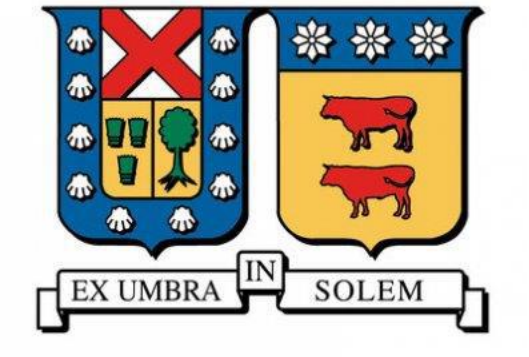


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INDUSTRIAS

PREDICCIÓN DE AUMENTO DE CUPO DE TARJETAS DE CRÉDITOS EN INSTITUCIÓN BANCARIA MEDIANTE MODELOS ESTADISTICOS Y REDES NEURONALES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

JOAQUÍN ENRIQUE COLLAO ASTROZA

MEMORIA PARA OPTAR AL TÍTULO
DE INGENIERA CIVIL INDUSTRIAL

PROFESOR GUÍA : SR. OSCAR SAAVEDRA R.
PROFESOR CORREFERENTE : SR. ELOY ALVARADO N.

ABRIL, 2024

AGRADECIMIENTOS:

Agradecido por todas las personas que estuvieron involucradas en mi carrera universitaria. Ha sido una aventura larga, llena de obstáculos que superar, sobre todo los internos. Ya no soy la misma persona que ingresó en primero año de universidad y agradezco mucho en la persona que me he transformado.

Agradezco mucho las experiencias que me ha generado mi trayectoria, incluyendo las práctica que he tenido, las nuevas responsabilidades que he adquirido y sobre todo, agradezco mucho la autodisciplina que me ha generado en mí.

Agradezco en especial a mi familia, mi pareja, mis compañeros de generación, a los profesores y en especial a mi profesor guía. Ustedes fueron la base principal. Gracias por la paciencia que me han tenido.

Agradezco mucho las dificultades, altos y bajos que he tenido y, por último, le dedico mi trabajo a mi hija Kida.

¡Gracias Totales!

CONTENIDO

RESUMEN EJECUTIVO:

En esta investigación se desarrollan modelos predictivos para datos relacionados con usuarios bancarios que solicitan aumentar el límite de sus tarjetas de crédito, lo que constituye un problema de clasificación binaria. Los modelos predictivos abarcan tanto enfoques estadísticos como de aprendizaje automático. Para ello, se emplea la metodología "Descubrimiento de Conocimiento en Bases de Datos" (KDD por sus siglas en inglés), que facilita la creación de modelos predictivos mediante técnicas de aprendizaje supervisado.

A partir de esta metodología, se generó un modelo logístico, representativo del enfoque estadístico, y cuatro modelos de redes neuronales: una red neuronal básica, dos redes neuronales profundas y una red neuronal convolucional. La evaluación de estos modelos se realizó utilizando la matriz de confusión para obtener métricas como exactitud (accurancy), precisión (precision), sensibilidad (recall) y F1 Score, permitiendo así una comparación cuantitativa entre ellos. Los resultados indican que el mejor rendimiento se atribuye al segundo modelo de red neuronal profunda.

Para evaluar el impacto potencial de estos modelos en la industria bancaria, se llevó a cabo un análisis de optimización que reveló que el segundo modelo de red neuronal profunda generó el mayor beneficio esperado, alcanzando un total de 117.226.658 CLP, mientras que el modelo de red neuronal básica presentó el menor beneficio esperado, con un monto de 46.450.565 CLP.

Desde un punto de vista cualitativo, se observaron diferencias significativas, como el hecho de que el modelo logístico permite identificar qué variables impactan más en la variable objetivo, mientras que en los modelos de redes neuronales, a mayor número de capas, mayor es la precisión, pero también aumenta el tiempo de entrenamiento. Por lo tanto, aunque el modelo con mejor rendimiento generó el mayor beneficio esperado, es importante considerar el tiempo necesario para su entrenamiento al decidir qué modelo implementar en una empresa.

CAPÍTULO I: INTRODUCCIÓN

1. PROBLEMA DE INVESTIGACIÓN

A lo largo de la historia, se ha evidenciado que los grandes avances y progresos de la humanidad han sido impulsados por las revoluciones industriales (Revista Empresarial, 2020). Cada una de estas revoluciones ha estado centrada en un aspecto fundamental. Por ejemplo, en la primera revolución industrial, el motor principal fue la máquina de vapor, seguido por la electricidad en la segunda y la información en la tercera (Sectorial, 2020).

En el contexto actual, marcado por la cuarta revolución industrial (Sectorial, 2020), cuyo foco principal es la digitalización, la información y la inteligencia artificial (Revista Empresarial, 2020), se ha creado una oportunidad sin precedentes para el crecimiento empresarial. Sin embargo, esto también ha llevado a que las empresas se vean en la necesidad de manejar grandes volúmenes de datos para mantener su competitividad y crecer financieramente, especialmente en una era donde los consumidores poseen un poder informativo considerable, lo que facilita su capacidad para cambiar de proveedores. (Suarez, 2020)

Por esta razón, el análisis de datos se ha convertido en un componente esencial de la estrategia empresarial, sobre todo con el crecimiento exponencial de los datos a nivel mundial. El objetivo del análisis de datos es obtener modelos que permitan tomar decisiones más informadas en el futuro de una empresa.

Una de las metodologías que ha potenciado el análisis de datos es el "descubrimiento de conocimiento en datos" (KDD por sus siglas en inglés), un proceso que consiste en encontrar patrones y otra información valiosa en grandes conjuntos de datos (Chiu, 2008), es decir, usa la estrategia de trabajar de forma eficiente con altos volúmenes de datos, con el fin de crear modelos que puedan ayudar a tomar decisiones, mejorar sistemas de información, aprender patrones, entender nuevos horizontes etc. La creación de nuevos modelos, mediante los datos, también se conoce como minería de datos (IBM, 2018) .

La ventaja de esta metodología es que se puede trabajar de forma libre, es decir, el investigador puede escoger el modelo estadístico con el cual desee trabajar. Es decir, con esto se puede trabajar con diversos modelos al mismo tiempo, todo esto con el fin de comparar modelos y escoger el más adecuado para la ocasión.
La capacidad de predecir el comportamiento de los clientes es una ventaja significativa para las empresas. Por lo tanto, muchas de ellas buscan crear modelos predictivos para mejorar su rendimiento. Entre estos modelos, destacan la regresión logística, utilizada para identificar las variables que influyen en la probabilidad de que ocurra un evento, y las redes neuronales, que han ganado popularidad recientemente en el campo de la inteligencia artificial por su capacidad para modelar el comportamiento humano.

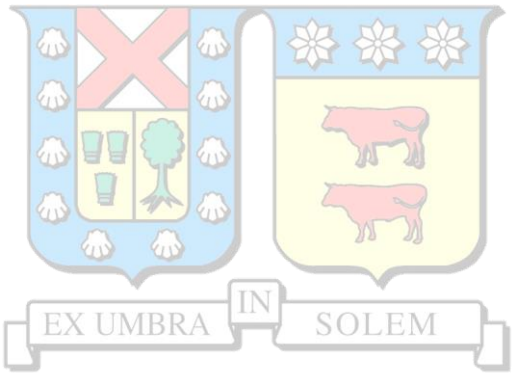
El modelo de regresión logística (SimpliRoute, 2023) es una metodología que trabaja con una variable dependiente dicotómica (es decir, que tomar valor 1 ó 0) y es útil para identificar que variables influyen en la probabilidad de que ocurra un evento, como, por ejemplo, una clasificación, decisión, enfermedad, acierto, etc. Esta herramienta (SimpliRoute, 2023) es utilizada en las empresas para identificar que influye en las decisiones de clientes, factibilidad de un proceso productivo, tiempos de entregas, etc. Además, también existe su versión "regresión logística múltiple", cuya diferencia es que la variable objetivo, se puede explicar con varias variables independientes.

Dentro de las ramas de la inteligencia artificial, últimamente se han hecho muy populares el modelo de red neuronal, cuyo modelo, busca moldear o simular el comportamiento de un cerebro humano. Las redes neuronales (Amazon AWS, 2018) pueden ayudar a crear diversos modelos, como por ejemplo, modelos de predicción de datos, simuladores de chat, detector de fallos, etc. Es decir, es una herramienta muy versátil.

Un desafío importante para las empresas es predecir el comportamiento de sus usuarios y clientes.

Por ejemplo, en el caso de los bancos, sería beneficioso poder predecir qué usuarios están dispuestos a aumentar su límite de crédito, siempre y cuando tengan un historial crediticio sólido. Para abordar este desafío, es necesario evaluar el rendimiento de herramientas

predictivas como la regresión logística y las redes neuronales en la predicción de estas decisiones. Esta investigación se propone precisamente eso: determinar la capacidad de estos modelos para predecir la decisión de un cliente de aumentar su límite de crédito y evaluar su impacto en el sector bancario.



2. OBJETIVOS

2.1. OBJETIVO GENERAL

Crear modelo de optimización utilizando modelos predictivos de redes neuronales y estadísticos, con el propósito de aumentar beneficios de una institución bancaria mediante la predicción de usuarios que deciden aumentar su cupo de tarjeta de créditos.

2.2. OBJETIVOS ESPECÍFICOS

- Limpiar conjunto de datos que serán usados para la implementación de modelos predictivos.
- Construir modelo predictivo de regresión logística múltiple.
- Construir modelo predictivo de red neuronal.
- Identificar las diferencias cualitativas y cuantitativas de los modelos predictivos.
- Identificar las ventajas y/o desventajas que pueda entregar estos dos modelos al estudiar un mismo caso en común.
- Identificar cual modelo predictivo es el más eficiente para predecir el aumento de cupo de tarjeta de crédito.
- Construir modelo de optimización.

CAPÍTULO II: ESTADO DEL ARTE Y ALCANCE

1. ESTADO DEL ARTE

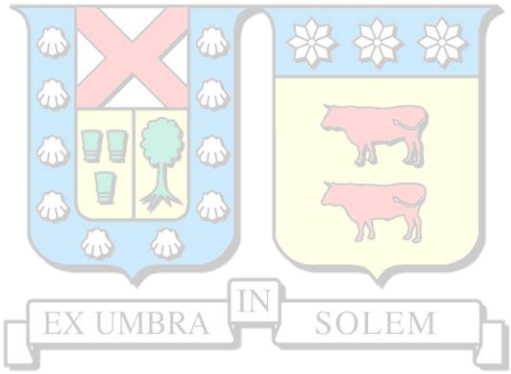
Mediante la investigación de otros casos de uso en donde se puede explicar la predicción de datos, se puede verificar que estas herramientas de predicción, tanto como regresión logística como redes neuronales, pueden beneficiar a las empresas y además a diferentes industrias.

A continuación, se adjunta algunos ejemplos de estudios que utilizaron herramientas de manejo de datos junto con redes neuronales para poder solucionar problemas.

Jorge Luis Morán Leal (2017) aborda en su investigación el pronóstico de la volatilidad del tipo de cambio mediante métodos de Inteligencia Artificial. En este trabajo, el autor compara modelos econométricos con modelos de redes neuronales utilizando datos de series temporales que abarcan más de 10 años de información sobre el tipo de cambio del dólar y el euro. Morán Leal desarrolla un modelo predictivo de series temporales utilizando únicamente un modelo de red neuronal, así como un modelo híbrido que combina ambos enfoques. Los resultados revelan que ninguno de los modelos por sí solos demuestra ser superior al otro. Sin embargo, la combinación de ambos modelos muestra un potencial de mejora significativo, lo que sugiere que su sinergia puede optimizar la precisión de las predicciones.

Por otro lado, Guera et al. (2021) exploran modelos de regresión logística multinomial ordinal y redes neuronales artificiales para la clasificación de tipos de madera. Utilizando un enfoque estadístico, desarrollan una regresión logística que incorpora 24 variables independientes, mientras que el modelo de redes neuronales cuenta con una capa oculta de 8 neuronas. Las pruebas de predicción revelan que el modelo de redes neuronales supera al modelo logístico en términos de precisión, demostrando una mayor tasa de aciertos. Sin embargo, el modelo logístico proporciona información valiosa sobre las variables que tienen un mayor impacto en el proceso de clasificación.

Finalmente, Quintana Reyes (2022) presenta un modelo predictivo de insolvencia financiera en PYMES utilizando redes neuronales artificiales. El autor desarrolla un modelo de red neuronal con una capa oculta de 4 neuronas y dos neuronas de salida, con el objetivo de identificar la solvencia o insolvencia de una empresa a partir de variables seleccionadas previamente y sometidas a un filtro de selección. A través de pruebas de clasificación, Quintana Reyes identifica las variables más influyentes en la solvencia de las PYMES estudiadas, destacando así la relevancia de su enfoque en la evaluación del riesgo financiero.



2. ALCANCE

2.1PROPUESTA DE VALOR

Con el fin de aprovechar las nuevas herramientas tecnológicas que ofrece la era actual, relacionadas con el análisis de datos, se propone crear un modelo que pueda aportar beneficios a las empresas, gracias a la predicción de datos. Para poder demostrar los beneficios que pueden entregar estas herramientas de análisis de datos, se propone aplicar un caso de uso correspondiente a la predicción de aumento de cupo de tarjeta de créditos, según el historial de comportamientos de los usuarios de estás mismas.

Para eso con la base de datos del historial de los clientes se pretenderá construir 5 modelos, uno correspondiente a una regresión logística, es decir, un modelo tradicional de estadístico y 4 correspondientes a redes neuronales, cuyos modelos neuronales corresponden a un modelo de red neuronal básico, dos modelos neuronales profundo y un modelo neuronal convolucional.

El objetivo es que los 5 modelos puedan predecir cual son los usuarios que tendrían una mayor posibilidad de tomar la decisión de aumentar el cupo de su tarjeta de crédito. Para este caso en específico, la aplicación de estos modelos predictivos podrían ayudar al área de marketing de un banco, con el fin de generar mayores ingresos a este mismo. Por lo mismo, se propondrá un modelo de optimización para verificar los beneficios que podrían entregar los modelos predictivos.

2.2ALCANCE DE LA INVESTIGACIÓN

Mediante el uso de software correspondiente a Rstudio, se trabajará en la base de datos entregada por el concurso de batalla de datos, concurso patrocinado por el banco Itaú. Cuya base de datos corresponde al historial de los clientes que utilizan tarjetas de créditos.

Con el análisis de datos, usando técnicas como minería de dato y KDD, se construirá 5 modelos predictivos. Uno correspondiente a regresión logística, otro correspondiente a una red neuronal básica, dos correspondiente a una red neuronal profunda y el último modelo correspondiente a una red neuronal convolucional.

Se realizará un análisis cualitativo y cuantitativo de los modelos predictivos, con el de identificar posibles ventajas o desventajas de estos modelos. Además, se realizará una evaluación de la predicción de los datos, con el fin de identificar que modelo es más eficiente y/o eficaz.

Para finalizar, se creará un modelo de optimización que permita identificar con cual modelo predictivo se puede obtener mayores beneficios, en un caso hipotético de que se llevase a cabo la aplicación del modelo predictivo.

CAPÍTULO III: MARCO TEÓRICO

3.1 MARCO TEÓRICO

3.1.1. INDUSTRIA BANCARIA

Los orígenes de la industria bancaria se datan de hace muchos siglos atrás. Se sabe que los primeros bancos consolidados, nacieron en la civilización de Mesopotamia en el siglo XX A.C. (Reinventa, 2021) estos bancos lo que hacían era prestar y comercializar granos a los agricultores de la época. El primer banco que pudo desarrollar un sistema de préstamo de capital como se conoce hoy en día, se creó en la civilización de Grecia en los siglos IV A.C. (Reinventa, 2021). Esta civilización fue la primera en crear sistemas bancarios que se dedican a prestar dinero, almacenar y cuidar dinero a personas particulares.

La industria bancaria ha evolucionado durante el transcurso de los siglos, aprovechando las nuevas tecnologías que se van desarrollando a medida que pasa el tiempo (Reinventa, 2021).

La industria bancaria es una de las principales que mueven la economía de un país, incluso del mundo, ya que esta es la que puede financiar nuevos proyectos y acelerar el crecimiento de otras industrias (Reinventa, 2021).

Esta industria ofrece tres tipos de productos: Sistemas de financiamiento, sistemas de ahorros y sistemas de Inversiones.

3.1.2. SISTEMA DE FINANCIAMIENTOS

Los sistemas de financiamientos que ofrecen los bancos se componen por dos grandes grupos: Prestamos y Productos bancarios.

Los prestamos corresponden principalmente a los créditos de consumo, hipotecarios y comerciales.

Y los productos bancarios corresponden a las tarjetas de créditos, líneas de créditos y productos enfocados sólo para empresas como el factoring, leasing, capital semilla, etc.

3.1.3. TARJETA DE CRÉDITOS

Según la CMF, las tarjetas de créditos son "*instrumentos que operan como un crédito, y que se utiliza para pagar bienes o servicios.*". Es decir, son instrumentos que pueden servir como créditos de consumo accesibles y personalizados, ya que, el usuario puede definir el monto necesario y la cantidad de cuotas que se estime a pagar, siempre y cuando este no supere el límite establecido por la institución financiera.

Este monto es conocido como cupo de tarjeta de crédito.

Estos cupos pueden ser aumentado siempre y cuando ambas partes estén de acuerdo, es decir, esté de acuerdo el usuario como la institución financiera (Quiroa, Economipedia.com, 2021).

3.2. ECONOMETRÍA

La econometría es una rama de la ciencia, la cual utiliza modelos estadísticos con el propósito de explicar situaciones económicas. Situaciones que pueden ser a nivel macro o micro. (Chiu, 2008)

Esta rama busca explicar el comportamiento de un mercado, industria, país sociedad, etc. Mediante los modelos matemáticos, económicos y estadísticos.

El objetivo de la econometría dependerá de quienes utilicen esta rama, es decir, que esto podría servir como herramienta para tomar decisiones, analizar situaciones, aprender nuevos comportamientos, etc. (Chiu, 2008)

3.2.1. MODELOS ESTADÍSTICOS

Los modelos estadísticos corresponden a ecuaciones y/o funciones que determinan un comportamiento de un fenómeno estadístico. Es decir, trata de simular el comportamiento de

un determinado caso. Con la aplicación de los modelos estadísticos, podría servir para predecir ciertos fenómenos y/o comportamientos. (Chiu, 2008)

Estos modelos pueden funcionar para fenómenos de dos variables o de multivariables. Todo esto va a depender del contexto en sí.

De los modelos estadísticos, se pueden encontrar la regresión lineal, la regresión logística y las series de tiempo.

3.2.2. REGRESIÓN LOGÍSTICA

La regresión logística es un método estadístico utilizado para modelar y analizar relaciones entre una variable binaria dependiente (también conocida como variable categórica dicotómica, que toma dos valores posibles, como 0 o 1) y una o más variables independientes (también llamadas variables predictoras o covariables). Aunque su nombre incluye la palabra "regresión", la regresión logística se utiliza principalmente para problemas de clasificación, donde el objetivo es asignar instancias a una de dos clases posibles. (Chiu, 2008)

$$f(x) = \frac{1}{1 + e^{-x}}$$

***Ecuación 1**: Función logística simple. Fuente: Elaboración propia.*

El gráfico que representa una regresión logística se conoce como curva “S” el cual se detalla a continuación:

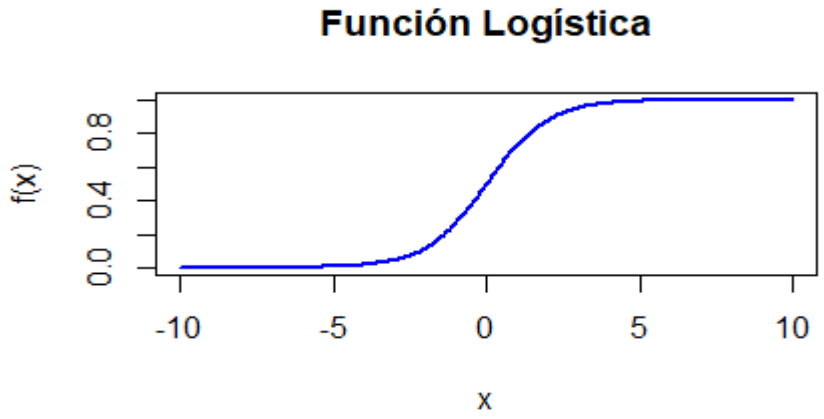


***Gráfico 1:** Ejemplo de gráfico de función logística. Fuente: Elaboración propia.*

Como se puede observar, es un modelo que entrega valores entre 0 y 1.

3.2.2.1 REGRESIÓN LOGÍSTICA MÚLTIPLE

La regresión logística múltiple es un modelo en donde la variable dependiente es dicotómica, pero tiene la cualidad de que la variable objetivo es posible explicarla por varias variables independientes .

La forma general de la regresión logística múltiple se expresa en la siguiente ecuación:

$$z = b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_k \cdot x_k + a$$

$$f(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_k \cdot x_k + a)}}$$

***Ecuación 2**: Ejemplo de cómo funcionaría una función logística múltiple. Fuente: Elaboración propia*.

Donde:

- e es la base del logaritmo natural.
- $b_1, \ldots, b_k$ son los coeficientes del modelo que se ajustan durante el proceso de entrenamiento.
- a corresponde al grado de incertidumbre que posee un modelo estadístico.
- $X_1 X_2, \ldots, X_k$ son los valores de las variables independientes.

El objetivo en la regresión logística es encontrar los valores de los coeficientes β que maximizan la verosimilitud de los datos observados.

Para cada valor estimado de beta, al aplicar Euler, este genera un valor que permite interpretar el impacto que influye la variable independiente. Esto se conoce como ods (Cardenas, 2015), que corresponde al ratio de posibilidad que pueda ocurrir una variable independiente relacionado con la variable dependiente (Chiu, 2008). Cuando el valor que toma es mayor a uno, quiere decir que, al aumentar la variable, la probabilidad de que la variable objetivo

tome un valor "1" aumenta. En caso de que el ratio tome un valor menor a uno, disminuye la probabilidad de que la variable objetivo tome el valor "1".

3.2.3. APLICACIÓN DE LOGARITMO

En algunos casos, aplicar la estrategia del logaritmo puede ser útil cuando los datos no siguen una distribución lineal o al comparar los datos con los datos objetivos no es posible observar un modelo.

Para mejorar el modelado (Chiu, 2008), se aplica logaritmos a los datos con el fin de:

Estabilizar la varianza: Es decir, la eliminación de la heterocedasticidad de una muestra de datos. La heterocedasticidad corresponde cuando la varianza de una muestra varía según la proporción de datos que se trabaja. Al aplicar logaritmo en algunos casos, puede ayudar a que la varianza se vuelva constante.

Linealizar: Corresponde en los casos cuando los datos se comportan como una función no lineal. Pero al momento de aplicar logaritmo, estos datos se empiezan a compartir de esa forma.

Manejar los datos atípicos: En algunos casos, la aplicación de logaritmo permite amortiguar la cantidad de datos atípicos que se puede encontrar en una muestra.

3.2.4. MULTICOLINEALIDAD

La multicolinealidad (Chiu, 2008) es un fenómeno que ocurre cuando dos o más variables independientes del modelo poseen una alta correlación entre sí. Es decir, estas variables poseen relaciones fuertes entre ellas.

Algunos efectos negativos (Chiu, 2008) de la multicolinealidad podrían corresponder a: Inestabilidad de los coeficientes, dificultad en la interpretación, menor precisión en la interpretación e inestabilidad numérica.

La forma de contrarrestar la multicolinealidad es identificando las variables que poseen una alta correlación y ver el caso de eliminar o combinar las variables, según el caso sea necesario.

Formas de contrarrestar la multicolinealidad (Chiu, 2008): Lo primero es detectar que variables independientes poseen una alta correlación entre sí.

La correlación es la relación que existe entre los datos de unas variables, y esta medida indica que tan alta es la relación entre ellas. Se recomienda la correlación Spearson (Cosio, 2021)y se sugiere que una alta correlación está asociada a un valor superior a 0.7.

Esa es la manera de identificar las variables que se encuentren relacionadas. Una vez que se identifican se puede realizar las siguientes estrategias.

Primero: Combinar las variables en caso de que sea conveniente

Segundo: Eliminar las variables que se relacionan y dejar la que se estime conveniente

Tercero: Disminuir el conjunto de datos con el cual se está trabajando.

3.2.5. TRANSFORMACIÓN DE VARIABLES

Para algunos casos, la transformación de variables puede ser útil para entender mejor los modelos o para mejorar la calidad de información respecto al problema.

La transformación de variable puede ser:

- ***Transformación de variables binarias.***

Corresponde a transformar variables discretas y numéricas a dicotómicas, es decir, clasificar según los parámetros que el investigador estime conveniente (Chiu, 2008).

- ***Transformación de variables numéricas a variables categóricas.***

Corresponde a asignarles una categoría a cierto grupo de datos. Por lo general esto pasa de tener datos numéricos a transformarlos en datos categóricos. Esto con el fin de tener información de mayor calidad al momento de crear un modelo estadístico.

En algunos casos cuando las variables poseen una alta relación entre ellas, se puede combinar las variables creando una nueva que pueda entregar más información para un modelo.

3.2.6. METODOLOGÍAS DE SELECCIÓN DE VARIABLES

A continuación, se definen 3 metodologías para seleccionar variables de un modelo logístico, las cuales corresponden a "*forward*", "*Backward*" y "*Stepwise*".

- ***Forward***

Es un método para seleccionar variables de un modelo estadístico (statisticalecology, 2012), el cual consiste en introducir secuencialmente en el modelo las variables. La primera variable que se introduce es la de mayor correlación (+ o -) con la variable dependiente. Dicha variable se introducirá en la ecuación solo si cumple el criterio de entrada. A continuación, se considerará la variable independiente cuya correlación parcial sea la mayor y que no esté en la ecuación. El procedimiento termina cuando ya no quedan variables que cumplan el criterio de entrada.

- ***Backward***

Es un método para seleccionar variables de un modelo estadístico en donde se introducen todas las variables en la ecuación (statisticalecology, 2012) y después se van excluyendo una tras otra. En cada etapa se elimina la variable menos influyente según el contraste individual.

- ***Stepwise***

Este método es una combinación de los procedimientos anteriores (statisticalecology, 2012). En cada paso se introduce la variable independiente que no se encuentre ya en la ecuación y que tenga la probabilidad para F más pequeña. Las variables ya introducidas en la ecuación de regresión pueden ser eliminadas del modelo. El método termina cuando ya no hay más variables candidatas a ser incluidas o eliminadas.

3.2.7. CRITERIO DE INFORMACIÓN DE AKAIKE (AIC)

Corresponde a un criterio (statologos, 2021) que sirve para poder comparar la calidad de modelos estadísticos. En donde a menor sea el valor del AIC, mejor será la calidad del modelo al compararlo con los demás.

La fórmula para calcular ese criterio corresponde a la siguiente:

$$AIC = -2 \cdot Ln(\hat{V}) + 2k$$

Ecuación 3: *Fórmula para calcular el AIC de un modelo. Fuente: Elaboración propia.*

En donde:

- $\hat{V}$ es la función de máxima verosimilitud del modelo.
- K es el número de parámetros en el modelo, incluyendo el intercepto.

3.3. INTELIGENCIA ARTIFICIAL

Según la rae, la definición de inteligencia corresponde a "*Facultad de la mente que permite aprender, entender, razonar, tomar decisiones y formarse una idea determinada de la realidad*." y para artificial corresponde a "*Que ha sido hecho por el ser humano y no por la naturaleza.*" (RAE, 2022). En otras palabras, se puede definir inteligencia artificial como Maquinas que pueden ser hardware y/o softwares que tienen la capacidad de aprender y tomar decisiones, es decir, solucionar problemas.

