividad (cootaylor)	Bodega	Tabla que contiene información de negocio de la cooperativa
fact Demografica	Bodega	Tabla con información demográfica
fact tenenciasalud	Bodega	Información de productos del sector salud
factTenenciaRecreacionTurismo	Bodega	Información de turismo
factAtentosCasosPQR	Bodega	Información de quejas y reclamos
factSegmentos	Bodega	Información de agrupación de variables, según negocio
factTenenciaProteccion	Bodega	Última partición
factTenenciaMicrocreditos	Bodega	Información de productos financieros
factTenenciaCrediSolidario	Bodega	Información de productos financieros cooperativos
factTenenciaEducacionFundacion	Bodega	Información de productos o programas con la fundación
factTurnos	Bodega	Información de turnos en las oficinas
factAprobacionesNegaciones	Bodega	Aprobaciones o negaciones de productos
factIndicadores	Bodega	Información de indicadores usados por la cooperativa.
factBeneficio	Bodega	Tabla que contiene los beneficios monetarios por los productos cooperativos
factIndicadoresCuotaCeroUno	Bodega	Información de pago en las primeras cuotas de los clientes
factCarterizacionProfundizacion	Bodega	Información de gestión comercial
factGestionComercialProfundizacion	Bodega	Información comercial de colocación de productos.
factDetalleProfundizacion	Bodega	Información de la gestión comercial a clientes y comentarios.
TmpHistoricoBanco	stage	Información histórica de productos de banco

Fuente: Elaboración propia.

Las tablas de la base de datos corporativa de la tabla 2, son extraídas, transformadas y

cargadas mediante un proceso de ETL (Extract, Transform and Load), en un file server, para luego ser cargado a la nube de GCP (Google Cloud Platform). El acceso a las bases de datos distribuidas en GCP, se hace a través de una API y las solicitudes de información se pueden realizar por medio de la terminal de Linux (Bash).

La adquisición de los datos se puede hacer de dos formas, haciendo una solicitud a la API de una muestra de clientes de manera aleatoria o a través de una solicitud de clientes particulares mediante el campo de identificación encriptado. En el desarrollo del proyecto se usó esta última, generando una lista de 411.600 clientes con llave de identificación en un archivo PKL. El proceso de extracción del sistema de base datos, realiza consultas por lotes de 2000 clientes. Estos lotes son exportados en archivos con formato .csv, en una ruta especificada por el usuario que realiza la consulta al sistema.

9.1.2. FICHA TÉCNICA DE LA INFORMACIÓN Y TIEMPO DE ANÁLISIS EN LA BASE DE DATOS

En la tabla 3, se muestra la información del sistema de base de datos distribuidos en GCP, que contiene la cantidad de variables por cada dimensión.

Tabla 3. Ficha técnica de la base de datos

Dimensión	Descripción	Cantidad de Variables
Demográfica	Variables demográficas y segmentos de los asociados	18
Interacciones de Compra	-Productos Vigentes y Cancelados en la historia recolectada de productos -Participación en eventos en la cooperativa	87
Comportamiento de pago	-Inactividad: Asociado entra en estado de inactividad en la cooperativa, por dejar de pagar la cuota en un lapso de 90 días -Cuota cero o cuota uno: No paga ni la primera cuota, y solo paga una	4
Beneficios como Asociado	Tenencia y uso de los productos	98
Servicio	Solicitudes, quejas y reclamos Negaciones de crédito	8
Estado	Define si el cliente está activo o retirado	1
Total		216

Fuente: Elaboración propia.

Dado lo anterior, con los resultados de 206 archivos en formato .csv que contiene toda la información de las dimensiones de la tabla 3, se realiza un programa en Python que transforma los archivos .csv, en una sábana de datos organizada con la información histórica de los últimos 10 años, para el periodo 2008-2018, de tal forma que quedan en

las filas los clientes y en las columnas las variables con una dimensión de 411.600 filas y 216 columnas.

9.1.3. DICCIONARIO DE VARIABLES

En la sección anterior se define cuantas variables componen la extracción del conjunto de datos por dimensión. En la tabla 4, se muestra el diccionario de variables.

Tabla 4. Diccionario de variables

Dimensión	Id	Variables
Beneficios	1	facttenenciacredisolidario_numcantidadcuotasmora_sum
	2	facttenenciacredisolidario_numvalorcuota_sum
	3	facttenenciacredisolidario_numvalorsaldoactual_sum
	4	facttenenciacredisolidario_numvalorsaldomora_sum
	5	facttenenciacredisolidario_numvalortasaefectivaanual_sum
	6	facttenenciaeducacionfundacion_count_sieeducacion-cancelado_sum
	7	facttenenciaeducacionfundacion_count_sieeducacion-inactivo_sum
	8	facttenenciaeducacionfundacion_count_sieeducacion-suspendido_sum
	9	facttenenciaeducacionfundacion_count_sieeducacion-terminado_sum
	10	facttenenciaeducacionfundacion_count_sieeducacion-vigente_sum
	11	facttenenciaeducacionfundacion_count_siefundacion-cancelado_sum
	12	facttenenciaeducacionfundacion_count_siefundacion-inactivo_sum
	13	facttenenciaeducacionfundacion_count_siefundacion-suspendido_sum
	14	facttenenciaeducacionfundacion_count_siefundacion-terminado_sum
	15	facttenenciaeducacionfundacion_count_siefundacion-vigente_sum
	16	facttenenciaeducacionfundacion_count_suecoeduc-suspendido_sum
	17	facttenenciaeducacionfundacion_count_suecoeduc-terminado_sum
	18	facttenenciaeducacionfundacion_count_suecoeduc-vigente_sum
	19	facttenenciaeducacionfundacion_count_suecofund-suspendido_sum
	20	facttenenciaeducacionfundacion_count_suecofund-terminado_sum
	21	facttenenciaeducacionfundacion_count_suecofund-vigente_sum
	22	facttenenciaeducacionfundacion_numvalorcuota_sum
	23	facttenenciamicrocreditos_count_al_dia_sum
	24	facttenenciamicrocreditos_count_en_mora_sum
	25	facttenenciamicrocreditos_numvalorcuota_sum
	26	facttenenciamicrocreditos_numvalordesembolso_sum
	27	facttenenciaproteccion_count_autos_comisionable-cancelado_sum
	28	facttenenciaproteccion_count_autos_comisionable-vencida_por_renovacin_sum
	29	facttenenciaproteccion_count_autos_comisionable-vigente_sum
	30	facttenenciaproteccion_count_bancaseguros-cancelado_sum
	31	facttenenciaproteccion_count_bancaseguros-vigente_sum
	32	facttenenciaproteccion_count_campo_nulo-cancelado_sum
	33	facttenenciaproteccion_count_campo_nulo-cotizado_sum
	34	facttenenciaproteccion_count_campo_nulo-vigente_sum

	35	facttenenciaproteccion_count_desempleo_banco-cancelado_sum
	36	facttenenciaproteccion_count_desempleo_banco-vigente_sum
	37	facttenenciaproteccion_count_hogar_mas_y_total_home_comisionable-cancelado_sum
	38	facttenenciaproteccion_count_hogar_mas_y_total_home_comisionable-vencida_por_renovacin_sum
	39	facttenenciaproteccion_count_hogar_mas_y_total_home_comisionable-vigente_sum
	40	facttenenciaproteccion_count_nan-vigente_sum
	41	facttenenciaproteccion_count_otras_polizas-cancelado_sum
	42	facttenenciaproteccion_count_otras_polizas-vencida_por_renovacin_sum
	43	facttenenciaproteccion_count_otras_polizas-vigente_sum
	44	facttenenciaproteccion_count_perdidas_parciales-cancelado_sum
	45	facttenenciaproteccion_count_perdidas_parciales-vigente_sum
	46	facttenenciaproteccion_count_plan_100_daños-cancelado_sum
	47	facttenenciaproteccion_count_plan_100_daños-vencida_por_renovacin_sum
	48	facttenenciaproteccion_count_plan_100_hurto-cancelado_sum
	49	facttenenciaproteccion_count_plan_100_hurto-vencida_por_renovacin_sum
	50	facttenenciaproteccion_count_rce_en_exceso-cancelado_sum
	51	facttenenciaproteccion_count_rce_en_exceso-vencida_por_renovacin_sum
	52	facttenenciaproteccion_count_soat-cancelado_sum
	53	facttenenciaproteccion_count_soat-vigente_sum
	54	facttenenciaproteccion_count_total_rc_medica-cancelado_sum
	55	facttenenciaproteccion_count_total_rc_medica-vencida_por_renovacin_sum
	56	facttenenciaproteccion_count_total_rc_medica-vigente_sum
	57	facttenenciaproteccion_count_vida_fsv-cancelado_sum
	58	facttenenciaproteccion_count_vida_fsv-vigente_sum
	59	facttenenciaproteccion_count_vida_grupo_prima_unica-cancelado_sum
	60	facttenenciaproteccion_count_vida_grupo_prima_unica-vigente_sum
	61	facttenenciaproteccion_numvalorcuota_sum
	62	facttenenciaproteccion_strcodtipomovimiento_sum
	63	facttenenciasalud_count_campo_nulo-activo_sum
	64	facttenenciasalud_count_cem-activo_sum
	65	facttenenciasalud_count_cem-retirado_sum
	66	facttenenciasalud_count_medicina_integral-activo_sum
	67	facttenenciasalud_count_medicina_integral-retirado_sum
	68	facttenenciasalud_count_nan-retirado_sum
	69	facttenenciasalud_count_plan_obligatorio_de_salud-activo_sum
	70	facttenenciasalud_count_plan_obligatorio_de_salud-fallecido_sum
	71	facttenenciasalud_count_plan_obligatorio_de_salud-retirado_sum
	72	facttenenciasalud_count_plan_obligatorio_de_salud-suspendido_sum
	73	facttenenciasalud_count_salud_oral-activo_sum
	74	facttenenciasalud_count_salud_oral-retirado_sum
	75	facttenenciasalud_numvalorcuota_sum
	76	tmpproductosbancohistoricotenencia_count_banco_sta_cof04001_otrosc_sum
	77	tmpproductosbancohistoricotenencia_count_banco_sta_cof04100_sum
	78	tmpproductosbancohistoricotenencia_count_banco_sta_cof04101_sum
	79	tmpproductosbancohistoricotenencia_count_banco_sta_cof04102_sum

	80	tmpproductosbancohistoricotenencia_count_banco_sta_cof04103_sum
	81	tmpproductosbancohistoricotenencia_count_banco_sta_cof04104_sum
	82	tmpproductosbancohistoricotenencia_count_banco_sta_cof041106_sum
	83	tmpproductosbancohistoricotenencia_count_banco_sta_cof041111_sum
	84	tmpproductosbancohistoricotenencia_count_banco_sta_cof0411116_sum
	85	tmpproductosbancohistoricotenencia_count_banco_sta_cof041113_sum
	86	tmpproductosbancohistoricotenencia_count_banco_sta_cof041115_sum
	87	tmpproductosbancohistoricotenencia_count_banco_sta_cof0411460_sum
	88	tmpproductosbancohistoricotenencia_count_banco_sta_cof041401_sum
	89	tmpproductosbancohistoricotenencia_count_banco_sta_cof042110_sum
	90	tmpproductosbancohistoricotenencia_count_banco_sta_cof0421112_sum
	91	tmpproductosbancohistoricotenencia_count_banco_sta_cof0421116_cupo_sum
	92	tmpproductosbancohistoricotenencia_count_banco_sta_cof0421116_saldo_sum
	93	tmpproductosbancohistoricotenencia_count_banco_sta_cof042130_sum
	94	tmpproductosbancohistoricotenencia_count_banco_sta_cof042170_sum
	95	tmpproductosbancohistoricotenencia_count_banco_sta_cof042250_sum
	96	tmpproductosbancohistoricotenencia_count_banco_sta_cof042280_sum
	97	tmpproductosbancohistoricotenencia_count_banco_sta_cofcmcm_sum
	98	tmpproductosbancohistoricotenencia_count_none_sum
Comportamiento_de_pago	99	factindicadorescuotacerouno_count_1196.0_sum
	100	factindicadorescuotacerouno_count_1197.0_sum
	101	factindicadorescuotacerouno_count_272.0_sum
	102	factindicadorescuotacerouno_count_91.0_sum
interacciones	103	factasociatividadsipas_count_asistencia_juridica-cancelado_sum
	104	factasociatividadsipas_count_asistencia_juridica-desembolsado_sum
	105	factasociatividadsipas_count_asistencia_juridica-inactivo_sum
	106	factasociatividadsipas_count_asistencia_juridica-pendiente_sum
	107	factasociatividadsipas_count_asistencia_juridica-vigente_sum
	108	factasociatividadsipas_count_asistencia_pensional-cancelado_sum
	109	factasociatividadsipas_count_asistencia_pensional-desembolsado_sum
	110	factasociatividadsipas_count_asistencia_pensional-inactivo_sum
	111	factasociatividadsipas_count_asistencia_pensional-pendiente_sum
	112	factasociatividadsipas_count_asistencia_pensional-vigente_sum
	113	factasociatividadsipas_count_auxilio_funerario-cancelado_sum
	114	factasociatividadsipas_count_auxilio_funerario-desembolsado_sum
	115	factasociatividadsipas_count_auxilio_funerario-inactivo_sum
	116	factasociatividadsipas_count_auxilio_funerario-pendiente_sum
	117	factasociatividadsipas_count_auxilio_funerario-vigente_sum
	118	factasociatividadsipas_count_campo_nulo-cancelado_sum
	119	factasociatividadsipas_count_campo_nulo-vigente_sum
	120	factasociatividadsipas_count_desempleo-cancelado_sum
	121	factasociatividadsipas_count_desempleo-desembolsado_sum
	122	factasociatividadsipas_count_desempleo-inactivo_sum
	123	factasociatividadsipas_count_desempleo-pendiente_sum
	124	factasociatividadsipas_count_desempleo-vigente_sum
	125	factasociatividadsipas_count_exequial-cancelado_sum
	126	factasociatividadsipas_count_exequial-desembolsado_sum
	127	factasociatividadsipas_count_exequial-inactivo_sum

	128	factasociatividadsipas_count_exequial-vigente_sum
	129	factasociatividadsipas_count_herencia-cancelado_sum
	130	factasociatividadsipas_count_herencia-desembolsado_sum
	131	factasociatividadsipas_count_herencia-inactivo_sum
	132	factasociatividadsipas_count_herencia-vigente_sum
	133	factasociatividadsipas_count_hospitalizacion-cancelado_sum
	134	factasociatividadsipas_count_hospitalizacion-desembolsado_sum
	135	factasociatividadsipas_count_hospitalizacion-inactivo_sum
	136	factasociatividadsipas_count_hospitalizacion-vigente_sum
	137	factasociatividadsipas_count_incrementos_plan_basico-cancelado_sum
	138	factasociatividadsipas_count_incrementos_plan_basico-desembolsado_sum
	139	factasociatividadsipas_count_incrementos_plan_basico-inactivo_sum
	140	factasociatividadsipas_count_incrementos_plan_basico-pendiente_sum
	141	factasociatividadsipas_count_incrementos_plan_basico-vigente_sum
	142	factasociatividadsipas_count_plan_educativo-cancelado_sum
	143	factasociatividadsipas_count_plan_educativo-inactivo_sum
	144	factasociatividadsipas_count_plan_educativo-suspendido_sum
	145	factasociatividadsipas_count_plan_educativo-vigente_sum
	146	factasociatividadsipas_count_prima_nivelada-inactivo_sum
	147	factasociatividadsipas_count_prima_nivelada-suspendido_sum
	148	factasociatividadsipas_count_prima_nivelada-vigente_sum
	149	factasociatividadsipas_count_recuperacion-cancelado_sum
	150	factasociatividadsipas_count_recuperacion-desembolsado_sum
	151	factasociatividadsipas_count_recuperacion-inactivo_sum
	152	factasociatividadsipas_count_recuperacion-vigente_sum
	153	factasociatividadsipas_count_segunda_opinion_medica-cancelado_sum
	154	factasociatividadsipas_count_segunda_opinion_medica-desembolsado_sum
	155	factasociatividadsipas_count_segunda_opinion_medica-inactivo_sum
	156	factasociatividadsipas_count_segunda_opinion_medica-pendiente_sum
	157	factasociatividadsipas_count_segunda_opinion_medica-vigente_sum
	158	factasociatividadsipas_count_solvencia-cancelado_sum
	159	factasociatividadsipas_count_solvencia-desembolsado_sum
	160	factasociatividadsipas_count_solvencia-inactivo_sum
	161	factasociatividadsipas_count_solvencia-suspendido_sum
	162	factasociatividadsipas_count_solvencia-vigente_sum
	163	factasociatividadsipas_count_tranquilidad-cancelado_sum
	164	factasociatividadsipas_count_tranquilidad-desembolsado_sum
	165	factasociatividadsipas_count_tranquilidad-inactivo_sum
	166	factasociatividadsipas_count_tranquilidad-vigente_sum
	167	factasociatividadsipas_count_vida_clasica-cancelado_sum
	168	factasociatividadsipas_count_vida_clasica-desembolsado_sum
	169	factasociatividadsipas_count_vida_clasica-inactivo_sum
	170	factasociatividadsipas_count_vida_clasica-pendiente_sum
	171	factasociatividadsipas_count_vida_clasica-vigente_sum
	172	factasociatividadsipas_count_vida-cancelado_sum
	173	factasociatividadsipas_count_vida-desembolsado_sum
	174	factasociatividadsipas_count_vida-inactivo_sum
	175	factasociatividadsipas_count_vida-vigente_sum

	176	factasociatividadsipas_numvalorcuota_sum
	177	factasociatividadsipas_numvalorproteccioninicialsolicitado_sum
	178	factasociatividadsipas_strcodtipomovimiento_sum
	179	factestadotac_numcupoasignadotac_sum
	180	factestadotac_numcupoutilizadotac_sum
	181	factestadotac_numtarjetasutilizadotac_sum
	182	facttenenciarecreacionturismo_count_campo_nulo-activo_sum
	183	facttenenciarecreacionturismo_count_convenios-activo_sum
	184	facttenenciarecreacionturismo_count_convenios-cancelado_sum
	185	facttenenciarecreacionturismo_count_eventos-activo_sum
	186	facttenenciarecreacionturismo_count_eventos-cancelado_sum
	187	facttenenciarecreacionturismo_count_turismo-activo_sum
	188	facttenenciarecreacionturismo_count_turismo-inactivo_sum
	189	facttenenciarecreacionturismo_numvalorcuota_sum
servicio	190	factaprobacionesnegaciones_count_anulada_sum
	191	factaprobacionesnegaciones_count_aprobada_sum
	192	factaprobacionesnegaciones_count_cupo_activado_sum
	193	factaprobacionesnegaciones_count_cupo_aprobado_sum
	194	factaprobacionesnegaciones_count_desembolsado_sum
	195	factaprobacionesnegaciones_count_en_estudio_sum
	196	factaprobacionesnegaciones_count_negada_sum
	197	factaprobacionesnegaciones_numvalormonto_sum
Demográfica	198	ingresos_num_valor_ingresos
	199	numcantidadhijos
	200	numcantidadpersonascargoadultas
	201	numcantidadpersonascargomenores18
	202	numcodtipovinculacion
	203	numcodunicocorte
	204	numedad
	205	segmentos_num_cantidad_productos_sum
	206	Segmento_Ciclo_de_Vida
	207	Genero
	208	Tipo_Persona
	209	Corte_Factura
	210	Actividad_Laboral
	211	Área_Conocimiento
	212	Estado_Civil
	213	Estrato
	214	Nivel_Academico
	215	Tipo_Vivienda
Estado	216	Label

Fuente: Elaboración propia.

