UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

Ciclo Optativo de Especialización y Profesionalización en Marketing y Finanzas

"SEGMENTACIÓN DE LA BASE DE DATOS DE UN CALL CENTER PARA LAS VENTAS DEL SERVICIO DE TELEFONÍA MÓVIL, USANDO EL MODELO DE REGRESIÓN LOGÍSTICA Y EL ALGORITMO DE ÁRBOL DE CLASIFICACIÓN CART"

Trabajo de Titulación para optar el Título Profesional de :

INGENIERO ESTADÍSTICO E INFORMÁTICO

PALOMINO QUISPE JACQUELINE ROXANA
PRADO PARIONA VANESA

Lima – Perú

2015

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

Ciclo Optativo de Especialización y Profesionalización en

Marketing y Finanzas

"SEGMENTACIÓN DE LA BASE DE DATOS DE UN CALL CENTER PARA LAS VENTAS DEL SERVICIO DE TELEFONÍA MÓVIL, USANDO EL MODELO DE REGRESIÓN LOGÍSTICA Y EL ALGORITMO DE ÁRBOL DE CLASIFICACIÓN CART"

Presentado por :

PALOMINO QUISPE JACQUELINE ROXANA
PRADO PARIONA VANESA

Trabajo de Titulación para optar el Título Profesional de :

INGENIERO ESTADÍSTICO E INFORMÁTICO

Sustentado y aprobado ante el siguiente jurado:

Mg. Sc. Clodomiro Miranda Villagómez
PRESIDENTE

Mg. Sc. Jesús Salinas Flores
MIEMBRO

M. S.c. Carlos López de Castilla Vásquez
MIEMBRO

M.S. Rino Sotomayor Ruiz
ASESOR

DEDICATORIA

A mis padres Adela y Alipio por su gran apoyo incondicional, día a día, sin perder las esperanzas de que lograría mucho más, como ellos dicen tengo un largo camino por recorrer y estoy segura que con su gran apoyo lo conseguiré, también agradezco de corazón a mis tíos Antonio y Martha quienes depositaron su confianza, creyendo que iba a alcanzar lo que ahora soy y gracias a Dios lo conseguí. Estoy segura que lograre mucho más.

Jacqueline Roxana Palomino Quispe

A mis padres Rafael y Rómula por que creyeron en mí en todo momento, por sus palabras de ánimo y motivación a lo largo de mi carrera, por el amor y el apoyo incondicional brindado, por los valores inculcados, por enseñarme a no rendirme jamás que con paciencia y esfuerzo se puede lograr todo, gracias a ustedes hoy puedo ver concretada mi meta, agradezco a Dios por darme a los mejores padres del mundo estoy muy orgullosa que sean mis padres los amo y este logro es para ustedes.

Vanesa Prado Pariona

AGRADECIMIENTOS

A nuestro asesor el M.S. Rino Sotomayor Ruiz, por la confianza otorgada, sus aportes, enseñanzas y sus consejos compartidos para la realización del trabajo de investigación.

A nuestros profesores miembros del jurado por compartir sus conocimientos, aportes y recomendaciones brindadas para el trabajo de investigación.

Finalmente un eterno agradecimiento a esta prestigiosa Universidad, la cual nos ha formado para ser grandes profesionales de éxito y habernos preparado para un futuro competitivo.

ÍNDICE

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

El trabajo, consistió en detallar paso a paso la metodología (**CRISP–DM)** para poder identificar grupos óptimos de clientes más propensos a migrar de un plan prepago a postpago con el fin de formular un plan de mejora en la gestión de llamadas mediante la clasificación de la base de datos.

Este trabajo ha sido motivado por que actualmente se ha visto una disminución de la tasa de efectividad y contactabilidad con los clientes, para esto se ha utilizado el software Rapid Miner ya que es más detallada la representación de flujos de manera gráfica y por su gran capacidad para trabajar con una amplia gama de bases de datos.

Se aplicaron modelos de clasificación para analizar las características que genera la compra de los diferentes servicios. Se realizó la comparación del modelo de Regresión Logística y el algoritmo de Árbol de Clasificación CART, quedando como modelo más óptimo la Regresión Logística ya que ofreció mejores resultados y mayor efectividad.

A partir de lo anterior, se encontraron grupos diferenciados por las probabilidades de éxito venta (Migrar de un plan prepago a postpago), segmentos que reflejan necesidades y características particulares, que permita diseñar acciones de marketing focalizado con el objetivo de incrementar la tasa de efectividad, contactabilidad e incrementar las ventas.

Se realizaron recomendaciones para futuras acciones de marketing, un ejemplo es identificar grupos que se debe intentar desarrollar y otros grupos que sólo que se debe tratar de fidelizar, ya que han alcanzado gran parte de su potencial dentro de la empresa.

Cómo trabajos futuros se recomienda replicar la metodología con mayor información demográfica, con el fin de aumentar los índices de desempeño de los modelos predictivos. Además de poder cuantificar el aumento de la efectividad debido a la aplicación de esta metodología, a través de una campaña real.

I. INTRODUCCIÓN

Hoy en día la industria de telemarketing es muy competitiva, es por esto, que para mantenerse en los primeros lugares los esfuerzos de marketing deben ser máximos y eficientes. Para esto será primordial entender y entregar los productos que desean los consumidores.