La inteligencia artificial puede estar relacionada desde la robótica más avanzada hasta el dispositivo más cotidiano como un smartphone, como, por ejemplo, el uso de mapas y traslado en los teléfonos actuales como el Google maps, entre otros.

El informático Jhon McCarthy fue el primer erudito que mencionó el concepto de inteligencia artificial en un seminario conocido como "conferencia de Dartmouth", en donde él acuñó por primera vez este concepto, definiendo la inteligencia artificial como "la ciencia y la ingeniería de hacer máquinas inteligentes" (bbvaopenmind, 2016)

3.3.1. MACHINE LEARNING

El aprendizaje automático, conocido como "*machine learning*" en inglés, es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender patrones y tomar decisiones a partir de datos, sin ser programadas explícitamente para realizar una tarea específica (Berry, 2004). En lugar de seguir reglas y programación rígida, el aprendizaje automático permite que las máquinas mejoren su rendimiento a medida que se les proporciona más información.

El proceso de aprendizaje automático implica entrenar un modelo utilizando datos históricos o ejemplos previos. El modelo utiliza estos datos para identificar patrones y relaciones en los datos, y luego puede aplicar ese conocimiento para hacer predicciones o tomar decisiones sobre nuevos datos no vistos previamente. El objetivo es que el modelo generalice de manera efectiva, lo que significa que pueda funcionar bien en nuevos datos a los que no ha sido expuesto anteriormente.

Hay varios enfoques y algoritmos dentro del aprendizaje automático, que se pueden clasificar en tres categorías principales (Berry, 2004):

- **Aprendizaje Supervisado:** En este enfoque, el modelo se entrena utilizando un conjunto de datos que contiene ejemplos etiquetados, es decir, datos con respuestas o resultados conocidos. El objetivo es que el modelo aprenda a mapear las entradas a las salidas correctas, de manera que pueda hacer predicciones precisas sobre nuevos datos.

- **Aprendizaje No Supervisado:** Aquí, el modelo se entrena con datos que no tienen etiquetas o respuestas predefinidas. El objetivo es que el modelo encuentre patrones y estructuras en los datos, como agrupamientos o relaciones subyacentes.

- **Aprendizaje por Refuerzo:** En este enfoque, un agente aprende a tomar decisiones en un entorno determinado para maximizar una recompensa acumulada. El agente realiza acciones y recibe retroalimentación positiva o negativa según el resultado de esas acciones. A lo largo del tiempo, el agente aprende a tomar decisiones más efectivas para maximizar la recompensa.

El aprendizaje automático tiene un amplio campo de aplicación, que va desde la clasificación de correos electrónicos no deseados (spam) hasta la detección de fraudes en transacciones financieras, la recomendación de productos en línea, la visión por computadora, el procesamiento del lenguaje natural, etc.

3.3.3. RED NEURONAL

Las redes neuronales es un tipo de IA, cuyo fin corresponde a imitar el comportamiento de un cerebro humano (Amazon AWS, 2018). Esta herramienta es capaz de aprender por ensayo y error, también puede recurrir de ayuda (intervención humana) que le pueda enseñar sobre algún determinado problema. El objetivo de las redes neuronales es que estas puedan analizar un problema, puedan crear una predicción del problema y no depender de la intervención humana.

Las redes neuronales están compuestas por varios nodos, cuyos nodos se conocen como capas, cuyas capas de conocen como unidades de procesamientos. En las capas iniciales se ingresa la información, en las capas medias se procesa la información y en las capas finales se entrega un resultado respecto a la información procesada.

Las capas se componen por capas iniciales, en donde se entrega la información, capas ocultas, en donde ocurre el razonamiento y la capa final es en donde se obtiene el resultado del modelo.

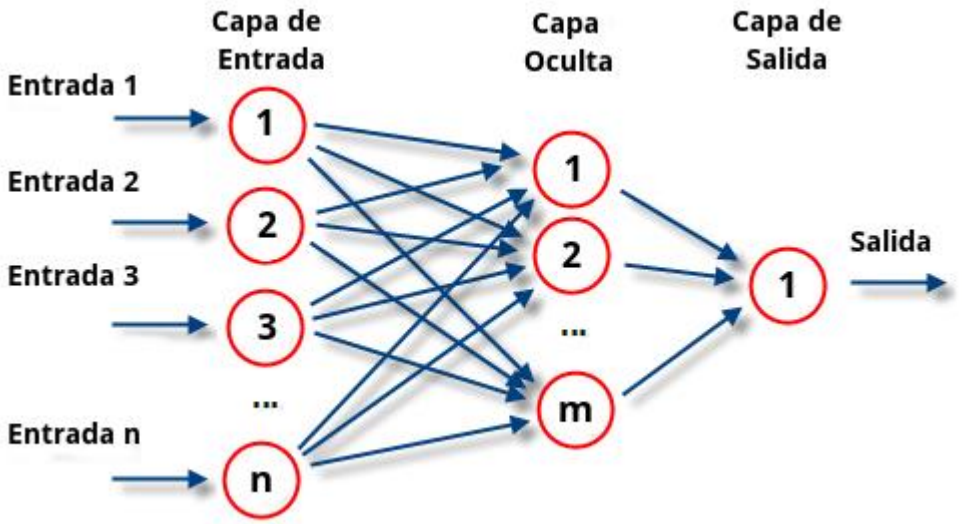


Imagen 1: *Esquema que corresponde a una red neuronal. Para este esquema corresponde a un modelo binario, dado que la capa final posee un nodo. Fuente:* https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/

A continuación, se describirá la composición de las neuronas.

3.3.3.1 CAPAS DE ENTRADAS

Corresponde a la capa inicial de las neuronas. En esta capa es por donde ingresan las variables de un conjunto de base (Amazon AWS, 2018). Cabe a destacar, que la cantidad de neuronas debe coincidir con la cantidad de variables que contenga la base de datos de estudio.

3.3.3.2 CAPAS OCULTAS

Las capas ocultas corresponden a todas las capas que hay entre la capa de entrada y la de salida. En las capas ocultas (Amazon AWS, 2018), se encuentra conformados por varios nodos o neuronas, cuyas neuronas lo que hacen es realizar una suma ponderada y aplicar una función de salida, que se conoce como función de activación. Entre cada neurona posee un peso entre cada neurona. Entre mayor sea la cantidad de neuronas y capas ocultas, mayor será el nivel de profundidad de la neurona.

En la siguiente ecuación se explica de forma matemática lo que ocurre dentro de las capas ocultas

$$y = f\left(\sum_i w_i x_i - T\right)$$

En donde y corresponde al resultado que sale de la función, w_i corresponde al peso estimado que el modelo les entrega a las variables, x_i corresponde a las variables que va ingresando a la neurona y T corresponde al sesgo, parámetro enfocado en corregir o disminuir errores.

Dentro de las funciones de activaciones que se aplican en los modelos de redes neuronales, estas dependerán del tipo de problema. Algunos ejemplos de estas funciones corresponden a la función tangente hiperbólica, función sigmoidal, función Relu, etc.

3.3.3.3 CAPA DE SALIDA

La capa de salida corresponde a la capa final del modelo (Amazon AWS, 2018), cuyo propósito es entregar el resultado final del razonamiento aplicado en las capas ocultas. En esta capa, también se aplica la función de activación.

La arquitectura de la capa final no es al azar (Amazon AWS, 2018), ya que esta va a depender del fin que se estime. Por ejemplo, para los modelos que buscan una predicción binaria, se recomienda que la capa final debe tener una sola neurona y su función de activación debe ser "sigmoide".

3.3.3.5 ENTRENAMIENTO

El entrenamiento (Ruiz, 17 de Marzo de 2019)consiste en ejecutar el modelo de red neuronal con los datos de entrenamiento y con los datos de pruebas. En esta sección se sugiere dividir la base de datos con la cual se está trabajando. Una parte para el uso de entrenamiento y otro por el uso de prueba.

Dentro del entrenamiento, existen las siguientes variables que hay que tener en cuenta para llevarlo a cabo, de los cuales corresponde a los siguientes:

Receta: Corresponde a los pasos (Ruiz, 17 de Marzo de 2019) previos que se deben aplicarles a los datos antes de ejecutarlo al modelo de red neuronal. Es decir, esto correspondería al procesamiento de datos antes de llevarlos al modelo en sí.

Modelo: Corresponde al modelo de red neuronal, es decir, la arquitectura de trabajo con la cual se va a trabajar.

Épocas: Las épocas corresponde a la cantidad (Ruiz, 17 de Marzo de 2019) de veces que se va a ejecutar el modelo. Por ejemplo, si el número de épocas corresponde a 10, el modelo de red neuronal realizará 10 actividades para desarrollarse.

Lote de datos: Dentro del entrenamiento, uno puede escoger la cantidad de datos con los cuales vaya a trabajar la red neuronal. Estos son escogidos al azar y son ejecutados por el modelo de red neuronal.

No existe una receta para saber cuánto es lo suficiente para entrenar el modelo. Simplemente se debe crear el plan de pruebas y dependiendo de los resultados, estos se van ajustando.

Un detalle que es importante a considerar sobre el entrenamiento es evitar el sobreajuste, por lo cual, se debe diseñar un entrenamiento equilibrado que pueda entregar un buen modelo.

3.3.3.6 SOBREAJUSTE

El sobreajuste (AWS, 2020) corresponde cuando un modelo de red neuronal empieza a memorizar datos, en vez de pensar cómo resolverlos. Esto ocurre por lo general cuando los modelos entrenan más de lo necesario. Es por esto que es necesario eliminar el racionamiento de "Entre más entrenamiento, más eficiente será el modelo". Una forma de verificar que hay sobreajuste, cuando la función de pérdida de validación empieza a ser más grande que la función de pérdida del entrenamiento.

3.4 RED NEURONAL CONVOLUCIONAL

Las redes neuronales convolucionales, son un tipo de red neuronal artificial que trabaja principalmente en base de datos enfocada en objetivos, como videos, imágenes, canciones, etc.

Las redes neuronales convolucionales (CNN por sus siglas en inglés, Convolutional Neural Networks) son un tipo de red neuronal artificial especialmente diseñada para el procesamiento de datos que tienen una estructura de cuadrícula, como imágenes. Estas redes son muy efectivas en tareas relacionadas con el reconocimiento visual, como la clasificación de imágenes y la detección de objetos.

La característica principal de las CNN es la capa de convolución, que aplica operaciones de convolución a la entrada y pasa el resultado a través de una función de activación no lineal. Esto permite que la red aprenda automáticamente características relevantes de los datos de entrada, como bordes, texturas y patrones, a diferentes escalas y niveles de abstracción.

Además de las capas de convolución, las CNN suelen incluir capas de agrupación (pooling) para reducir la dimensionalidad de las características extraídas y capas completamente conectadas al final para la clasificación o regresión final.

Las CNN han demostrado ser muy exitosas en una amplia gama de aplicaciones de visión por computadora, como reconocimiento de objetos, segmentación semántica, detección de

rostros, entre otras, y también se han utilizado en tareas que no son de visión por computadora, como procesamiento de secuencias temporales en datos de audio y texto.

3.5 EVALUACIÓN DE MODELOS

Existen diversas formas de como evaluar modelos, todo eso va a depender del contexto del cual se está trabajando.

Para los modelos encargados de predecir situaciones binarias, es decir, variables objetivas que corresponden a valores por ejemplo “si” o “no”, “1” ó “0”, “bueno” o “malo”, etc. Es decir, en donde la variable objetivo sea dicotómica. Se puede utilizar las siguientes herramientas (Amazon AWS, 2018):

- Matriz confusión
- Exactitud
- Precisión
- Sensibilidad
- F1-Score

Estas herramientas sirven para evaluar los modelos de predicción binarias, independientes si el modelo es estadístico tradicional o de inteligencia artificial.

3.5.1. MATRIZ CONFUSIÓN

La matriz confusión corresponde a una tabla que entrega la relación entre los datos que predijo el modelo, versus los reales (Arce, 2019). El objetivo de la tabla es identificar de los valores que predijo el modelo, cuales corresponden a falsos positivos y verdaderos positivos.

En esta tabla se encuentra los “Verdaderos positivos”, “Verdaderos negativos”, “Falsos positivos” y “Falsos negativos”

- ***Verdadero positivos***: Corresponde a los datos que fueron predichos de forma positiva por un modelo. Es decir, la predicción coincide con la realidad.
- ***Verdadero negativo***: Corresponde a los datos que fueron predichos de forma negativo en la clasificación de un modelo. Es decir, la predicción dice que es negativo y en la realidad también lo es.
- ***Falso Positivo***: Ocurre cuando un modelo predice que un dato es falso, pero en la realidad es positivo. Es decir, la predicción no concuerda con el real.
- ***Falso Negativo***: Ocurre cuando un modelo de predicción predice que la variable objetivo es falsa, pero resulta que en la realidad la variable es verdadera.

Además, con esa información permite obtener otras herramientas de análisis como la exactitud, la precisión, la sensibilidad y el F1-Score.

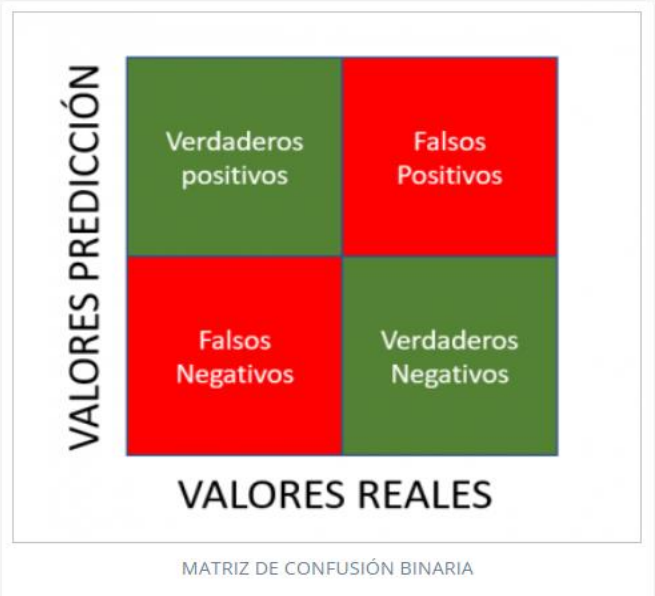


Imagen 3: *Ejemplo de matriz confusión. Fuente: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/ .*

3.5.2. PRECISIÓN (*ACCURACY*)

La precisión es una métrica que permite evaluar el rendimiento de un modelo de predicción dicotómico (Arce, 2019). Esta herramienta mide la proporción de predicciones correctas versus el total de datos.

A continuación, se adjunta la fórmula para calcular la precisión:

$$Precisión = \frac{Verdaderos\ positivos + Verdaderos\ negativos}{Total\ de\ los\ datos}$$

Ecuación 4: *Fórmula correspondiente a la Precisión de un modelo binominal. Fuente: Elaboración propia.*

3.5.3. PRECISIÓN (*PRECISION*)

El valor predictivo positivo (VPP) corresponde a la proporción de los datos acertado por el modelo versus la proporción de verdaderos reales que se encuentra en la muestra del estudio (Arce, 2019). Es decir, del total de datos positivos con los cuales se trabajó con el modelo, cuales si predijo el modelo. Esta métrica permite medir la calidad del modelo a la hora de predecir.

A continuación, se adjunta la fórmula para calcular la precisión de un modelo predictivo.

$$Precision = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ positivos}$$

Ecuación 5: *Fórmula correspondiente al valor predictivo positivo de un modelo binominal. Fuente: Elaboración propia.*

3.5.4. SENSIBILIDAD (*RECALL*)

La sensibilidad ("*recall"* conocida en inglés) es una métrica que se utiliza para medir la capacidad de un modelo de clasificación para identificar correctamente todos los casos positivos en un conjunto de datos (Arce, 2019). Es decir, mide cuántos de los casos positivos reales el modelo es capaz de capturar.

A continuación, se adjunta la fórmula para calcular la sensibilidad de un modelo:

$$Sensibilidad = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ negativos}$$

Ecuación 6: *Fórmula correspondiente a la Sensibilidad de un modelo binominal. Fuente: Elaboración propia.*

3.5.5. F1-SCORE (*F SCORE*)

La métrica f1-score es una métrica que busca combinar la precisión con la sensibilidad de un modelo en una sola puntuación (Arce, 2019). Esta métrica es el resultado de la media armónica de la precisión y la sensibilidad. Esta medida ayuda a identificar si el modelo predictivo cumple con el objetivo de tener un equilibrio entre la Precisión y la Sensibilidad.

A continuación, se adjunta la fórmula para calcular el F1-Score:

$$Puntaje\ F1 = \ 2 * \frac{Precision * Sensibilidad}{Precision + Sensibilidad}$$

Ecuación 7: *Fórmula correspondiente al puntaje F1 Score de un modelo binominal. Fuente: Elaboración propia.*

Por último, no existe la manera exacta determinar que valores puede demostrar si el modelo de predicción es bueno o no. Todo va a depender del contexto del problema.

Lo que si se recomienda es utilizar estas cuatro métricas en conjunto para poder concluir si un modelo es aceptable o no. También estas métricas pueden ayudar a comparar.

3.6. METODOLOGÍA DE ANÁLISIS DE DATOS

3.6.1 KDD

La metodología conocida como KKD (*Knowlegde Discovery*), es una metodología de trabajo enfocado en el análisis de datos, cuyo fin es obtener patrones, modelos, inferencias, etc. (Berry, 2004). Que permitan generar nuevos conocimientos.

Esta metodología se compone de fases de trabajo de las cuales se detallan a continuación:

1. Identificación del problema y del dominio de trabajo.
2. Creación del conjunto de datos.
3. Pre-procesamiento de los datos.
4. Reducción de datos y proyección.
5. Formulación de los objetivos del PROCESO KDD (Descubrimiento de Conocimiento en Bases de Datos)
6. Exploración de análisis, modelo y selección de la hipótesis.
7. Minería de datos
8. Interpretación de los patrones encontrados
9. Descubrimiento de nuevo conocimiento

El objetivo principal de esta metodología es obtener conocimiento e interpretación de los datos

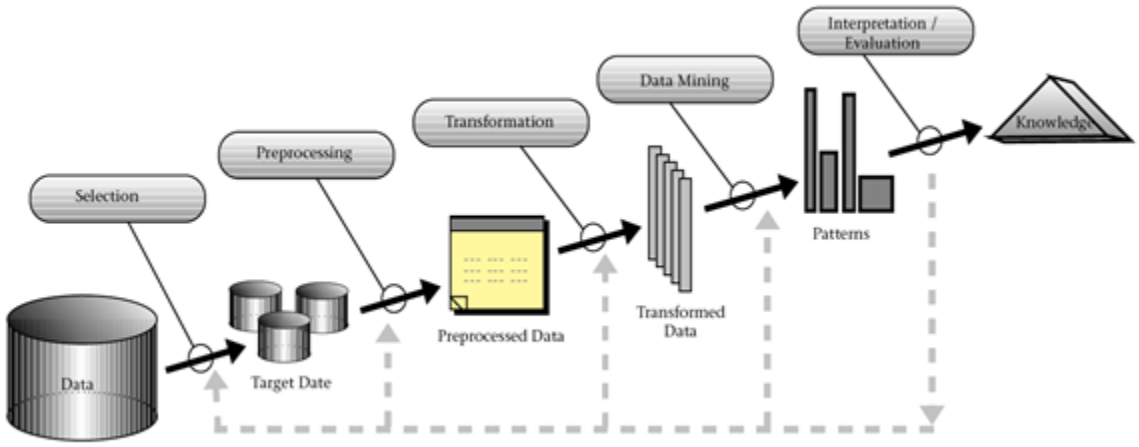


Imagen 4: *Ejemplo procedimiento KDD. Fuente: https://www.researchgate.net/figure/Figura-2-El-proceso-KDD-CRISP-DM-por-su-parte-corresponde-a-una-metodologia-que_fig1_268687479.*

3.6.1.1 IDENTIFICAR EL PROBLEMA Y ENTENDER EL NEGOCIO

Cada dato y base de datos, posee información que puede ayudar a comprender un contexto más grande. Es por esto, que antes de iniciar, es recomendable entender de que se trata el negocio que hay detrás de los datos e identificar cual es el problema que se desea comprender al momento de llegar a utilizar esta metodología. Además, identificar cuáles son los modelos predictivos, o modelo que puedan explicar y/o predecir el comportamiento que se desee estudiar.

3.6.1.2 SELECCIÓN DE INFORMACIÓN

Una vez que se tenga listo el conocimiento del negocio, se procede a seleccionar las bases de datos con las cuales se van a llevar a cabo el trabajo.

3.6.1.3 PREPROCESAMIENTO

Luego que se tenga la información respecto al trabajo de datos, se procede a preprocesar los datos, es decir, estudiarlos, verlos, identificar cuales influyen, cuales son adecuados para trabajar, cuales no, identificar el tipo de dato, etc.

3.6.1.4 TRANSFORMACIÓN

Con el fin de mejorar la información que entregue cada dato, si el contexto lo requiere, se realiza una transformación de datos.

3.6.1.5 MINERÍA DE DATOS

La minería de datos es una metodología iterativa, que busca encontrar un modelo de datos que permita predecir o explicar cómo solucionar el problema que se plantea al principio del estudio. Este método se explicará con más detalles más adelante.

3.6.1.6 INTERPRETACIÓN DE PATRONES Y EVALUACIÓN DE MODELOS

Esta etapa culmine, corresponde a las conclusiones que generaron la creación de un modelo en donde se puede aprender de él o se puede utilizar el modelo que se creó dentro del desarrollo. Todo esto con el fin de generar valor en el negocio, en el cual se está trabajando.

3.6.2 MINERÍA DE DATOS

La metodología (Berry, 2004) con la cual se va a trabajar corresponde a la, CRISP-DM (*Cross-Industry Standard Process for Data Mining*, su nombre en inglés) es decir, análisis de minería de datos. Cuyo proceso consta con la realización de 5 pasos para poder realizar un análisis. De los cuales corresponden a los siguientes:

1. Leer Datos
2. Pre-Procesar datos
3. Entrenar modelo
4. Evaluar modelo
5. Mostrar resultados

Cuya descripción de pasos se encuentra en los siguientes tópicos.

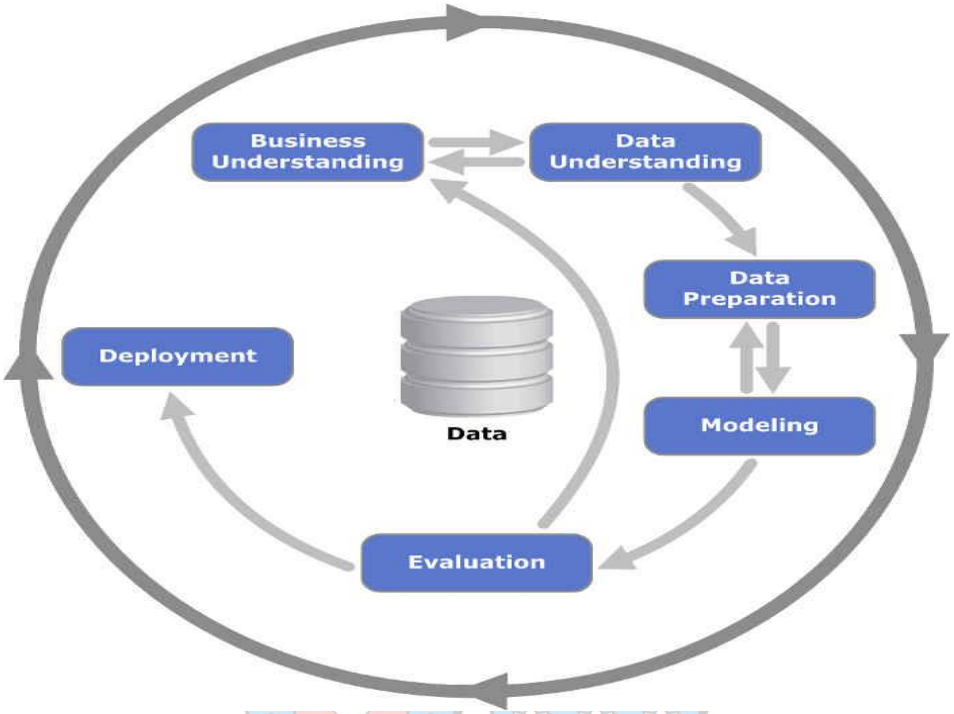


Imagen 5: *Esquema Cross-Industry Standard Process for Data Mining. Fuente: https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/.*

3.6.2.1. LEER DATOS

Este proceso corresponde a la lectura de la base de datos que se va a trabajar. En este paso se prioriza verificar como pueden interactuar las variables, identificar que variables son primordiales y hacer un descarte lógico según el contexto en donde se trabaje.

Es importante destacar que antes de realizar este procedimiento es necesario comprender el negocio (o el contexto). Esto con el fin de poder tomar decisiones más fáciles respecto a que hay que hacer con los datos.

3.6.2.2. PREPROCESAR DATOS

En este paso, se procede a construir una "receta", es decir, el código o los pasos previos necesarios para entrenar un modelo según sea el contexto. La estrategia que se recomienda siempre hacer corresponde a la división de los datos. Es decir, un porcentaje importante corresponderá para el entrenamiento del modelo y otra parte corresponde para probar el modelo.

Por lo general este es uno de los pasos más importante, el que lleva más tiempo en hacer y de los cuales puede provocar mayores errores según un estudio de "AWS Amazon" (Amazon AWS, 2018)

3.6.2.3. ENTRENAR MODELO

En este paso corresponde al entrenamiento del modelo el cual se puede utilizar diferentes técnicas y herramientas para trabajar en la minería de datos. Cuyas herramientas de trabajos pueden ser modelos estadísticos hechos con softwares enfocados en eso o con modelos de machine learning.

Cabe a destacar que se recomienda iterar hasta encontrar el modelo que pueda satisfacer las necesidades que se busca o por lo menos no quedarse con el primer intento.

3.6.2.4. EVALUAR MODELO

Corresponde a la evaluación del modelo según el contexto que se haya definido el propósito de estos. El objetivo es que pueda responder con las demandas o necesidades que se hayan establecidos al principio.

Existen diversas maneras de como evaluar un modelo, todo va a depender de cuál es el contexto del problema.

3.6.2.5. COMPARTIR RESULTADOS O INTEGRACIÓN DE MODELO

Una vez finalizada la evaluación de los modelos, se comparte los resultados y se dialoga los aprendizajes y/o conclusiones que estos puedan entregar. La idea principal es implementar el modelo de trabajo que se haya generado o encontrar abrir la oportunidad de crear otro modelo que pueda ayudar mejor con el contexto inicial.

3.7.0 BENEFICIOS DE LA MINERÍA DE DATOS

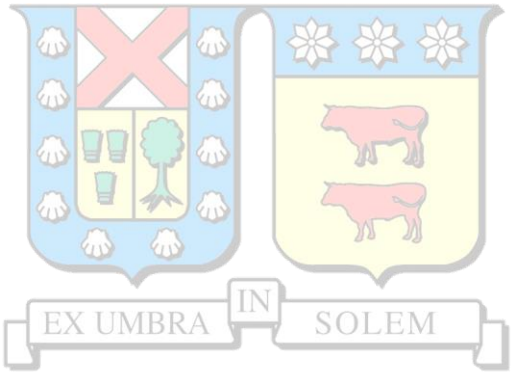
Son diversos factores en los cuales puede beneficiar (apiumhub, 2022) la minería de datos en las industrias y/u organizaciones. Los principales corresponden a los siguientes:

- Puede ayudar a las empresas a obtener información basada en el conocimiento.
- Puede implantarse tanto en sistemas nuevos como en plataformas existentes
- La minería de datos ayuda a las organizaciones a realizar ajustes rentables en el funcionamiento y la producción.
- Facilita la predicción automática de tendencias y comportamientos, así como el descubrimiento automático de patrones ocultos.
- La minería de datos ayuda en el proceso de toma de decisiones.
- Es un proceso rápido que facilita a los usuarios el análisis de grandes cantidades de datos en menos tiempo.