El diccionario de variables que muestra la tabla 4, dimensiona la cantidad de información que se extrae en el proceso de extracción de la información consolidada para su posterior análisis descriptivo.

9.1.4. ANÁLISIS DESCRIPTIVO E IDENTIFICACIÓN DE VARIABLES RELEVANTES

El resultado del Dataframe construido hasta el momento contiene 411.600 filas y 216 columnas, y en primera instancia se realiza una agrupación de variables con un significado similar según el proceso operativo en la cooperativa, por ejemplo, los productos inactivos y cancelados, se suman para agrupar la variable. En la tabla 5, se puede observar un resumen de las variables por dimensión cuando están agrupadas.

Tabla 5. Agrupación de variables

Dimensión	Variables	
	No agrupadas	Agrupadas
Demográfica	18	18
Interacciones de Compra	87	69
Comportamiento de pago	4	4
Beneficios como Asociado	98	85
Servicio	8	5
Estado (activo o retirado)	1	1
Total	216	182

Fuente: Elaboración propia.

Estas 182 variables agrupadas que contienen información de cada variable y con análisis de completitud de información que muestra en la tabla 6, solo 28 variables tienen información completa y las otras contienen más del 94% de información incompleta.

Tabla 6. Variables agrupadas y valores vacíos

Dimensión	No. Variable	Variables agrupadas	% Vacíos
Beneficios	1	autos_comisionableActivo	97%
	2	autos_comisionableCancelado	96%
	3	bancasegurosActivo	95%
	4	bancasegurosCancelado	0,0%
	5	facttenenciacredisolidario_numcantidadcuotasmora_sum	98%
	6	facttenenciacredisolidario_numvalorcuota_sum	0,0%
	7	facttenenciacredisolidario_numvalorsaldoactual_sum	98%

	8	facttenenciacredisolidario_numvalorsaldomora_sum	100%
	9	facttenenciacredisolidario_numvalortasaefectivaanual_sum	96%
	10	facttenenciaeducacionfundacion_numvalorcuota_sum	0,0%
	11	facttenenciamicrocreditos_count_al_dia_sum	98%
	12	facttenenciamicrocreditos_count_en_mora_sum	95%
	13	facttenenciamicrocreditos_numvalorcuota_sum	100%
	14	facttenenciamicrocreditos_numvalordesembolso_sum	97%
	15	facttenenciaproteccion_count_campo_nulo-cancelado_sum	95%
	16	facttenenciaproteccion_count_campo_nulo-cotizado_sum	97%
	17	facttenenciaproteccion_count_campo_nulo-vigente_sum	98%
	18	facttenenciaproteccion_count_desempleo_banco-cancelado_sumCancelado	95%
	19	facttenenciaproteccion_count_desempleo_banco-vigente_sumActivo	97%
	20	facttenenciaproteccion_count_nan-vigente_sum	99%
	21	facttenenciaproteccion_numvalorcuota_sum	97%
	22	facttenenciaproteccion_strcodtipomovimiento_sum	98%
	23	facttenenciasalud_count_campo_nulo-activo_sum	95%
	24	facttenenciasalud_count_cem-activo_sum	96%
	25	facttenenciasalud_count_cem-retirado_sum	97%
	26	facttenenciasalud_count_medicina_integral-activo_sum	100%
	27	facttenenciasalud_count_medicina_integral-retirado_sum	96%
	28	facttenenciasalud_count_nan-retirado_sum	100%
	29	facttenenciasalud_numvalorcuota_sum	98%
	30	hogar_mas_y_total_homeActivo	96%
	31	hogar_mas_y_total_homeCancelado	97%
	32	otras_polizasActivo	95%
	33	otras_polizasCancelado	0,0%
	34	perdidas_parciales-canceladoCancelado	99%
	35	perdidas_parciales-vigenteActivo	97%
	36	plan_100_dañosCancelado	99%
	37	plan_100_hurtoCancelado	95%
	38	plan_obligatorio_de_saludActivo	95%
	39	plan_obligatorio_de_saludCancelado	97%
	40	rc_medicaActivo	99%
	41	rc_medicaCancelado	97%
	42	rce_en_excesoCancelado	99%
	43	salud_oralActivo	95%
	44	salud_oralCancelado	96%
	45	sieeducacionActivo	95%
	46	sieeducacionCancelado	95%

	47	sieeducacionterminado	0,0%
	48	siefundacionActivo	97%
	49	siefundacionCancelado	98%
	50	siefundacionterminado	0,0%
	51	soatActivo	98%
	52	soatCancelado	0,0%
	53	suecoeducActivo	100%
	54	suecoeducCancelado	0,0%
	55	suecoeducterminado	96%
	56	suecofundActivo	96%
	57	suecofundCancelado	96%
	58	suecofundterminado	100%
	59	tmpproductosbancohistoricotenencia_count_banco_sta_cof04001_otrosc_sum	96%
	60	tmpproductosbancohistoricotenencia_count_banco_sta_cof04100_sum	0,0%
	61	tmpproductosbancohistoricotenencia_count_banco_sta_cof04101_sum	97%
	62	tmpproductosbancohistoricotenencia_count_banco_sta_cof04102_sum	99%
	63	tmpproductosbancohistoricotenencia_count_banco_sta_cof04103_sum	98%
	64	tmpproductosbancohistoricotenencia_count_banco_sta_cof04104_sum	96%
	65	tmpproductosbancohistoricotenencia_count_banco_sta_cof041106_sum	99%
	66	tmpproductosbancohistoricotenencia_count_banco_sta_cof041111_sum	0,0%
	67	tmpproductosbancohistoricotenencia_count_banco_sta_cof0411116_sum	97%
	68	tmpproductosbancohistoricotenencia_count_banco_sta_cof041113_sum	96%
	69	tmpproductosbancohistoricotenencia_count_banco_sta_cof041115_sum	95%
	70	tmpproductosbancohistoricotenencia_count_banco_sta_cof0411460_sum	99%
	71	tmpproductosbancohistoricotenencia_count_banco_sta_cof041401_sum	100%
	72	tmpproductosbancohistoricotenencia_count_banco_sta_cof042110_sum	99%
	73	tmpproductosbancohistoricotenencia_count_banco_sta_cof0421112_sum	97%
	74	tmpproductosbancohistoricotenencia_count_banco_sta_cof0421116_cupo_sum	0,0%
	75	tmpproductosbancohistoricotenencia_count_banco_sta_cof0421116_saldo_sum	97%
	76	tmpproductosbancohistoricotenencia_count_banco_sta_cof042130_sum	96%
	77	tmpproductosbancohistoricotenencia_count_banco_sta_cof042170_sum	96%
	78	tmpproductosbancohistoricotenencia_count_banco_sta_cof042250_sum	98%
	79	tmpproductosbancohistoricotenencia_count_banco_sta_cof042280_sum	99%
	80	tmpproductosbancohistoricotenencia_count_banco_sta_cofcmcm_sum	96%
	81	tmpproductosbancohistoricotenencia_count_none_sum	100%
	82	vida_fsvActivo	99%
	83	vida_fsvCancelado	95%
	84	vida_grupo_prima_unicaActivo	100%
	85	vida_grupo_prima_unicaCancelado	97%

Comportamiento_de_pago	86	factindicadorescuotacerouno_count_1196.0_sum	97%
	87	factindicadorescuotacerouno_count_1197.0_sum	96%
	88	factindicadorescuotacerouno_count_272.0_sum	98%
	89	factindicadorescuotacerouno_count_91.0_sum	100%
interacciones	90	asistencia_juridicaActivo	95%
	91	asistencia_juridicaCancelado	100%
	92	asistencia_juridicadesembolsado	99%
	93	asistencia_juridicaPendiente	98%
	94	asistencia_pensionalActivo	100%
	95	asistencia_pensionalCancelado	97%
	96	asistencia_pensionaldesembolsado	99%
	97	asistencia_pensionalPendiente	99%
	98	auxilio_funerarioActivo	100%
	99	auxilio_funerarioCancelado	97%
	100	auxilio_funerariodesembolsado	99%
	101	auxilio_funerarioPendiente	100%
	102	desempleoActivo	97%
	103	desempleoCancelado	98%
	104	desempleodesembolsado	95%
	105	desempleoPendiente	98%
	106	exequialActivo	96%
	107	exequialCancelado	100%
	108	exequialdesembolsado	98%
	109	factasociatividadsipas_count_campo_nulo-cancelado_sum	98%
	110	factasociatividadsipas_count_campo_nulo-vigente_sum	96%
	111	factasociatividadsipas_numvalorcuota_sum	0,0%
	112	factasociatividadsipas_numvalorproteccioninicialsolicitado_sum	0,0%
	113	factasociatividadsipas_strcodtipomovimiento_sum	95%
	114	factestadotac_numcupoasignadotac_sum	97%
	115	factestadotac_numcupoutilizadotac_sum	98%
	116	factestadotac_numtarjetasutilizadotac_sum	99%
	117	facttenenciarecreacionturismo_count_campo_nulo-activo_sum	96%
	118	facttenenciarecreacionturismo_count_convenios-activo_sum	0,0%
	119	facttenenciarecreacionturismo_count_convenios-cancelado_sum	95%
	120	facttenenciarecreacionturismo_count_eventos-activo_sum	0,0%
	121	facttenenciarecreacionturismo_count_eventos-cancelado_sum	97%
	122	facttenenciarecreacionturismo_count_turismo-activo_sum	100%
	123	facttenenciarecreacionturismo_count_turismo-inactivo_sum	97%
	124	facttenenciarecreacionturismo_numvalorcuota_sum	0,0%
	125	herenciaActivo	98%
	126	herenciaCancelado	98%

	127	herenciadesembolsado	95%
	128	hospitalizacionActivo	99%
	129	hospitalizacionCancelado	100%
	130	hospitalizaciondesembolsado	97%
	131	incrementos_plan_basicoActivo	96%
	132	incrementos_plan_basicoCancelado	95%
	133	incrementos_plan_basicodesembolsado	98%
	134	incrementos_plan_basicoPendiente	97%
	135	plan_educativoActivo	98%
	136	plan_educativoCancelado	96%
	137	prima_niveladaActivo	99%
	138	prima_niveladaCancelado	95%
	139	recuperacionActivo	96%
	140	recuperacionCancelado	95%
	141	recuperaciondesembolsado	96%
	142	segunda_opinion_medicaActivo	96%
	143	segunda_opinion_medicaCancelado	97%
	144	segunda_opinion_medicadesembolsado	98%
	145	segunda_opinion_medicaPendiente	97%
	146	solvenciaActivo	99%
	147	solvenciaCancelado	97%
	148	solvenciadesembolsado	97%
	149	tranquilidadActivo	96%
	150	tranquilidadCancelado	96%
	151	tranquilidaddesembolsado	96%
	152	vida_clasicaActivo	99%
	153	vida_clasicaCancelado	98%
	154	vida_clasicadesembolsado	99%
	155	vida_clasicaPendiente	95%
	156	vidaActivo	100%
	157	vidaCancelado	99%
	158	vidadesembolsado	97%
Servicio	159	aprobacionesnegacionesActivo	99%
	160	aprobacionesnegacionesCancelado	95%
	161	aprobacionesnegacionesdesembolsado	99%
	162	aprobacionesnegacionesPendiente	0,0%
	163	factaprobacionesnegaciones_numvalormonto_sum	99%
Demográfica	164	ingresos_num_valor_ingresos	95%

	165	numcantidadhijos	100%
	166	numcantidadpersonascargoadultas	98%
	167	numcantidadpersonascargomenores18	98%
	168	numcodtipovinculacion	95%
	169	numcodunicocorte	97%
	170	numedad	0,0%
	171	segmentos_num_cantidad_productos_sum	0,0%
	172	Segmento_Ciclo_de_Vida	0,0%
	173	Genero	98%
	174	Tipo_Persona	96%
	175	Corte_Factura	0,0%
	176	Actividad_Laboral	0,0%
	177	Área_Conocimiento	0,0%
	178	Estado_Civil	0,0%
	179	Estrato	0,0%
	180	Nivel_Academico	0,0%
	181	Tipo_Vivienda	0,0%
Estado	182	Label	0,0%

Fuente: Elaboración propia.

En la tabla 5 se muestra que solo 28 variables tienen información completa, y las otras variables más del 94% son vacías. Dado este resultado, se trabaja con las 28 variables descritas en la tabla 7.

Tabla 7. Variables seleccionadas

Id	Variables	Tipo de variable	Descripción
1	cooperativa_valor_cuota	Continua	Monto acumulado que pagó el asociado, por estar activo en la cooperativa
2	valor_proteccion	Continua	Hace referencia al monto valor del seguro en dinero, por el cual el asociado está o estuvo protegido
3	valor_cuota_credito_solidario	Continua	Monto acumulado que pagó por las cuotas del crédito solidario
4	valor_cuota_fundacion_educacion	Continua	Monto acumulado que pagó por las cuotas de los servicios de fundación en educación
5	usos_turismo	Discreta	Cantidad de usos que realizó el asociado en el sector de turismo
6	usos_eventos	Discreta	Cantidad de usos que realizó el asociado en el sector de eventos

7	valor_cuota_turismo	Continua	Monto acumulado que pagó por las cuotas de los servicios de la agencia de turismo
8	tarjeta_debito	Discreta	Cantidad de veces que ha tenido la tarjeta debito del banco de la cooperativa
9	cuenta_deposito	Discreta	Cantidad de veces que ha tenido la cuenta deposito del banco de la cooperativa
10	tarjetavisasaldo_cupo	Discreta	Cantidad de veces que ha tenido la tarjeta visa con un cupo asignado por el banco
11	aprobacionesnegacionesPendiente	Discreta	Cantidad de aprobaciones y negaciones de los productos de la cooperativa que están pendientes
12	bancaseguros	Discreta	Cantidad de veces que aseguró los productos financieros de la cooperativa
13	otras_polizas	Discreta	Cantidad de veces que ha tenido una póliza en el sector asegurador de la cooperativa
14	educacion_terminado	Discreta	Cantidad de veces que tomó un producto en el sector educación y culminó el uso
15	fundacion	Discreta	Cantidad de veces que tomó un producto en el sector de fundación
16	soat	Discreta	Cantidad de veces que tomó un producto de Soat
17	educacion_cancelado	Discreta	Cantidad de veces que tomó un producto de educación y lo canceló
18	numedad	Discreta	Edad del asociado
19	cantidad_productos	Discreta	Cantidad de productos que ha tomado con la cooperativa
20	ciclo_vida	Ordinal	Segmentación del ciclo de vida de un asociado: - Joven asociado: Asociados con edades entre los 18 y 25 años, hombres o mujeres sin hijos - En formación: Mayores a 25 años hasta los 45 años, solo hombres sin hijos - Mujer independiente: Mayores a 25 años hasta los 45 años, solo mujeres sin hijos - Consolidación: Mayores a 25 años hasta los 60 años, hombre o mujeres con hijos - Transición: Mayores a 45 hasta los 60 años, hombre o mujeres sin hijos - Maduro: Mayores a 60 años con hijos o sin hijos, hombres o mujeres
21	unicocorte	Ordinal	unicocorte es una categoría del tiempo en que le llega la factura al asociado -05_del_mes -25_del_mes -30_del_mes -20_del_mes -15_del_mes -10_del_mes
22	actividad_laboral	Nominal	actividad_laboral es una categoría que define el estado laboral de un asociado:

			-asalariado -independiente -pensionado_-_jubilado -otro_tipo_de_actividad -socio_sociedad -estudiante -ama_de_casa -rentista_capital
23	area_conocimientos	Nominal	Categoría del área de conocimiento a la que pertenece el asociado: -economía, _ administración, contaduria_y_afines -ingeniería, _arquitectura, _urbanismo_y_afines -ciencias_de_la_salud -ciencias_sociales, _derecho_y_ciencia_politica -ciencias_de_la_educacion -agronomia, _veterinaria_y_afines -matematicas_y_ciencias_naturales -bellas_artes -humanidades_y_ciencias_religiosas
24	estado_civil	Nominal	Categoría del estado civil del asociado: -soltero -casado -union_libre -separado -viudo -divorciado
25	estrato	Ordinal	Categoría del estrato socio económico del asociado: -bajo-bajo -bajo -medio-bajo -medio -medio-alto -alto
26	nivel_academico	Ordinal	Categoría del nivel académico del asociado: -primaria -tecnólogo -técnico -profesional -especialización
27	tipo_vivienda	Nominal	Categoría del tipo de vivienda del asociado: -familiar -propia -alquiler
28	label	Binaria	Asigna 1 al asociado retirado y 0 al asociado activo

Fuente: Elaboración propia.