La industria de telemarketing ha recibido una detenida atención académica en estos últimos años debido a su crecimiento numérico, su diseminación geográfica y debido a su extensión en término de operaciones, teniendo cada vez mayores actividades y de creciente complejidad. Como parte de este interés, distintas dimensiones han sido exploradas en estos espacios laborales.

Actualmente el Call Center posee una baja efectividad en las campañas de cross y up selling, es por esto que se busca encontrar una clasificación de la base de clientes, que permita diseñar acciones de marketing focalizado con el objetivo de aumentar la tenencia de productos por cliente y aumentar la efectividad de las campañas.

El objetivo de este trabajo de investigación fue realizar una metodología de clasificación que permita identificar grupos óptimos de clientes más propensos a migrar de un plan prepago a postpago usando el modelo de Regresión Logística y el algoritmo de Árbol de Clasificación CART, de tal manera que uno de ellos sea aplicado a la base de clientes del Call center y defina las características que generan la compra de los diferentes productos, con ello se busca aumentar el nivel de eficacia de las acciones comerciales.

Para este trabajo de investigación no fue necesario realizar levantamiento de información socio demográfica y de estilos de vida de los clientes, debido a que se contó con una base de datos proporcionada por el Call Center, lo cual facilitó la implementación y desarrollo del respectivo proyecto.

1.1. EL PROBLEMA DE INVESTIGACIÓN

El Call Center o Centro de atención de llamadas ofrece diferentes servicios y uno de ellos es realizar ventas out a usuarios de Telefonía Móvil (Movistar Perú), ha operado normalmente en el pasado, sin embargo actualmente por el crecimiento de la demanda se está presentando una deficiente atención al cliente, esto se da debido al incremento de empresas con servicios muy similares, algunos de ellos con diversos niveles de calidad, marcadores eficientes (aplicativos para realizar llamadas MOSAIX). Adicionalmente las necesidades de los clientes en este mercado van evolucionando rápidamente, donde cada vez se ve reflejada la insatisfacción de este servicio.

La cantidad mínima de ventas logradas cada mes de gestión (llamadas a los clientes) no cubre los gastos realizados por el Call Center, no logra identificar clientes que estén más propensos a migrar de un plan prepago a postpago sin necesidad de realizar un número elevado de llamadas y minimizar costos, es por ello que se vio conveniente realizar una clasificación a la base de datos, identificando las variables más influyentes para encontrar grupos de clientes propensos a adquirir el producto ofrecido.

- **FORMULACIÓN DEL PROBLEMA**

¿Es posible clasificar la base de datos del Call Center para identificar clientes más propensos a migrar de un plan prepago a postpago al momento de una llamada telefónica, usando el modelo de Regresión Logística y el algoritmo de Árbol de Clasificación CART?

1.2. OBJETIVOS DE LA INVESTIGACIÓN

Objetivo General:

- Identificar grupos óptimos de clientes más propensos a migrar de un plan prepago a postpago mediante la clasificación de la base de datos, usando un modelo de Regresión Logística y el algoritmo de Árbol de Clasificación CART.

Objetivos Específicos:

- Identificar los factores asociados hacia la posibilidad de migración de un cliente prepago a postpago mediante la aplicación de técnicas estadísticas multivariadas como la Regresión Logística y Árboles de Clasificación.
- Diferenciar, con la clasificación realizada, a los clientes esporádicos para mantener relaciones continuas en el tiempo, de los clientes potenciales que generan valor al Call Center obteniendo la migración de un cliente prepago a postpago.
- Determinar las variables que mejor explican las diferencias encontradas en los diversos grupos de clientes en la decisión de adquirir un nuevo plan, para ser consideradas en el Modelo de Regresión Logística y Árboles de Clasificación.
- Comparar la técnica paramétrica (Regresión Logística) y no paramétrica (Árbol de Clasificación usando el algoritmo CART), y así analizar sus ventajas y desventajas, para decidir que técnica ofrece mejores resultados y mayor efectividad.
- Incentivar al asesor de ventas mediante los merchandising ofrecidos por Telefónica, para enfocarse en los grupos óptimos que se encontrarán en el proyecto de investigación, con el fin de incrementar las ventas obteniendo mayor efectividad en el Call Center.
- Definir una estrategia implementando el mejor modelo en el trabajo de investigación para contrarrestar los problemas de efectividad y contactabilidad en la gestión de llamadas.

1.3. JUSTIFICACIÓN

En un mercado tan competitivo como el de las telecomunicaciones, donde las necesidades de los clientes evolucionan y los desarrollos tecnológicos avanzan constantemente es necesario contar con un canal de comunicación efectivo y eficiente con el cliente, para poder brindar una atención oportuna a sus requerimientos.

Debido a las características de consumo, nivel socioeconómico, comportamiento, etc. que puede presentar un cliente de telefonía celular prepago, la probabilidad que un cliente

acepte en un eventual ofrecimiento, migrarse a un plan postpago, depende básicamente del tipo de necesidades que puede tener el cliente al cual se le ofrece la promoción.