Si se habla de casos de uso en las industrias puede impactar a diversas áreas, todo va a depender al contexto que se desea aplicar. A continuación, algunos beneficios que puede impactar en el mundo industrial:

- Marketing: La minería de datos puede increm,entar la eficiencia al momento de segmentar clientes. Puede predecir los comportamientos de los clientes y determinar cuáles son los factores que pueden influir en la toma de decisión de compras.
- La banca: Puede ayudar a comprender los factores que influyen en la calificación de riesgo de un individuo que desee solicitar un crédito. Además, puede ayudar a detectar fraudes bancarios y detectar potenciales clientes.
- Educación: En las universidades, la minería de datos ha ayudado a identificar los estudiantes que son más propenso a abandonar sus estudios. Por lo cual, ha sido una herramienta necesaria para mejorar las decisiones de los profesores con el fin de aumentar la retención de estudiantes.
- Medicina: Ha permitido ayudar a la predicción de diagnósticos de pacientes y de mejorar el rendimiento de tratamientos contra enfermedades.

Son muchos los ejemplos en donde la minería de datos ha aportado en las industrias, en donde es la razón del porque cada vez son más empresas las que toman la decisión de aplicar esta estrategia para ser más competitivas.



CAPÍTULO IV: DESARROLLO

4.0HERRAMIENTAS DE TRABAJOS.

A continuación, se procederá aplicar el modelo de trabajo de KKD, el cual corresponde a los siguientes pasos:

- Identificación del problema y del dominio de trabajo.
- Creación del conjunto de datos.
- Pre-procesamiento de los datos.
- Reducción de datos y proyección.
- Formulación de los objetivos del PROCESO KDD (Descubrimiento de Conocimiento en Bases de Datos)
- Exploración de análisis, modelo y selección de la hipótesis.
- Minería de datos
- Interpretación de los patrones encontrados
- Descubrimiento de nuevo conocimiento

DEFINICIÓN DEL PROBLEMA

El objetivo es crear un modelo de optimización, a partir de un análisis de datos correspondientes al historial de clientes de un banco. La idea es poder identificar cuales de los clientes tiene una mayor posibilidad de escoger un aumento de cupo de tarjeta de crédito, mediante la información entregada por el historial. Clientes que deciden aumentar su cupo de tarjeta de crédito, implica que puede aumentar los ingresos del banco a través de intereses (CMF, 2020).

Para poder desarrollar esta problemática, se pretenderá crear modelos predictivos a través de algoritmos correspondientes a la regresión logística y a 3 diferentes tipos de redes neuronales, con el fin de poder visualizar las diferencias que estas puedan tener.

La herramienta con la cual se va a desarrollar este estudio corresponde al software Rstudio, en donde se les instaló librerías que se pueda trabajar con redes neuronales. De las cuales corresponde a las siguientes:

- ***Reticulate***:

  La librería reticulate (Kalinowski, 2022) permite la integración de código en Python dentro del entorno de R. Esto es útil cuando se quiere utilizar funcionalidades específicas de Python en un proyecto de R. Puede ser especialmente útil cuando se trabaja con librerías de aprendizaje profundo escritas en Python, como TensorFlow y Keras (Rosa, johan-rosa.com/, 2020).

- ***Tensorflow***:

  Tensorflow es una librería de código abierto (Rosa, johan-rosa.com/, 2020) para cómputo numérico que se utiliza principalmente para construir y entrenar modelos de aprendizaje profundo. TensorFlow proporciona herramientas para implementar modelos de redes neuronales, tanto para investigación como para producción. La integración de TensorFlow en R a través de la librería tensorflow permite a los usuarios de R aprovechar las capacidades de TensorFlow directamente desde R.

- ***keras:***

  keras (Rosa, johan-rosa.com/, 2020) es una interfaz de alto nivel para construir y entrenar modelos de aprendizaje profundo. En R, la librería keras actúa como una interfaz para TensorFlow, permitiendo a los usuarios de R construir y entrenar modelos de manera más sencilla utilizando funciones de alto nivel. Esto facilita la construcción de modelos complejos de aprendizaje profundo sin tener que escribir código de bajo nivel en TensorFlow.

- ***Yardstick:***

  Yardstick (Kuhn, 2020) es una librería que proporciona herramientas para evaluar el rendimiento de modelos estadísticos y de aprendizaje automático. Se centra en la evaluación de modelos a través de diversas métricas, como

precisión, sensibilidad, especificidad, entre otras. Esta librería es útil para medir el rendimiento de modelos de clasificación y regresión.

CONJUNTO DE DATOS

Para el desarrollo de este estudio, se utilizó la base de datos correspondiente al concurso "Batalla de datos", entregado por el banco Itaú (Banco ITAU, 2019). Esta base de datos posee 1.583.258 observaciones y un total de 52 variables. Cabe de destacar que esta información corresponde a la interacción de los usuarios de un determinado mes. Información que se puede encontrar más detallada en el concurso.

Para el desarrollo de los modelos de predicciones se utilizó el programa Rstudio y se descargó las librerías que se mencionaron anteriormente para trabajar en modelo logísticos y modelos de Redes neuronales.

Para el modelo de red neuronal se utilizó las librerías de Tensorflow y Keras.

A continuación, se adjunta tabla correspondiente a todos las variables de la base de datos:

Variable	Descripción
id	Numero identificador de cliente
Edad	Edad
RentaAltamira	Renta del cliente
RentaEstiamda	Renta Estimada
Q_Prods	Cantidad de productos vigentes
SegmentoRiesgo	Segmento de riesgo
MIX_PROD	Codigo de productos vigentes
CCTE_Q_OP	Cantidad de operaciones de cuenta corriente
CCTE_SALDO_CLP	Saldo de cuenta corriente
CCUOTA_Q_OP	Cantidad de crédito de consumo
CCUOTA_MONTO_SOLICITADO	Monto solicitado crédito de consumo
CCUOTA_SALDO_CAPITAL	Saldo capital crédito de consumo
CRENEG_Q_OP	Cantidad de créditos renegociados
CRENEG_MONTO_SOLICITADO	Monto solitidado crédito renegociado
CRENEG_SALDO_CAPITAL	Saldo capital crédito renegociado
DAP_Q_OP	Cantidad de operaciones de DAP
DAP_SALDO_CLP	Saldo depósito a plazo
FFMM_Q_OP	Cantidad operaciones fondos mutuos
FFMM_SALDO_PESOS	Saldo fondo mutuos
HIPO_Q_OP	Cantidad de operaciones de hipotecario
HIPO_SALDO_CLP	Saldo de hipotecario

LCRED_Q_OP	Cantidad de operaciones de línea de crédito
LCRED_CUPO	Cupo de línea de crédito
LCRED_SALDO_CLP	Saldo línea de crédito
SEG_Q_OP	Cantidad operaciones de seguro
TC_Q_OP	Cantidad de operaciones de tarjeta de crédito
TC_CUPO	Cupo tarjeta de crédito
TC_SALDO_CLP	Saldo de tarjeta de crédito
GIROATM_Q_OP_MES	Cantidad operaciones giros cajero automático
GIROATM_MONTO_MES_CLP	Montos de giro cajero automático
COMPRADEB_Q_OP_MES	Cantidad de operaciones compra debito
COMPRADEB_MONTO_MES_CLP	Monto operaciones debito
AVNC_TC_Q_OP_MES	Cantidad de operaciones de avance de tarjeta de crédito
AVNC_TC_MONTO_MES_CLP	Monto operaciones avance de tarjeta de crédito
COMPRA_TC_Q_OP_MES	Cantidad de operaciones tarjeta de crédito
COMPRA_TC_MONTO_MES_CLP	Monto compra de tarjeta de crédito
PAS_Q_OP	Cantidad de operaciones de abono de remuneración en cuenta
CVISTA_Q_OP	cantidad de operaciones de cuenta vista
CVISTA_SALDO_CLP	saldo operaciones cuenta vista
CVISTA_SALDO_AVG_CLP	saldo promedio operaciones cuenta vista
PAT_Q_OP	cantidad de pago automático tarjeta de crédito
PAC_Q_OP	cantidad de pago automático cuenta corriente
DIGITAL_CLASE	score digital
IIR_CLASE	índice de relacionamiento
SBIF_COMER_MONTO_CLP	Deuda de créditos comerciales en sistema financiero
SBIF_CONSUMO_MONTO_CLP	Deuda de créditos de consumo en sistema financiero
SBIF_Q_ACREED_CONSUMO	Cantidad de acreedores banco en sistema financiero
SBIF_HP_MONTO_CLP	Deuda de créditos hipotecarios en sistema financiero
SBIF_LDISP_MONTO_CLP	Disponibilidad de cupo de tarjeta de crédito y líneas de crédito en sistema financiero
aumento_cupo	Decisión de si el cliente aumentó el cupo. 1 de si y 0 no
sexo_id	Sexo del cliente. 1 hombre, 0 mujer

Tabla 1: *Correspondiente al diccionario de variables de la base de datos. Fuente: Elaboración propia*.

PREPROCESAMIENTO

Lo primero que se hizo en esta etapa de la metodología de trabajo, se hicieron los primeros ajustes de coherencia según el contexto. Dado que la base de datos posee 52 columnas y 1.583.258 filas, eso implica, que la base de datos posee 82.329.416. Lo cual implica que el software tendría que hacer "mucho trabajo". Es por esto que, con el fin de optimizar el análisis del software, se filtraron los datos correspondientes a un mes determinado. Reduciendo las filas en 229.652.

Por último, mediante un análisis cualitativo, se eliminaron variables que no son relevantes para este estudio, quedando 36 variables para poder crear los modelos predictivos. Cuyas variables de adjuntan a continuación:

Variable	Descripción
Edad	Edad
RentaAltamira	Renta del cliente
RentaEstiamda	Renta Estimada
Q_Prods	Cantidad de productos vigentes
SegmentoRiesgo	Segmento de riesgo
MIX_PROD	Codigo de productos vigentes
CCTE_Q_OP	Cantidad de operaciones de cuenta corriente
CCTE_SALDO_CLP	Saldo de cuenta corriente
CCUOTA_Q_OP	Cantidad de crédito de consumo
CCUOTA_SALDO_CAPITAL	Saldo capital crédito de consumo
CRENEG_Q_OP	Cantidad de créditos renegociados
CRENEG_MONTO_SOLICITADO	Monto solitidado crédito renegociado
CRENEG_SALDO_CAPITAL	Saldo capital crédito renegociado
DAP_Q_OP	Cantidad de operaciones de DAP
FFMM_Q_OP	Cantidad operaciones fondos mutuos
HIPO_SALDO_CLP	Saldo de hipotecario
LCRED_CUPO	Cupo de línea de crédito
LCRED_SALDO_CLP	Saldo línea de crédito
TC_Q_OP	Cantidad de operaciones de tarjeta de crédito
TC_CUPO	Cupo tarjeta de crédito
TC_SALDO_CLP	Saldo de tarjeta de crédito
GIROATM_MONTO_MES_CLP	Montos de giro cajero automático
COMPRADEB_MONTO_MES_CLP	Monto operaciones debito

AVNC_TC_Q_OP_MES	Cantidad de operaciones de avance de tarjeta de crédito
COMPRA_TC_Q_OP_MES	Cantidad de operaciones tarjeta de crédito
COMPRA_TC_MONTO_MES_CLP	Monto compra de tarjeta de crédito
CVISTA_Q_OP	cantidad de operaciones de cuenta vista
PAT_Q_OP	cantidad de pago automático tarjeta de crédito
PAC_Q_OP	cantidad de pago automático cuenta corriente
DIGITAL_CLASE	score digital
IIR_CLASE	índice de relacionamiento
SBIF_CONSUMO_MONTO_CLP	Deuda de créditos de consumo en sistema financiero
SBIF_Q_ACREED_CONSUMO	Cantidad de acreedores banco en sistema financiero
SBIF_HP_MONTO_CLP	Deuda de créditos hipotecarios en sistema financiero
aumento_cupo	Decisión de si el cliente aumentó el cupo. 1 de si y 0 no
sexo_id	Sexo del cliente. 1 hombre, 0 mujer

Tabla 2: *Correspondiente a las variables seleccionadas. Fuente: Elaboración propia*.

De las variables, 35 corresponderán a las variables independientes que explicarán la predicción de la variable objetivo (aumento_cupo).

4.1REALIZACIÓN DE METODOLOGÍA MINERÍA DE DATOS.

Se comenzó a trabajar con los datos, con el fin de analizarlos. Mediante el uso de histogramas, gráficos de cajas y matriz de correlación, se procedió a realizar análisis de los datos con el fin de realizar los ajustes necesarios para llevar a cabo los modelos de predicción. Dado que, se trabajó con varias variables, los gráficos fueron dejados en el Anexo de la investigación con el fin de no generar espacio.

Dentro del análisis de datos, se pudo descartar datos que son atípicos y datos que no pertenecen al contexto del estudio. Por lo cual, se pudo reducir la base de datos de trabajo, quedando una base de datos de 36 columnas con 107.628 filas.

Esto debido a que se descartaron todos los datos que no cumplieran con el contexto del problema, cuyas cualidades se describen en la siguiente tabla:

Variable	Restricción de ella
Renta	mayor a $100.000
Renta	Menor a $6.000.000
Cuenta corriente	mayor a $0
Cupo disponible para crédito	Mayor a $0
Línea de crédito	Mayor a $0
Saldo tarjeta de crédito	Mayor a $0

Saldo cuenta vista	Menor a $8.000.000
Saldo de deuda total	Menor a $60.000.000

Tabla 3: *Correspondiente a la restricción de variables. Fuente: Elaboración propia*.

Luego de todo este análisis se trabajó con los modelos predictivos de regresión logística, modelo red neuronal simple, modelo red neuronal profunda y modelo de red neuronal convolucional. En donde se dividió la base de datos en dos conjuntos. El primer conjunto corresponde a la base de datos de entrenamiento, cuya base de datos corresponde el 80% de los datos totales y el 20% restante, corresponde a los datos de pruebas, cuya finalidad será evaluar el modelo que se obtuvo del entrenamiento de los modelos predictivos. Esto se aplicó para todos los modelos de predicción.

Cabe a destacar que en todos los modelos se hizo un ajuste y se estuvo realizando mejoras a los modelos predictivos, hasta conseguir valores aceptables dentro de los gráficos de entrenamiento, es decir evitar el sobreajuste.

4.2 VALORES MATRIZ CORRELACIÓN

A continuación se adjunta los resultados de la evaluación de los modelos en la siguiente tabla:

Modelo	Accurancy	Precision	Recall	F1 Score
Regresión Logística	0,9253	0,9148	0,7470	0,8225
Red neuronal simple	0,8479	0,6851	0,5835	0,6302
Red neuronal profunda	0,9910	**0,9921**	0,9617	0,9767
Red neuronal profunda 2	**0,9936**	0,9877	**0,9786**	**0,9831**
Red neuronal convolucional	0,9814	0,9784	0,9272	0,9521

Tabla 4: *Resumen de la evaluación de los modelos predictivos. Fuente: Elaboración propia*.

Es posible observar que el mejor de los casos podría corresponder al segundo modelo correspondiente a red neuronal profunda. Se puede decir que ese podría ser el mejor modelo para la aplicación de este modelo.

Evaluaciones

Matriz confusión:

Matriz de Confusión

	Predicción: "no aumentó cupo de tarjeta"	Predicción: "aumentó cupo de tarjeta"
No aumentó cupo de tarjeta	17364	88
Aumentó cupo de tarjeta	50	4024

Tabla 5: *Correspondiente a la matriz confusión obtenida del modelo logístico. Fuente: Elaboración propia*.

Con la información entregada por la matriz confusión, se desprende la siguiente información respecto al modelo:

El modelo predijo 17.364 Falsos positivos, es decir, del total de las personas que no aumentaron su cupo de tarjeta de crédito, el modelo pudo predecir 17.364 usuarios.

4.024 verdaderos positivos, es decir, del total de las personas que no aumentaron su cupo de tarjeta de crédito, el modelo pudo predecir 4.024 usuarios.

50 falso negativo, es decir, el modelo dice que 50 personas no aumentarán su cupo de tarjeta de crédito, sin embargo, en la realidad si lo hicieron.

88 falso positivos. Es decir, el modelo dice que 88 personas aumentarán su cupo de tarjeta de crédito. Sin embargo, en la realidad no lo hicieron.

Esto quiere decir que el modelo es capaz de predecir el 99,36% de los datos que se le entrega, de los cuales pueden ser tanto para el caso que la variable objetivo tome el valor positivo o negativo.

La información que se entrega es de los datos que predijo de forma positiva la variable objetivo, el 98,77% corresponde a la realidad. Si se sitúa bajo el contexto en el cual se está trabajando. Quiere decir que de los valores positivos que predijo el modelo (usuarios que hayan aumentado su cupo de la tarjeta de crédito) el 98,77% realmente lo hizo y el otro 1,23% fue un caso erróneo de predicción.

La sensibilidad tomó el valor del 97,86%. Eso quiere decir que de los casos reales de las personas que, si aumentaron su cupo de tarjeta de crédito, el modelo pudo predecir el 97,86%.

4.3COMPARACIÓN DE MODELOS

Para estos dos modelos de predicciones se comparará de dos puntos de vista. Un punto de vista cualitativo enfocado en cómo se ven los modelos en sí. Y otro punto de vista cuantitativo enfocado en los resultados que entregaron ambos modelos.

Comparación cualitativa

Dentro del punto de vista cualitativo, se puede observar una gran diferencia entre el modelo logístico y el modelo de red neuronal.

El modelo logístico se puede observar cuales son las variables que podrían afectar de forma positiva o negativa la probabilidad de que ocurra el aumento de cupo de tarjeta de crédito de un usuario.

En cambio, para el modelo de predicción neuronal, no es posible identificar que variable a simple vista podría influir en aumentar o disminuir la probabilidad de que ocurra el evento.

Esto se debe a que el modelo de red neuronal se puede asociar como una caja negra, es decir, se sabe que ingresan los datos por la capa de entrada, se sabe que estos de procesan en las capas ocultas y luego sale un resultado en la capa de salida.

Dentro de este punto de vista, el modelo predictivo logístico puede identificar cual es la variable que tiene mayor peso al momento de aumentar la probabilidad de que ocurra el suceso.

No obstante, el modelo de red neuronal puede ayudar perfectamente a tomar decisiones de forma rápida. Lo cual, eso podría aumentar la competencia de una empresa. Siempre y cuando, se tenga un modelo confiable.

Creación de modelo de optimización

Con el fin de identificar cuanto es lo que puede ganar un banco utilizando este modelo, se procedió a buscar información relevante correspondiente al banco Itau, ya que, la base de datos corresponde a ese banco en particular. De la memoria anual, correspondiente al año 2022, se puede conocer que el banco ganó aproximadamente unos $141.667.000.000 al año en concepto de interés generado por el uso de tarjeta de crédito por los clientes. Teniendo en cuenta este valor y que además, la empresa posee aproximadamente 400.000 clientes, se puede deducir que en promedio un cliente puede generar en promedio de $29.514 más para el banco por cada uso de tarjeta de crédito en un mes. Por lo tanto, se puede usar este dato para obtener la ganancia que podría generar el aumento de cupo de una tarjeta de crédito correspondiente a un determinado cliente. Con este dato se puede construir una función de optimización, cuya función permitirá identificar cual modelo podría generar un mayor beneficio esperado a la hora de genera un aumento de cupo de un cliente.

Entonces a continuación se adjunta una función de optimización, correspondiente al beneficio esperado al aplicar estos modelos:

$$E = V_P * C_1 + V_N * C_2 + F_P * C_3 + F_N * C_4$$

En donde se describe los componentes de la ecuación:

Coeficiente	valor
Vp	Verdadero positivo
Vn	Verdadero negativo
Fp	Falso positivo
Fn	Falso negativo

Tabla 6: *Correspondiente a los coeficientes de la ecuación. Fuente: Elaboración propia*.

Para los casos en donde el modelo predictivo se haya equivocado en la predicción, es decir, haya generado un falso positivo, se generará una penalización al beneficio esperado,

correspondiente al 45% de lo que se gana en promedio por cada cliente que haya adquirido un aumento de cupo de tarjeta de crédito. Es decir, se asignará un valor de $-\$13.281$. Para los casos que se hayan generado un falso negativo, se le asignará al modelo una penalización de un 25%, es decir un $-\$7.379$. Esto con el fin de evaluar la precisión de un modelo predictivo. Por lo tanto, los coeficientes quedan de la siguiente manera:

Coeficiente	valor
C1	$29.514
C2	$0
C3	$-7.379
C4	$-13.281

Tabla 7: *Correspondiente a los coeficientes de la ecuación. Fuente: Elaboración propia.*

Se Reemplaza los valores en la función optimización del beneficio esperado en cada modelo y se obtiene lo siguiente:

$$E_{reg\ log} = 3.727 * \$29.514 + 16.190 * 0 - 347 * \$7.379 - 1.262 * \$13.281 = \$90.677.543$$

$$E_{red\ neuronal\ bas} = 2.791 * \$29.514 + 15.460 * 0 - 1.283 * \$7.379 - 1.992 * \$13.281 = \$46.450.565$$

$$E_{red\ neuronal\ profunda\ 1} = 4042 * \$29.514 + 17.291 * 0 - 32 * \$7.379 - 161 * \$13.281 = \$116.921.219$$

$$E_{red\ neuronal\ profunda\ 2} = 4024 * \$29.514 + 17.364 * 0 - 50 * \$7.379 - 88 * \$13.281 = \$\mathbf{117.226.658}$$

$$E_{red\ convolucional} = 3.986 * \$29.514 + 17.139 * 0 - 88 * \$7.379 - 313 * \$13.281 = \$112.836.499$$

Gracias a este modelo de optimización, se puede deducir que el mejor modelo para predecir las ganancias que puede generar la predicción de aumento de cupos corresponde al Segundo modelo de red neuronal profunda, dado que este es el que genera un mayor beneficio esperado. Por otro punto, también hay que tener en cuenta, que este modelo predictivo es el que tiene un mayor tiempo de entrenamiento al momento de emplearse.

5.0 CONCLUSIONES

Con el avance tecnológico en constante crecimiento, este estudio ha destacado los beneficios del análisis de datos y su potencial para impulsar diversas industrias. La aplicación de metodologías como KDD ha demostrado su versatilidad y aplicabilidad en una amplia gama de sectores.

En el caso específico abordado en esta investigación, el uso de la metodología KDD condujo a la generación de cinco modelos predictivos. Mediante una función de optimización diseñada para evaluar el beneficio esperado de estos modelos, se confirmó que el segundo modelo de red neuronal profunda era el más adecuado para resolver el problema planteado.

Es importante destacar que, en este caso particular, se observó una correlación entre la profundidad del modelo de red neuronal y su eficiencia. Sin embargo, es crucial considerar aspectos cualitativos que podrían influir en la predicción, como los costos asociados a la implementación.

Durante la investigación, se identificó que la ejecución del segundo modelo de red neuronal profunda conllevó un tiempo de procesamiento significativo en comparación con otros modelos, lo que podría resultar en costos más altos.

Por lo tanto, es fundamental tener en cuenta los costos asociados a la implementación de modelos predictivos, que pueden incluir tanto el tiempo de ejecución como los recursos necesarios, como el hardware adecuado y la experiencia de profesionales especializados.

En conclusión, el análisis de datos emerge como una herramienta valiosa para las empresas, ya que permite la creación de modelos que mejoran su desempeño y generan valor. Esta tendencia hacia la demanda de profesionales en el campo del análisis de datos refleja el reconocimiento de su importancia en el mercado laboral actual.

REFERENCIAS

Agenciab12. (11 de Noviembre de 2019). *Agencia b12*. Obtenido de B12 Admark: https://agenciab12.com/noticia/origen-concepto-inteligencia-artificial

Amazon AWS. (21 de Mayo de 2018). *¿Que es una red neuronal?* Obtenido de aws.amazon.com: https://aws.amazon.com/es/what-is/neural-network/

APD. (4 de Abril de 2019). *Asociación para Progreso de la Dirección*. Obtenido de Asociación para Progreso de la Dirección: https://www.apd.es/algoritmos-del-machine-learning/

APD. (29 de Julio de 2021). *APD*. Obtenido de Asociación para el Progreso de la Dirección: https://www.apd.es/el-gran-impacto-de-la-inteligencia-artificial-en-las-empresas/

apiumhub. (18 de Febrero de 2022). *apiumhub*. Obtenido de https://apiumhub.com/es/tech-blog-barcelona/mineria-de-datos-casos-de-uso-beneficios/

Arce, J. I. (26 de Julio de 2019). *https://www.juanbarrios.com/*. Obtenido de La matriz de confusión y sus métricas: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Arias, P. S. (23 de Abril de 2015). *Marketing / Mercadotecnia.* Obtenido de Economipedia.com: https://economipedia.com/definiciones/mercadotecnia-marketing.html

Arias, P. S. (15 de Octubre de 2020). *4 C's del marketing*. Obtenido de economipedia.com: https://economipedia.com/definiciones/4-cs-del-marketing.html#:~:text=Las%204%20C's%20del%20marketing,%2C%20comunicaci%C3%B3n%2C%20conveniencia%20y%20coste.

AWS. (12 de Febrero de 2020). *¿Que es sobre ajuste?* Obtenido de AWS.amazon: https://aws.amazon.com/es/what-is/overfitting/

Banco ITAU. (15 de enero de 2019). *Batalla de datos*. Obtenido de https://batalladedatos.firstjob.me/: https://batalladedatos.firstjob.me/

bbvaopenmind. (4 de septiembre de 2016). *bbvaopenmind*. Obtenido de www.bbvaopenmind.com: https://www.bbvaopenmind.com/tecnologia/inteligencia-artificial/el-verdadero-padre-de-la-inteligencia-artificial/

Berry, M. J. (2004). *Data Mining Techniques* . Indianapolis: Second Edition.

buhoagenciadigital. (16 de Diciembre de 2019). *buhoagenciadigital.com*. Obtenido de ¿Como funciona el Posicionamiento SEM? Vende a través de google ADS: https://buhoagenciadigital.com/como-funciona-el-posicionamiento-sem-vende-a-traves-de-google-ads/#:~:text=El%20concepto%20de%20POSICIONAMIENTO%20SEM,Google%20Ads%20o%20Bing%20Ads.

Cardenas, J. (1 de Diciembre de 2015). *Odd ratio: qué es y cómo se interpreta.* Obtenido de networkianos: https://networkianos.com/odd-ratio-que-es-como-se-interpreta/#toc-1

CESUMA. (12 de Octubre de 2020). *Cesuma.mx*. Obtenido de https://www.cesuma.mx/: https://www.cesuma.mx/blog/el-departamento-de-marketing-objetivos-funciones-y-tareas.html#:~:text=Funciones%20del%20departamento%20de%20Marketing&text=Investigar%20la%20situaci%C3%B3n%20del%20mercado,proceso%20de%20fijaci%C3%B3n%20de%20precios

Chiu, S. (2008). *Data Mining and Market .* Burlington: First edition.

CMF. (29 de Julio de 2020). *¿De donde obtienen dinero los bancos?* Obtenido de cmfchile.cl: https://www.cmfchile.cl/educa/621/w3-article-27138.html#:~:text=Los%20bancos%20obtienen%20beneficios%20de,t%C3%A9rmino%20ingl%C3%A9s%20de%20%22spread%22.