Dado los resultados anteriores, se procede a realizar el análisis exploratorio de datos.

Este análisis se realizará para entender algunos patrones de separabilidad de las clases que se quiere predecir.

Ahora bien, con respecto a la definición de la variable a predecir, en la tabla 3 se define la variable label del asociado, que a continuación se codifica en:

$$Label = \begin{cases} 1, si\ el\ asociado\ se\ retira\ de\ la\ cooperativa \\ 0, el\ asociado\ esta\ activo \end{cases}$$

Esta variable contiene 226.446 (55%), asociados categorizados con label 1 y 185.154 (45%), con label en 0, lo que demuestra un balanceo de clases a la hora del entrenamiento de los modelos.

En primera instancia, se trabajó el análisis de las distribuciones de cada una de las cinco variables de los montos de cuotas que paga el asociado por los productos o servicios, las cuales hace evidente la necesidad de una transformación (logaritmo), debido a la presencia de colas muy alargadas que puedan influenciar la adecuada descripción de los asociados. En las figuras 8 a 12 se muestran los histogramas de las variables originales y transformadas.

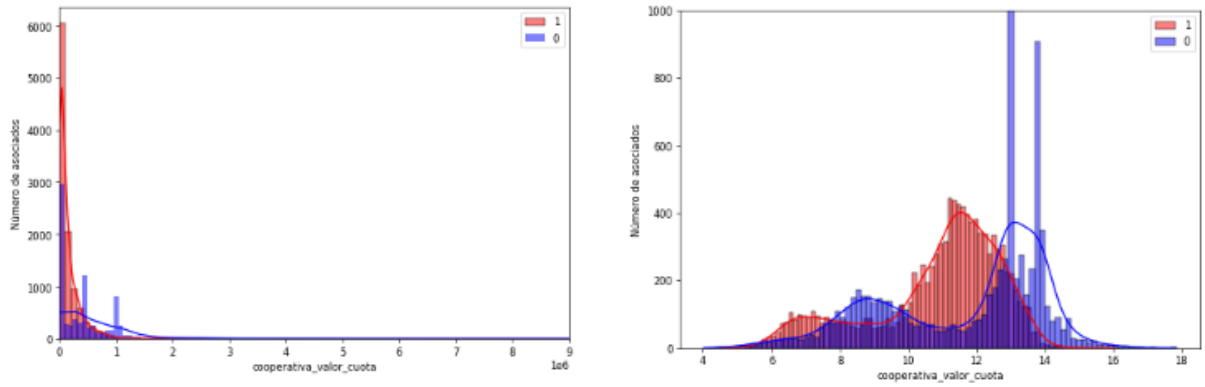


Figura 8. Histograma del valor de la cuota de la cooperativa y logaritmo del valor de la cuota de la cooperativa.

En la figura 8, se observa como los asociados activos (label=0), tienen una distribución bimodal de los datos, pero teniendo una mayor frecuencia de datos por encima de los asociados retirados (label=0).

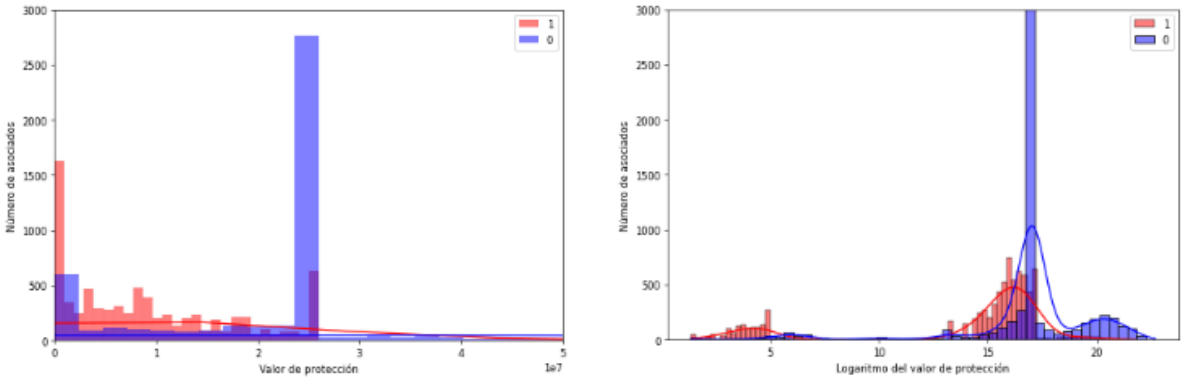


Figura 9. Histograma del valor de protección de la cooperativa y logaritmo del valor de protección.

En la figura 9, se muestra el comparativo de asociados activos(label=0) y asociados retirados(label=1), para la variable valor de protección en la cooperativa, con un valor aproximado de 24 millones más repetido en los activos y valores menores con más frecuencia en los retirados.

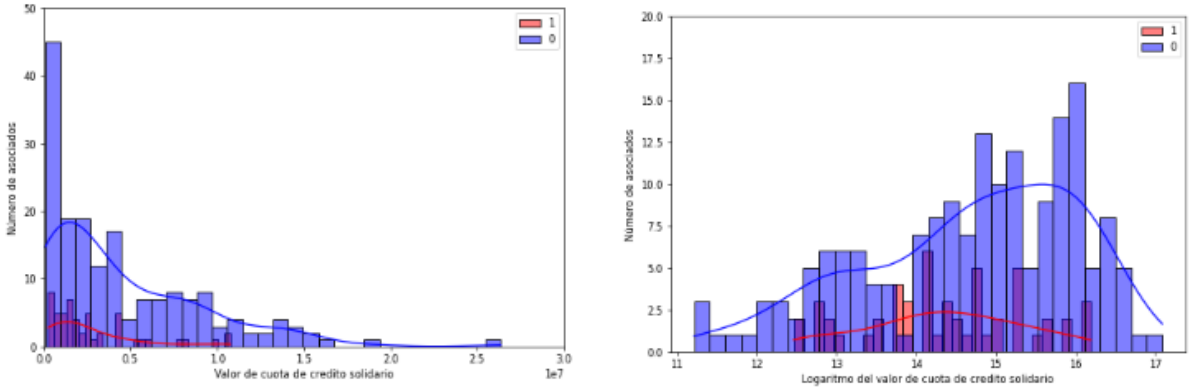


Figura 10. Histograma del valor cuota crédito solidario de la cooperativa y el logaritmo del valor cuota crédito solidario.

En la figura 10, el histograma del valor de la cuota muestra como los asociados tienen una distribución más simétrica a valores más altos, comparados con los asociados retirados mientras estuvieron activos en la cooperativa.

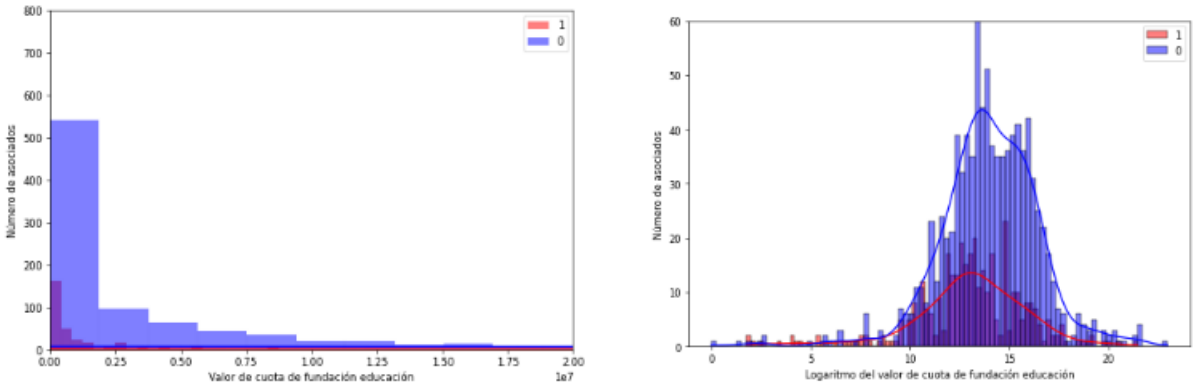


Figura 11. Histograma del valor de la cuota de fundación educación de la cooperativa y logaritmo del valor de la cuota de fundación educación.

En la figura 11, el comportamiento del valor de la cuota para productos de la fundación en educación se observa con distribuciones similares, pero teniendo mayor frecuencia en los asociados que están activos comparado con los retirados en su momento que estaban activos en la organización.

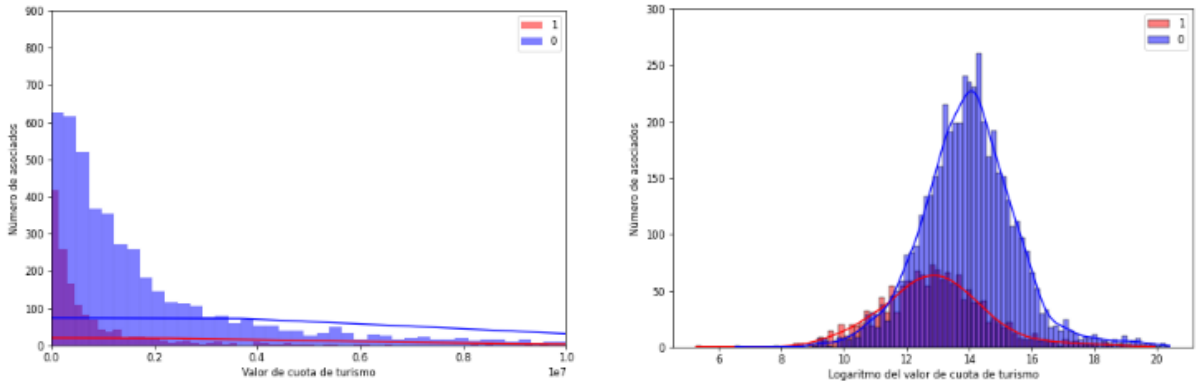


Figura 12. Histograma del valor de cuota de turismo de la cooperativa y logaritmo del valor de cuota de turismo.

El histograma de la figura 12, muestra el comparativo del valor de la cuota de turismo de los asociados activos(label=0) y retirados(label=1).

En las siguientes variables, se trabajará gráficos de puntos y boxplot, para analizar los comportamientos de los asociados retirados y activos.

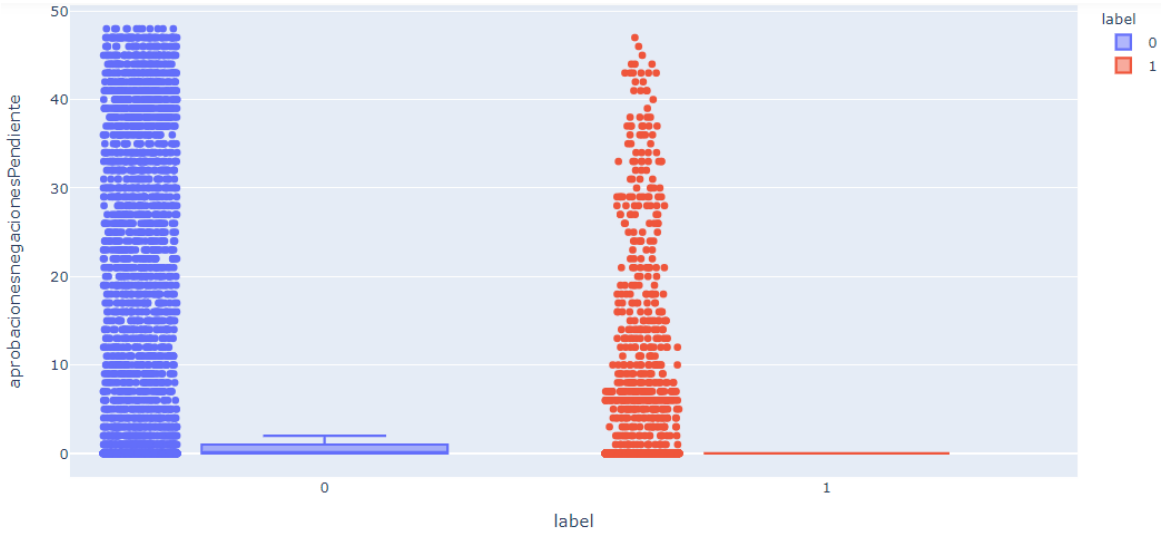


Figura 13. Box plot y puntos de las aprobaciones y negaciones pendiente de la cooperativa.

En la figura 13, se muestra el grafico de puntos y boxplot, en las poblaciones de asociados activos(label=0) y retirados(label=0), para la variable de aprobaciones y negaciones pendientes, con una mayor concentración en la población de activos, en valores mayores a 40.

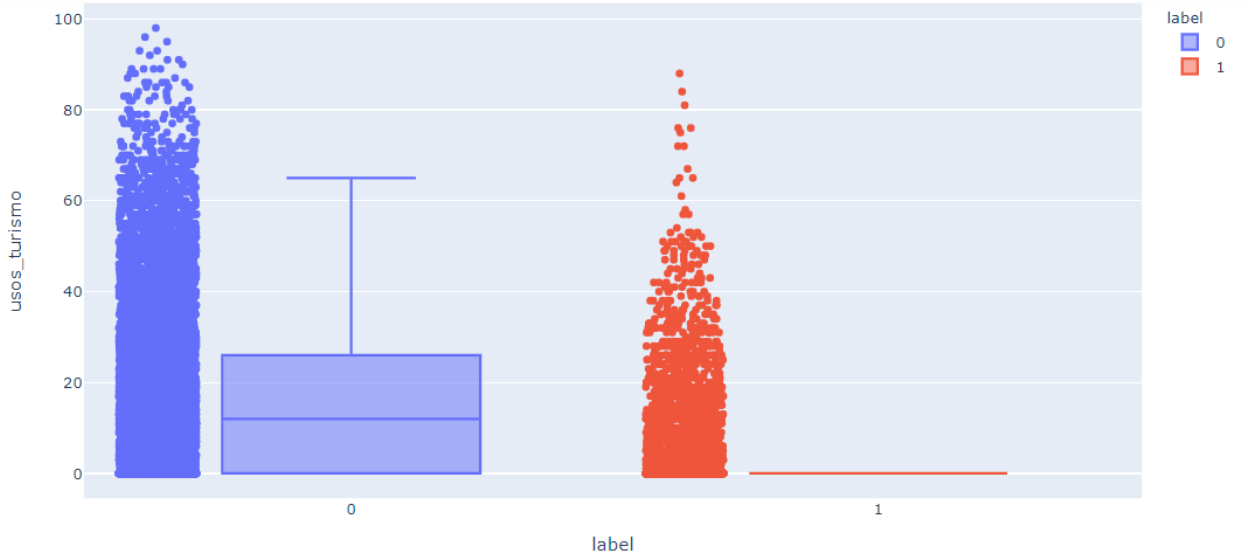


Figura 14. Box plot y puntos del uso de turismo de la cooperativa.

En la figura 14, se observa como los asociados activos(label=0), presentan una mayor concentración en el uso de turismo, en valores mayores a 40 comparado con los retirados. El boxplot tiene una mediana de 12 en asociados activos, comparado con mediana en 0 para asociados retirados, representado en mayor valor del turismo en los asociados activos.