En el Perú hoy en día se está incrementando el tema de la gestión de grandes volúmenes de base de datos para poder detectar perfiles y comportamientos de los clientes, y de esta manera enfocar los esfuerzos del Marketing hacia clientes que si son más propensos a adquirir una promoción, ya sea creando productos que pueden ser del interés de dichos clientes, o en su defecto detectando perfiles que son potenciales a ofrecerles nuestro producto actual por su comportamiento en usos de telefonía celular.

Si bien es cierto los estudios para la administración de base de datos se desarrollan en diferentes campos, poco o casi nada se puede encontrar acerca de cómo se puede optimizar la gestión de una cartera de cliente prepago y ofrecerles migrar a postpago a través de llamadas telefónicas (Call Center – Gestion Outbound).

Por esta razón, se decidió desarrollar el presente trabajo de investigación que busca la manera de poder clasificar a los clientes potenciales que desean adquirir el producto ofrecido, logrando de esta manera minimizar costos de gestión, llamadas, operadores, tiempo, etc. y cumplir con el objetivo principal del negocio que es incrementar el volumen de migraciones.

Para el análisis y obtención de resultados se utilizaron las técnicas de clasificación de base de datos: Regresión Logística y Arboles de Clasificación usando el algoritmo CART las cuales permitieron un panorama más amplio acerca de los diferentes tipos de clientes que tiene el Call Center.

II. REVISIÓN DE LITERATURA

2.1. ANTECEDENTES DE LA INVESTIGACIÓN

La evolución de la telefonía móvil en los últimos 10 años ha generado un proceso acelerado de adaptación de dispositivos móviles en el mundo.

En la actualidad se estima que existen igual cantidad de dispositivos móviles que habitantes en el planeta. Según la ITU (International Telecommunication Union) de las Naciones Unidas, a principios del 2013 existían 6.8 billones de equipos celulares.

De acuerdo al estudio de Ipsos Apoyo (2013), dentro de los principales usos que se le da al dispositivo móvil en el Perú (aparte de hacer llamadas), está el enviar mensajes de texto (SMS), chatear a través de aplicativos de mensajería instantánea tipo BBM, WhatsApp, Facebook, Twitter, etc., otros usos adicionales son el de la cámara de fotos y conectarse a Internet para buscar información.

Los esfuerzos del Marketing en la telefonía móvil deben ser cada vez más agresivos, los consumidores buscan diariamente ofertas y promociones para adquirir un teléfono móvil así mismo un plan de datos adicional y todo esto al menor costo posible, la encuesta anual de Opsitel demuestra que en el Perú todavía estamos atrasados tanto en la penetración de celulares en general como de Smartphone respecto a países desarrollados. Una de las principales razones para no contar con un equipo móvil es el factor económico, habiendo un 45 % de personas que no lo tienen porque le resulta muy caro. Se necesita una mayor y mejor oferta de operadores móviles para masificar el acceso a la telefonía móvil.

La evolución de la telefonía móvil en el Perú, está impactando el marketing y la publicidad, ya que conforme la gente pasa más tiempo con su dispositivo móvil, interactúa más con él y se conecta más tiempo a internet, aumenta su necesidad de comunicación y de esta manera se puede aprovechar para impactar en clientes y potenciales clientes para ofrecerles un determinado plan o servicio.

Sin embargo, todavía hay una gran brecha a nivel mundial entre lo que se invierte en marketing móvil versus el tiempo que cada vez más le dedican las personas a esta herramienta. Es fundamental que las empresas comiencen a desarrollar estrategias de marketing móvil y comenzar a integrarlo dentro del "mix" de herramientas que están utilizando para alcanzar a su público objetivo.

Según, José Miguel Gamero (director de Marketing de Telefónica Móviles Movistar), "El mercado de la telefonía móvil sigue creciendo en el Perú y ya reporta una penetración a nivel nacional de 80%, actualmente existen casi 23 millones de líneas celulares en servicio gracias a la mayor capacidad de consumo de los peruanos. El mercado está evolucionando y si el consumidor peruano es muy cuidadoso con el precio también lo es respecto al servicio y valora mucho las mejoras en su calidad de vida gracias a poder estar conectado, por ello lo más importante para el consumidor peruano es estar conectado."
Gamero consideró que actualmente el consumidor peruano se conecta por telefonía móvil a niveles bastante económicos y competitivos con cualquier mercado en el mundo y en la región.

Según Gustavo Kitazono Sugahara (director del segmento masivo de Movistar), "Los clientes prepago podrán ser ahora parte de la Red Privada Móvil, que les permitirá estar conectados ilimitadamente a una tarifa preferencial con empresas, personas, etc. Dentro del Segmento Masivo Movistar, se busca incrementar la comunidad RPM para que sean más los usuarios que se beneficien comunicándose a un menor precio. El objetivo es dar a los clientes prepago con más meses de antigüedad la posibilidad de aprovechar todas las oportunidades que brinda ser parte de la red privada móvil más grande del país".

El artículo titulado "Modelos Churn para clientes prepago de una empresa de telecomunicaciones de celulares que utilizan grandes mercados de datos" (2011), presenta el análisis y comparación de dos conocidas técnicas (Modelo de Regresión Logística y Arboles de decisión) para predecir el churn (fuga) de clientes de una compañía de telecomunicación móvil de Polonia. Con estos modelos, el operador de telecomunicaciones podría ofrecerles, por ejemplo, una promoción de tiempo adicional gratuito (o cualquier otro incentivo) y en consecuencia, se las arregla para retener al 25% de los potenciales churners. También se logró demostrar que la Regresión Logística es una buena opción cuando se está modelando la rotación de clientes prepago.