Corrales, J. A. (19 de Agosto de 2020). *rockcontent.* Obtenido de rockcontent.com: https://rockcontent.com/es/blog/segmentacion-de-clientes/

Cosio, N. A. (10 de diciembre de 2021). *medium*. Obtenido de medium.com/@nicolasarrioja: https://medium.com/@nicolasarrioja/covarianza-y-correlaci%C3%B3n-7f16e59445b4

cuadernodemarketing. (30 de diciembre de 2018). *cuadernodemarketing*. Obtenido de cuadernodemarketing.com: https://cuadernodemarketing.com/el-valor-es-lo-importante-pero-de-que-hablamos-cuando-hablamos-de-valor/

d'Arc, T. (25 de Mayo de 2022). *Que es la inteligencia artificial: 16 ejemplos en tu vida diaria.* Obtenido de Smart Hint: https://www.smarthint.co/es/que-son-ejemplos-de-inteligencia-artificial/

DATAtab. (Marzo de 28 de 2021). *DATAtab Team.* Obtenido de datatab.es: https://datatab.es/tutorial/pearson-correlation

desouttertools. (18 de mayo de 2018). *desouttertools Industrial*. Obtenido de desouttertools: https://www.desouttertools.mx/industria-4-0/noticias/1015/revolucion-industrial-de-industria-1-0-a-industria-4-0

FERNÁNDEZ, L. S. (2018). *Piano Marketing.* Obtenido de https://www.pianomarketing.es/.

finanzasparamortales. (11 de Mayo de 2020). *finanzasparamortales*. Obtenido de finanzasparamortales.es: https://finanzasparamortales.es/henry-ford-el-padre-de-la-produccion-en-masa/

Galán, J. S. (4 de Agosto de 2017). *Posicionamiento*. Obtenido de Economipedia: https://economipedia.com/definiciones/posicionamiento.html

García, A. (5 de Diciembre de 2018). *usellcrm*. Obtenido de https://www.usellcrm.net/: https://www.usellcrm.net/4-pasos-segmentacion-optima-clientes/

Habla el mercado. (17 de Agosto de 2020). *technologyreview.* Obtenido de technologyreview: https://www.technologyreview.es/s/12486/la-inteligencia-artificial-en-chile-una-industria-en-crecimiento

https://rstudio.github.io/reticulate/. (18 de Marzo de 2015). *rstudio.github.io*. Obtenido de rstudio.github.io: https://rstudio.github.io/reticulate/

IBM. (12 de Febrero de 2018). *IBM.* Obtenido de ¿Qué es la minería de datos?: https://www.ibm.com/es-es/topics/data-mining

Kalinowski, T. (18 de Julio de 2022). *rstudio.github.io*. Obtenido de rstudio.github.io: https://rstudio.github.io/reticulate/

Krawicki, J. (13 de Abril de 2020). *observatoriorh*. Obtenido de observatoriorh.com: https://www.observatoriorh.com/actualidad/las-personas-y-la-cuarta-revolucion-industrial-somos-uno-y-debemos-integrarnos.html

Kuhn, M. (Abril de 30 de 2020). *yardstick.tidymodels.org*. Obtenido de yardstick.tidymodels.org: https://yardstick.tidymodels.org/

Lastra, E. F. (6 de Febrero de 2018). *Artyco*. Obtenido de artyco.com: https://artyco.com/como-realizar-segmentacion-de-clientes-exito/

Marketing Digital Madrid. (24 de Abril de 2021). *Marketing Digital Madrid*. Obtenido de https://www.mastermarketingdigital-madrid.com/: https://www.mastermarketingdigital-madrid.com/blog/mkt/marketing-historia-evolucion/#:~:text=El%20concepto%20moderno%20del%20Marketing,UU.

masquenegocio. (26 de Agosto de 2015). *masquenegocio.* Obtenido de masquenegocio.com: https://www.masquenegocio.com/2015/08/26/segmentacion-mercados/

Minitab. (15 de Marzo de 2021). *Soporte Minitab*. Obtenido de Minitab Statistical Software: https://support.minitab.com/es-mx/minitab/20/help-and-how-to/statistical-modeling/regression/how-to/fit-binary-logistic-model/interpret-the-results/all-statistics-and-graphs/receiver-operating-characteristic-roc-curve/

Na8. (3 de Marzo de 2020). *Sets de Entrenamiento, Test y Validación.* Obtenido de aprendemachinelearning: https://www.aprendemachinelearning.com/sets-de-entrenamiento-test-validacion-cruzada/

Quiroa, M. (1 de Noviembre de 2019). *economipedia*. Obtenido de Haciendo fácil la Economía: https://economipedia.com/definiciones/cliente.html

Quiroa, M. (13 de Octubre de 2021). *Economipedia.com*. Obtenido de https://economipedia.com/definiciones/mercado-en-marketing.html

RAE. (20 de abril de 2022). *www.rae.cl*. Obtenido de https://dle.rae.es/cliente

Reinventa. (30 de Septiembre de 2021). *https://www.banbif.com.pe/.* Obtenido de https://www.banbif.com.pe/: https://www.banbif.com.pe/Portals/0/blog-reinventa/noticias/entrada-40.html

Revista Empresarial. (20 de Mayo de 2020). *El Poder de la Revolución Industrial*. Obtenido de https://revistaempresarial.com/tecnologia/el-poder-de-la-revolucion-industrial/

Rosa, J. (9 de marzo de 2020). *johan-rosa.com/*. Obtenido de johan-rosa.com/: https://www.johan-rosa.com/post/tensorflow-y-keras-con-r/

Rosa, J. (29 de Marzo de 2020). *Tensorflow y keras con R.* Obtenido de Johan Rosa: https://www.johan-rosa.com/post/tensorflow-y-keras-con-r/

rstudio.github.io. (s.f.). *rstudio.github.io*. Obtenido de rstudio.github.io: https://rstudio.github.io/reticulate/

Ruiz, E. (17 de Marzo de 2019). *Aprendizaje Automático .* Ciudad de Mexico: datadaymx.

Sectorial. (20 de julio de 2020). *Cuatro revoluciones industriales*. Obtenido de https://www.sectorial.co/articulos-especiales/item/220049-las-cuatro-revoluciones-industriales-de-la-historia-infograf%C3%ADa

SimpliRoute. (11 de Enero de 2023). *SimpliRoute.* Obtenido de Regresión Logística: Qué Es y Cómo Funciona: https://simpliroute.com/es/blog/regresion-logistica

statisticalecology. (18 de Agosto de 2012). *Estadística computacional*. Obtenido de statisticalecology: https://statisticalecology-ec.blogspot.com/2012/08/regresion-seleccion-de-variables.html

statologos. (12 de Septiembre de 2021). *Stalogos* . Obtenido de Criterio de información de Akaike: definición, fórmulas: https://statologos.com/criterio-de-informacion-de-akaikes/#google_vignette

Suarez, F. (20 de mayo de 2020). *Revista Empresarial*. Obtenido de https://revistaempresarial.com/tecnologia/el-poder-de-la-revolucion-industrial/

ANEXO

Modelo logístico multivariable

Matriz confusión

Matriz de Confusión

	Predicción: "no aumentó cupo de tarjeta"	Predicción: "aumentó cupo de tarjeta"
No aumentó cupo de tarjeta	16190	1262
Aumentó cupo de tarjeta	347	3727

Modelo red neuronal básica

Matriz confusión:

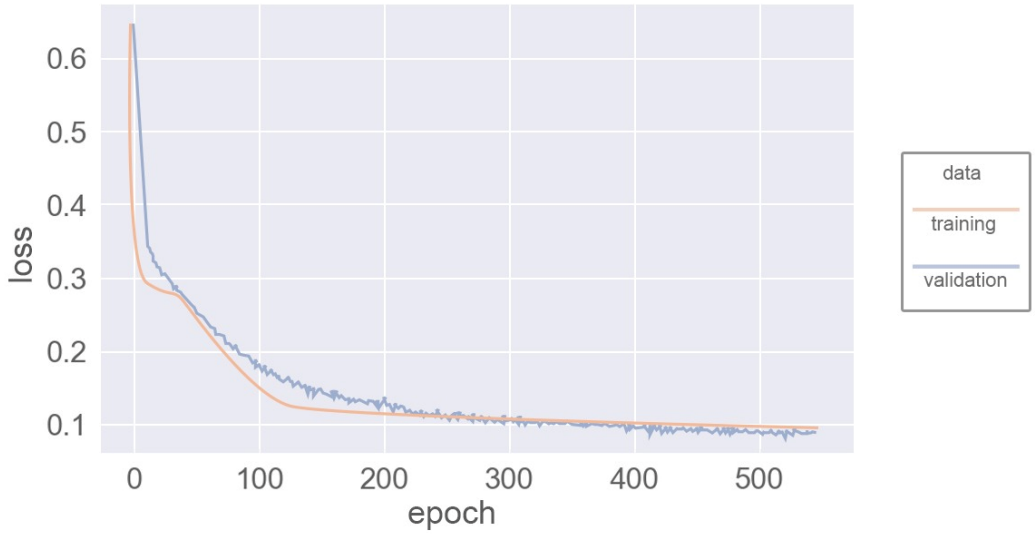
Matriz de Confusión

	Predicción: "no aumentó cupo de tarjeta"	Predicción: "aumentó cupo de tarjeta"
No aumentó cupo de tarjeta	15460	1992
Aumentó cupo de tarjeta	1283	2791

Arquitectura del modelo:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_4 (Dense)             (None, 35)                1155
 dense_3 (Dense)             (None, 32)                1152
 dense_2 (Dense)             (None, 16)                528
 dropout (Dropout)           (None, 16)                0
 dense_1 (Dense)             (None, 32)                544
 dense (Dense)               (None, 1)                 33
=================================================================
Total params: 3,412
Trainable params: 3,412
Non-trainable params: 0
_________________________________________________________________
```

Gráfico de entrenamiento



Modelo red neuronal profunda

Matriz confusión:

Matriz de Confusión

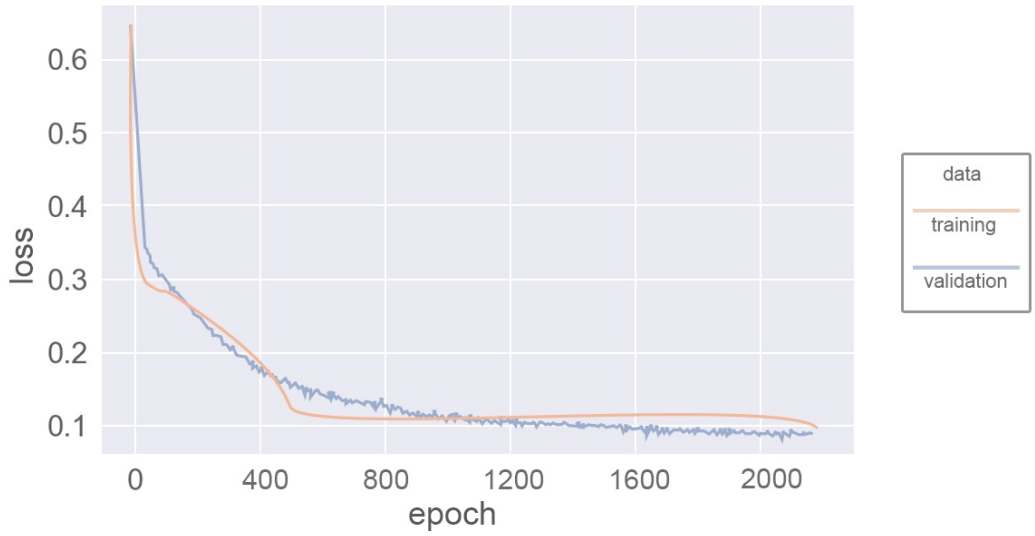
	Predicción: "no aumentó cupo de tarjeta"	Predicción: "aumentó cupo de tarjeta"
No aumentó cupo de tarjeta	17291	161
Aumentó cupo de tarjeta	32	4042

Arquitectura del modelo:

Layer (type)	Output Shape	Param #
dense_12 (Dense)	(None, 35)	1155
dense_11 (Dense)	(None, 32)	1152
dropout_5 (Dropout)	(None, 32)	0
dense_10 (Dense)	(None, 16)	528
dropout_4 (Dropout)	(None, 16)	0
dense_9 (Dense)	(None, 32)	544
dropout_3 (Dropout)	(None, 32)	0
dense_8 (Dense)	(None, 16)	528
dropout_2 (Dropout)	(None, 16)	0
dense_7 (Dense)	(None, 32)	544
dropout_1 (Dropout)	(None, 32)	0
dense_6 (Dense)	(None, 16)	528
dense_5 (Dense)	(None, 1)	17

Total params: 4,996
Trainable params: 4,996
Non-trainable params: 0

Gráfico de entrenamiento:



Segundo modelo red neuronal profunda

Matriz confusión:

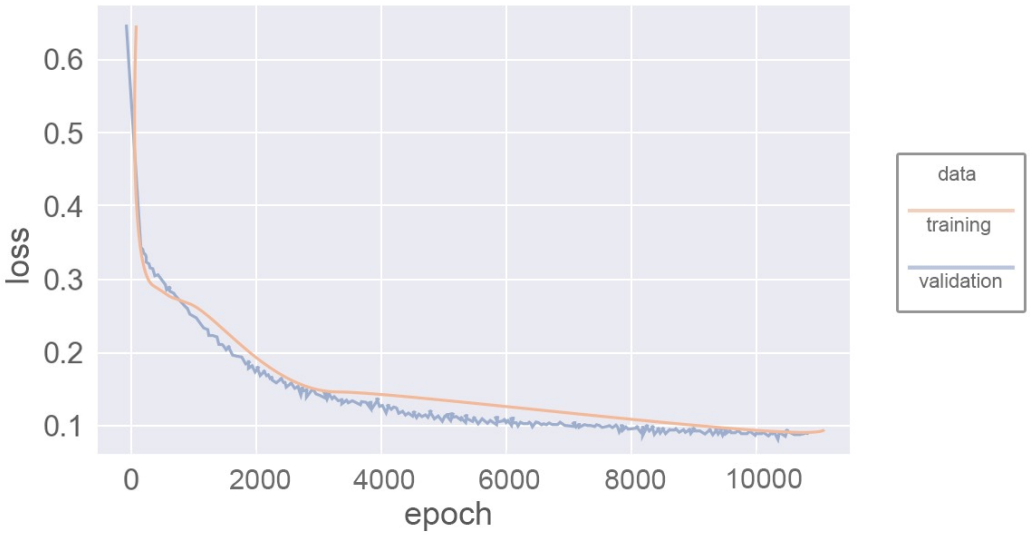
Matriz de Confusión

	Predicción: "no aumentó cupo de tarjeta"	Predicción: "aumentó cupo de tarjeta"
No aumentó cupo de tarjeta	17364	88
Aumentó cupo de tarjeta	50	4024

Arquitectura del modelo:

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_29 (Dense)             (None, 35)                1155
 dense_28 (Dense)             (None, 32)                1152
 dropout_15 (Dropout)         (None, 32)                0
 dense_27 (Dense)             (None, 16)                528
 dropout_14 (Dropout)         (None, 16)                0
 dense_26 (Dense)             (None, 32)                544
 dropout_13 (Dropout)         (None, 32)                0
 dense_25 (Dense)             (None, 16)                528
 dropout_12 (Dropout)         (None, 16)                0
 dense_24 (Dense)             (None, 32)                544
 dropout_11 (Dropout)         (None, 32)                0
 dense_23 (Dense)             (None, 16)                528
 dense_22 (Dense)             (None, 32)                544
 dropout_10 (Dropout)         (None, 32)                0
 dense_21 (Dense)             (None, 16)                528
 dense_20 (Dense)             (None, 32)                544
 dropout_9 (Dropout)          (None, 32)                0
 dense_19 (Dense)             (None, 16)                528
 dense_18 (Dense)             (None, 1)                 17
=================================================================
Total params: 7,140
Trainable params: 7,140
Non-trainable params: 0
_________________________________________________________________
```

Gráfico de entrenamiento



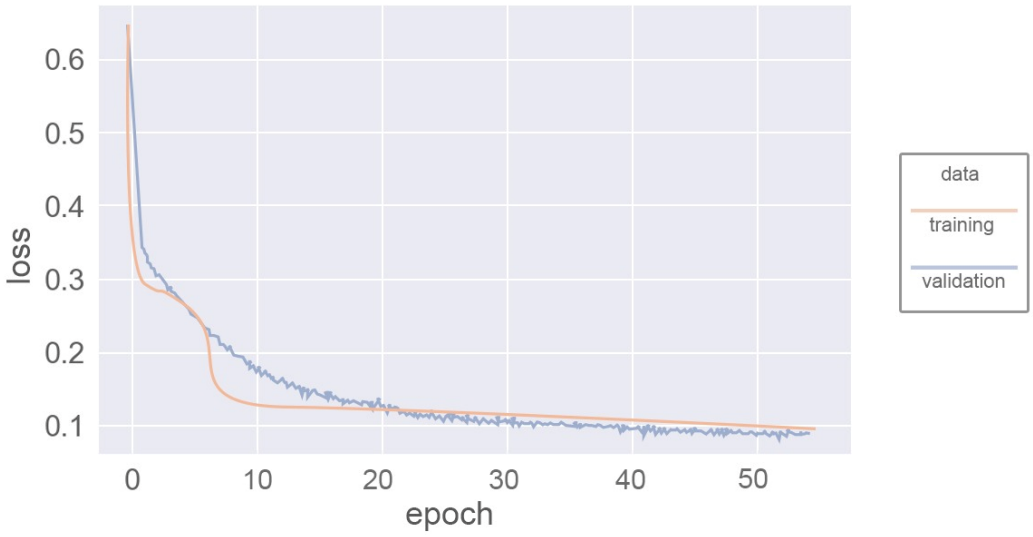
Modelo Red neuronal convolucional

Matriz confusión

Matriz de Confusión

	Predicción: "no aumentó cupo de tarjeta"	Predicción: "aumentó cupo de tarjeta"
No aumentó cupo de tarjeta	17139	313
Aumentó cupo de tarjeta	88	3986

Gráfico de entrenamiento

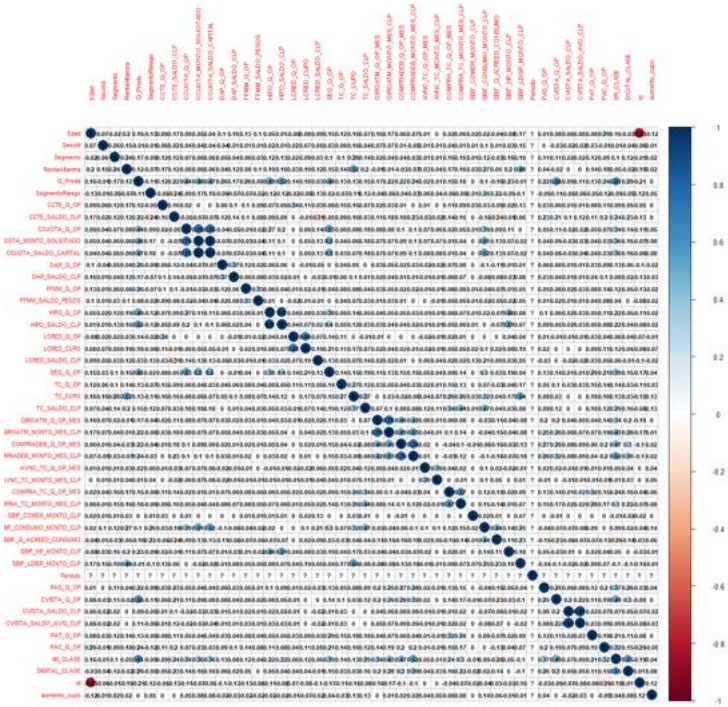


Arquitectura del modelo

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 30, 35)	140
flatten (Flatten)	(None, 1050)	0
dense_17 (Dense)	(None, 16)	16816
dropout_8 (Dropout)	(None, 16)	0
dense_16 (Dense)	(None, 32)	544
dropout_7 (Dropout)	(None, 32)	0
dense_15 (Dense)	(None, 16)	528
dropout_6 (Dropout)	(None, 16)	0
dense_14 (Dense)	(None, 32)	544
dense_13 (Dense)	(None, 1)	33

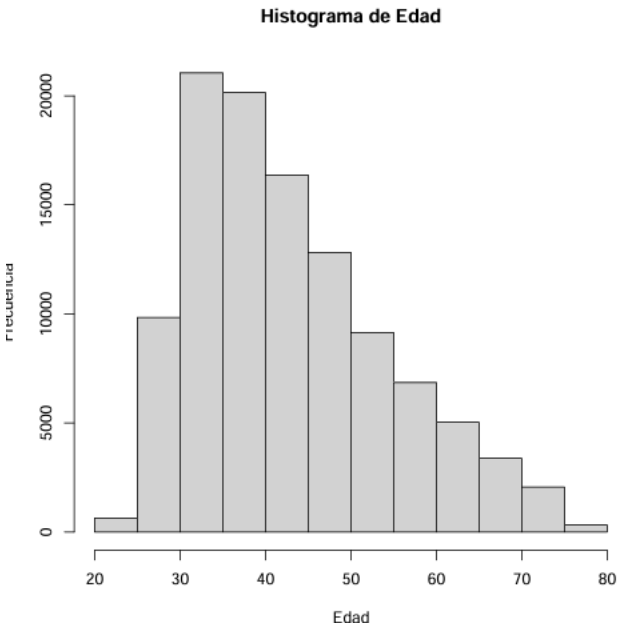
Total params: 18,605
Trainable params: 18,605
Non-trainable params: 0

Matriz correlación

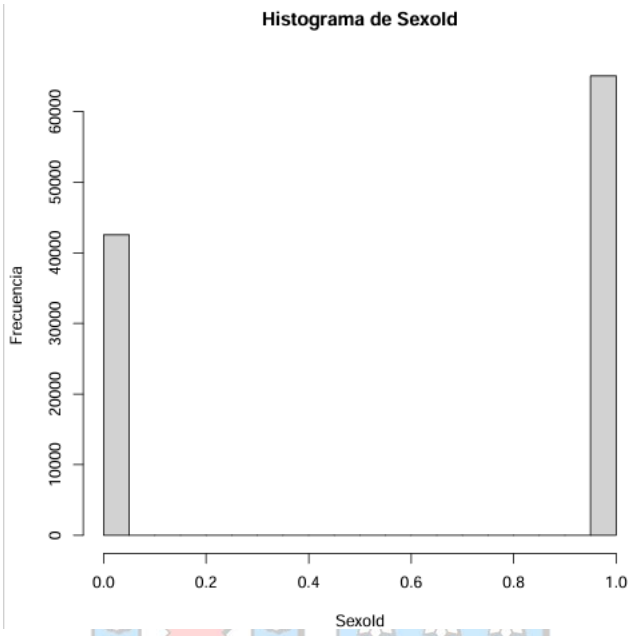


Visualización de datos:

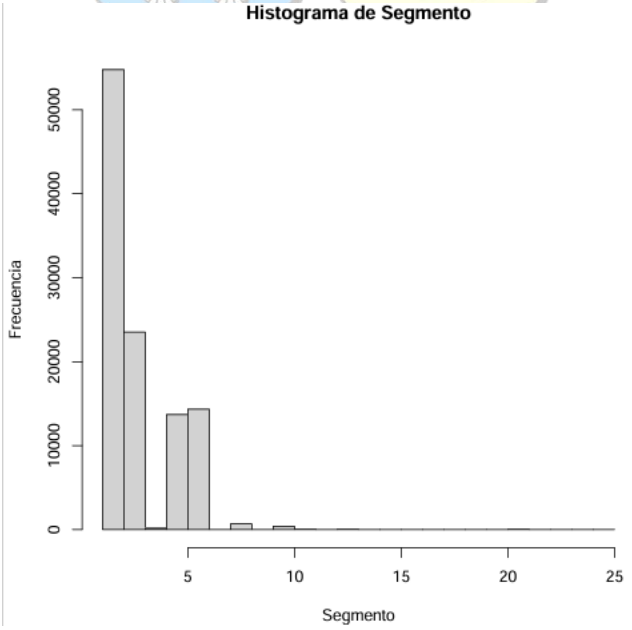
A continuación, se realizó un histograma a cada dato de las variables con el fin de verificar de estudiar sus comportamientos e identificar los arreglos necesarios.