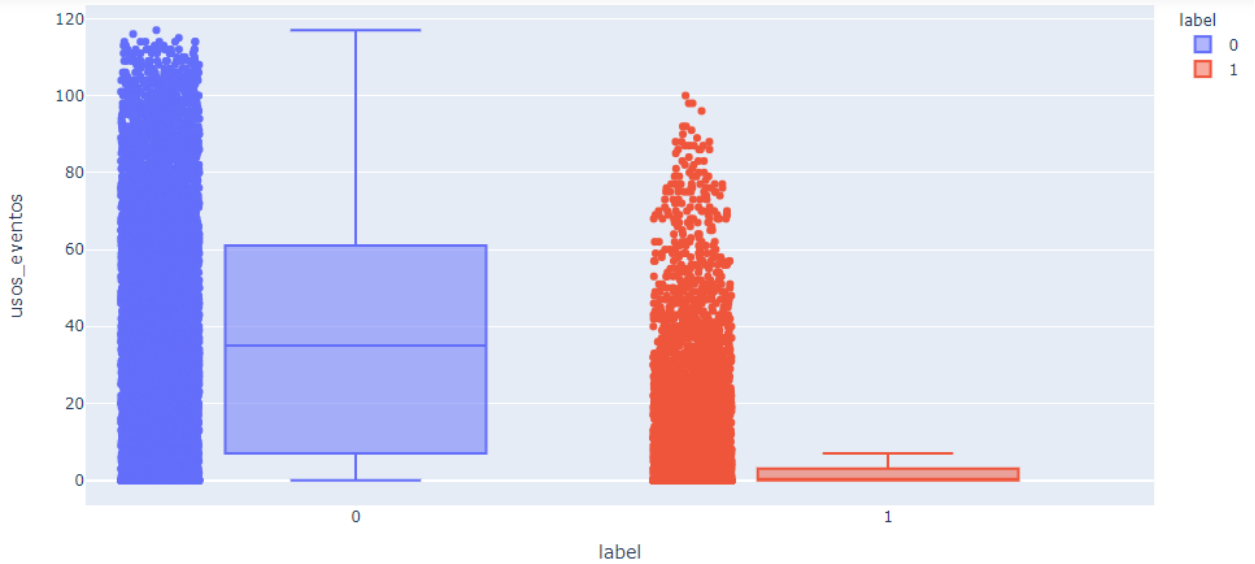


Figura 15. Box plot y puntos del uso de eventos de la cooperativa.

En la figura 15, se observa una mayor concentración en uso de eventos para los asociados activos(label=0), comparado con los asociados retirados(label=1). El comportamiento demuestra que los asociados logran tener una conexión con los eventos mientras están activos.

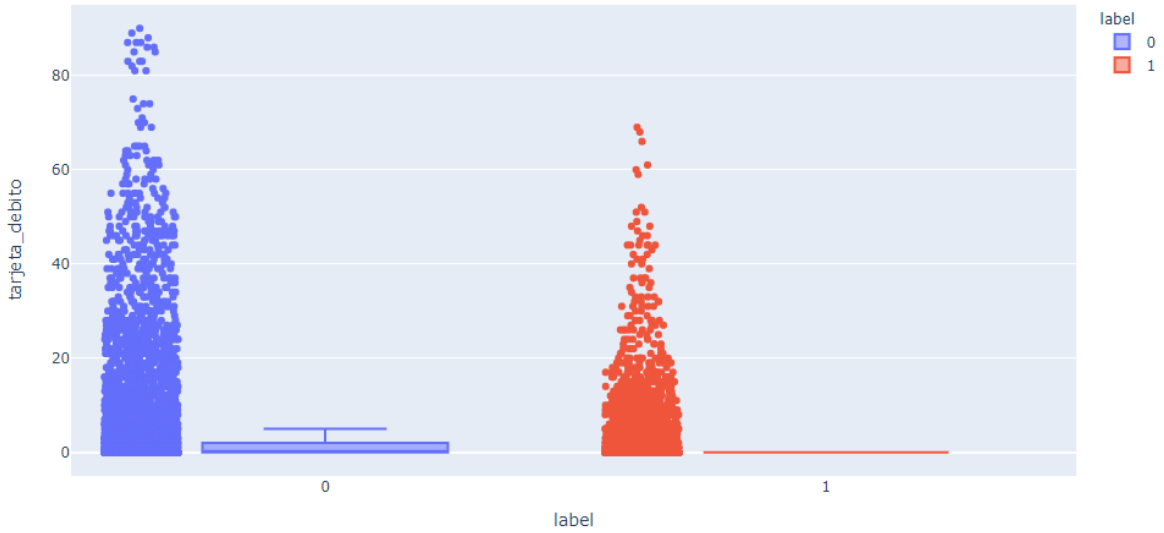


Figura 16. Box plot y puntos tenencia de la tarjeta débito.

En la figura 16, se analiza la tenencia acumulada de la tarjeta débito, a partir del grafico de puntos y box plot, se tiene comportamientos similares entre ambas poblaciones objeto de estudio.

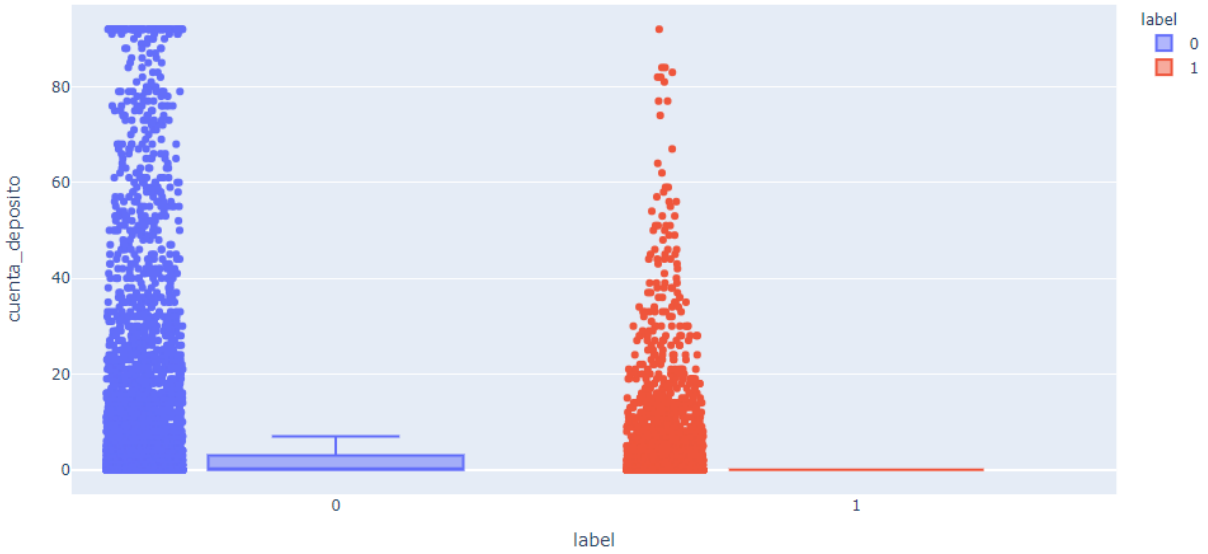


Figura 17. Box plot y puntos tenencia de la cuenta deposito.

La tenencia acumulada de la cuenta deposito, se muestra en la figura 17 mayor cantidad de puntos en la población activa de asociados(label=0), en valores mayores a 20, pero con concentraciones muy similares en las poblaciones.

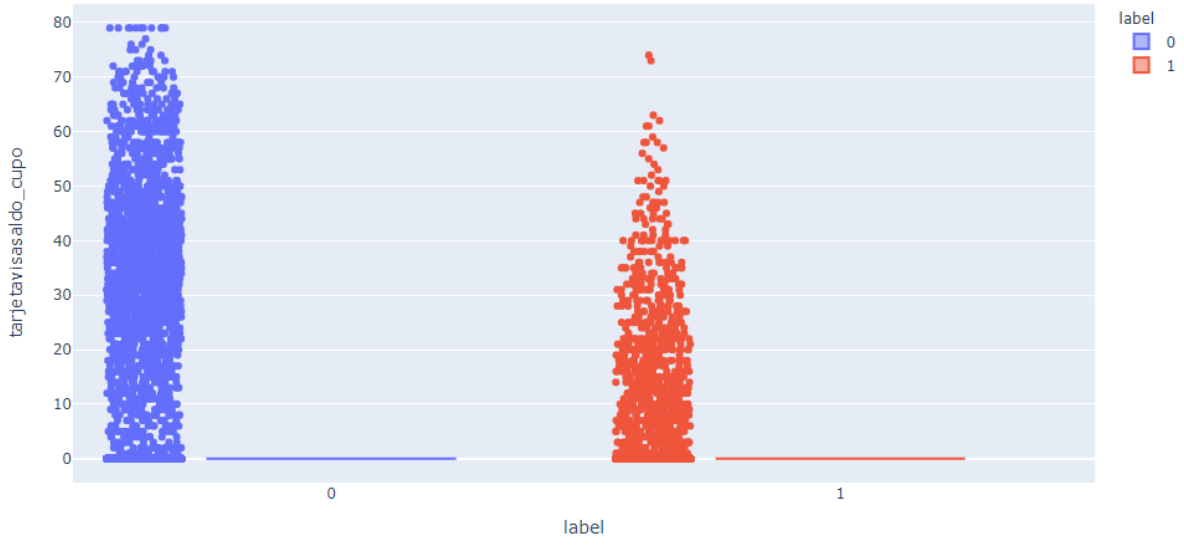


Figura 18. Box plot y puntos tenencia de la tarjeta visa con cupo.

En la figura 18, la población activa de asociados(label=0), para la tenencia de la tarjeta visa con cupo, se evidencia con mayor concentración en la nube de puntos en valores mayores a 30, comparado en la población de asociados retirados(label=1) mientras

estuvo activo en la organización.

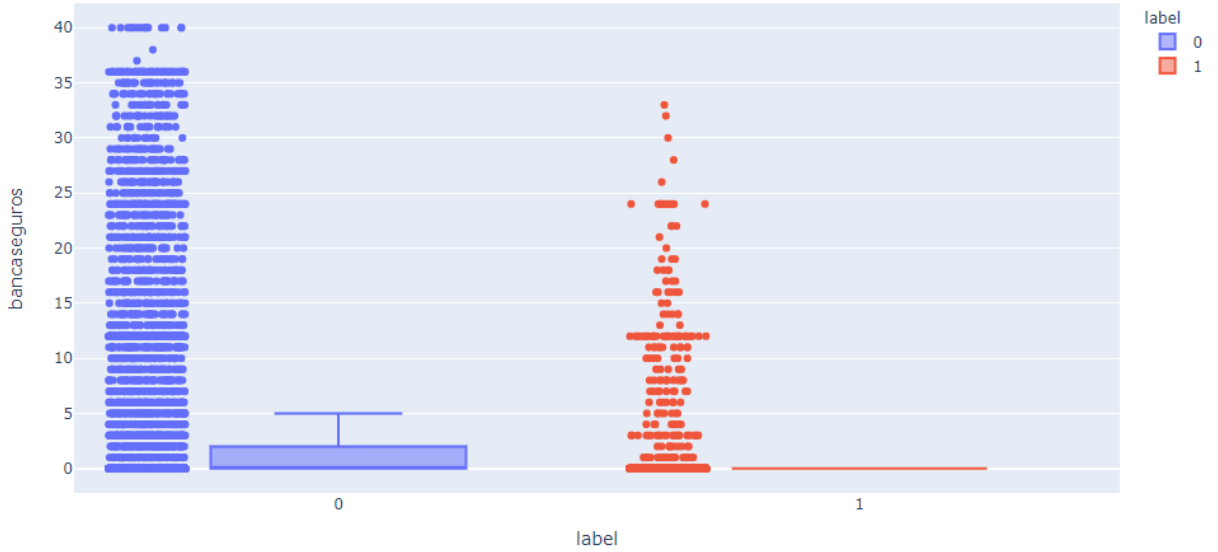


Figura 19. Box plot y puntos de producto de banca seguros.

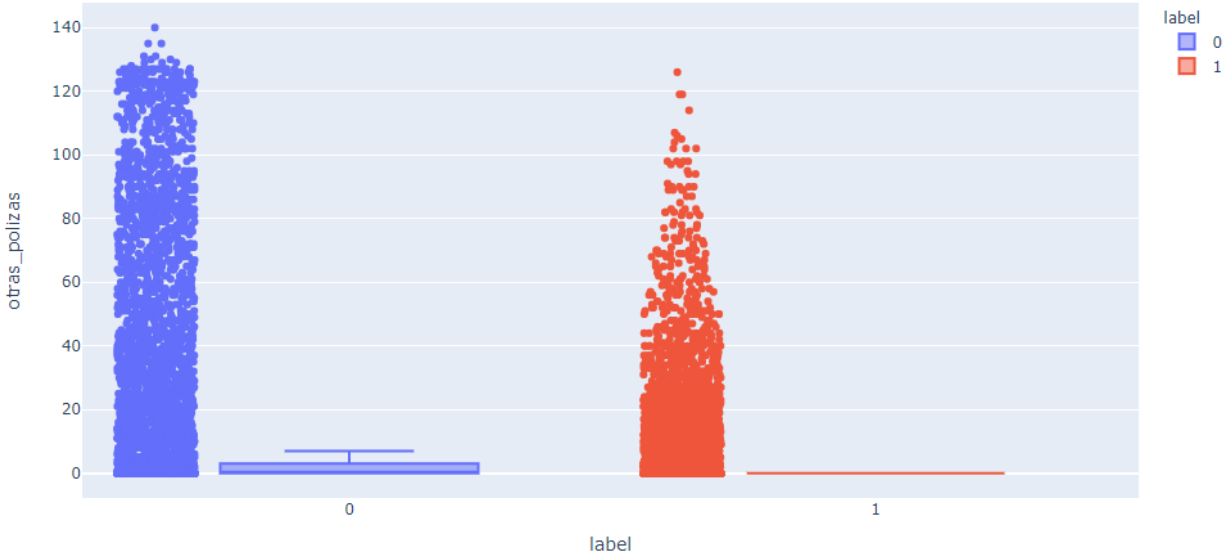


Figura 20. Box plot y puntos de producto de otras pólizas.

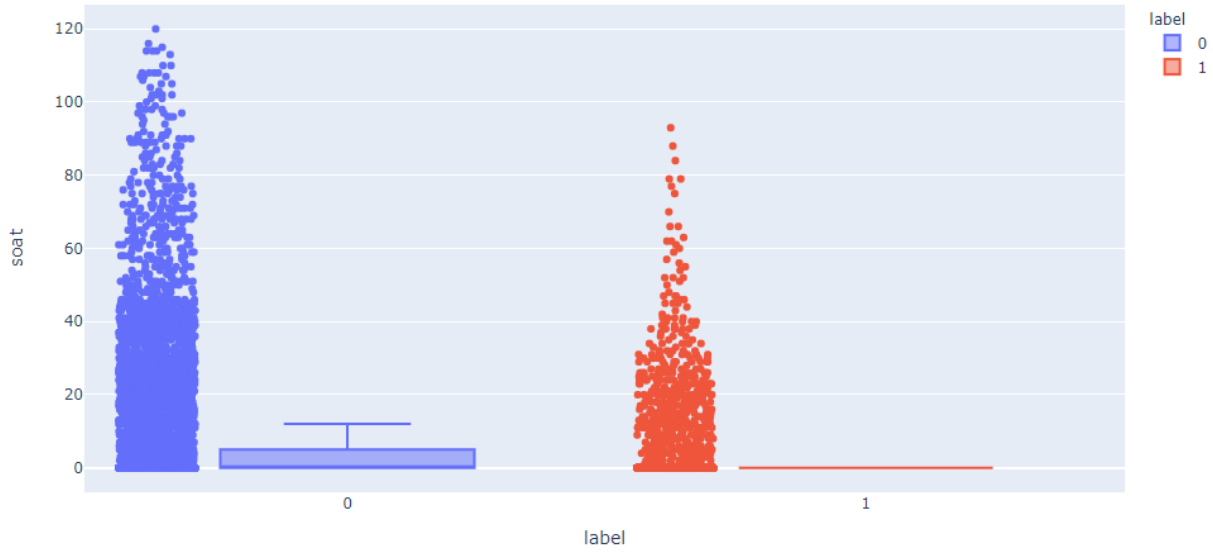


Figura 21. Box plot y puntos de producto de soat.

En el sector asegurador, se muestran las figuras 19,20 y 21. La tenencia del producto como banca seguros, otras pólizas y soat, se establece distribuciones de puntos concentrados en su mayoría en valores pequeños, pero se diferencia en los valores grandes, al comparar los asociados activos(label=0) con retirados(label=1).

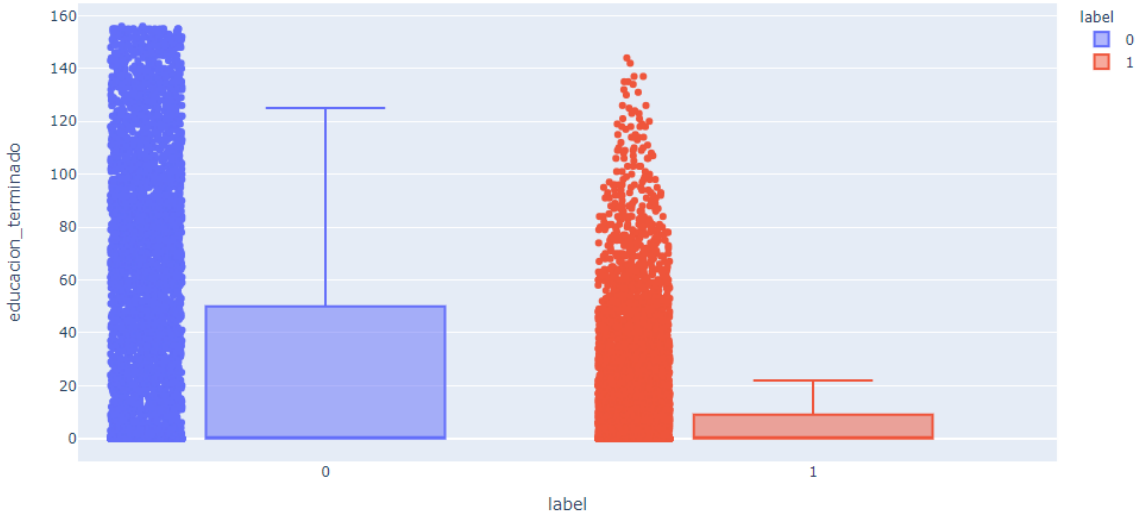


Figura 22. Box plot y puntos créditos de educación terminados.