En la Universidad Nacional de Ingeniería; Jack Lazo (2012) realizó un "Modelo de Propensión para la identificación de Clientes fraude en el servicio de Telefonía Móvil para clientes postpago aplicando la Regresión Logística". La situación en la que se desarrolló la investigación fue en la que la reventa de servicios de telefonía móvil fue alta. Esto ocurrió hacia el año 2008, año en el que era común hallar personas con chalecos verdes fosforescentes en los comercios, universidades y principales avenidas ofreciendo el servicio de llamadas a distintos operadores. El presente estudio tiene en consideración a los clientes post pago de Telefónica del Perú en sus diversos planes, los cuales ascienden según OSIPTEL 1.6 millones de líneas a nivel nacional. El resultado del modelo de Regresión Logística, permite identificar aquellos clientes que son más probables a ser un cliente fraude, para aquellos que tienen una probabilidad mayor a 0.9, los cuales se sugiere otro tipo de acción. Los chalequeros se concentran en 80% en líneas Post pago esto es porque no es tan rentable las líneas de control.

El artículo titulado "Modelos de Oportunidad de Mejora para el Cliente en la elección de su plan celular mediante el diseño de una datawarehouse y árboles de Clasificación" de una empresa de telecomunicaciones de celulares presenta un análisis para la toma de decisiones en los dos tipos de productos que ofrece la empresa (Plan Controlado y Plan Prepago) considerados los de mayor captación en la empresa. En donde se le ofrece a cada cliente una mejor opción contra el producto que posee. Para esto se realizó un análisis en los clientes activos sobre los últimos tres meses de consumo y así poder presentarle opciones que le permitan obtener un mejor beneficio según sus hábitos de consumo, es decir ver si sus consumos son mayores al plan que tiene actualmente, se le puede brindar un plan que cubra con las expectativas del cliente y con esto atraerlo con un buen servicio y a su vez aumentar el mercado. La empresa celular basándose en la información que ha recogido en sus bases de datos, deberá tomar la mejor decisión con respecto a dichos planes, para así presentarles a sus clientes una mejor opción mostrándole mayor beneficio al obtener el nuevo plan. El Resultado del Modelo de árboles de Clasificación permite identificar aquellos clientes que tienen el mismo patrón de comportamiento para poder agruparlos y de esta manera ofrecerles un determinado tipo de plan que se adecue a sus necesidades y que genere mayor rentabilidad para la empresa.

2.2. BASE TEORICA

Frecuentemente la investigación estadística se ve enfrentada a manipular grandes cantidades de datos complejos que incluyen un gran número de variables, de los cuales es necesario obtener información, encontrar patrones y definir tendencias. Como se mencionó anteriormente, el objetivo del presente trabajo de investigación fue identificar grupos óptimos de clientes usando el modelo de Regresión Logística y Árboles de clasificación CART como una herramienta de clasificación, teniendo siempre en cuenta que la clasificación se aplicó para obtener una predicción más precisa de un evento (en este caso los clientes propensos a migrar de un plan prepago a postpago al momento de una llamada telefónica) en el universo analizado.

2.2.1. Caracterización del Call Center

El Call Center o Centro de Atención de Llamadas es un conjunto de herramientas de Informática y de Telecomunicaciones que, puestas a disposición de un grupo de operadores encargados de atender llamadas telefónicas masivas, eleva la productividad de los recursos tecnológicos y de los recursos humanos. El mayor valor agregado que proporciona un Call Center bien equipado es registrar la historia de los contactos potenciando una mejor atención a sus clientes.

El Call Center constituye un fenómeno que está dando lugar a un cambio radical en la forma de operar de las empresas. Entre los servicios que se suelen prestar figuran:

- Atención al cliente.
- Encuestas telefónicas (estudios de mercado, sondeos de opinión, calidad y satisfacción de clientes).
- Creación y actualización de bases de datos.
- Seguimiento de acciones de marketing.
- Recepción de pedidos, etc.

2.2.2. Fundamentos del Marketing

Los clientes son diferentes entre sí, tienen necesidades diferentes y el valor de unos y otros es diferente, es por ello que se tiene la necesidad de segmentarlos de forma que podamos agrupar clientes con el mismo comportamiento, esta segmentación se basa en la existencia de base de datos de clientes, reales o potenciales, y el uso de técnicas de análisis estadístico de estos datos.

La segmentación de clientes es uno de los procesos estratégicos que se desarrollan en el marketing, que divide al mercado en grupos homogéneos con características similares, para aplicarle una estrategia diferenciada, satisfaciendo de forma más eficiente a cada grupo de clientes.

- **Se definen algunos términos básicos:**

  1. **Segmento:** Grupo homogéneo de consumidores en cuanto a deseos, preferencias de compra, uso de productos, estilos de vida similares del mismo segmento al cual pertenecen.

  2. **Mercado Meta:** Grupo de clientes (segmento seleccionado) que la empresa decide captar y satisfacer más eficientemente que la competencia, dirigiéndole su programa de marketing.