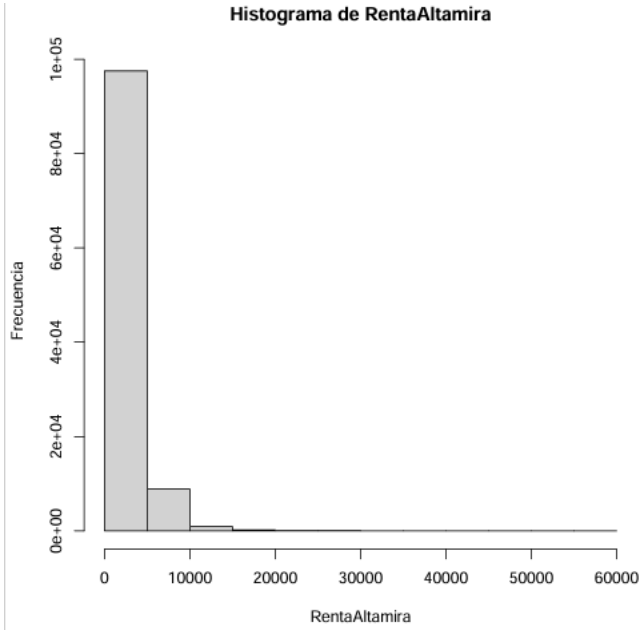
A1: *Histograma correspondiente a la variable edad. Fuente: Rstudio.*



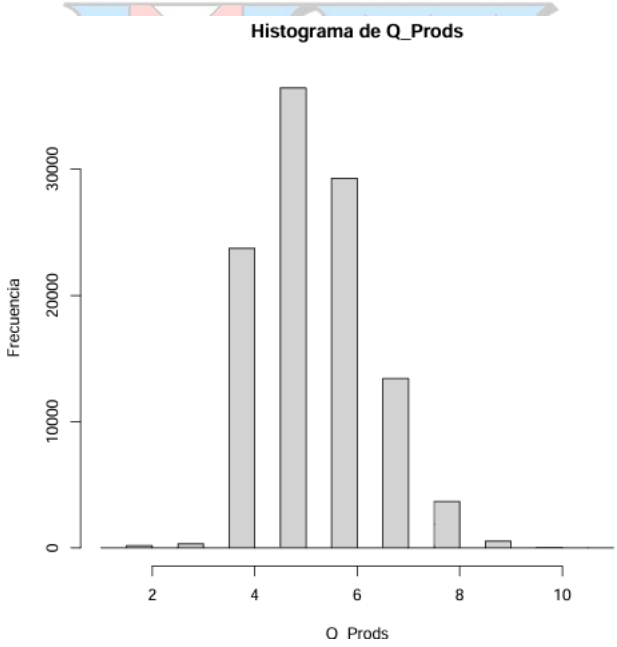
A2: *Histograma correspondiente a la variable Sexold. Fuente: Rstudio.*



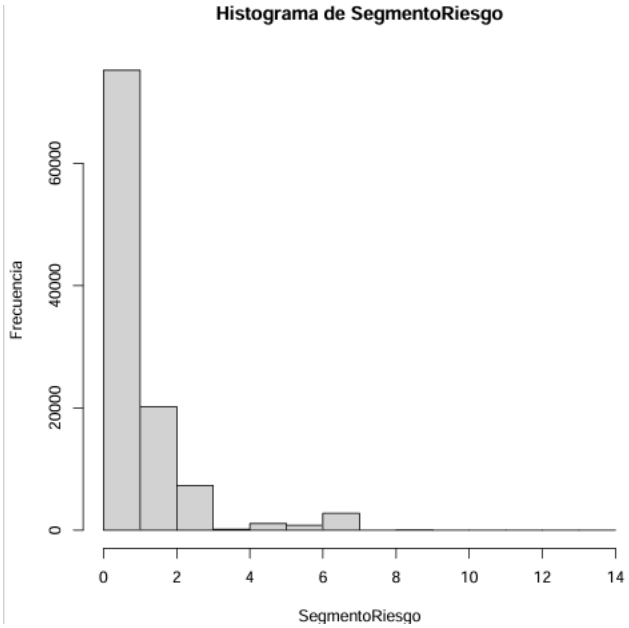
A3: *Histograma correspondiente a la variable Segmento. Fuente: Rstudio.*



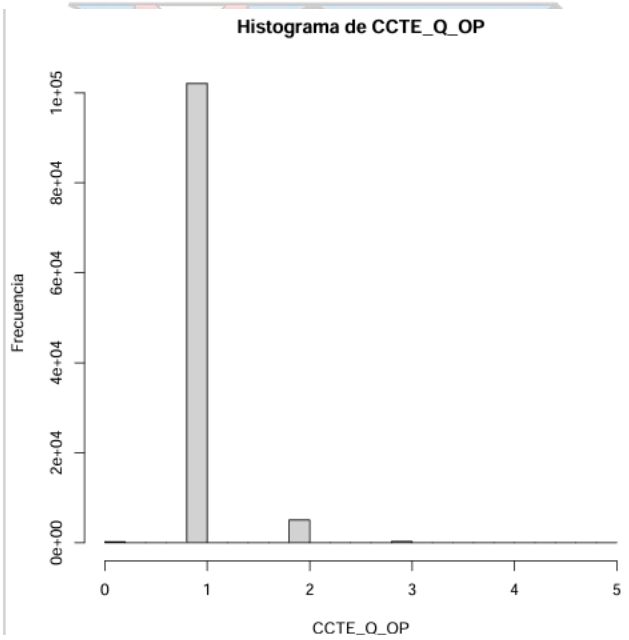
A4: *Histograma correspondiente a la variable RentaAltamira. Fuente: Rstudio.*



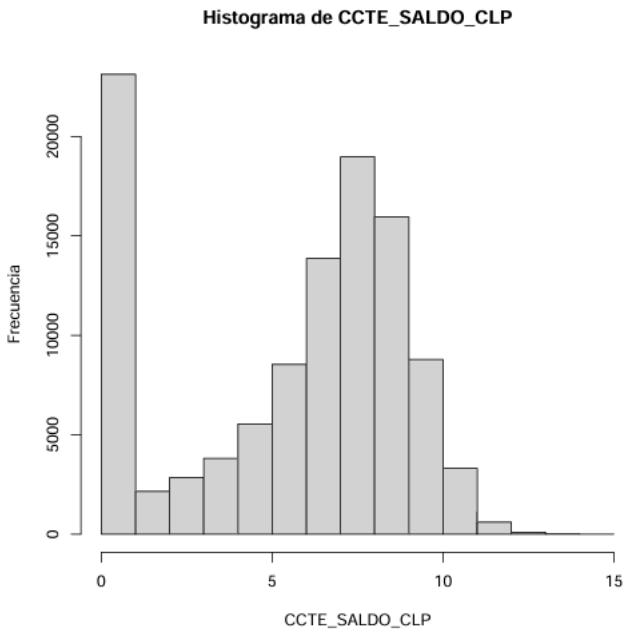
A5: *Histograma correspondiente a la variable Q_prods. Fuente: Rstudio.*



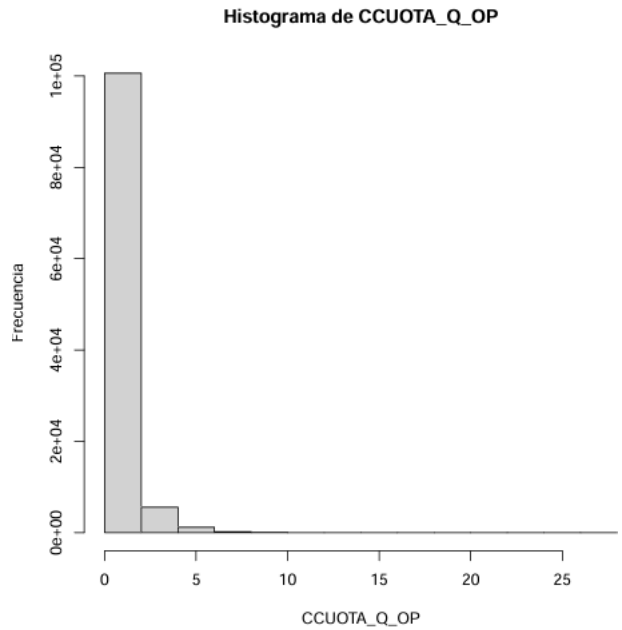
A6: *Histograma correspondiente a la variable SegmentoRiesgo. Fuente: Rstudio.*



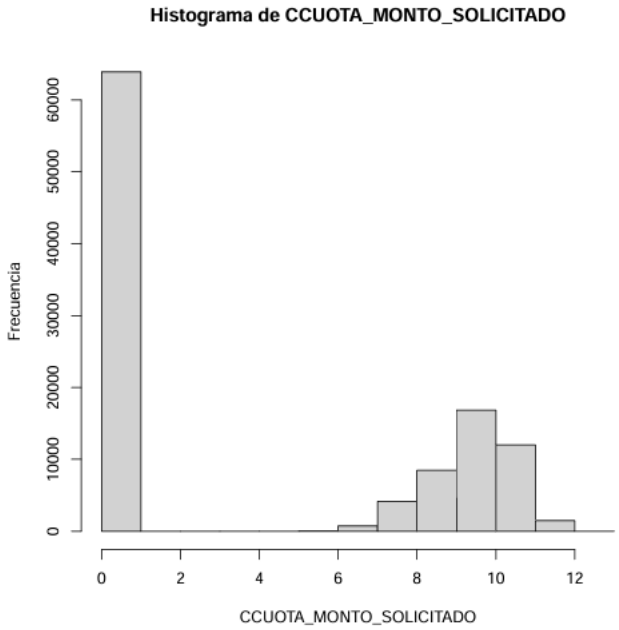
A7: *Histograma correspondiente a la variable Q_CCTE_Q_OP. Fuente: Rstudio.*



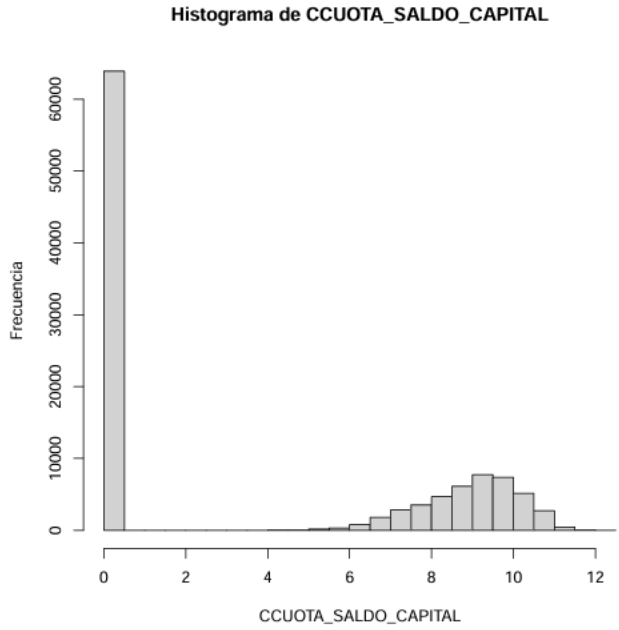
A8: *Histograma correspondiente a la variable CCTE_SALDO_CLP. Fuente: Rstudio.*



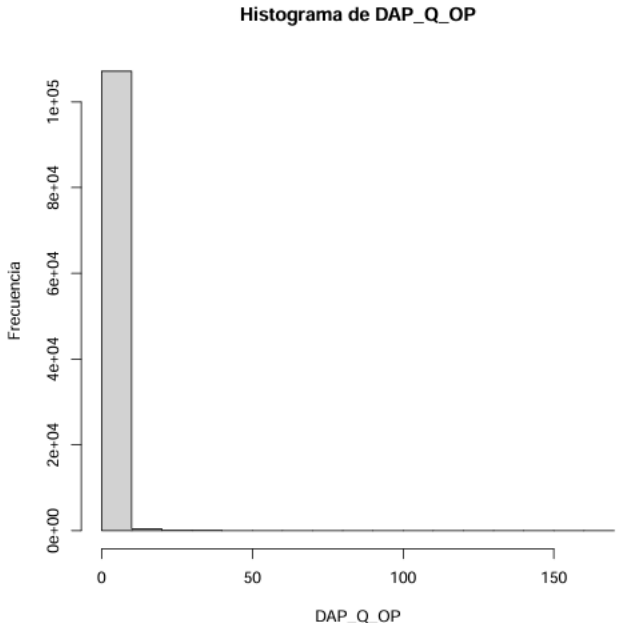
A9: *Histograma correspondiente a la variable CCUOTA_Q_OP. Fuente: Rstudio.*



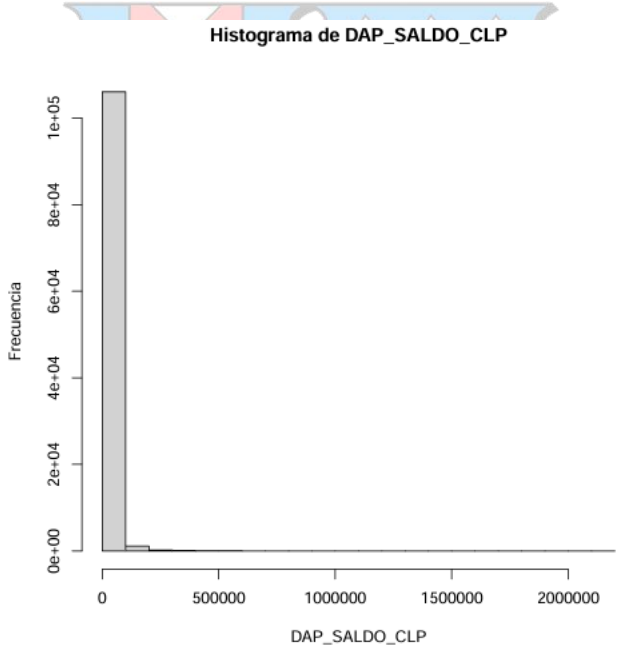
A10: *Histograma correspondiente a la variable CCUOTA_MONTO_SOLICITADO. Fuente: Rstudio.*



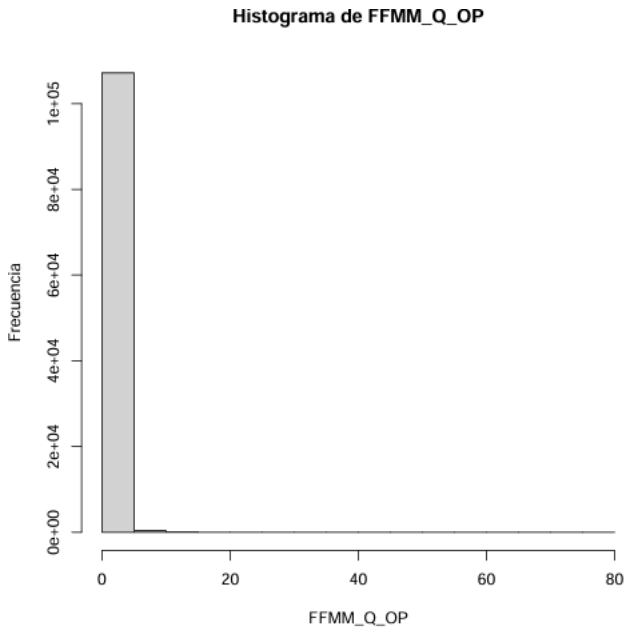
A11: *Histograma correspondiente a la variable CCUOTA_SALDO_CAPITAL. Fuente: Rstudio.*



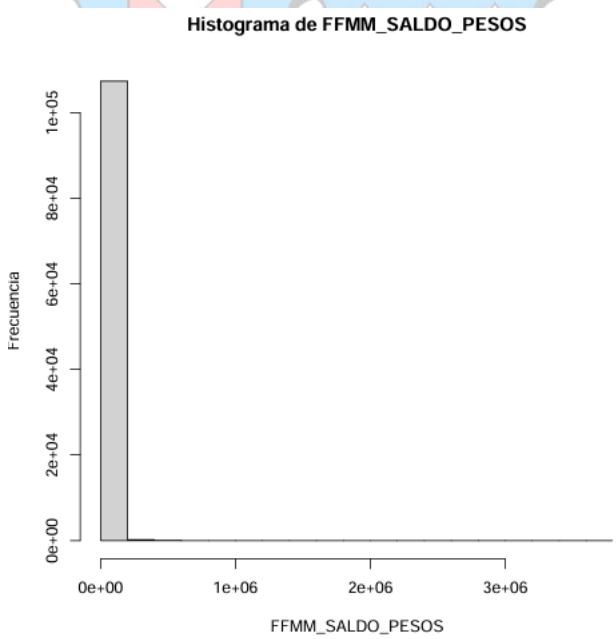
A12: *Histograma correspondiente a la variable DAP_Q_OP. Fuente: Rstudio.*



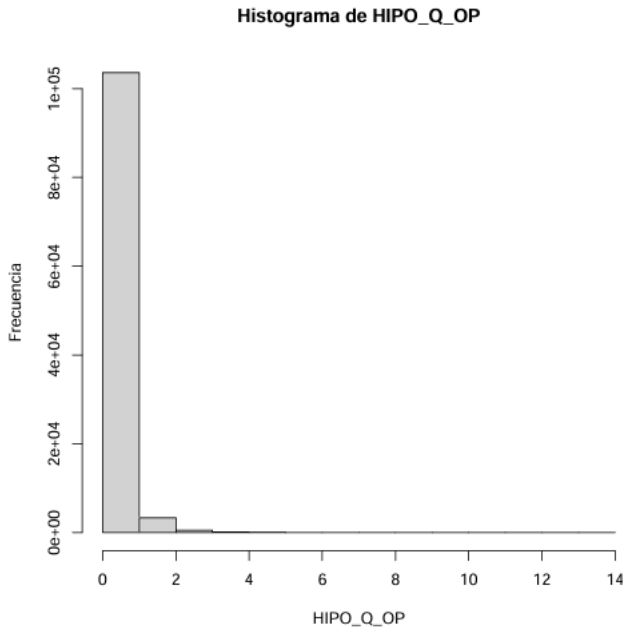
A13: *Histograma correspondiente a la variable DAP_SALDO_CLP. Fuente: Rstudio.*



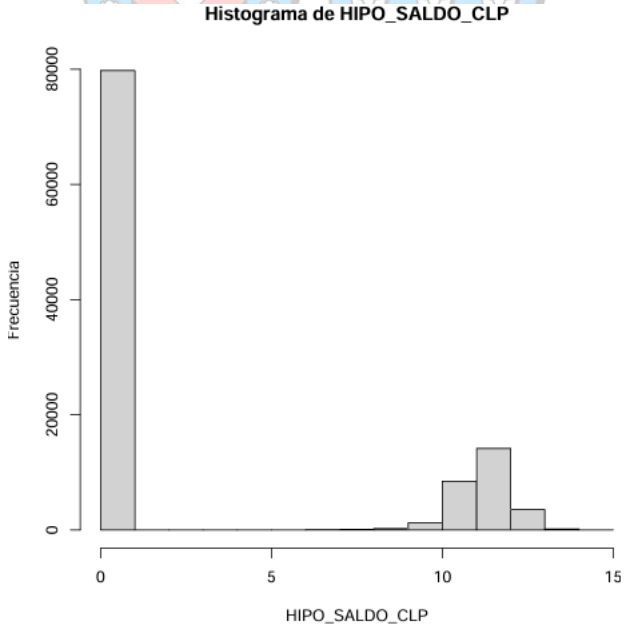
A14: *Histograma correspondiente a la variable FFMM_Q_OP. Fuente: Rstudio.*



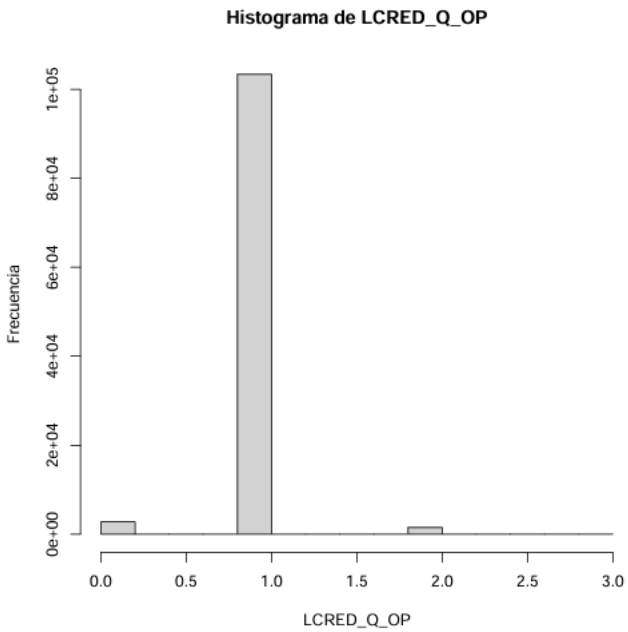
A15: *Histograma correspondiente a la variable FFMM_SALDO_PESOS. Fuente: Rstudio.*



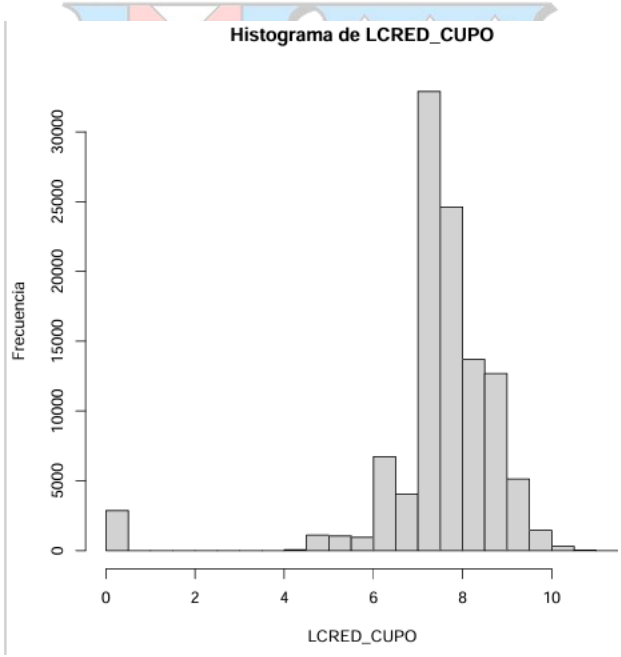
A16: *Histograma correspondiente a la variable HIPO_Q_OP. Fuente: Rstudio.*



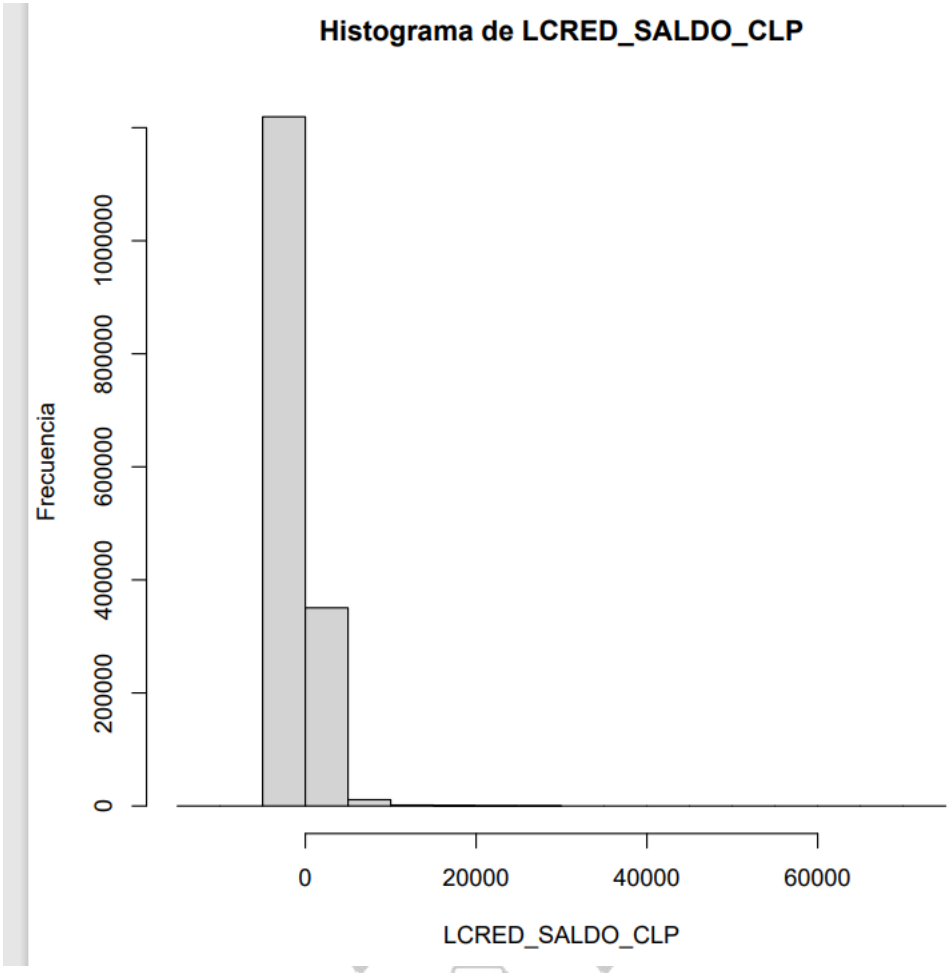
A17: *Histograma correspondiente a la variable HIPO_SALDO_CLP. Fuente: Rstudio.*



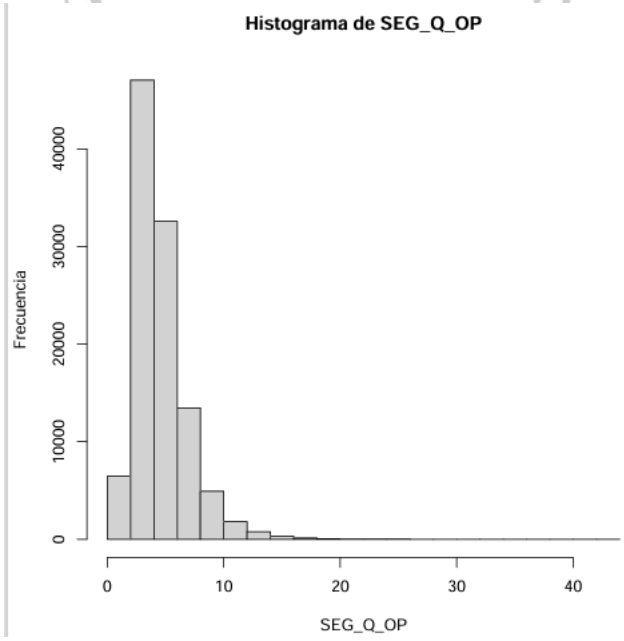
A18: *Histograma correspondiente a la variable LCRED_Q_OP. Fuente: Rstudio.*



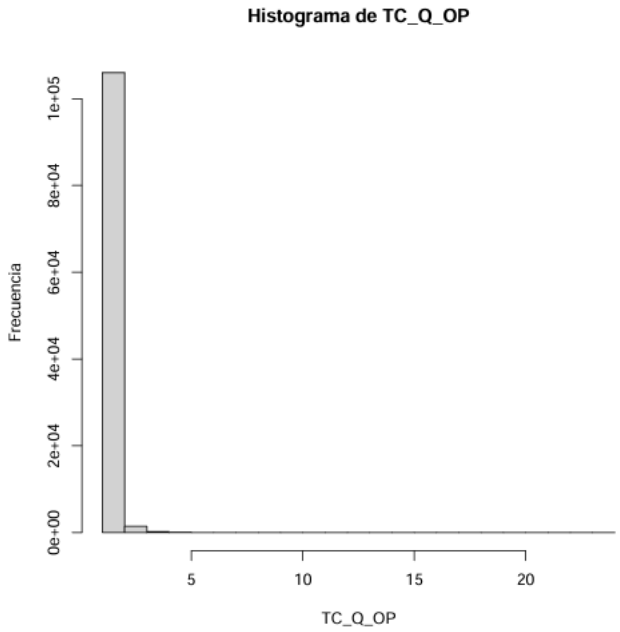
A19: *Histograma correspondiente a la variable LCRED_CUPO. Fuente: Rstudio.*



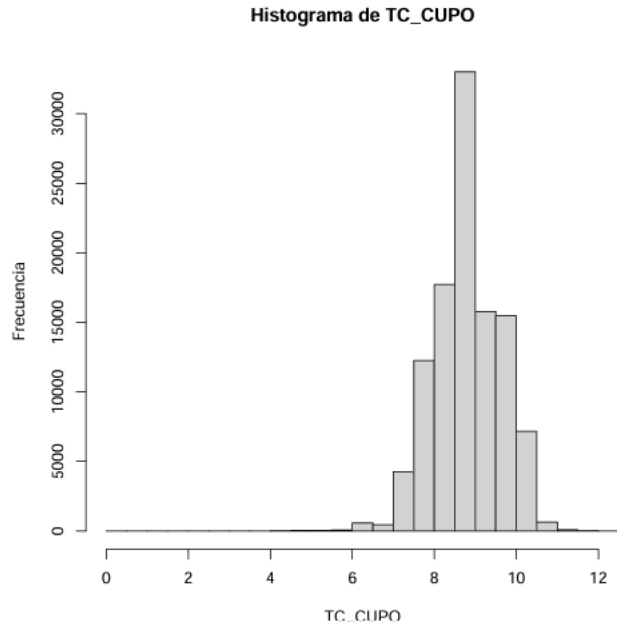
A20: *Histograma correspondiente a la variable LCRED_SALDO_CLP. Fuente: Rstudio.*



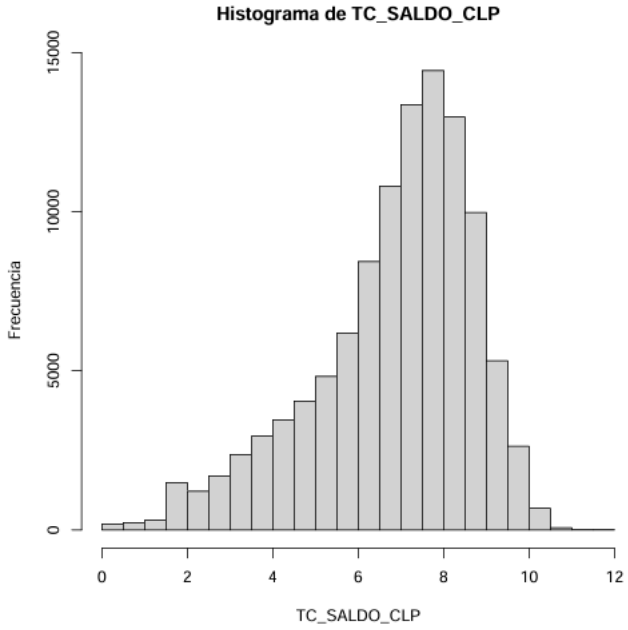
A21: *Histograma correspondiente a la variable SEG_Q_OP. Fuente: Rstudio.*



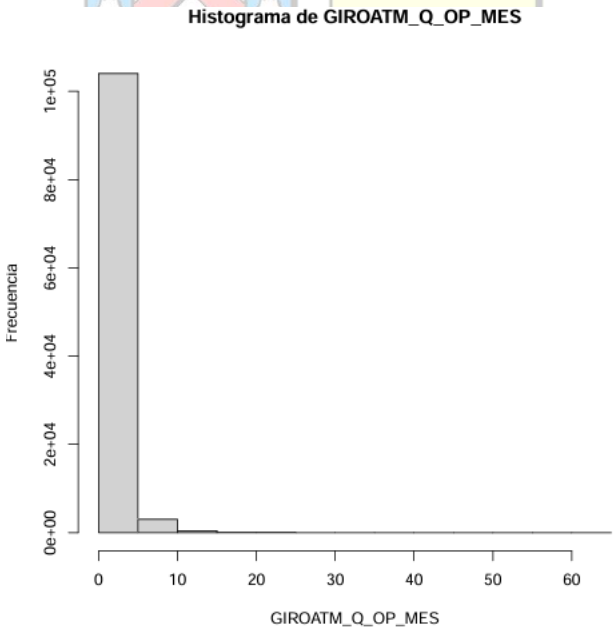
A22: *Histograma correspondiente a la variable TC_Q_OP. Fuente: Rstudio.*