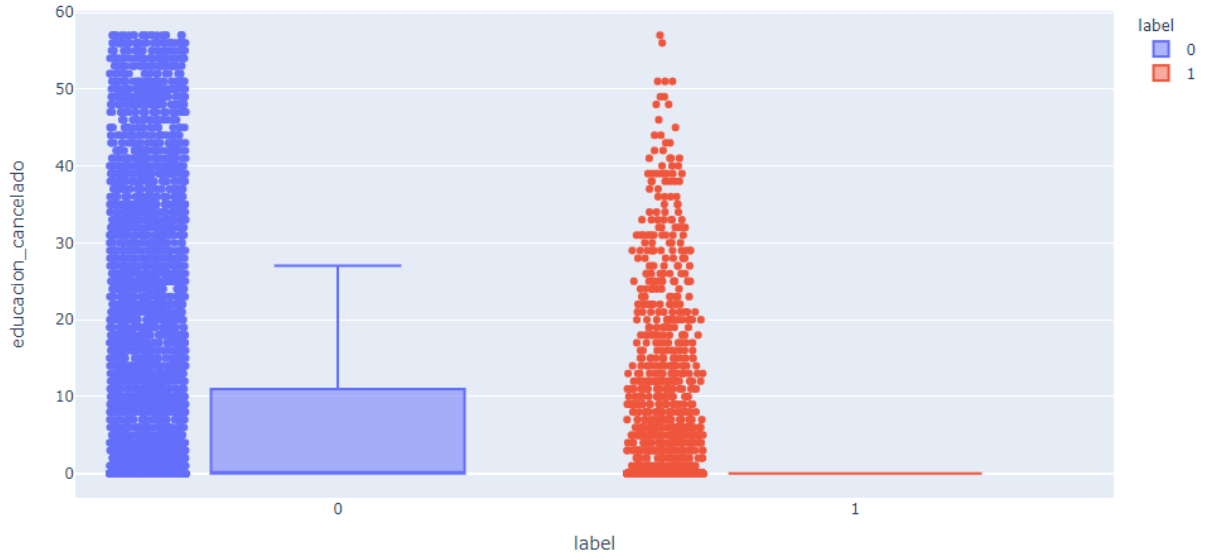


Figura 23. Box plot y puntos créditos de educación cancelado.

En las figuras 22 y 23, se evidencia que los usos de los productos del sector educativo tienen mayor participación en la población de asociados activa (label=0). Los boxplot con una distribución asimétrica a la derecha, evidenciando valores más grandes y concentrados si es comparado con los box plot de la población retirada(label=1).

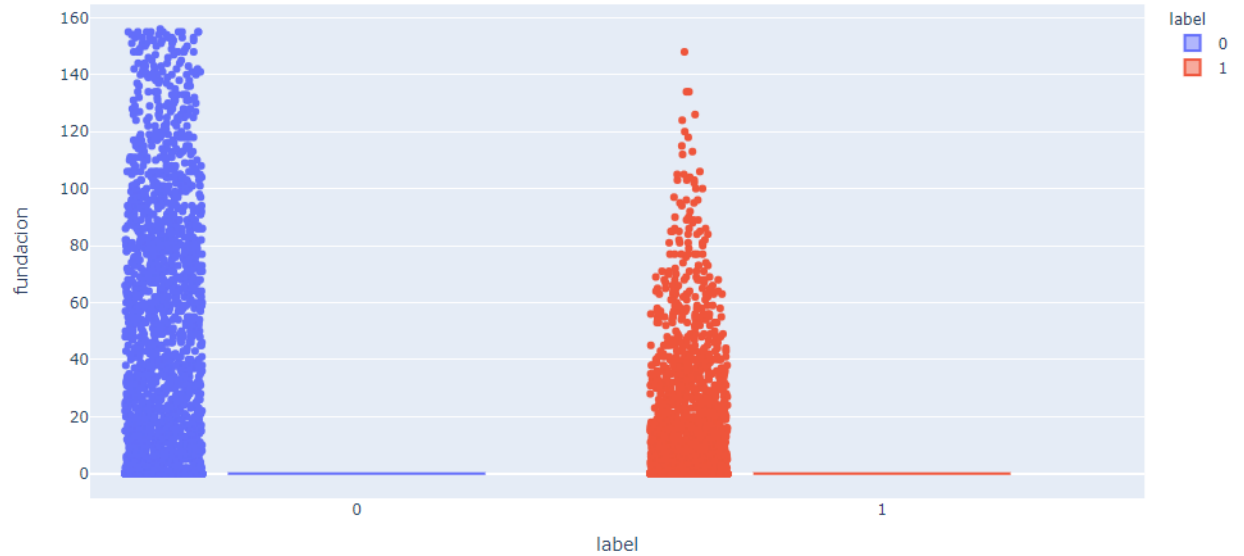


Figura 24. Box plot y puntos participación en la fundación.

En la figura 24, la participación en el sector fundación, aunque se evidencia mayor participación en por la densidad de puntos en la población activa(label=0), de asociados. Los boxplot son muy similares y solo se observa diferencias en las concentraciones de

puntos en valores mayores a 40.

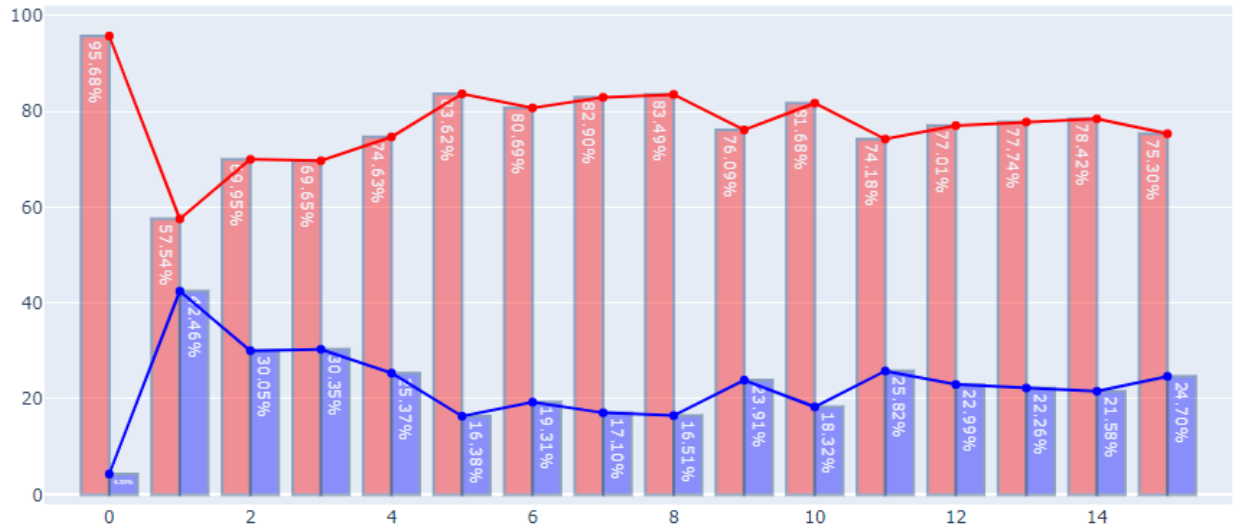


Figura 25. Comparativo en porcentaje de participación por cantidad de productos.

En la figura 25, se observa la cantidad de productos en el eje x y la participación por la población activa(label=0), comparado con la retirada(label=1). Se evidencia que no tener productos la participación de los asociados retirados es del 95%, comparado con los asociados activos que solo es del 5%.

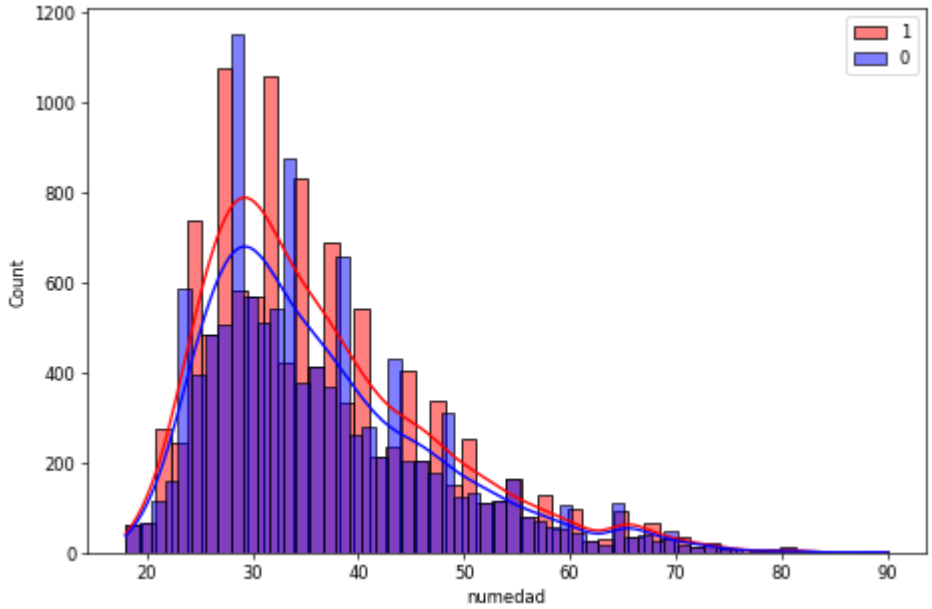


Figura 26. Histograma de la edad de los asociados.

En el histograma, de la figura 26 se muestra la distribución de la edad de asociados activos(label=0) y retirados(label=1), en este caso es muy similar teniendo casi distribuciones similares.

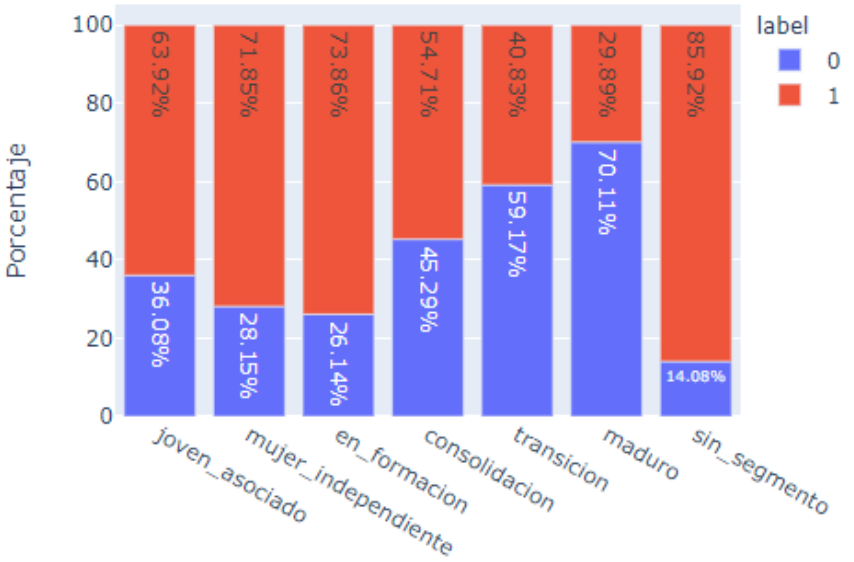


Figura 27. Comparativo en porcentaje de participación por segmento ciclo de vida.

En la segmentación del ciclo vida de los asociados que realiza la cooperativa, se muestra en la figura 27, que la participación de la población activa(label=0), en segmentos como transición y maduro es mayor en un 50% comparado con los retirados(label=1). En este caso, categorías como el joven, en formación y mujer independiente sus participaciones en población retirada(label=1), esta entre el 63% y 73%, evidenciando que la población con edad menores a 45 años, no logran conectarse con la cooperativa si es comparado con otros segmentos de mayor edad.

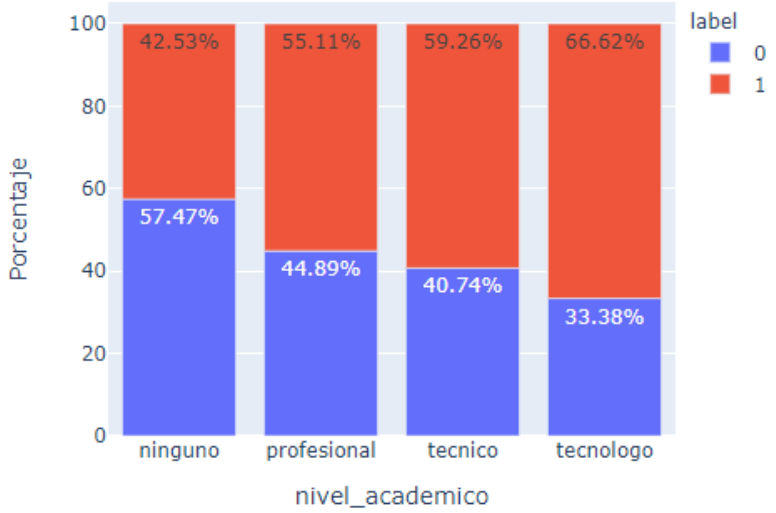


Figura 28. Comparativo en porcentaje de participación por nivel académico.

En la figura 28, se muestra la participación por la categoría de nivel académico, con un 55.11% de profesionales están retirados(label=1), comparado con un 44.89% de activos(label=0), se evidencia que los tecnólogos tienen 66.63% de población retirada, mientras que los activo son 33.38%.

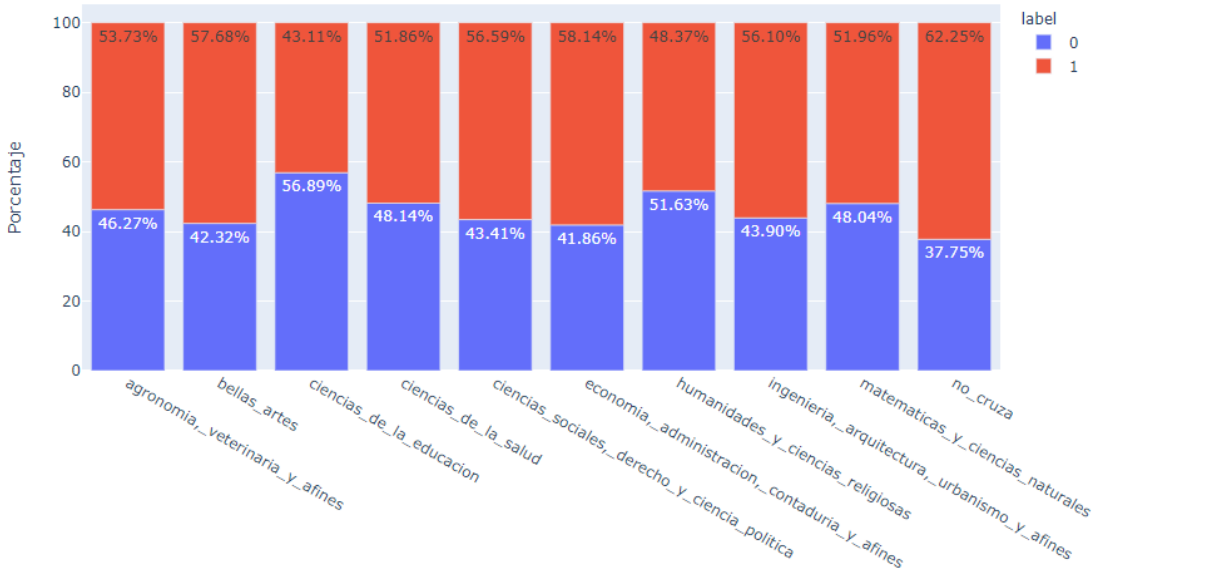


Figura 29. Comparativo en porcentaje de participación por área de conocimiento.

En la categoría de área de conocimiento, en la figura 29 las participaciones en las poblaciones son muy uniforme. Solo en área como la educación y humanidades superan el 50% de población activa (label=0).