- **Tipos de Segmentación:**

  1. **Según el objetivo**
     - Segmentación estratégica.
     - Segmentación táctica.
  2. **Según la dimensión del cliente**
     - Dimensión de valor vs. Necesidad.
     - Customer lifetime value.
     - Dimensión geográfica.
     - Dimensión comportamental.
     - Dimensión relacional.
     - Dimensión Social.

2.2.2.1. Segmentación y dimensión del cliente

El tipo de información de cliente usado en la segmentación ha evolucionado en paralelo al desarrollo de los sistemas de información.

Desde las segmentaciones sociodemográficas generales, más propias de la segmentación de mercados, pasando por el análisis comportamental basado en el valor se llega a los modelos de valor – necesidad, dominantes en la actualidad.

Actitudes, prescripción, vinculación, y análisis comportamental online constituyen las nuevas dimensiones que deben enriquecer los modelos de valor – necesidad.

2.2.2.2. Objetivos de la Segmentación[1]

- Determinan conjuntos de clientes que poseen el mismo comportamiento, para hacer llegar ofertas especialmente diseñadas al perfil de dichos clientes.
- Aplica estrategias comerciales diferenciadas para cada segmento, consiguiendo una mayor rentabilidad de las acciones de marketing.
- Localiza nichos e identifica mercados nula o escasamente mercados atendidos.
- Facilita el análisis de la competencia.

2.2.2.3. Tipos de campañas de segmentación

- Identificación de clientes más rentables, estimación de la cuota de cliente, simulación de sendas de abandono y alertas ante eventos de riesgo de abandono – reclamaciones, incidencias no resueltas, periodos de inactividad.
- Son campañas altamente dependientes del motivo del abandono, a menudo requieren una investigación de estas motivaciones de los clientes perdidos. Es clave conocer el valor de vida o valor futuro previsto del cliente, para dimensionar la oferta de recuperación, y actuar inmediatamente tras la deserción. Obviamente, siempre es preferible trabajar en la retención de un cliente que tener que hacerlo en su recuperación.

[1] **Lévy – Varela (2008)**

- Es definido el análisis de potencial de demanda por división, en sectores de retail los análisis de asociación permiten generar cestas de la compra y patrones secuenciales de compra. Son campañas muy rentables en compañías o grupos empresariales altamente diversificados. Los motores de recomendación supone una variante de cross – selling donde la campaña se lanza online, durante el proceso de compra.
- De nuevo es clave estimar correctamente la demanda total de cliente en la categoría, buscando maximizar tu cuota de cliente. En distribución minorista, suelen dividirse en acciones de incremento de ticket medio y acciones de incremento de frecuencia. En ambos casos a menudo asociaciones a análisis RFM-Recencia, Frecuencia, valor monetario.
- El potencial de demanda se estima mediante la búsqueda de gemelos – clientes similares a los que son más rentables o modelización sociodemográfica modelos predictivos de demanda basados en características sociodemográficas, generalmente provenientes de fuentes públicas como censos, padrones, estudios sectoriales.

2.2.3. Data Mining (minería de datos)

La minería de datos es el conjunto de técnicas y tecnologías que permiten explotar grandes bases de datos, con el objetivo de encontrar patrones repetitivos, tendencias, reglas que expliquen el comportamiento de los datos en un determinado contexto para la toma de decisiones de manera inteligente pero automatizada, se centra en llenar la necesidad de descubrir el porqué, para luego predecir y pronosticar las posibles acciones con cierto factor de confianza para cada predicción.

La minería de datos puede ayudar a encontrar cual es el mejor paquete de productos que le puede ofrecer a sus clientes existentes para que su relación sea más rentable. Por ejemplo se pueden encontrar afinidades entre los productos que se consumen para realizar promociones y ofertas focalizadas (**Marketing**), detecta patrones de uso fraudulento de tarjetas de crédito (Banca), detección de fraude telefónico (telecomunicaciones), identificación de terapias medicas satisfactorias para diferentes enfermedades (medicina), etc.

- **Fases de un proyecto de minería de datos**

Los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica especificada de extracción de conocimiento usada.

El proceso de minería de datos pasa por las siguientes fases:

1. **Selección de datos:** En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar.
2. **Preprocesamiento:** Esta etapa consiste en la preparación y limpieza de datos extraídos desde las fuentes de datos de una forma manejable, necesaria para las fases posteriores en esta etapa se utilizan estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.
3. **Transformación:** Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.
4. **Data Mining:** Es la fase de modelamiento propiamente tal, en donde los métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u "ocultos" en los datos.
5. **Interpretación y Evaluación:** Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