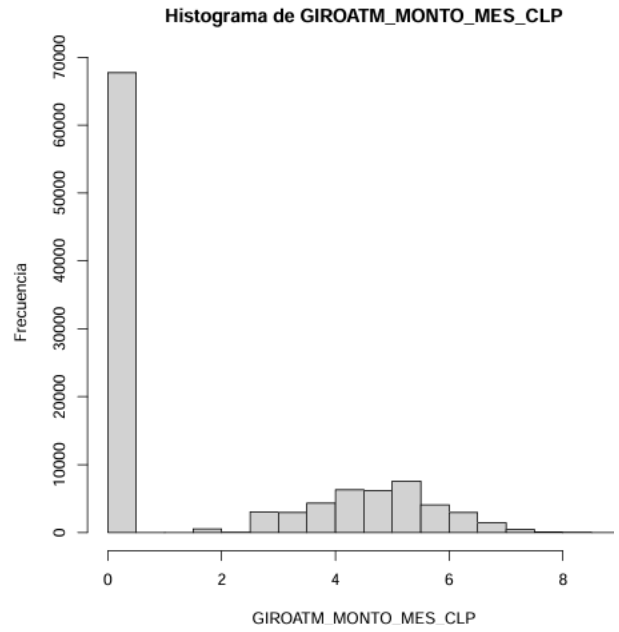
A23: *Histograma correspondiente a la variable TC_CUPO. Fuente: Rstudio.*



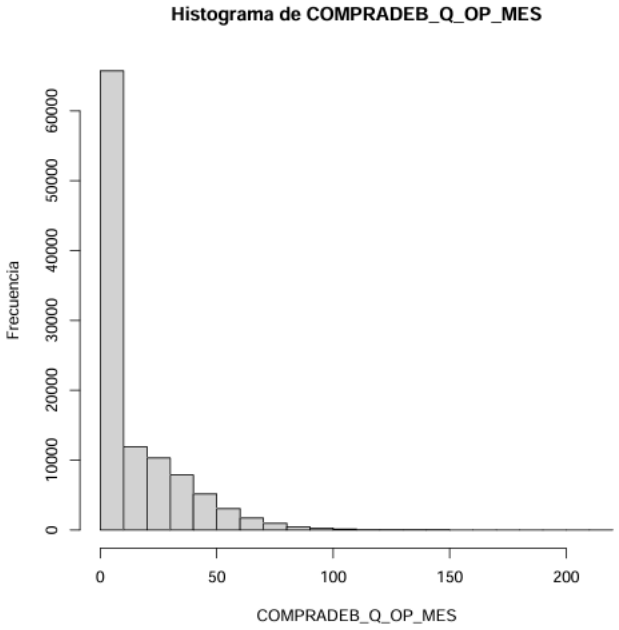
A24: *Histograma correspondiente a la variable TC_SALDO_CLP. Fuente: Rstudio.*



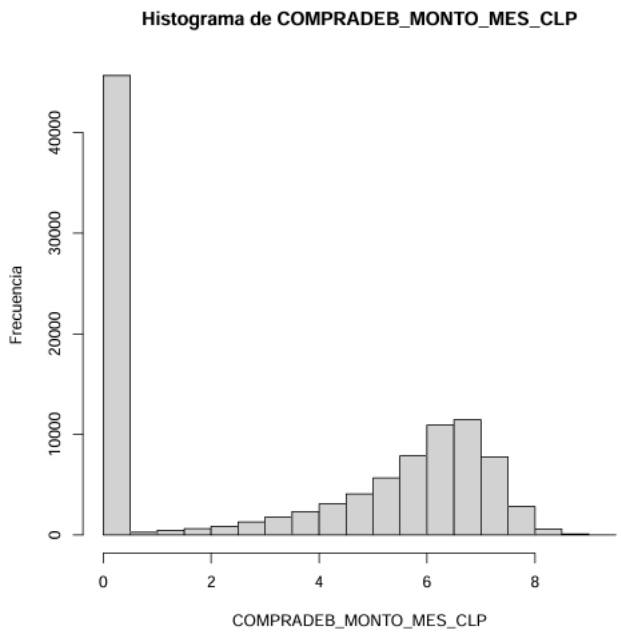
A25: *Histograma correspondiente a la variable GIROATM_Q_OP_MES. Fuente: Rstudio.*



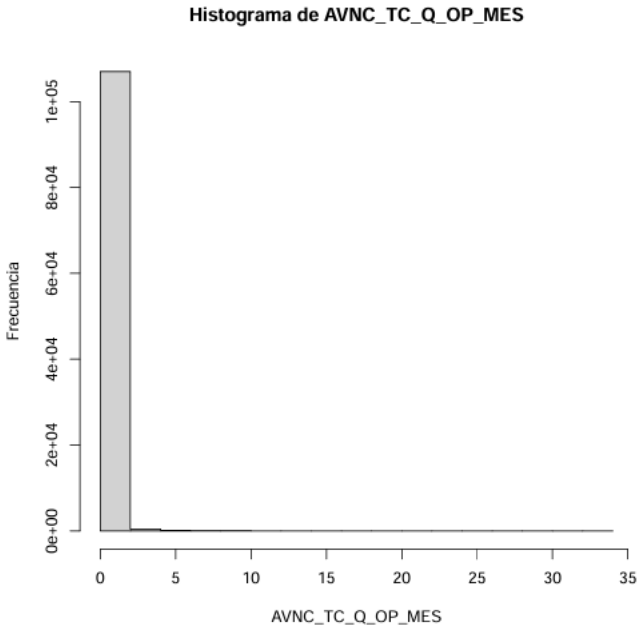
A26: *Histograma correspondiente a la variable GIROATM_MONTO_MES_CLP. Fuente: Rstudio.*



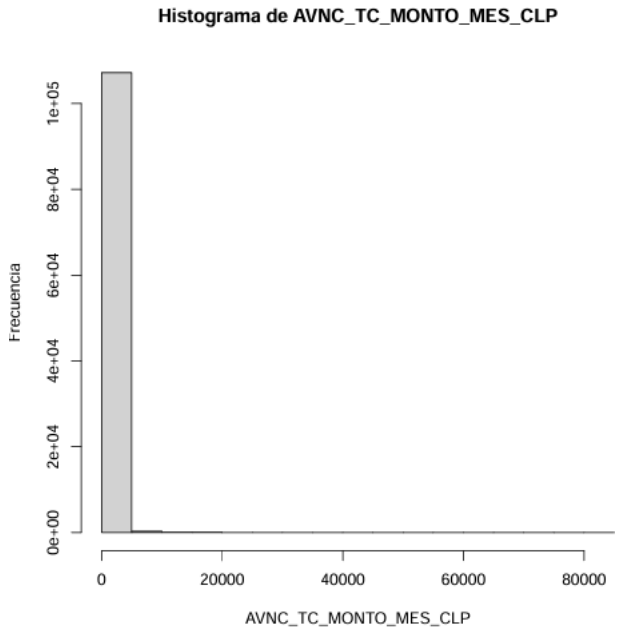
A27: *Histograma correspondiente a la variable COMPRADEB_Q_OP_MES. Fuente: Rstudio.*



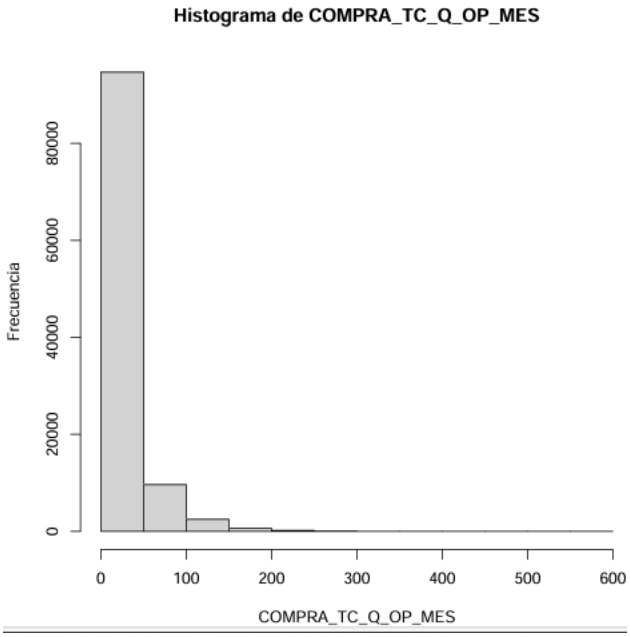
A28: *Histograma correspondiente a la variable COMPRADEB_MONTO_MES_CLP. Fuente: Rstudio.*



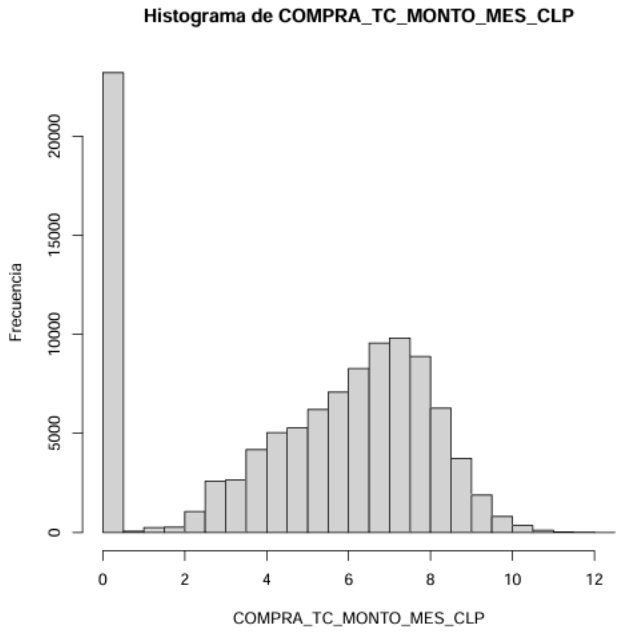
A29: *Histograma correspondiente a la variable AVNC_TC_Q_OP_MES. Fuente: Rstudio.*



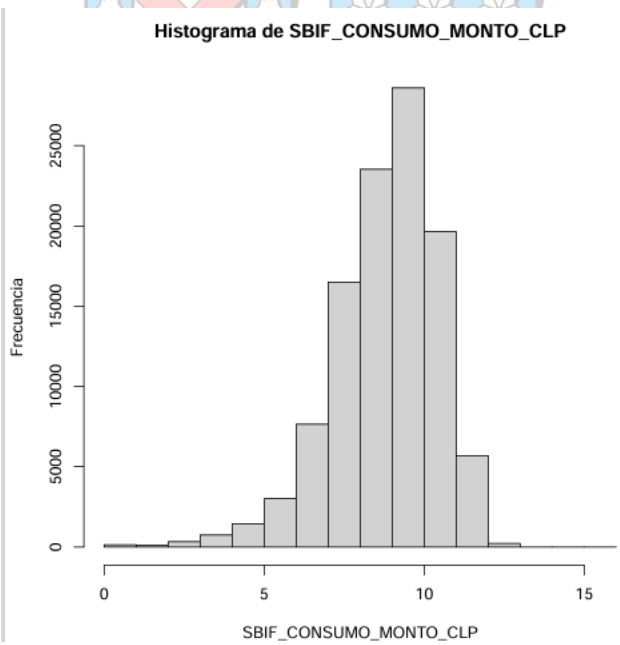
A30: *Histograma correspondiente a la variable AVNC_TC_MONTO_MES_CLP. Fuente: Rstudio.*



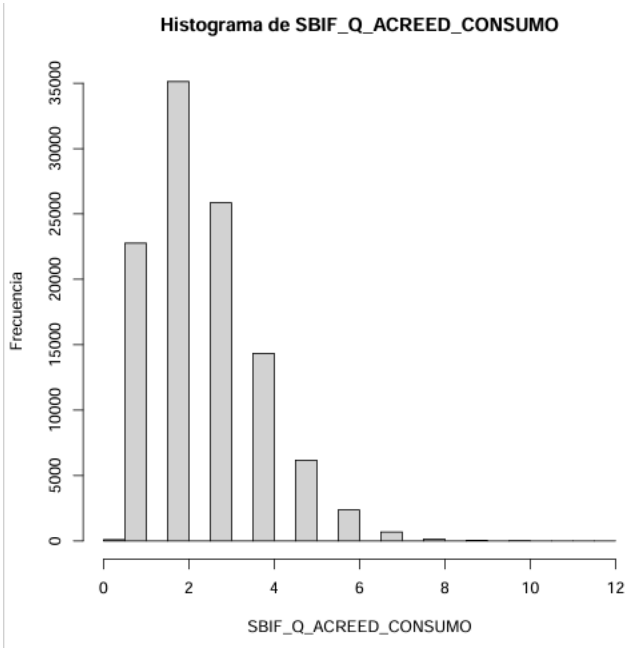
A31: *Histograma correspondiente a la variable COMPRA_TC_Q_OP_MES. Fuente: Rstudio.*



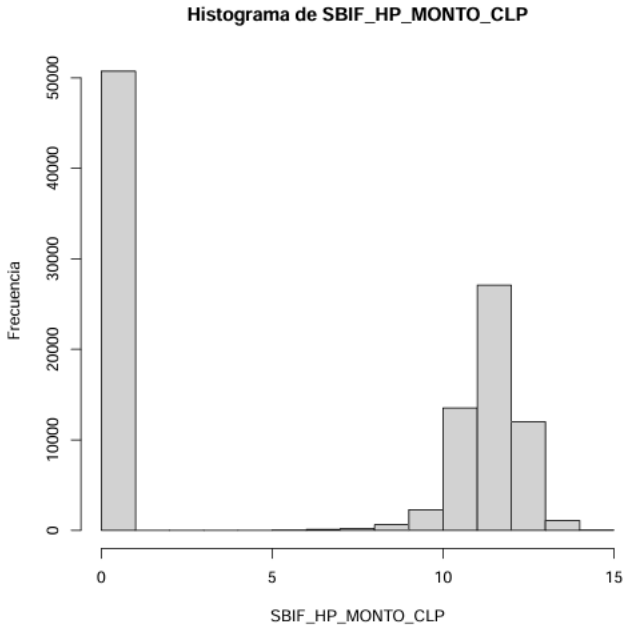
A32: *Histograma correspondiente a la variable COMPRA_TC_MONTO_MES_CLP. Fuente: Rstudio.*



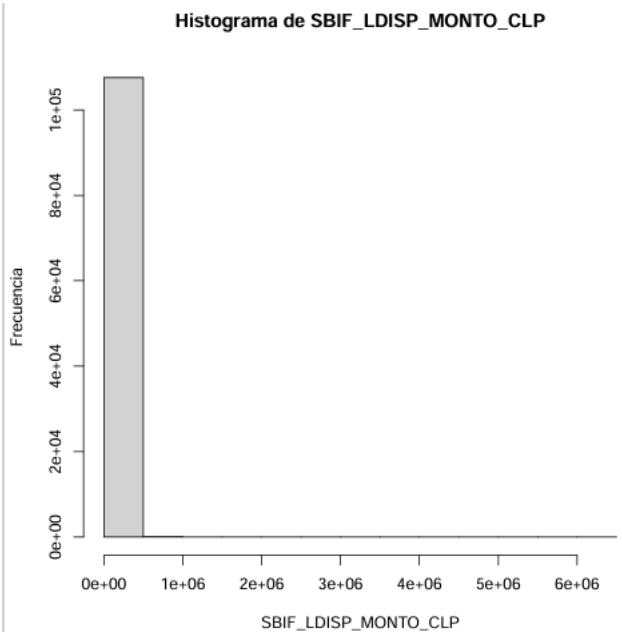
A33: *Histograma correspondiente a la variable SBIF_CONSUMO_MONTO _CLP. Fuente: Rstudio.*



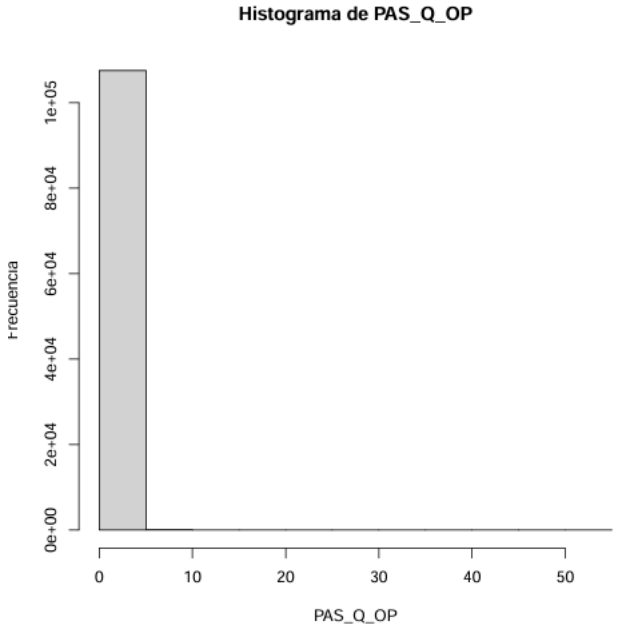
A34: *Histograma correspondiente a la variable SBIF_Q_ACREED_CONSUMO. Fuente: Rstudio.*



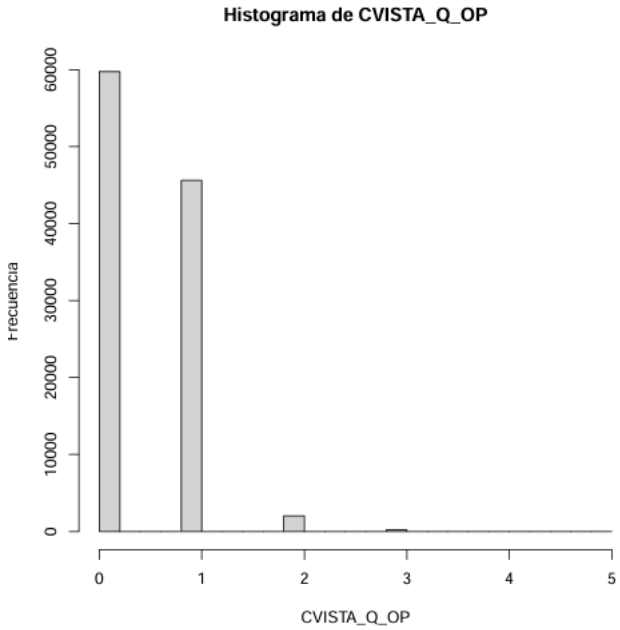
A35: *Histograma correspondiente a la variable SBIF_HP_MONTO_CLP. Fuente: Rstudio.*



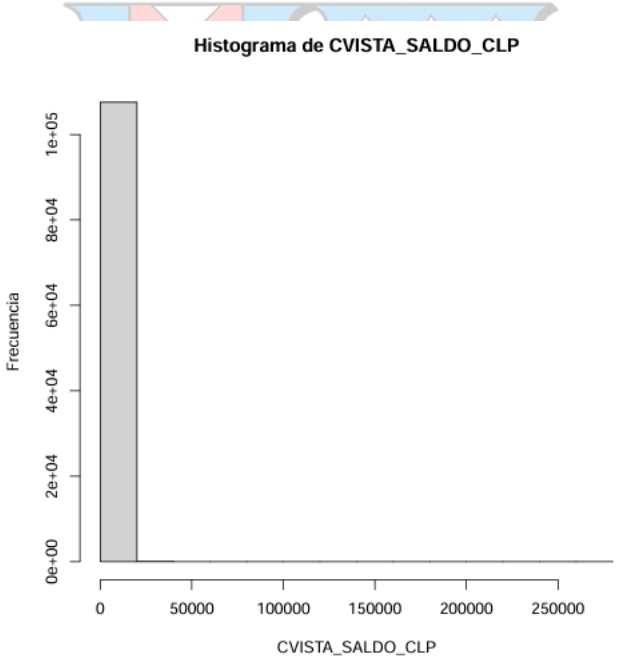
A36: *Histograma correspondiente a la variable SBIF_LDISP_MONTO_CLP. Fuente: Rstudio.*



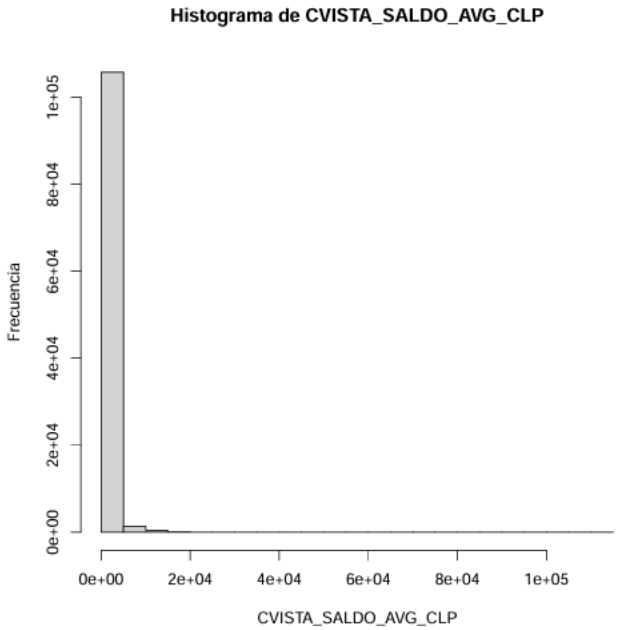
A37: *Histograma correspondiente a la variable PAS_Q_OP. Fuente: Rstudio.*



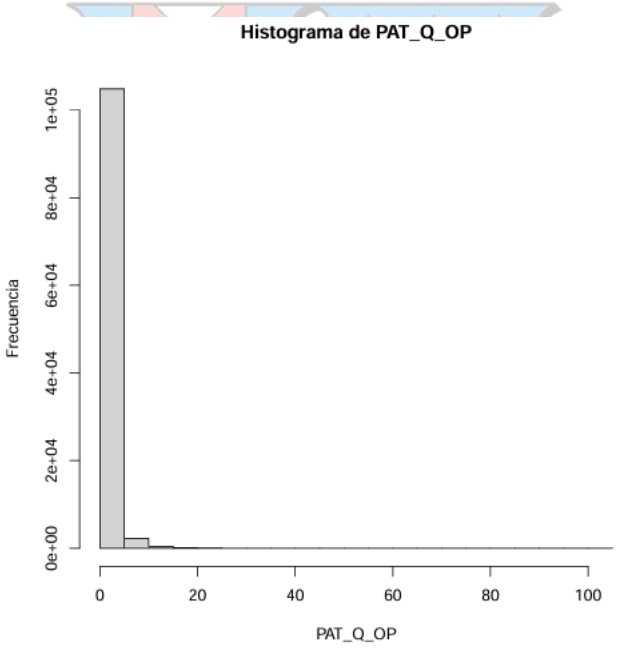
A38: *Histograma correspondiente a la variable CVISTA_Q_OP. Fuente: Rstudio.*



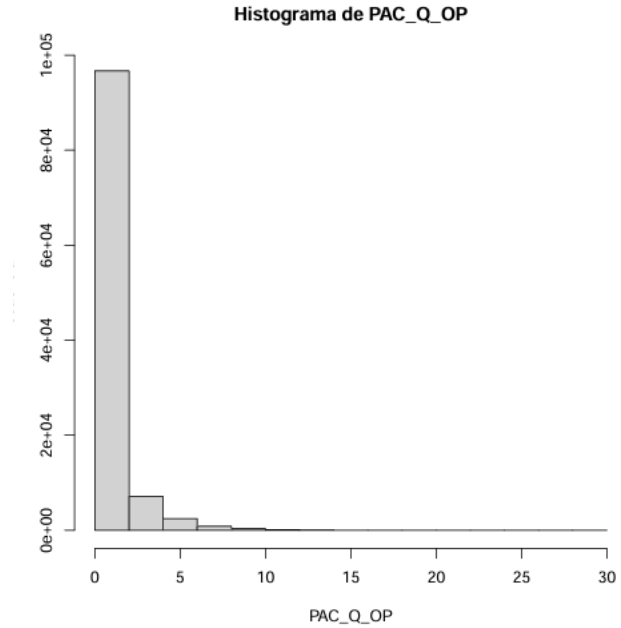
A39: *Histograma correspondiente a la variable CVISTA_SALDO_CLP. Fuente: Rstudio.*



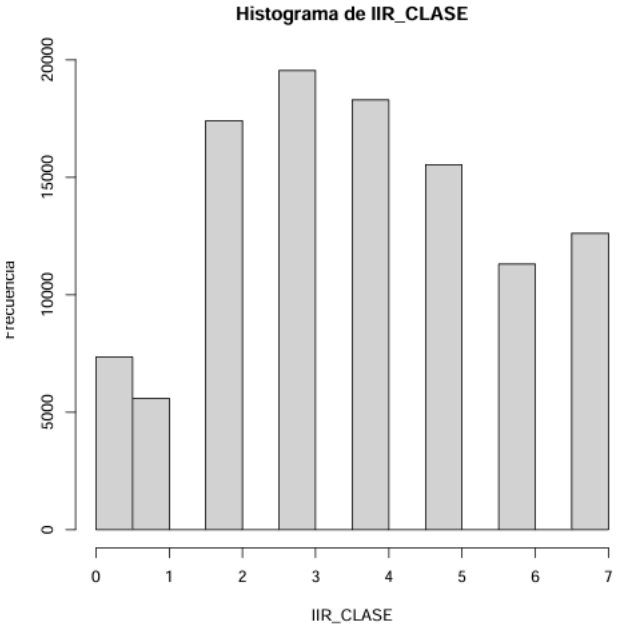
A40: *Histograma correspondiente a la variable CVISTA_SALDO_AVG_CLP. Fuente: Rstudio.*



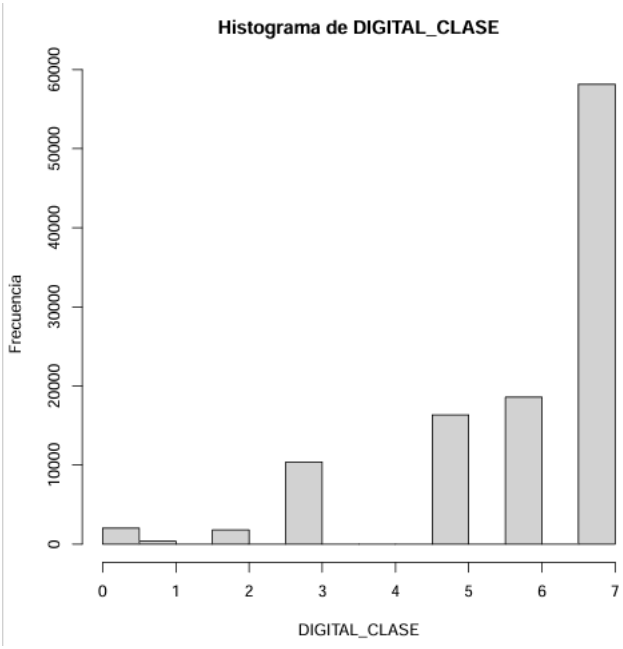
A41: *Histograma correspondiente a la variable PAT_Q_OP. Fuente: Rstudio.*