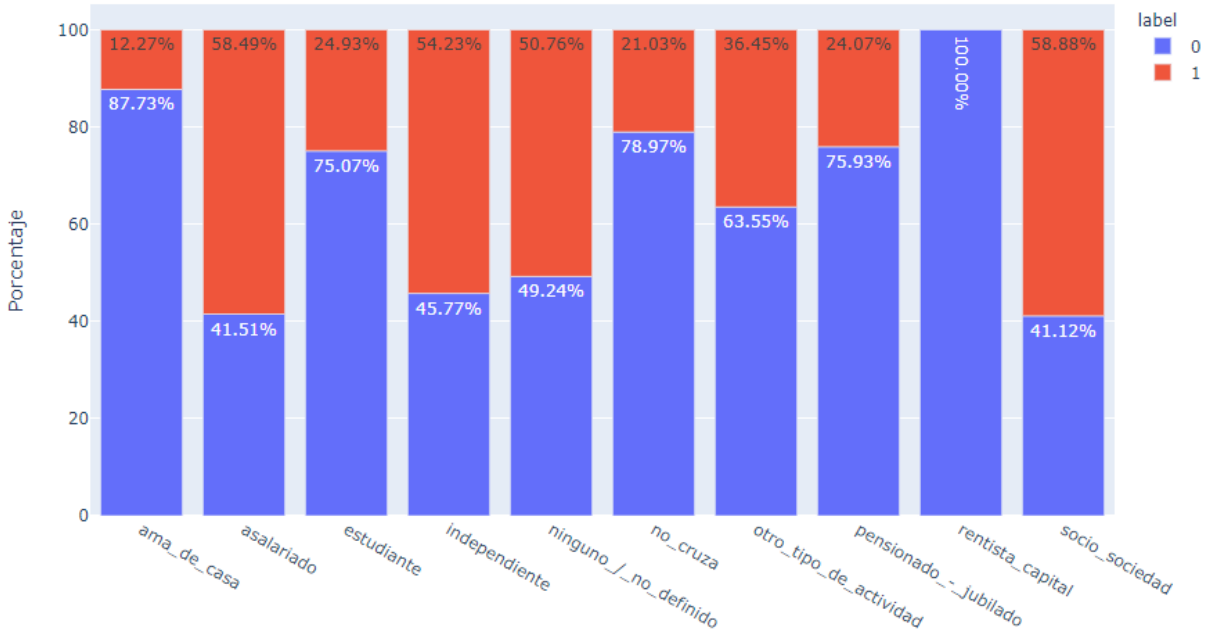


Figura 30. Comparativo en porcentaje de participación por actividad laboral.

En la actividad laboral, figura 30. Se muestra la distribución de participaciones de las categorías, para el caso de ama de casa, estudiante y pensionado y jubilado, el porcentaje supera el 70% de población activa(label=0).

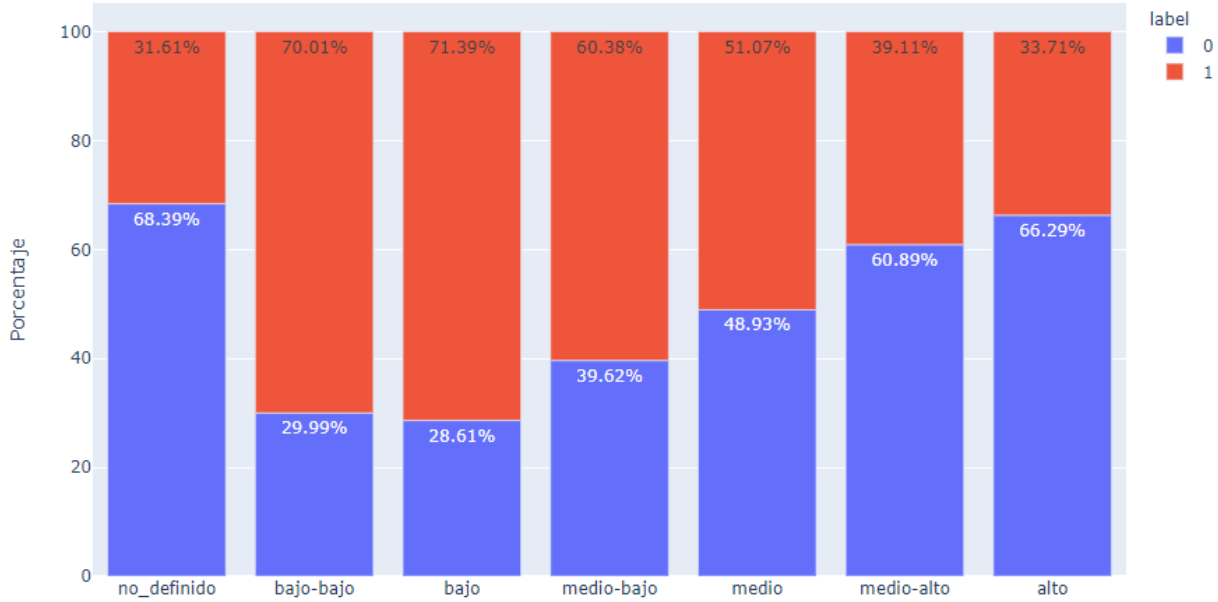


Figura 31. Comparativo en porcentaje de participación por estrato socio económico.

En la figura 31, el estrato socioeconómico, tiene un patrón de participación creciente a medida que el estrato es mayor. En estrato como bajo-bajo, el 29% están activos(label=0), mientras que en estrato altos es del 66%.

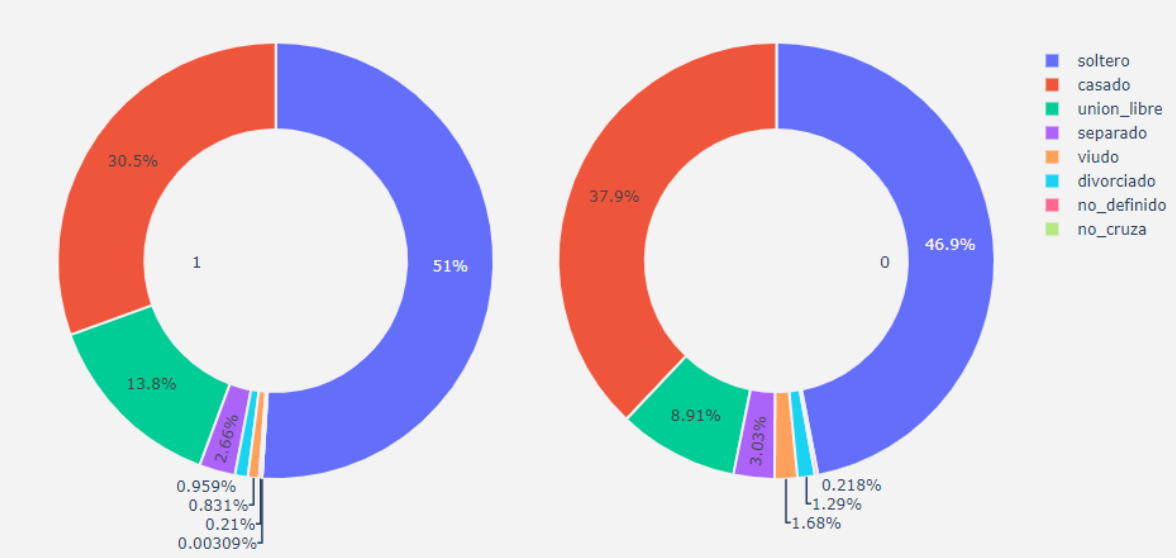


Figura 32. Comparativo en porcentaje de participación por estado civil.

En la figura 32, se muestra la participación por categoría del estado civil con porcentajes similares en las poblaciones de retiros(label=1) y activos(label=0).

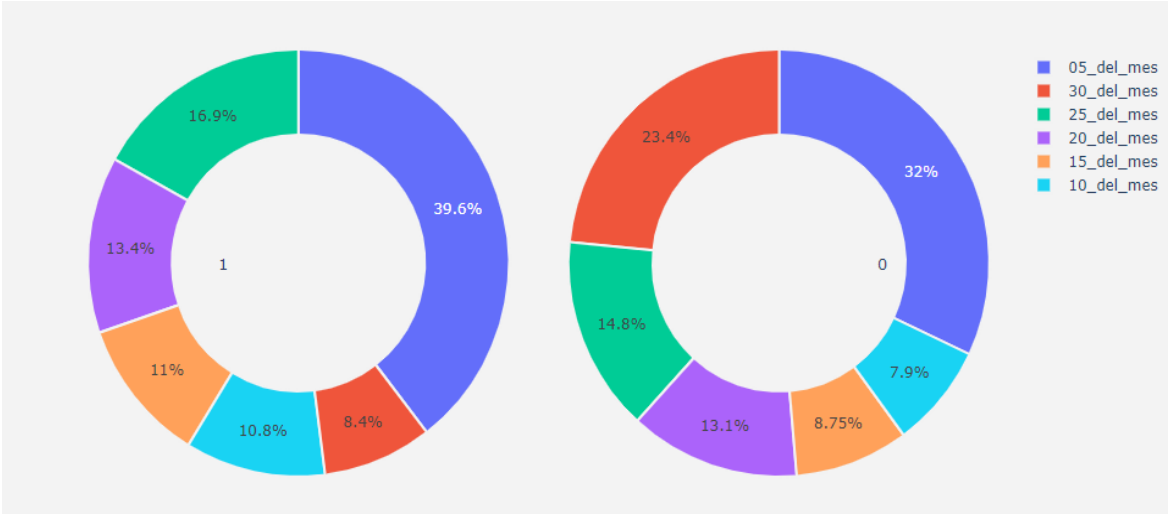


Figura 33. Comparativo en porcentaje de participación por corte de la factura.

Figura 33, se muestra el corte de la factura en las fechas establecidas por el proceso de facturación de la cooperativa. En las participaciones se destaca que los asociados de corte 30, el 16.9% están retirados(label=1), mientras que el 23.4% son activos(label=0).

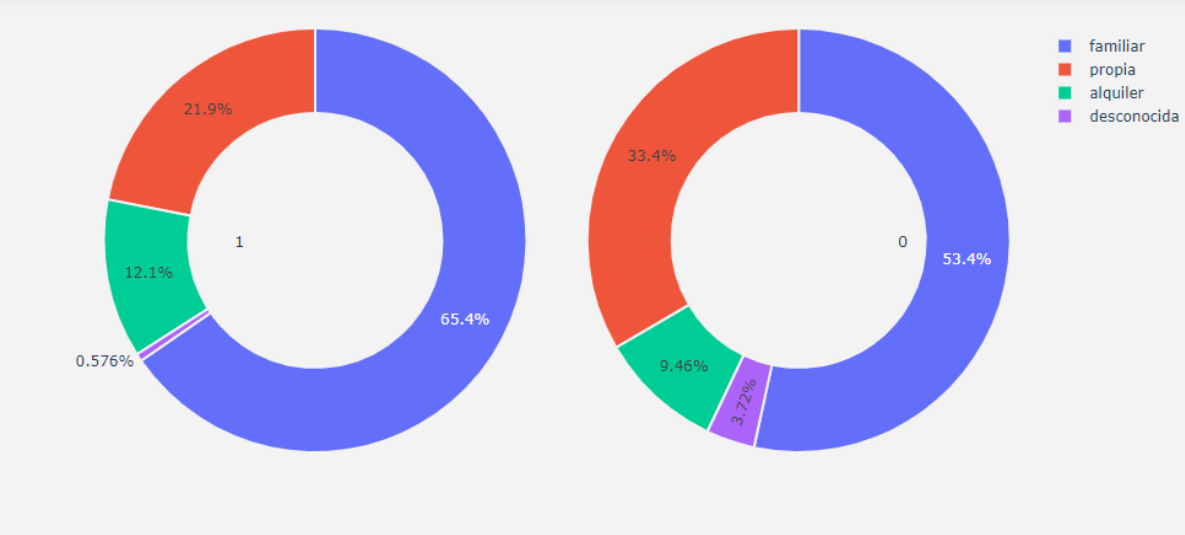


Figura 34. Comparativo en porcentaje de participación por tipo de vivienda.

En la figura 34, se muestra por población activa (label = 0) y retirada(label=1), los porcentajes por categoría, con un resultado destacado del 33.4% con vivienda propia en la población activa, mientras un 21.9% en la población retirada.

De acuerdo con la información presentada anteriormente, se evidenció algunos patrones de separabilidad en el label que caracteriza el estado de retiro o activo en la cooperativa. Las variables como el valor de la cuota, valor de protección y valor de cuota en crédito solidario, presentan comportamientos bimodales en su distribución, comparado con un comportamiento simétrico en las distribuciones de los datos en productos como fundación, recreación y turismo, evidenciando que existen variaciones en los comportamientos de cada variable con respecto al estado del asociado a la cooperativa. Por otro lado, la variable cantidad de productos presenta que los asociados sin producto tienen mayor participación en los retiros comparado con los activos, lo cual incentiva a proponer estrategias de colocación de productos para mitigar los retiros en la cooperativa.

En resumen, para las variables categóricas, se presentan participaciones que ayudan a caracterizar a los asociados, como se describe a continuación:

- En el segmento de ciclo de vida para las categorías jóvenes asociado, en formación y mujer independiente se tiene una participación mayor al 60% para el estado de retiro (label=1). Esto demuestra que para edades entre los 18 a 45 años, sea hombre o mujer y sin hijos, los asociados tienen un mayor perfil riesgo en retirarse, comparado con las siguientes categorías como consolidación, transición y maduro, donde las proporciones en el estado activo(label=0), empieza a cambiar los porcentajes de participación perfilando los asociados con edades entre 25 a 45

años, sea hombre o mujer con hijos.

- En la variable estrato socio económico, se evidencia que un mayor estrato genera menor proporción de asociados con estado de retiro.
- Corte de la factura, es una variable que identifica en qué momento se genera la factura de cada asociado y se evidencia que para el corte 30, existe un 23% de asociados activos comparado con un 8% de retirados.

Los análisis de variables univariados anteriormente demuestra la importancia de entender algunos patrones en la población objeto de estudio, logrando evidenciar cuales variables pueden tener una mayor propensión al riesgo de abandono en la cooperativa.

9.2 SELECCIÓN DE VARIABLES A PARTIR DE ANÁLISIS DE CORRRELACIÓN PARA LA CONTRUCCIÓN DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO

Con el objetivo de analizar patrones de correlaciones bivariados, en la figura 35 se muestra la correlación de Pearson entre variables numéricas. Un resultado a resaltar es la correlación positiva en variables como uso eventos y cantidad productos, seguido por usos turismo y cantidad productos, lo que evidencia la importancia en generar un mayor uso en sectores como turismo y recreación, para lograr que la cantidad de productos aumente y logre mitigar el riesgo de abandono en la cooperativa.

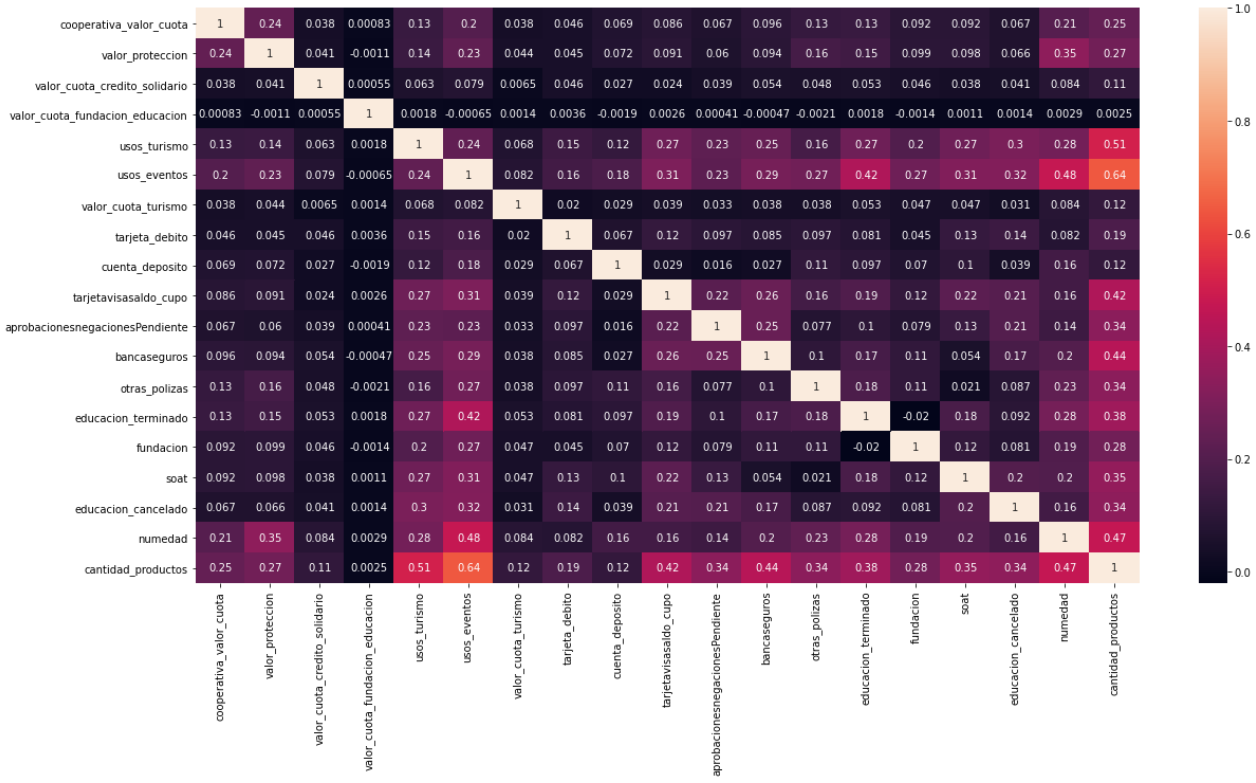


Figura 35. Correlación de variables numéricas.

De acuerdo con la información presentada anteriormente, las 28 variables que tienen información completa y además que según los gráficos realizados presentan separabilidad en los histogramas, cuando se comparan los activos y retirados, en las variables cuantitativas son relevantes para el análisis de la predicción del riesgo de abandono.

9.3 PREPARACIÓN DE LOS DATOS DE FORMA ADECUADA PARA EL ENTRENAMIENTO DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO

9.3.1. CODIFICACIÓN DE VARIABLES CATEGÓRICAS Y NORMALIZACIÓN DE VARIABLES CUANTITATIVAS.

De acuerdo con [18], un paso previo para la construcción de los modelos es el preprocesamiento de la información, con el conjunto de datos descritos anteriormente.