2.2.3.1. Funciones de la Minería de datos

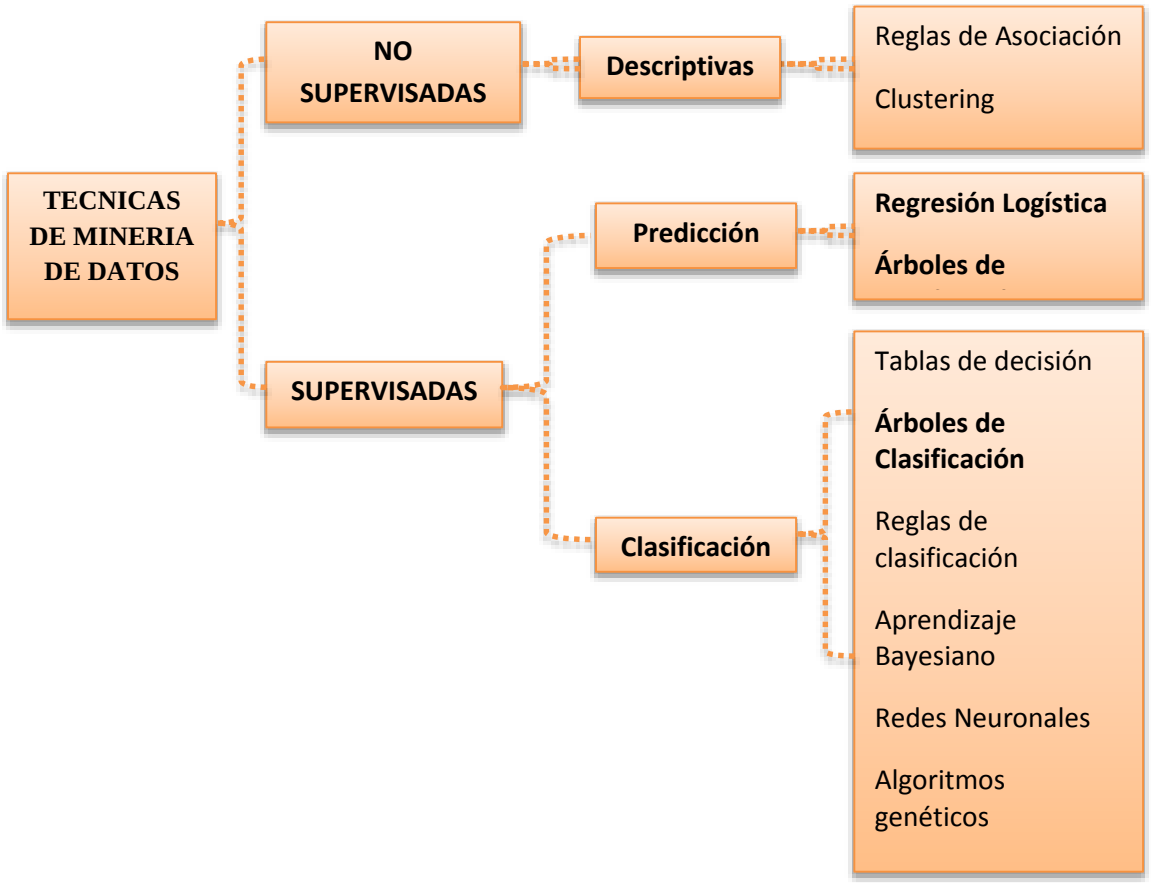
Data mining utiliza los datos existentes para:

- **Predecir:** Predice la pertenencia a una categoría.
- **Agrupar:** Descubre grupos de clientes homogéneos basados en sus características.
- **Identificar:** Identifica casos que no siguen un comportamiento esperado.
- **Asociar:** Encuentra eventos que ocurren simultáneamente o en una secuencia

Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ser automático o semiautomático (asistido) y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización.

Existen muchas formas diferentes de representar los modelos y cada una de ellas determina el tipo de técnica que puede usarse para inferirlos. En la práctica, los modelos pueden ser de dos tipos: predictivos y descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de interés, que denominan como variables objetivos o dependientes, usando otras variables o campos de la base de datos, como variables independientes o predictivas. Dentro de estos modelos se encuentran los **árboles de Clasificación.**

Figura N° 1: Clasificación de las técnicas de Minería de datos.



Fuente: Elaboración Propia.

2.2.4. Modelos Lineales Generalizados

Los modelos lineales (Regresión Simple, ANOVA, ANCOVA), se basan en los siguientes supuestos:

- Los errores se distribuyen normalmente.
- La varianza es constante.
- La variable dependiente se relaciona linealmente con las variables independientes.

En muchas ocasiones, sin embargo, nos encontramos con que uno o varios de estos supuestos no se cumplen por la naturaleza de la información. Al presentarse esta situación nos hallamos ante el uso de los modelos lineales generalizados.

Según Dobson (2002) se debe usar estos modelos cuando:

- La variable respuesta tienen una distribución no normal, pudiendo ser incluso categóricas.
- Las relaciones entre la variable respuesta y las variables explicativas no es lineal.

2.2.4.1. Características de los Modelos Lineales Generalizados

El Modelo Lineal Generalizado es una unificación de los modelos de regresión lineal y no lineal, que también permite incorporar distribuciones de respuesta no normales. En un modelo lineal generalizado la distribución de la variable respuesta solo necesita ser un miembro de la familia exponencial, que comprende las distribuciones normales, de Poisson, binomial, exponencial y gamma.

Además, el modelo lineal con error normal no es más que un caso especial del modelo lineal generalizado, por lo que en muchos aspectos se puede considerar que el modelo lineal generalizado es un método unificador de muchos aspectos del modelado y análisis empírico de datos.

Los Modelos Lineales Generalizados, extienden los modelos de regresión de mínimos cuadrados ordinarios a una condición de no normalidad en la distribución de la variable respuesta, modelando las funciones de la media.

Según Ato - López (1996), "los modelos estadísticos que forman los Modelos lineales Generalizados se caracterizan por distinguir entre elementos empíricos y teóricos, así como por poseer una estructura definida".