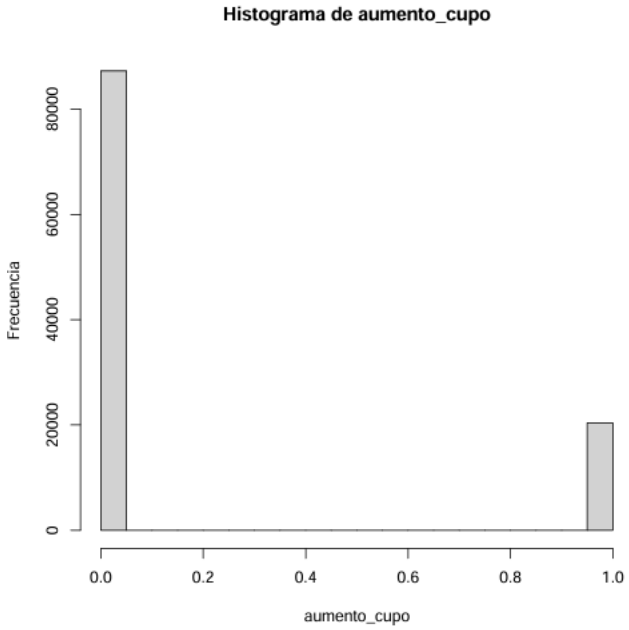
A42: *Histograma correspondiente a la variable PAC_Q_OP. Fuente: Rstudio.*



A43: *Histograma correspondiente a la variable IIR_CLASE. Fuente: Rstudio.*

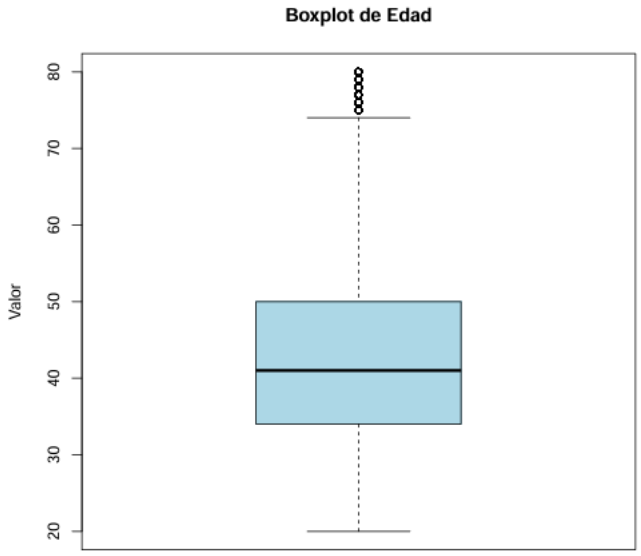


A44: *Histograma correspondiente a la variable IIR_CLASE. Fuente: Rstudio.*

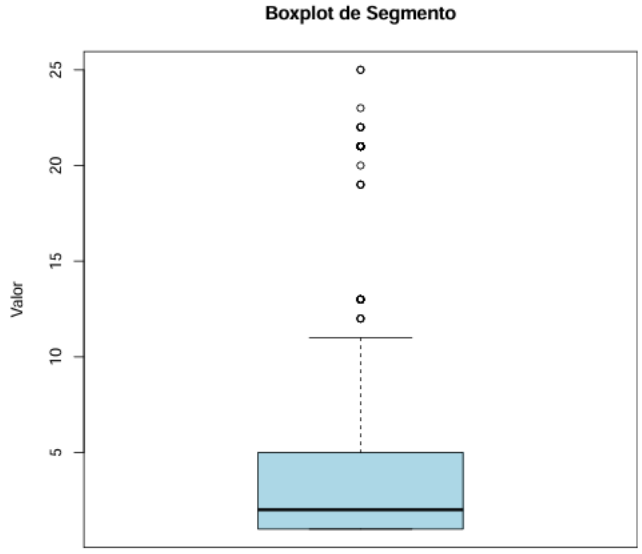


A45: *Histograma correspondiente a la variable aumento_cupo. Fuente: Rstudio.*

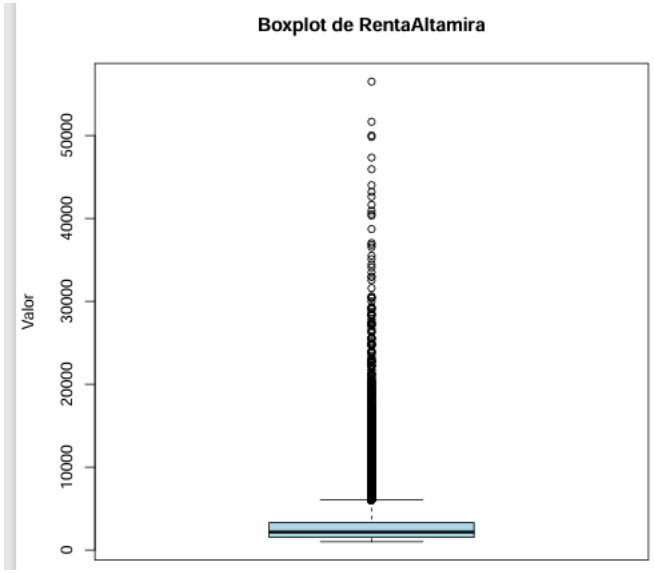
Diagrama de caja:



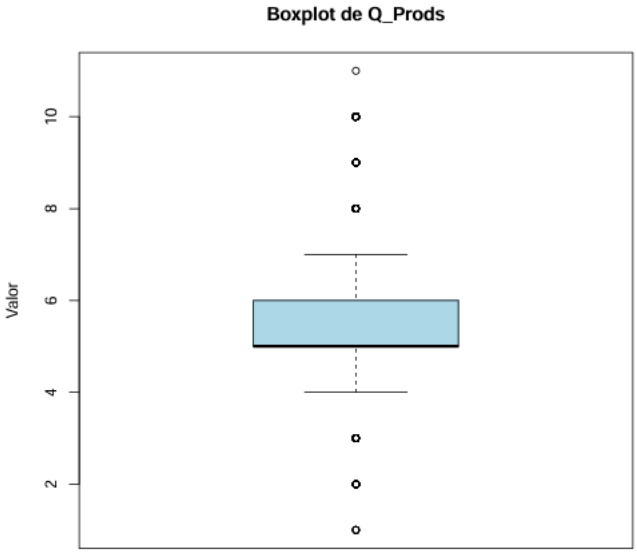
A46: *Gráfico de caja correspondiente a Edad. Fuente: Rstudio.*



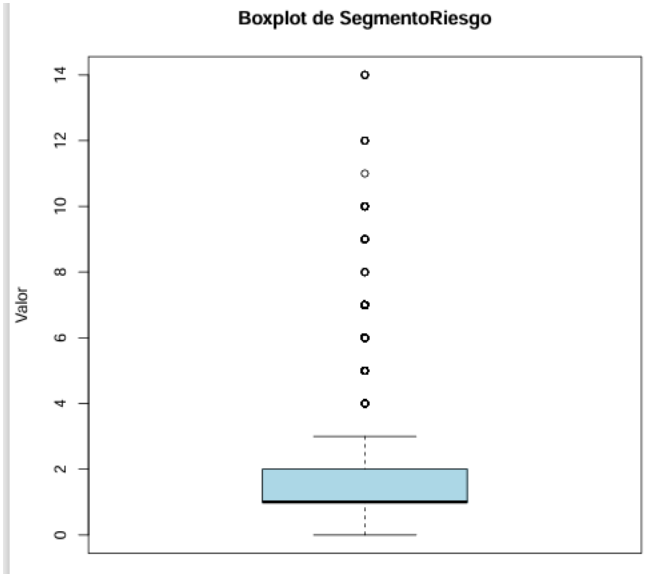
A47: *Gráfico de caja correspondiente a Segmento. Fuente: Rstudio.*



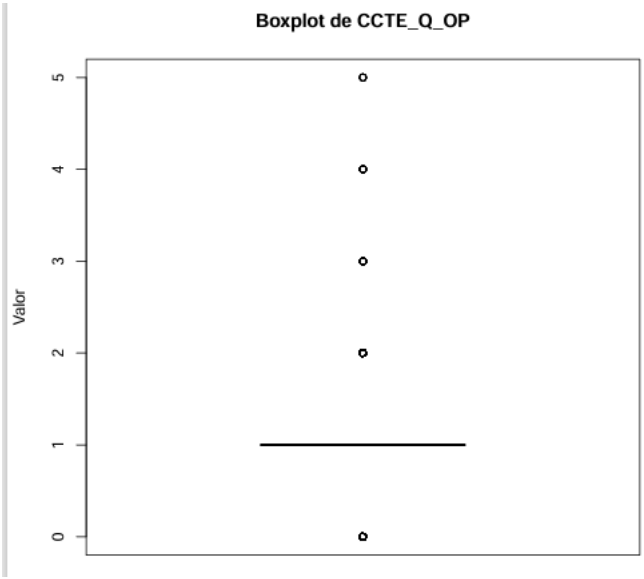
A48: *Gráfico de caja correspondiente a RentaAltamira. Fuente: Rstudio.*



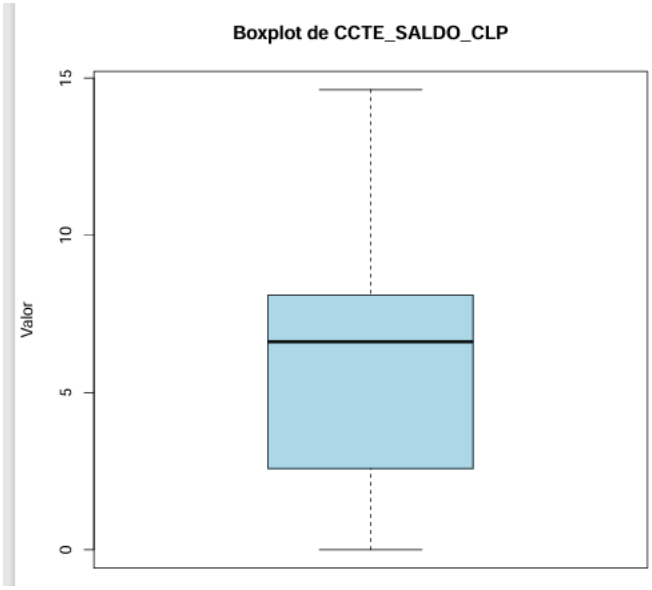
A49: : *Gráfico de caja correspondiente a Q_Prods. Fuente: Rstudio.*



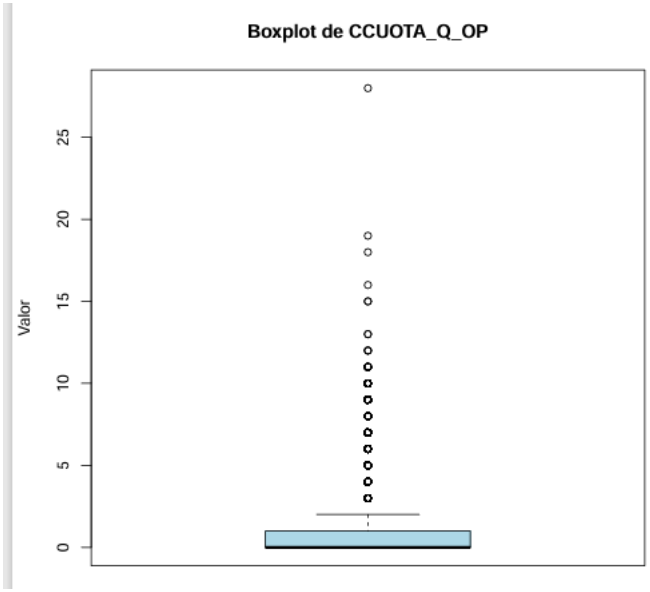
A50: *Gráfico de caja correspondiente a SegmentoRiesgo. Fuente: Rstudio.*



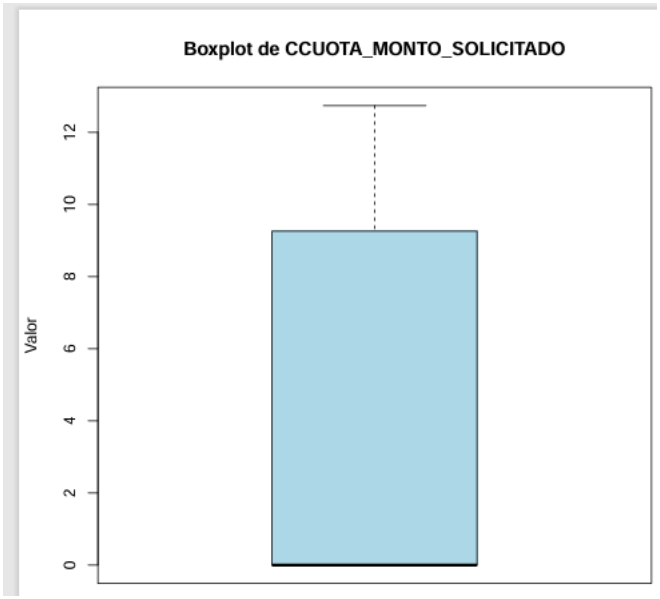
A51: : *Gráfico de caja correspondiente a CCTE_Q_OP. Fuente: Rstudio.*



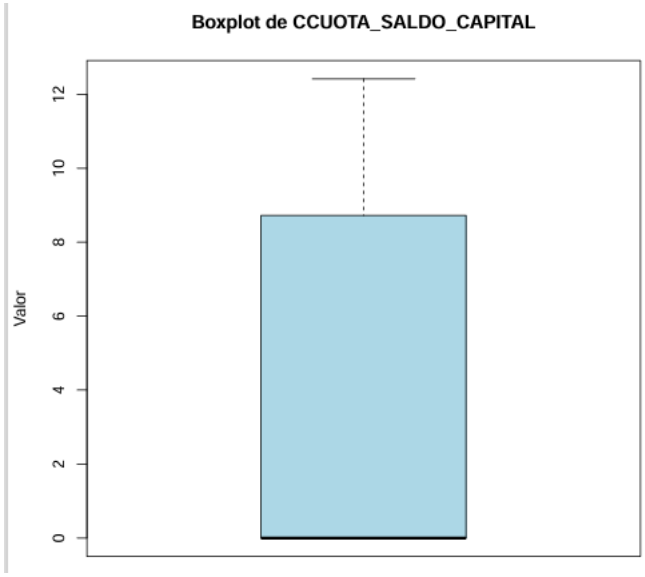
A52: : *Gráfico de caja correspondiente a CCTE_SALDO_CLP. Fuente: Rstudio.*



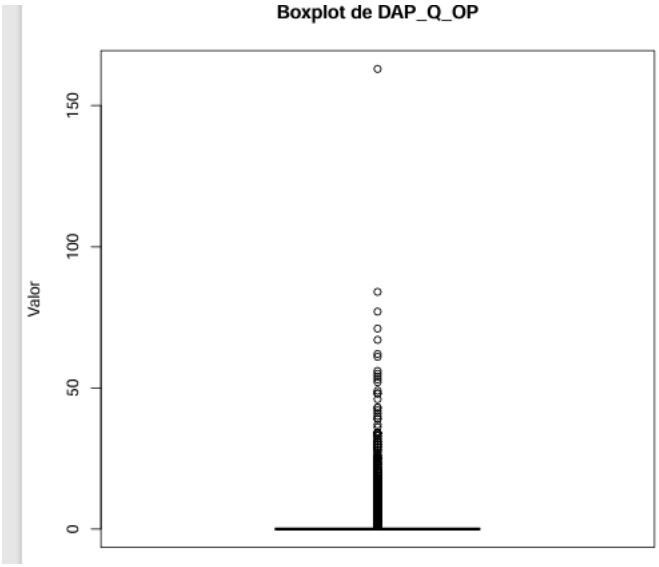
A53: : *Gráfico de caja correspondiente a CCUOTA_Q_OP. Fuente: Rstudio.*



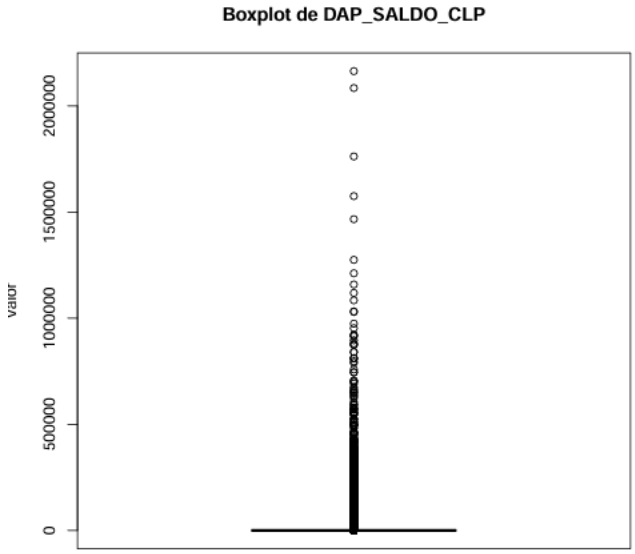
A54: *Gráfico de caja correspondiente a CCUOTA_MONTO_SOLICITADO. Fuente: Rstudio.*



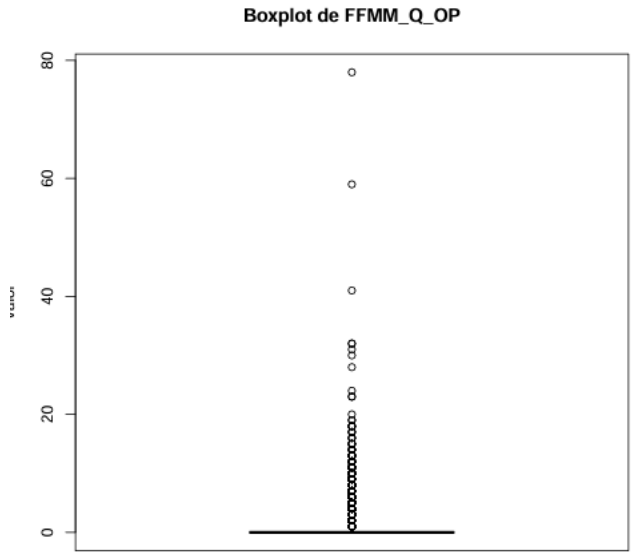
A55: : *Gráfico de caja correspondiente a CCUOTA_SALDO_CAPITAL. Fuente: Rstudio.*



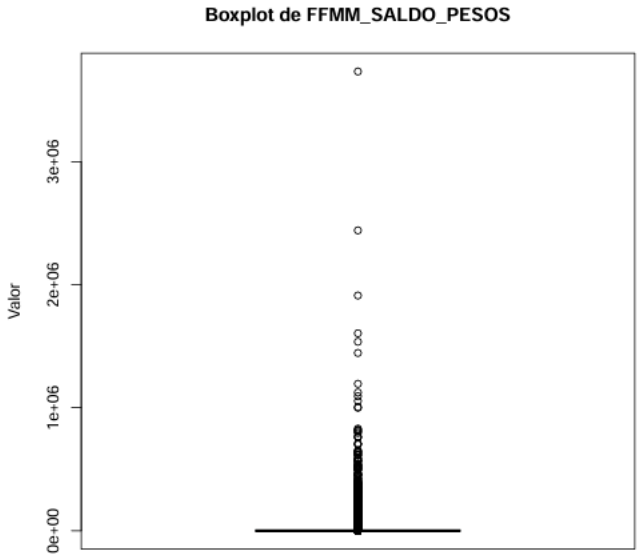
A56: : *Gráfico de caja correspondiente a DAP_Q_OP. Fuente: Rstudio.*



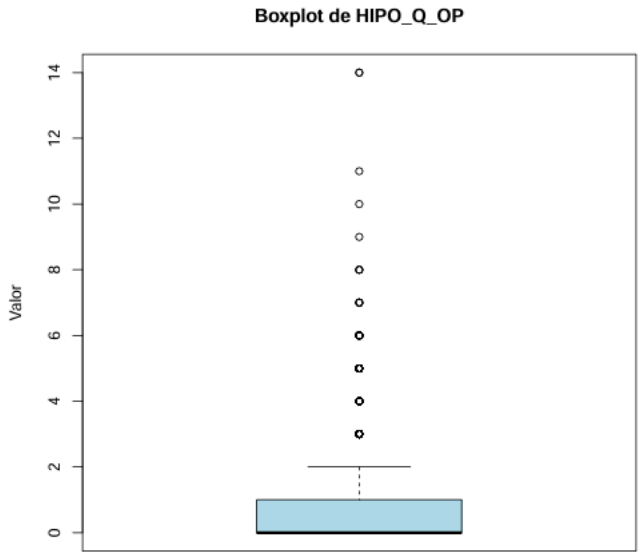
A57: : *Gráfico de caja correspondiente a DAP_SALDO_CLP. Fuente: Rstudio.*



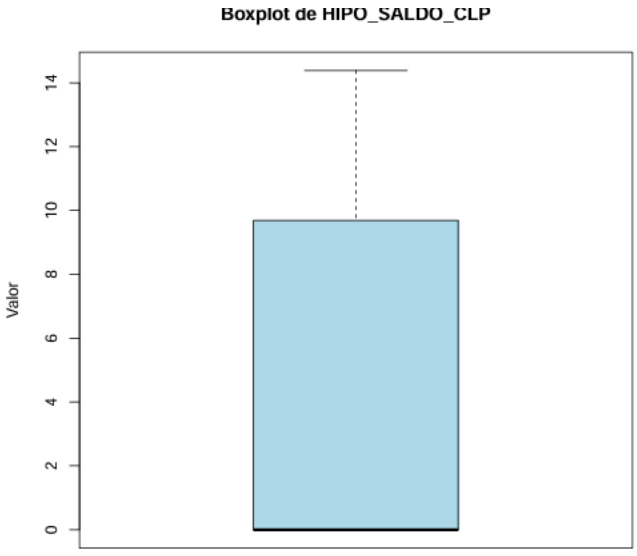
A58: : *Gráfico de caja correspondiente a FFMM_Q_OP. Fuente: Rstudio.*



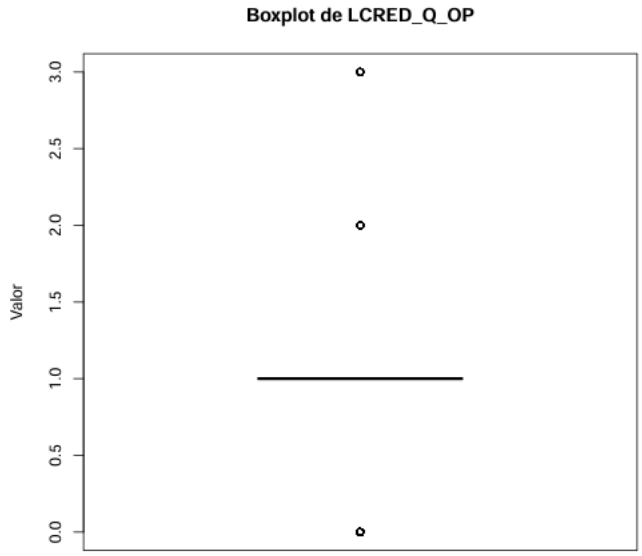
A59: : *Gráfico de caja correspondiente a FFMM_SALDO_PESOS. Fuente: Rstudio.*



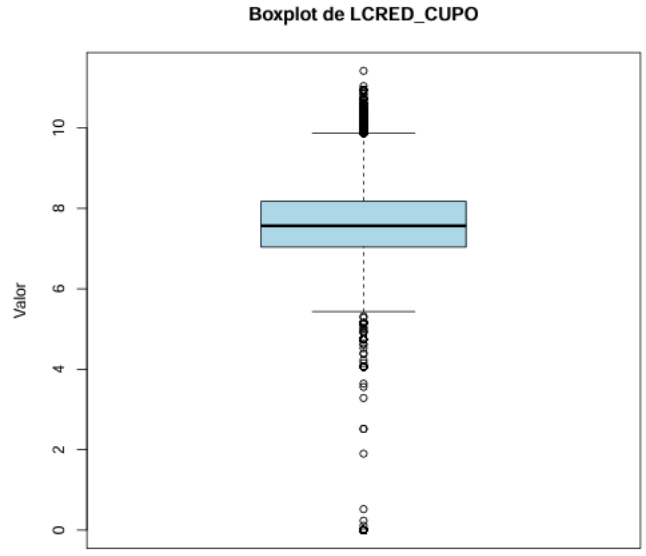
A60: : *Gráfico de caja correspondiente a HIPO_Q_OP. Fuente: Rstudio.*



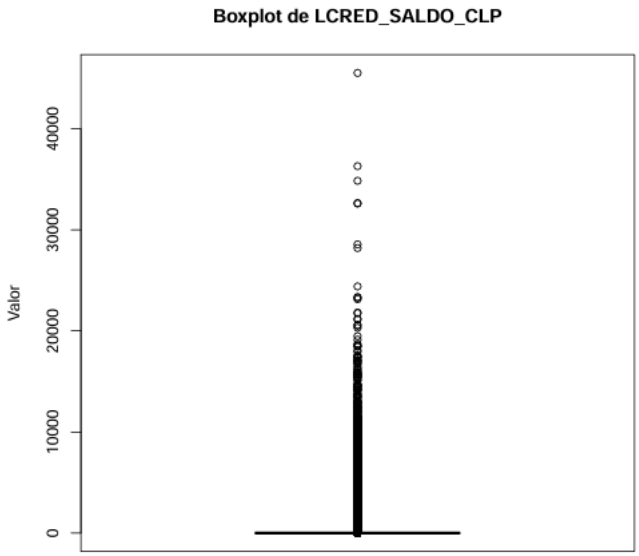
A61: : *Gráfico de caja correspondiente a HIPO_SALDO _CLP. Fuente: Rstudio.*



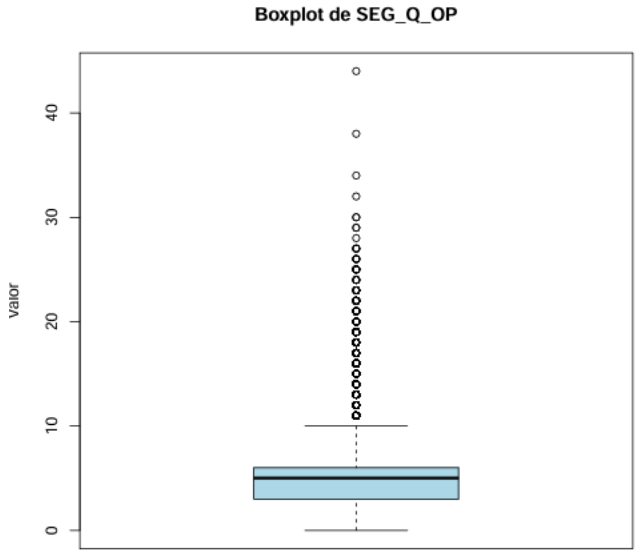
A62: : *Gráfico de caja correspondiente a LCRED_Q_OP. Fuente: Rstudio.*



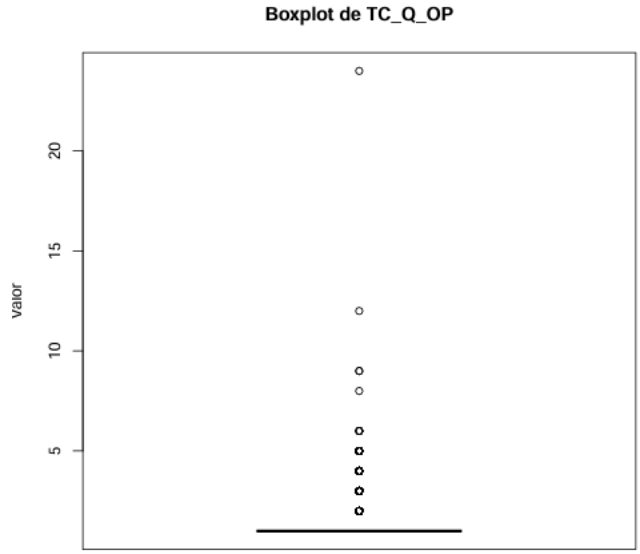
A63: : *Gráfico de caja correspondiente a LCRED_CUPO. Fuente: Rstudio*