Las tareas se realizaron en las siguientes actividades y se describen a continuación:

- De acuerdo con [17], la magnitud de las variables puede influir en el desempeño de los modelos basados en aprendizaje automático. Por tanto, se procedió a escalar las variables continuas en un rango de 0 a 1.

- Otra actividad como se define en [19], es transformar las variables categóricas en variables ficticias binarias que codifican la presencia o ausencia de una categoría particular. Por lo tanto, se crean v – 1 variables ficticias para representar los v valores únicos de cada variable categórica.

Al realizar las tareas mencionadas se obtuvo el conjunto de datos preprocesado que se utiliza para el entrenamiento de los modelos de aprendizaje automático.

9.3.2. SELECCIÓN DE LAS TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Una actividad importante es la selección de las técnicas de aprendizaje automático, ya teniendo la estructura de los datos construido como se trabajó en la sección 9.3.1 y de acuerdo con lo revisado en la sección 6, existen muchas técnicas de aprendizaje automático que son trabajadas para resolver un problema de clasificación, por tanto, algunas de ellas se seleccionaron con base en un análisis descrito en [20]. En la tabla 8 se muestra el análisis de las técnicas, donde se usaron diferentes conjuntos de datos en algunos sectores de la industria, resaltando la capacidad que poseen para resolver problemas que no son linealmente separable y su buen desempeño en las métricas de evaluación de las técnicas de aprendizaje automático.

Tabla 8. Ventajas y desventajas de las técnicas de aprendizaje automático

Técnicas	Data set	Ventajas	Desventajas
SVM (support vector machine)	* Insurance dataset * UCI-Telecom, Operator 1,Cell2Cell * UCI data & Home telecommunication carry dataset	* Alto accuracy y reduce el sistema de complejidad. * Mejora el accuracy usando priorización en el contexto del problema. * Mayor accuracy incluso en la presencia de muchos atributos, gran tasa de abandono, etc.	* Más aplicable para la detección fraude que predicción abandono * En sistemas regulatorios no se ajustan para su uso. * Selección de funciones y pesos del kernel no correcto, para problemas de alta dimensionalidad
Random Forest	* Categorical type churn data * Data from Wireless telecom company publically available by SGI	* Valores altos de accuracy y F-score. * Alta precisión y sensibilidad. * Rendimiento fiable en datos de la industria de las telecomunicaciones	* El proceso de submuestreo más general es empleado. * No se aplica en datos de tiempo real
Xgboost	* Six real-life proprietary European churn modeling datasets. * Data from Wireless telecom company publically available by SGI	* Aumenta el churn en el acurracy de la predicción. *Alto acurracy y sensibilidad. Rendimiento fiable en datos de la industria de las telecomunicaciones	* Los problemas individuales reducen el rendimiento general. * No se aplica en datos de tiempo real
Redes Multicapa	* European telecommunications company data * Churn data from UCI	* Alta predicción con la mejor reducción dimensionalidad * 92% accuracy de predicción en modelo de abandono	* Problema con datos desbalanceados * Inconvenientes con multicolinealidad de variables

Fuente: Elaboración propia.

Estas 4 técnicas de aprendizaje automático son las seleccionadas, para el entrenamiento y validación en cada experimento, con el objetivo de lograr de seleccionar la mejor en el desempeño de las métricas de evaluación en la predicción del riesgo de abandono.

9.3.3. DIVISIÓN DE LA BASE DE DATOS EN ENTRENAMIENTO Y PRUEBA

El siguiente proceso que se debe tener en cuenta es dividir la base de entrenamiento y prueba, para la construcción de las técnicas de aprendizaje automático. Este se realizó

mediante un muestreo aleatorio, el cual se selecciona de manera aleatoria una muestra que corresponde al 80% (329.280), del total de la información como entrenamiento y el 20% (82.320) de prueba. Esta información se guarda en archivos .csv, para los entrenamientos y pruebas de validación de las técnicas de aprendizaje automático.

9.4 TÉCNICAS DE APRENDIZAJE AUTOMÁTICO QUE PERMITEN PREDECIR EL RIESGO DE ABANDONO

9.4.1. DEFINICIÓN DE EXPERIMENTOS PARA LA CONSTRUCCIÓN DE LAS TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

En la construcción de las técnicas de aprendizaje automático, se generaron 2 experimentos que se describen a continuación:

- **Experimento 1:** Generar la estimación de hiperparámetros con los modelos por defecto que contienen las librerías de sklearn en Python y evaluar su rendimiento.

- **Experimento 2:** Generar el entrenamiento, optimización de hiperparámetros y evaluación final de los modelos seleccionados, con la técnica de búsqueda por grilla (Grid Search), de la librería sklearn en Python, donde se establecen una serie de subconjuntos de valores para los hiperparámetros a optimizar, teniendo como métricas de desempeño la eficiencia del modelo con cada combinación de parámetros.

En cada experimento en el ajuste de hiperparámetros se implementó la validación cruzada. Esta técnica consiste en dividir el conjunto de entrenamiento en K iteraciones o (K-fold), los datos se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de entrenamiento. Finalmente se realiza el promedio de los resultados de cada iteración y así analizar las métricas de rendimiento en el entrenamiento, luego con el modelo entrenado se realiza la validación con los datos de prueba.

9.4.2. RESULTADOS DEL EXPERIMENTO 1 (TÉCNICAS DE APRENDIZAJE AUTOMÁTICO POR DEFECTO)

El proceso de los entrenamientos y la estimación de los hiperparámetros de las técnicas se realizan mediante la validación cruzada, como se mencionó anteriormente. Los resultados de la tabla 10, muestran los indicadores de desempeño de las técnicas de aprendizaje automático obteniendo en general un buen desempeño en el rendimiento de la predicción del riesgo de abandono en la cooperativa. Los algoritmos que resaltan son

el random forest y Xgboost con el F1-score más alto en cada estado(retiro=1 y activo=0), de los asociados de la cooperativa con un valor mayor o igual al 92%.

Tabla 9. Resultados por clase en los Modelos Defecto experimento 1

Técnica	Experimento 1					
	Precisión		Recall		F1-score	
	0= No fuga	1=Fuga	0= No fuga	1=Fuga	0= No fuga	1=Fuga
SVM (support vector machine)	88%	86%	83%	91%	85%	89%
Random Forest	89%	96%	96%	91%	92%	93%
Xgboost	89%	93%	91%	90%	90%	92%
Redes Multicapa	88%	93%	92%	89%	90%	91%

Fuente: Elaboración propia.

Tabla 10 Resultados Modelos Defecto, experimento 1

Técnica	Experimento 1			
	Accuracy	Precisión	Recall	F1-score
SVM (support vector machine)	87%	86%	90%	88%
Random Forest	92%	96%	90%	93%
Xgboost	90%	92%	90%	92%
Redes Multicapa	90%	93%	89%	91%

Fuente: Elaboración propia.

En la tabla 11, se resalta que las técnicas de random forest y Xgboost, tienen el F1-score más alto, seguido por la red multicapa.

9.4.3. RESULTADOS DEL EXPERIMENTO 2 (OPTIMIZACIÓN DE HIPERPARÁMETROS)

En la tabla 12 se muestran la grilla de hiperparámetros que son utilizados para optimización para cada modelo y el conjunto de valores o rangos mediante la función gridsearch:

Tabla 11 Descripción y definición de hiperparámetros experimento 2

Técnica	Descripción de Hiperparámetros
Random Forest	**Número de estimadores:** Referente a cuántos modelos individuales se usan en el *ensemble,* los valores utilizados fueron *10, 50, 100, 500, 900 y 1500* **Criterio:** La función para medir la calidad de una división. Los criterios admitidos son "gini" para la impureza de Gini y "log_loss" y "entropía", ambos para la ganancia de información de Shannon. **Maxima Features:** El número de características a considerar al buscar la mejor división, se trabajaron la 'auto', 'sqrt' y 'log2'
Xgboost	**Número de estimadores:** Referente al número de rondas de refuerzo. *10, 50, 100, 200 y 500* **Máxima profundidad:** Profundidad máxima del árbol 3, 6, 10 y 20
SVM (support vector machine)	**Penalización:** El coste o parámetro "C" se debe ajustar en una SVM al igual que el tipo de kernel a emplear. "C" es un parámetro de regularización que controla la compensación entre maximizar el margen y minimizar el término de error de entrenamiento. Si "C" es demasiado pequeño, se colocará un esfuerzo insuficiente para ajustar los datos de entrenamiento. Si "C" es demasiado grande, entonces el algoritmo se ajustará a los datos de entrenamiento lo que es conocido como overfitting [21], los valores utilizados fueron 0.1, 10, 100,500 y 1000. **Kernel** Referente al tipo de transformación que se le aplica a los datos de entrada para que el problema se pueda resolver por medio de un clasificador lineal, los valores utilizados fueron *poly, rbf, sigmoid.*
Perceptrón Multicapa	**Épocas**: Hace referencia al número de iteraciones que el modelo realiza en su proceso de entrenamiento, los valores fueron 50 y 100 **Tipo de optimizador:** Hace referencia al método por el cual se realiza la optimización de parámetros internos de la red en el proceso de entrenar, para este caso se usó el adam **Tasa de aprendizaje**: Hace referencia a la sensibilidad con que se modifican los parámetros internos de la red en cada una de sus iteraciones. En este caso trabajó con 0.001 y 0,01 **Número de capas ocultas**: Hace referencia a cuantas capas ocultas tendrá la red, para el entrenamiento las capas utilizadas fueron 2 y 3. **Número de neuronas:** Hace referencia a cuantas neuronas tendrá cada una de las capas ocultas, para 2 capas ocultas se trabajaron 14 y 7, y para las 3 capas ocultas 14,7 y 3

Fuente: Elaboración propia

Al realizar los entrenamientos y la optimización de las técnicas mediante validación cruzada con k particiones igual a 10, se tiene los siguientes hiperparámetros en la tabla 12, y los resultados de las métricas de las mejores combinaciones de hiperparámetros en la tabla 13.

Tabla 12 Hiperparámetros seleccionados experimento 2

Técnica	Hiperparámetros Seleccionados
Random Forest	**Número de estimadores:** 900 **Criterio:** "entropía" **Maxima Features:** 'auto',
Xgboost	**Número de estimadores:** *200* **Máxima profundidad:** 20
SVM (support vector machine)	**Penalización: 1000** **Kernel:** *rbf*
Perceptrón Multicapa	**Épocas**: 100 **Tipo de optimizador**: adam **Tasa de aprendizaje**: 0.001 **Número de capas ocultas**: 3 **Número de neuronas**: 14,7 y 3

Fuente: Elaboración propia

Con el resultado anterior, de la mejor combinación de hiperparámetros, se obtiene los resultados de las métricas de rendimiento de las técnicas de aprendizaje automático, mostradas en la tabla 14 y 15, por clase a predecir y en general de ambas clases.

Tabla 13 Resultados por clase en los Modelos optimizados experimento 2

Técnica	Experimento 2					
	Precisión		Recall		F1-score	
	0= No fuga	1=Fuga	0= No fuga	1=Fuga	0= No fuga	1=Fuga
SVM (support vector machine)	90%	91%	90%	92%	90%	91%
Random Forest	96%	98%	97%	96%	96%	97%
Xgboost	97%	98%	97%	98%	97%	98%
Redes Multicapa	93%	95%	94%	94%	94%	95%

Fuente: Elaboración propia

Tabla 14 Resultados generales en los Modelos optimizados experimento 2

Técnica	Experimento 2			
	Accuracy	Precisión	Recall	F1-score
SVM (support vector machine)	91%	91%	90%	91%
Random Forest	96%	97%	96%	96%
Xgboost	97%	98%	97%	98%
Redes Multicapa	94%	95%	94%	94%

Fuente: Elaboración propia

En resumen, de la tabla 14 y 15 las técnicas de aprendizaje automático con mejor rendimiento de acuerdo con el F1-score, son las redes multicapa, random forest y xgboost. Por otro lado, el SVM (support vector machine), a pesar de que mejoró sus indicadores realizando la optimización de hiperparámetros, si se compara con los hiperparámetros por defecto no logra superar las otras técnicas.

9.5 EVALUACIÓN DEL DESEMPEÑO DE LOS MODELOS DE APRENDIZAJE AUTOMÁTICO ENTRENADOS Y SELECCIÓN DE LA MEJOR TÉCNICA

Para la evaluación de los modelos en los diferentes experimentos se utilizaron las métricas de evaluación descritas en la sección 6.5 y se trabajó con los resultados de los hiperparámetros seleccionados, entrenando nuevamente y evaluando los resultados en los diferentes experimentos, en la tabla 16 se presenta un comparativo de los experimentos en cada una de las métricas:

Tabla 15. Resultados generales comparativos de los experimentos en la evaluación de los modelos

	Accuracy		Precisión		Recall		F1-score	
	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2
SVM (support vector machine)	87%	91%	**86%**	91%	90%	90%	88%	91%
Random Forest	92%	96%	96%	97%	90%	96%	93%	96%
Xgboost	90%	97%	92%	98%	90%	97%	92%	98%
Redes Multicapa	90%	94%	93%	95%	89%	94%	91%	94%

Fuente: Elaboración propia

Al comparar los resultados de los modelos evaluados en cada experimento, existió una mejoría en la optimización de hiperparámetros del experimento 2 con respecto a los modelos ejecutados con los hiperparámetros por defecto del experimento 1. Comparando el SVM, pasó de un F1-score del 88% a 91%, pero es la técnica Xgboost que presenta el mejor rendimiento en métricas en el experimento 2, como se resalta en la tabla.

A continuación, en la tabla 17 se presenta un comparativo de la evaluación de los experimentos por las clases a predecir en cada uno de los experimentos y modelos:

Tabla 16 Resultados por clase comparativos de los experimentos en la evaluación de los Modelos

	Precisión				Recall				F1-score			
	0= No fuga		1=Fuga		0= No fuga		1=Fuga		0= No fuga		1=Fuga	
	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2
SVM (support vector machine)	88%	90%	86%	91%	83%	90%	91%	92%	85%	90%	89%	91%
Random Forest	89%	96%	96%	98%	96%	97%	91%	96%	92%	96%	93%	97%
Xgboost	89%	97%	93%	98%	91%	97%	90%	98%	90%	97%	92%	98%
Redes Multicapa	88%	93%	93%	95%	92%	94%	89%	94%	90%	94%	91%	95%

Fuente: Elaboración propia

De acuerdo con lo observado en la tabla 17, la interpretación en la cooperativa y la importancia que tiene la respuesta de la siguiente pregunta de negocio: ¿cuál proporción de asociados que se retiraron, se identificaron correctamente por el modelo?; se seleccionó como métrica de decisión la Precisión y F1-score. Así las cosas, el modelo Xgboost con un 98% de precisión y F1-score, como se resalta en la tabla 17, es el modelo elegido para implementar en el despliegue del prototipo de visualización de los resultados.

9.6 APLICATIVO WEB PARA VISUALIZACIÓN DE LAS PREDICCIONES DE LOS ASOCIADOS ACTIVOS A LA COOPERATIVA

El enfoque propuesto proporciona una visualización de los resultados que se realiza mediante un marco web escrito en Python conocido como streamlit. El objetivo principal de la aplicación es diseñar una interfaz interactiva donde los usuarios puedan fácilmente realizar la predicción del riesgo de abandono de los asociados a la cooperativa usando el modelo xgboost seleccionado. La visualización de los resultados brinda a los usuarios tres funcionalidades, que se describen a continuación:

- **EDA (exploratory data analysis)**: panel de entendimiento de patrones de variables y descripción de la estructura del dataframe, como entrada para la predicción del modelo entrenado.
- **Predicción lista**: los usuarios pueden cargar un archivo .csv, que contiene una lista de usuarios con los campos requeridos para realizar el cálculo de la predicción de abandono, descarga de los resultados y la visualización por usuario de variables demográficas, uso de servicios y tenencia de productos.
- **Predicción del asociado**: un panel con las variables que ingresan al modelo para realizar la predicción, el cual el usuario puede interactuar con cada una de ellas.

El despliegue de la aplicación web se realizó con recursos suministrados por el proyecto en Amazon Web Services (AWS), con el servicio de EC2, que es una computación en la nube elástica que actúa como un servidor virtual, donde tiene guardado el modelo xgboost seleccionado con los hiperparámetros ajustados en un archivo en formato pkl. Este modelo es cargado en el marco de Streamlit, que construye la interfaz de usuario, y que permite interactuar con las funcionalidades descritas anteriormente. Luego, esta visualización se empaqueta en un contenedor Docker, y se comparte para los usuarios que interactúan con la visualización, como se muestra en la figura 36.