2.2.4.1.1. Elementos empíricos:

Dos son los elementos empíricos que intervienen en la formulación de un modelo estadístico:

1. La variable dependiente o variable respuesta, su naturaleza de medida es lo que va a distinguir entre los modelos para datos numéricos y los modelos para datos categóricos.
2. Las variables explicativas, también llamadas covariables en estos modelos según (Hosmer y Lemeshow, 2000), son las que se proponen como causas probables de la variable respuesta.

Cuadro N° 1: Principales métodos de análisis estadístico para la respuesta y las variables explicativas medidas en diferentes escalas.

Variable Respuesta	Variable Explicativas	Métodos
Continua	Binaria	t-test
	Nominal, >2 categorías	Análisis de varianza
	Ordinal	Análisis de varianza
	Continua	Regresión Múltiple
	Nominal y algunas continuas	Análisis de Covarianza
	Categóricas y continuas	Regresión Múltiple
Binaria	Categórica	Tablas de Contingencia
		Regresión Logística
	Continua	Regresión Logística
		Probit y Modelos de Respuesta de dosis
	Categóricas y continuas	Regresión Logística
	Nominal	Tablas de Contingencia

Nominal con más de 2 categorías	Categóricas y continuas	Regresión Logística Nominal
Ordinal	Categóricas y continuas	Regresión Logística Ordinal
Conteo	Categóricas	Modelo Log-Lineal
	Categóricas y continuas	Regresión Poisson
Tiempo de fallo	Categóricas y continuas	Análisis de Supervivencia (paramétrico)
Respuesta correlacionada	Categóricas y continuas	Ecuaciones de Estimación Generalizada
		Modelos Multinivel

Fuente: Dobson (2002)

2.2.4.1.2. Elementos teóricos:

Los elementos teóricos que intervienen en la formulación de un modelo estadístico son el vector de la respuesta media, que contiene los valores $E(Y_i) = \mu$ de la variable respuesta, y el vector del predictor lineal, que es una función lineal aditiva de k variables explicativas X_j y de sus respectivos parámetros β_j.

$$E(Y_i) = \sum_{j=0}^{k} \beta_j X_{ij}$$

2.2.4.1.3. Estructura del Modelo:

La mayoría de los métodos estadísticos fueron demostrados por Nelder and Wederburn (1972) usando la idea de modelos lineales generalizados. Este modelo es definido en términos de un conjunto de variables aleatorias independientes $Y_1,\ldots..Y_N$ cada una con una distribución que pertenece a una familia exponencial y tiene los siguientes componentes:

1. **Componente Aleatorio**: Identifica la variable respuesta Y, y su distribución de probabilidad. Consiste en las observaciones independientes ($Y_1,\ldots.Y_N$) de una distribución de la familia exponencial cuya función de probabilidad o densidad es:

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)exp[y_i Q(\theta_i)]$$

Varias distribuciones importantes son casos especiales, incluyendo la Poisson y Binomial. El valor del parámetro θ_i puede variar para *i=1,..., N*, dependiendo de los valores de las variables explicativas. El término $Q(\theta_i)$ es llamado el parámetro natural.

2. **Componente sistemático:** Especifica las variables explicativas usadas en un modelo en una función lineal del predictor. Relaciona un vector $(\eta_1, \dots \dots, \eta_N)$ con las variables explicativas a través de un modelo lineal. Si x_{ij} denota el valor del predictor *j (j=1,....,k)* para el sujeto *i*. Entonces

$$\eta_i = \sum_j \beta_j x_{ij} , \qquad i=1,..., N.$$

Esta combinación lineal de variables explicativas es llamada predictor lineal. Usualmente, una de las $x_{ij} = 1$ para todo *i*, para el coeficiente del intercepto, frecuentemente denotado por β_0 en el modelo.

3. **Función enlace,** especifica la función de *E(Y)* en la que el modelo iguala al componente sistemático. Enlaza el componente aleatorio y sistemático. Siendo $\mu_i = E(Y_i)$, *i=1,..., N*. El modelo enlaza μ_i con η_i mediante $\eta_i = g(\mu_i)$, donde la función enlace g es una función monótona derivable. Así, g enlaza $E(Y_i)$ a las variables explicativas a través de la fórmula:

$$g(\mu_i) = \sum_j \beta_j x_{ij} \qquad\qquad i=1,..., N.$$

La función $g(\mu) = \mu$ llamada enlace identidad, tiene $\eta_i = \mu_i$. Especifica un modelo lineal para la media misma. Es la función enlace para la regresión ordinaria donde *Y* tiene distribución normal. La función enlace que transforma la media en el parámetro natural es llamada enlace canónico. Esto es $g(\mu_i) = Q(\theta_i)$ y $Q(\theta_i) = \sum_j \beta_j x_{ij}$

Las tres principales funciones de enlace son:

1. **Función logit**: $\eta = log\left\{\frac{\mu}{1-\mu}\right\}$
2. **Función probit** : $\eta = \phi^{-1}(\mu)$
3. **Función log-log complementaria**: $\eta = \log\{-\log(1-\mu)\}$

Cuadro N° 2: Tipos de Modelos Lineales Generalizados para el Análisis Estadístico.