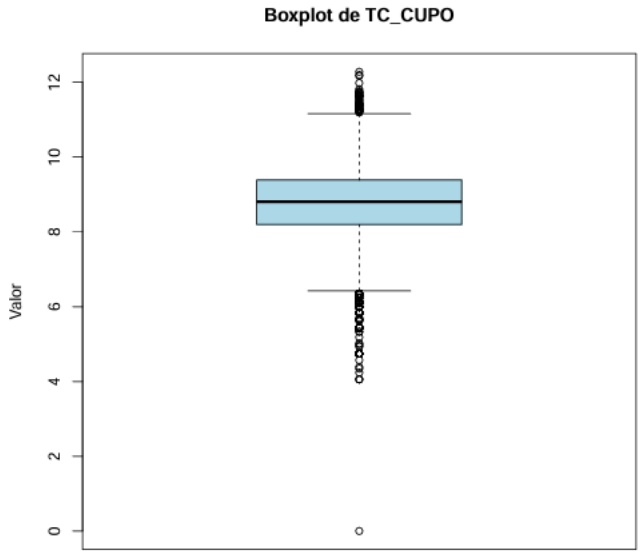
A64: : *Gráfico de caja correspondiente a LCRED_SALDO_CLP. Fuente: Rstudio*



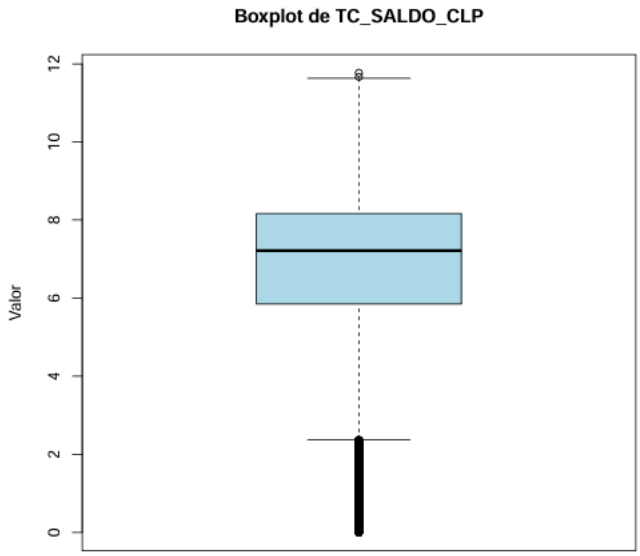
A65: : *Gráfico de caja correspondiente a SEG_Q_OP. Fuente: Rstudio*



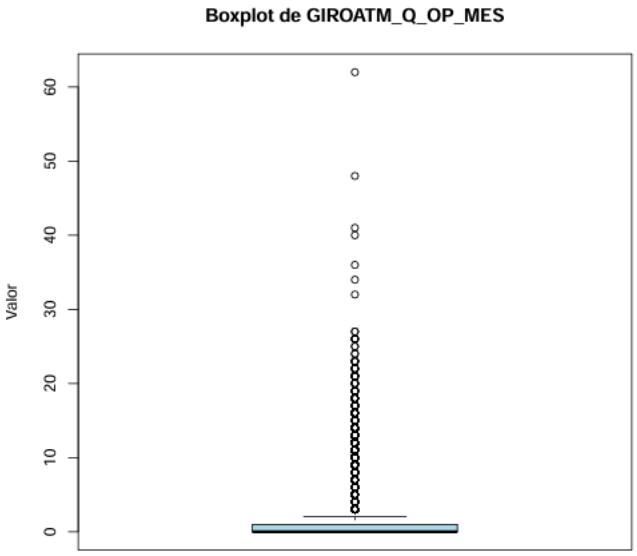
A66: : *Gráfico de caja correspondiente a TC_Q_OP. Fuente: Rstudio*



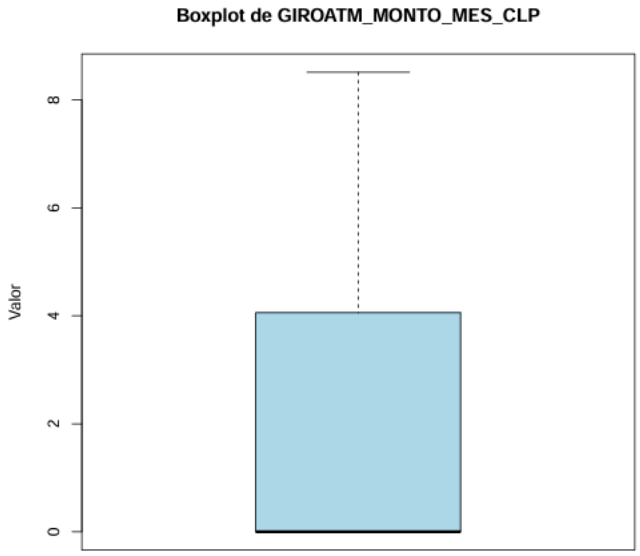
A67: : *Gráfico de caja correspondiente a TC_CUPO. Fuente: Rstudio*



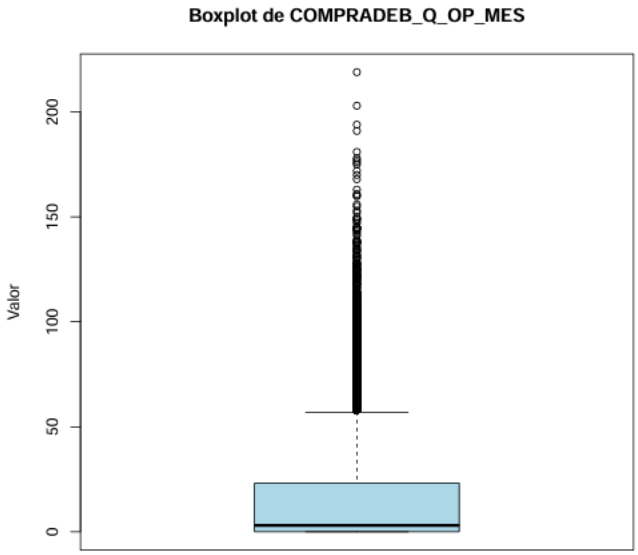
A68: : *Gráfico de caja correspondiente a TC_SALDO_CLP. Fuente: Rstudio*



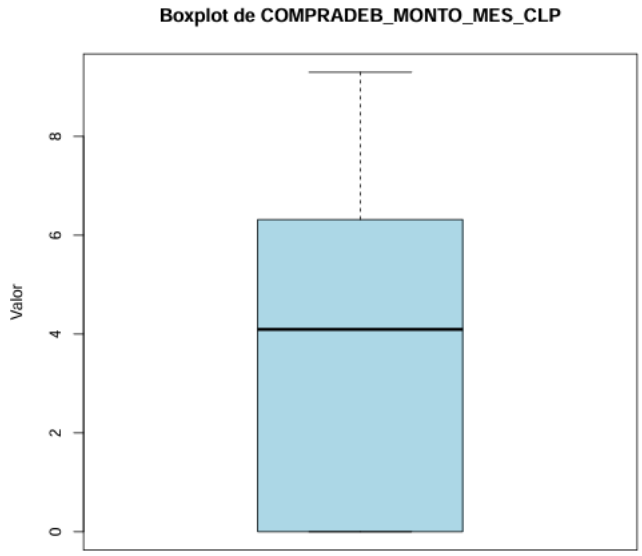
A69: : *Gráfico de caja correspondiente a GIROATM_Q_OP_MES. Fuente: Rstudio*



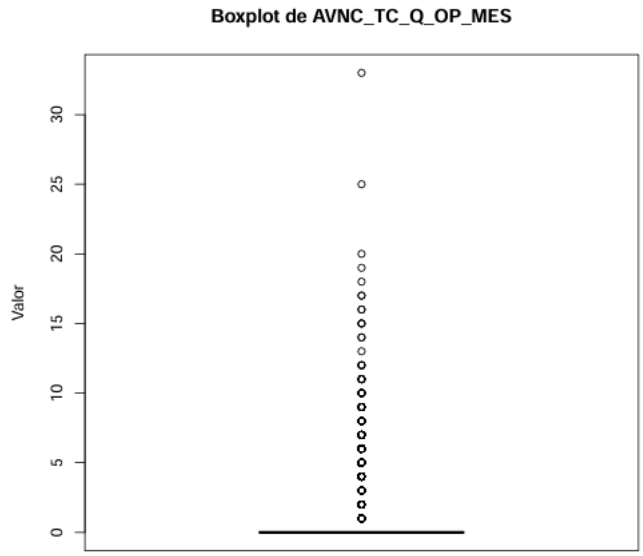
A70: : *Gráfico de caja correspondiente a GIROATM_MONTO_MES_CLP. Fuente: Rstudio*



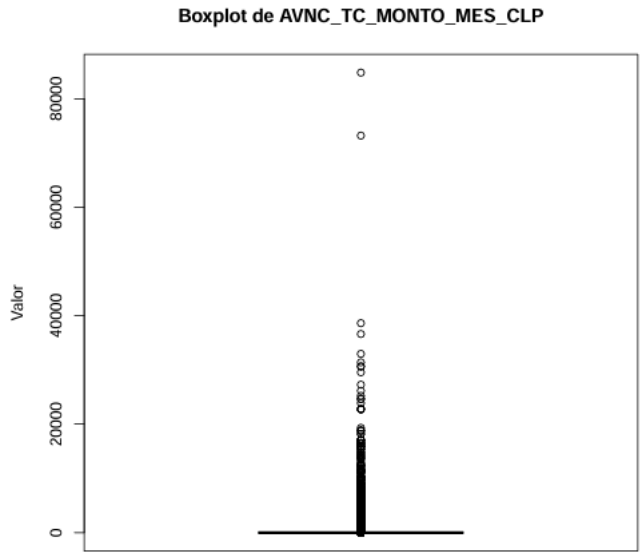
A71: : *Gráfico de caja correspondiente a COMPRADEB_Q_OP_MES. Fuente: Rstudio*



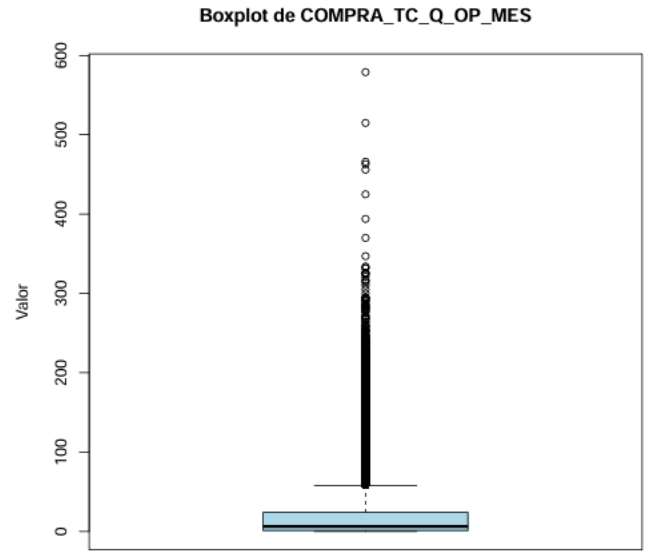
A72: : *Gráfico de caja correspondiente a COMPRADEB_MONTO_MES_CLP. Fuente: Rstudio*



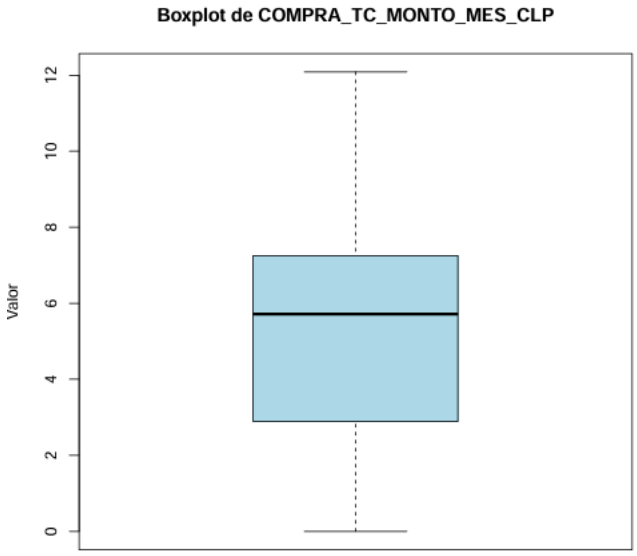
A73: : *Gráfico de caja correspondiente a AVNC_TC_Q_OP_MES. Fuente: Rstudio*



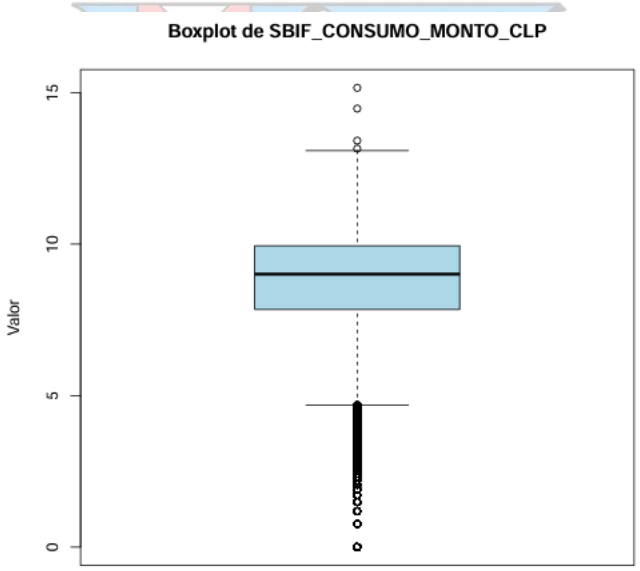
A74: *Gráfico de caja correspondiente a AVNC_TC_MONTO_MES_CLP. Fuente: Rstudio*



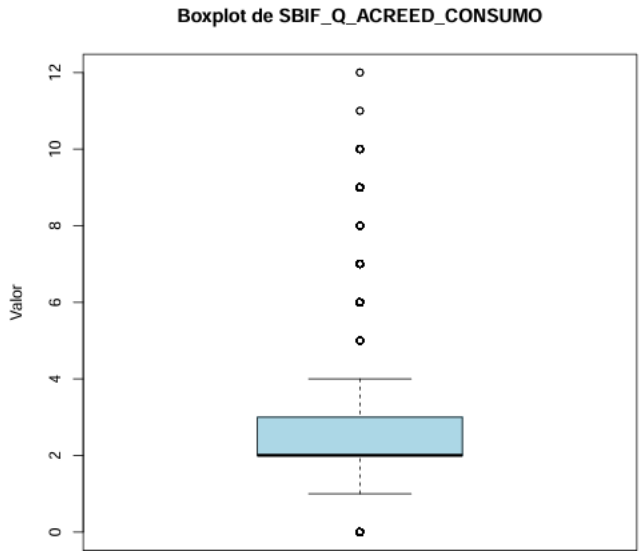
A75: : *Gráfico de caja correspondiente a COMPRA_TC_Q_OP_MES. Fuente: Rstudio*



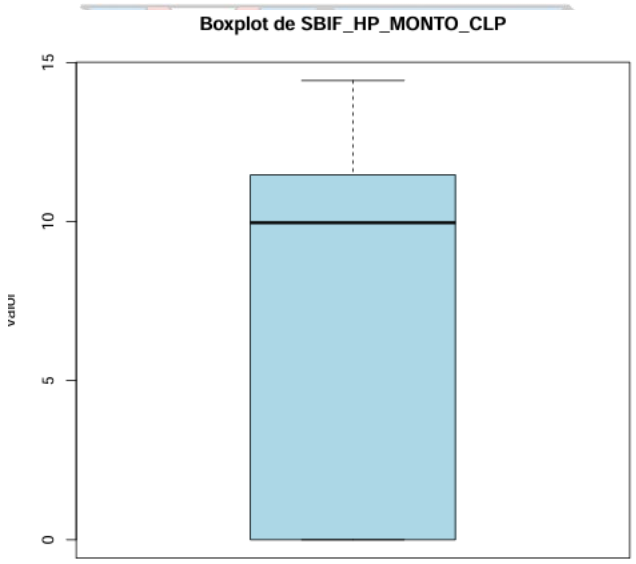
A76: : *Gráfico de caja correspondiente a COMPRA_TC_MONTO_MES_CLP. Fuente: Rstudio*



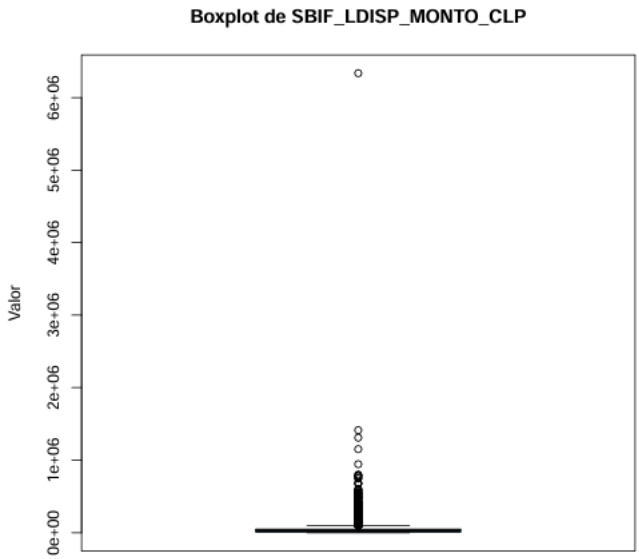
A77: *Gráfico de caja correspondiente a SBIF_CONSUMO_MONTO_CLP. Fuente: Rstudio*



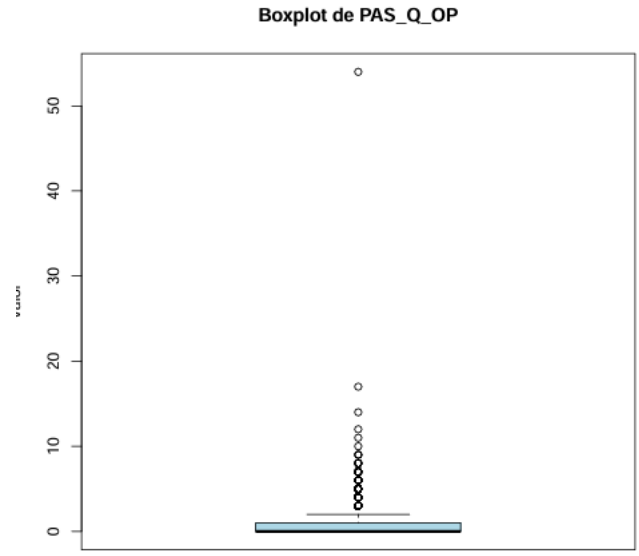
A78: *Gráfico de caja correspondiente a SBIF_Q_ACREED_CONSUMO. Fuente: Rstudio*



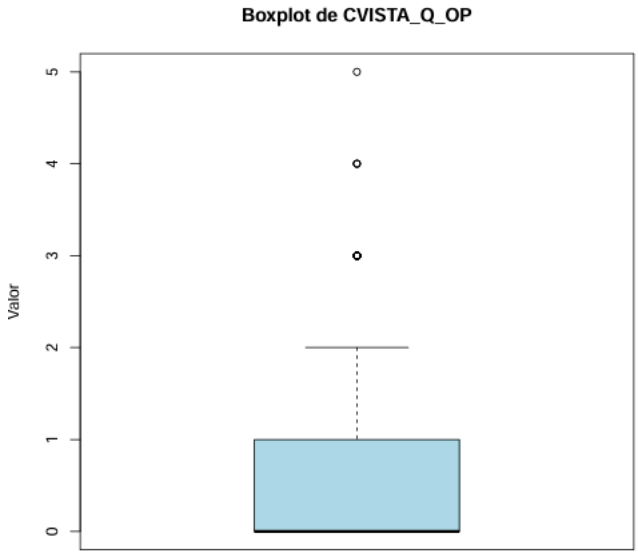
A79: *Gráfico de caja correspondiente a SBIF_HP_MONTO_CLP. Fuente: Rstudio*



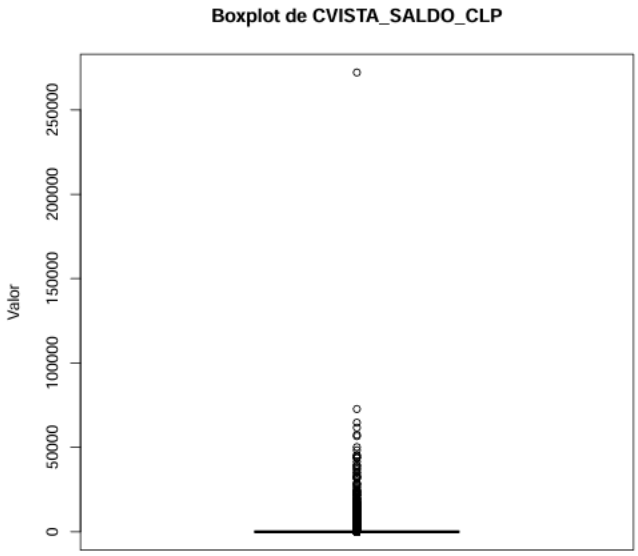
A80: *Gráfico de caja correspondiente a SBIF_LDISP_MONTO_CLP. Fuente: Rstudio*



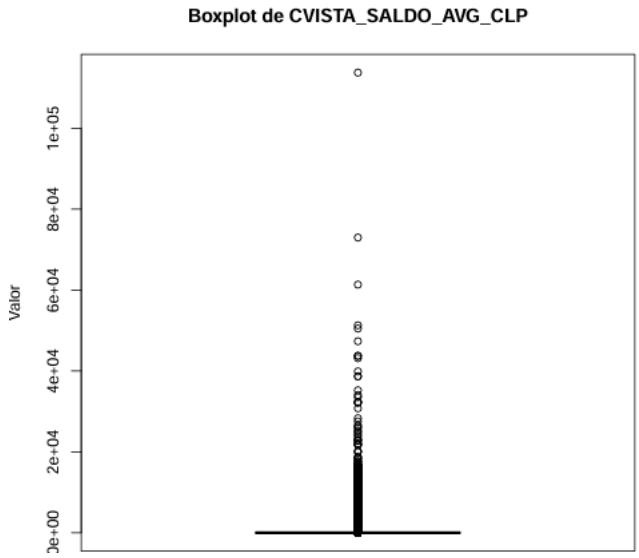
A81: *Gráfico de caja correspondiente a PASS_Q_OP. Fuente: Rstudio*



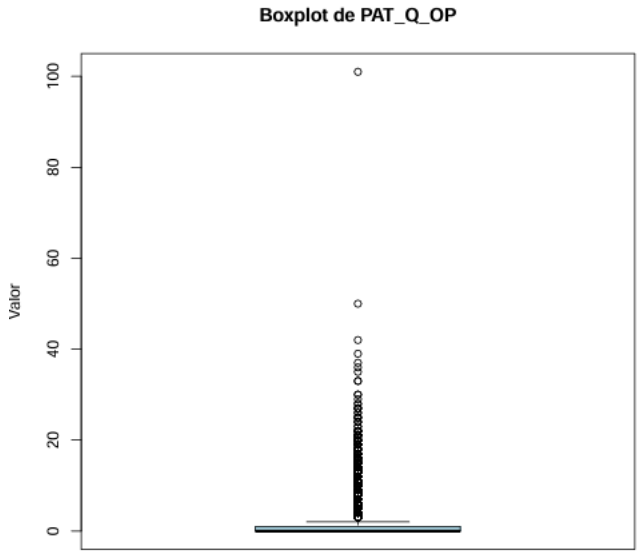
A82: *Gráfico de caja correspondiente a CVISTA_OP. Fuente: Rstudio*



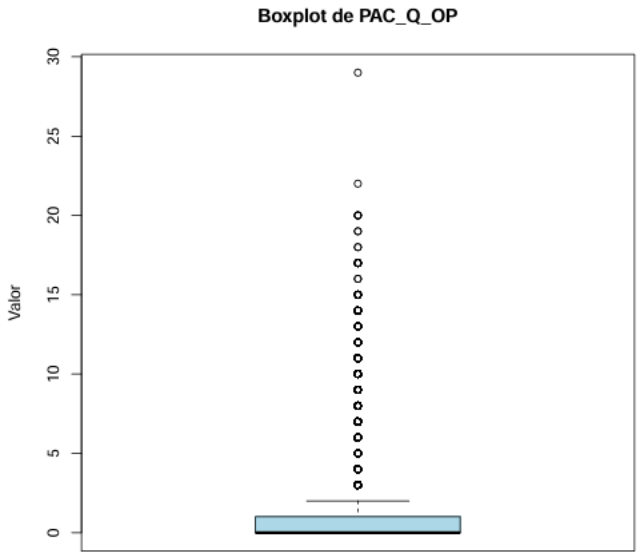
A83: *Gráfico de caja correspondiente a CVISTA_SALDO_CLP. Fuente: Rstudio*



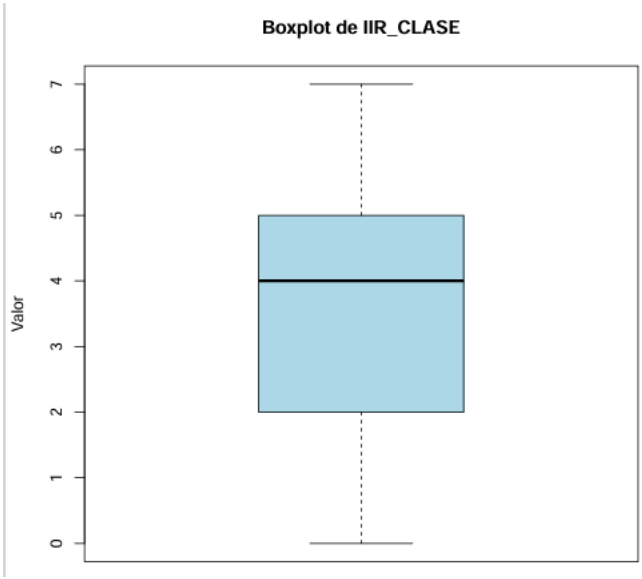
A84: *Gráfico de caja correspondiente a CVISTA_SALDO_AVG_CLP. Fuente: Rstudio*



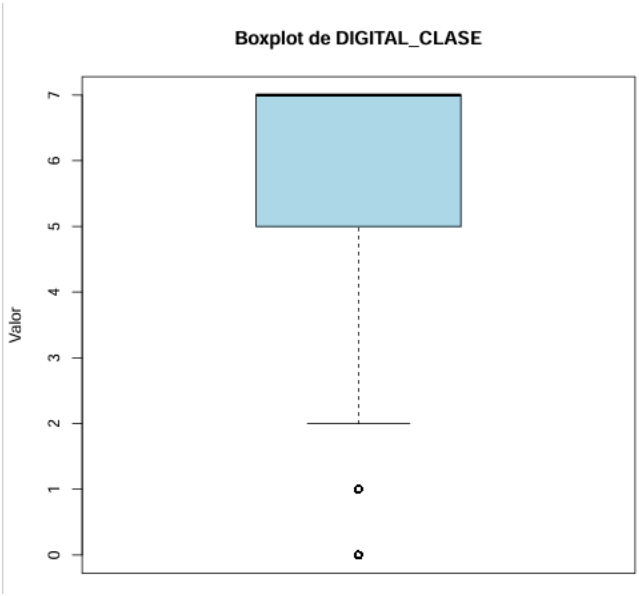
A85: *Gráfico de caja correspondiente a PAT_Q_OP. Fuente: Rstudio*



A86: *Gráfico de caja correspondiente a PAC_Q_OP. Fuente: Rstudio*



A87: *Gráfico de caja correspondiente a IIR_CLASE. Fuente: Rstudio*



A88: *Gráfico de caja correspondiente a DIGITAL_CLASE. Fuente: Rstudio*