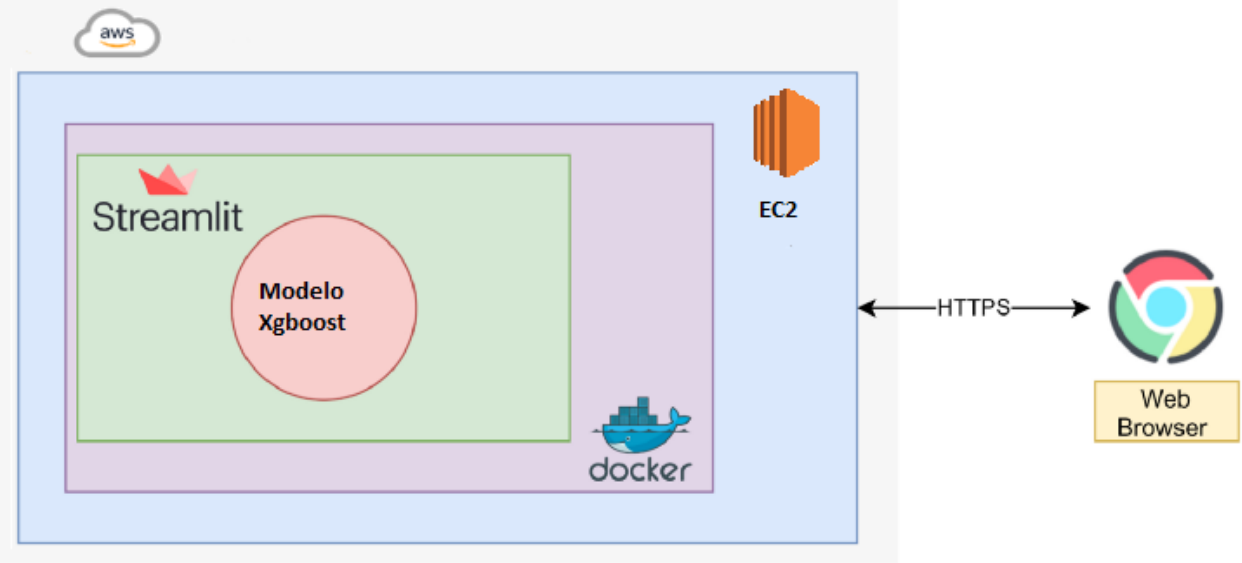


Figura 36. Implementación de la visualización de resultados.

Esta arquitectura descrita anteriormente, tendrá la siguiente interfaz de visualización de las figuras 37,38 y 39 en Streamlit.

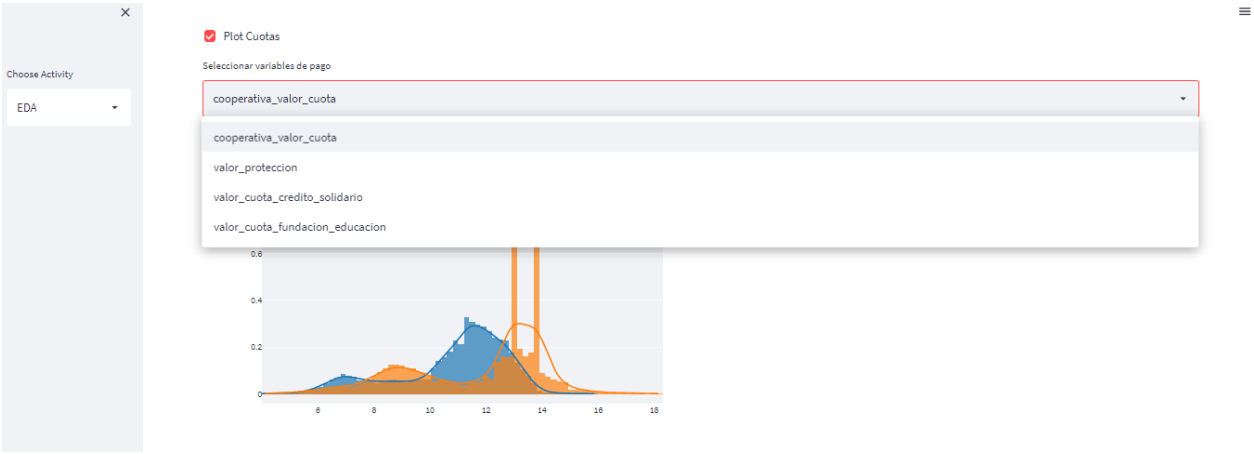


Figura 37. Interfaz de usuario para EDA.

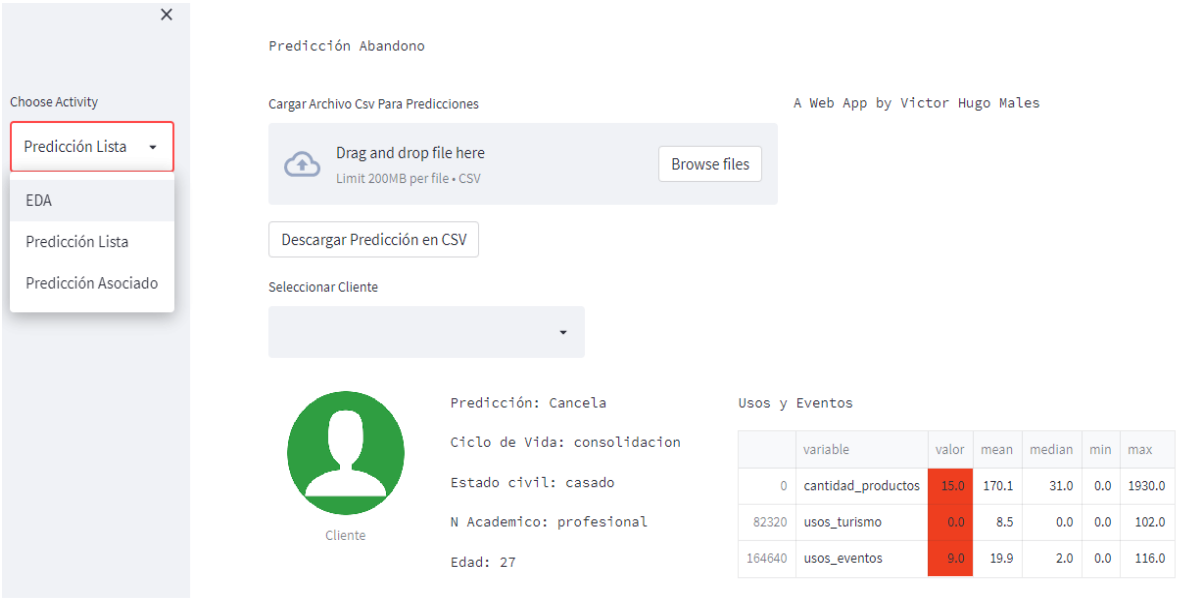


Figura 38. Interfaz de usuario para la predicción de lista de usuarios.

Figura 39. Interfaz de usuario para la predicción del asociado.

Finalmente, este enfoque ayuda al área comercial a identificar posibles asociados que inicien el proceso de retiro de la cooperativa a través de una interfaz muy intuitiva y fácil de usar y puede servir como base para tomar decisiones en la estrategia de reducir indicadores de abandono.

10. DISCUSIÓN DE RESULTADOS

El proceso metodológico aplicado permitió un aprendizaje valioso para entrenar técnicas de aprendizaje automático en la cooperativa. En primer lugar, se realizó el procesamiento de los datos, identificando todas las fuentes de la bodega corporativa en el proceso de adquisición de los datos. Además, se construyó un proceso en el sistema de base de datos distribuido en la nube (GCP) para orquestar todas las salidas en un único dataframe, que puede ser insumo para la construcción de otros modelos, en otras áreas de la cooperativa.

En comparación con el escenario inicial establecido por la cooperativa, donde el modelo solo incluía información demográfica, interacciones históricas de consumo de productos y beneficios adquiridos a lo largo de la historia conjunta en la cooperativa, ahora se consideró un conjunto más completo de variables, como las cuotas pagadas por servicios y otros usos relevantes. Este enfoque ampliado permitió mejorar la precisión y exactitud del modelo.

El modelo inicial se entrenó con una muestra aleatoria de 10 mil clientes y se validó con otros 3 mil. Los indicadores de exactitud, precisión y sensibilidad obtenidos en cada uno de los experimentos planteados en este proyecto superaron los resultados del modelo inicial. Además, se recopiló información de un total de 411.600 clientes, lo cual es considerablemente mayor al escenario inicial trabajado por la cooperativa. Las reglas de agrupamiento de las variables también jugaron un papel fundamental en la obtención de estos resultados significativos en los diferentes experimentos.

El análisis descriptivo realizado ha permitido obtener un mejor conocimiento de las variables asociadas a los miembros de la cooperativa. Se han identificado patrones significativos en relación con el estado de retiro o activo de los asociados, particularmente en cuanto al uso de servicios y la tenencia de productos. Estas variables se han revelado como factores importantes para visualizar grupos de interés, lo cual resulta fundamental para comprender y mejorar la toma de decisiones.

De acuerdo con los resultados obtenidos en cada uno de los experimentos evaluados en las secciones anteriores, se encontró que los modelos basados en métodos de ensamble estuvieron siempre entre los modelos con mejor desempeño para los experimentos.

Es relevante resaltar la notable diferencia en sensibilidad entre el escenario inicial, donde la cooperativa obtuvo un valor de 4.7%, y el modelo entrenado en este proyecto, el cual logró alcanzar un valor del 98%. Esto implica que, considerando los indicadores de retiro mensuales en la cooperativa, con un promedio de 2500 clientes por mes, el modelo inicial sería capaz de identificar aproximadamente 118 clientes, en contraste con los 2450 clientes que se espera detectar utilizando el modelo actualizado. Estos resultados adquieren una gran importancia, ya que permiten priorizar eficientemente la gestión de retención de clientes.

Finalmente, el experimento con mejores resultados es el experimento 2 con un promedio en el indicador f1 de todas las técnicas del 95%, superando en un 4% al experimento 1, y que demuestra que la optimización de los hiperparámetros en las técnicas es importante porque se optimizan los indicadores de desempeño en la evaluación de la predictibilidad de una variable dependiente.

11. CONCLUSIONES

El proyecto ha alcanzado el desarrollo de la aplicación web desplegado en infraestructura propia, que incluye un modelo entrenado de xgboost con una alta precisión del 97%. La visualización interactiva brinda a los usuarios la capacidad de analizar los patrones de las variables y realizar predicciones propias, lo que enriquece tanto al equipo técnico como al equipo estratégico de la cooperativa. Este logro representa un avance significativo en el uso de la tecnología para mejorar la iteración con los datos y la facilidad

para disponer un servicio (aplicativo web), que ayudará la toma decisiones en la cooperativa.

El análisis de la información revela una falta de completitud en los datos históricos de los asociados, con solo el 15% de las variables teniendo información completa. Se recomienda realizar un análisis profundo del proceso de ETL para identificar y corregir las posibles causas de esta falta de completitud. Garantizar una captura adecuada de la historia de todas las fuentes corporativas en la base de datos distribuida será fundamental para mejorar la calidad y la disponibilidad de los datos.

En la exploración y el análisis descriptivo de la información, se han identificado perfiles de asociados con mayor participación en la clase de retiro, como aquellos con segmento de ciclo de vida joven, sin productos y de estrato socioeconómico bajo. Además, se han analizado las correlaciones entre variables de uso de servicios, eventos y turismo, lo que ha llevado a la identificación de relaciones relevantes para la toma de decisiones en el riesgo de abandono. Estos resultados brindan una base sólida para implementar estrategias dirigidas a retener a los asociados y mejorar la gestión del riesgo de abandono en la cooperativa.

Se realizaron 2 experimentos con 4 técnicas de aprendizaje automático, a los cuales se les evaluó el desempeño mediante métricas definidas para problemas de clasificación. Se encontró un buen rendimiento de las métricas en ambos experimentos, siendo la optimización de hiperparámetros el experimento con mejores resultados en los indicadores de desempeño de las técnicas seleccionadas, si se compara con los resultados de las mismas técnicas sin realizar dicho proceso. Al evaluar los resultados de los cuatro modelos seleccionados, se encontró que la técnica xgboost fue el que cumplió de manera más integral con los criterios de capacidad predictiva.

En la Sección 1, de definición del problema, se expone un escenario de mejora en los indicadores de desempeño de las técnicas de aprendizaje automático que fueron construidas en la cooperativa. Si se compara los resultados de las técnicas en los escenarios de evaluación, se tiene que las técnicas aprendizaje automático construidas por la cooperativa tiene métricas de desempeño en la exactitud, precisión y sensibilidad del 88%, 18.7% y 4.7% respectivamente. Si es comparado con la técnica con mejor desempeño xgboost que obtuvo indicadores de exactitud, precisión y sensibilidad del 97%, 98% y 97% se evidencia una mejora en cada indicador.

12. TRABAJOS FUTUROS

Un trabajo futuro, crear un proceso de integración de fuentes de datos sin tanta perdida de datos históricos de los asociados a la cooperativa e incorporar información relacionada con variables psicográficas (Personalidad, estilos de vida, intereses, gustos, inquietudes,

opiniones, valores) y conductuales (Lealtad de marca, beneficios buscados (precio, calidad, servicio) etc.), mediante las actividades de actualización de datos en las empresas de cada sector y eventos que se realizan.

Evaluar y ajustar la arquitectura de la aplicación web por parte del equipo técnico de la cooperativa es fundamental para lograr una implementación exitosa, asegurando su compatibilidad con las necesidades tecnológicas y la conexión con bases de datos en ambientes productivos. Además, esta evaluación permitirá identificar mejoras y optimizaciones, asegurando un rendimiento óptimo, escalabilidad y confiabilidad en el entorno de producción.

13. REFERENCIAS BIBLIOGRÁFICAS

[1] Hurwitz, J., & Kirsch, D. (2018). Machine Learning for dummies. In Journal of the American Society for Information Science (Vol. 35). https://doi.org/10.1002/asi.4630350509

[2] Nadia Alboukaey, Ammar Joukhadar and Nada Ghneim, “Dynamic behavior-based churn prediction in mobile telecom Nadia,” Expert Systems with Applications .Elsevier, vol. 162, July 2020.

[3] Farid Shirazi and Mahbobeh Mohammadi, “A big data analytics model for customer churn prediction in the retiree segment,” International Journal of Information Management.Elsevier, vol. 48, pp. 238–253,October 2019.

[4] María Óskarsdóttir, Tine Van Calster, Bart Baesens ,Wilfried Lemahieu and Jan Vanthienen, “Time series for early churn detection: Using similarity based classification for dynamic networks,”Expert Systems with Applications. Elsevier, vol. 106, pp. 55–65, September 2018.

[5] Ying Huang and Tahar Kechadi, “An effective hybrid learning system for telecommunication churn prediction,” Expert Systems with Applications.Elsevier, vol. 40, pp. 5635–5647, October 2013.

[6] Shalev-Shwartz, S., & Ben-David, S. (2013). Understanding machine learning: From theory to algorithms. In Understanding Machine Learning: From Theory to Algorithms (Vol. 9781107057). https://doi.org/10.1017/CBO9781107298019

[7] Baccini, A, N Laporte, SJ Goetz, M Sun, and H Dong. 2008. “A First Map of Tropical Africa’s Above-Ground Biomass Derived from Satellite Imagery.” *Environmental Research Letters* 3 (4): 045011.

[8] Trevor Hastie, Robert Tibshirani and Jerome Friedman, Statistics, Data Mining,

Inference, and Prediction. California, IL: Springer Series in Statistics, 2008.

[9] Varela Arregoces, Ernesto, and Edwin Campbells Sanchez. 2011. “Redes Neuronales Artificiales: Una Revision Del Estado Del Arte, Aplicaciones Y Tendencias Futuras.”

[10] Lutins, E. (2017). Ensemble Methods in Machine Learning: What are They and Why Use Them? Towards Data Science. Retrieved from https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f

[11] R. Jindal and M. D. Borah, “A SURVEY ON EDUCATIONAL DATA MINING AND RESEARCH TRENDS" Rajni vol. 5, no. 3, pp. 53–73, 2013.

[12] Wang, John. 2003. *Data Mining: Opportunities and Challenges*. IGI Global.

[13] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. “From Data Mining to Knowledge Discovery in Databases.” *AI Magazine* 17 (3): 37–37.

[14] Gonzales Gutierrez Diego, “TÉCNICAS DE MACHINE LEARNING EN EL ANÁLISIS DEL CHURN RATE,” Tesis Master., Univ. Cantabria, 2019.

[15] DOMINGUEZ.D y HERNO.G. 2018. Retención y ‘Churn Rate’ ESIC [Consulta 14 de julio de 2019] Disponible en: https://www.esic.edu/documentos/editorial/resenas/9788473567183_Esic%20Alumni_01-04-08.pdf

[16] Vigneau Cesari Jean Paul, “Modelo predictivo para determinar la Tasa Churn en pacientes de un centro médico” 2018. [En línea]. Disponible en https://www.mti.cl/wp-content/uploads/2018/12/Tesina_2018_Cesari-Jean.pdf

[17] Hele, J. (2019). Scale, Standardize, or Normalize with Scikit - Learn. Towards Data Science. Retrieved from https://towardsdatascience.com/scale-standardize-or-normalize-with-scikitlearn-6ccc7d176a02

[18] Li, K. G., & Marikannan, B. P. (2019). Hybrid particle swarm optimization-extreme learning machine algorithm for customer churn prediction. Journal of Computational and Theoretical Nanoscience, 16(8), 3432-3436.

[19] De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. International Journal of Forecasting, 36(4), 1563-1578.

[20] Ahmed, A., & Linen, D. M. (2017, January). A review and analysis of churn prediction

methods for customer retention in telecom industries. In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 1-7). IEEE.

[21] “Análisis comparativo de algoritmos de aprendizaje para predecir la evolución de pacientes con Daño Cerebral Adquirido.” [Online]. Available: https://oa.upm.es/12231/1/INVE_MEM_2011_99803.pdf