Componente Aleatorio	Enlace	Componente Sistemático	Modelo
Normal	Identidad	Continuo	Regresión
Normal	Identidad	Categórico	Análisis de Varianza
Normal	Identidad	Mixto	Análisis de Covarianza
Binomial	Logit	Mixto	Regresión Logística
Poisson	Log	Mixto	Loglineal
Multinomial	Logit Generalizado	Mixto	Respuesta Multinomial

Fuente: Dobson (2002)

2.2.5. Regresión Logística

Los métodos de regresión se han convertido en un componente integral de cualquier análisis de datos, que consiste en la descripción de la relación entre una variable respuesta y una o más variables explicativas. A menudo es el caso de que la variable respuesta es dicotómica teniendo dos valores posibles. En la última década el modelo de Regresión Logística se ha convertido, en muchos campos en el método estándar de análisis en esta situación.

Según Hosmer y Lemeshow (2000), "lo que distingue a un modelo de Regresión Logística a partir del modelo de regresión lineal es que la variable respuesta en la Regresión Logística es binaria o dicotómica".

En general la Regresión Logística es adecuada cuando la variable respuesta es politómica (admite varias categorías de respuesta, tales como mejora mucho, empeora, se mantiene, mejora); pero es especialmente útil en particular cuando solo hay dos posibles respuestas (dicotómica), que es el caso común. En esta situación el investigador está

interesado en la predicción y explicación de las relaciones que influyen en la categoría en que un objeto está situado, y es en estas condiciones donde el modelo de Regresión Logística permite obtener una combinación lineal que representa una única relación multivariante con coeficientes como los de la Regresión Múltiple que indican la influencia relativa de las variables predictoras.

2.2.5.1. Uso del modelo de Regresión Logística

Este modelo es una generalización del modelo de regresión lineal clásico para variables dependientes categóricas dicotómicas (Ato y García 1996). Tiene la ventaja de no requerir supuestos como el de normalidad y el de homocedasticidad (igualdad de varianzas), que son difíciles de verificar. Además, es más potente que el análisis discriminante cuando estos supuestos no se cumplen.

La regresión logística está limitada, a la predicción de tan solo la medida dependiente de dos grupos. Por tanto, en casos donde la medida dependiente está formada por dos o más grupos se adecua mejor al análisis discriminante.

Otra ventaja radica en su similitud con la Regresión Múltiple: permite el uso de variables independientes continuas y categóricas (estas últimas por medio de su codificación a variables ficticias), cuenta con contrastes estadísticos directos, tiene capacidad de incorporar efectos no lineales y es útil para realizar diagnósticos.

- **Supuestos de la Regresión Logística:**

  1. Las variables explicativas son incorrelacionadas.

  2. La distribución Binomial, describe la distribución de los errores.

  3. No linealidad de la variable respuesta.

2.2.5.2. Modelo de Regresión Logística

Sea *Y* una variable de respuesta dicotómica (o binaria), con dos resultados posibles, por ejemplo 1 = acierto y 0 = fracaso, el interés se centra en describir el efecto de una o más variables explicativas $X= (X_1, X_2,..., X_k)$ sobre la variable respuesta, la cual seguirá una distribución binomial o Bernoulli con probabilidades (Abraira y Pérez de Vargas, 1996):

$$p(Y = 1|X) = \pi(x); P(Y = 0|X) = 1 - \pi(x)$$

La media o valor esperado de la variable *Y* será:

$$E(Y) = \pi(x)$$

y su varianza:

$$V(Y) = \pi(x)[1 - \pi(x)]$$

Supóngase que $\pi(x)$depende de los valores que tome cada una de las variables independientes $X= (X_1, X_2,..., X_k)$, para reflejar esta dependencia, y simplificar la notación, se seguirá la explicación con una sola variable explicativa *X* y se utilizará la notación $\pi(x)$.

- **Modelo de probabilidad lineal**

Una posibilidad a la hora de estimar el efecto de la variable explicativa es la utilización de un modelo estándar de regresión lineal, según el cual el valor esperado de *Y* es una función lineal de *X*, según la expresión:

$$E(Y) = \pi(x) = \beta_0 + \beta_1 x$$

El modelo resultante se denomina modelo de probabilidad lineal, puesto que la probabilidad de acierto cambia linealmente en *x*.

Sin embargo, existen serios inconvenientes en el modelo que lo hacen inapropiado (Ato y López, 1996):

1. El modelo podría arrojar valores externos al rango [0,1]
2. La varianza de *Y* no es constante para todo el rango de valores de cada variable independiente.

3. El coeficiente de determinación tiene un valor limitado en estos modelos, ya que suele ser muy bajo y, en consecuencia, es recomendable evitarlo en los modelos que contengan variables dependientes cualitativas.

Debido a todo lo anteriormente expresado, cuando la variable es dicotómica es más apropiado utilizar un modelo probabilístico que:

1. Aumente *E(Y)* a medida que aumentan las variables explicativas X_i, pero sin situarse fuera del intervalo [0,1].
2. Exprese una relación curvilínea entre *E(Y)* y cada una de las variables explicativas $X_{i.}$

Los modelos cuyas funciones de distribución acumulativa se ajustan a estas características y que más se utilizan con variables dicotómicas son:

1. Las funciones de distribución acumulativa logística, que da lugar a los modelos logit (para el caso de datos agrupados) y modelos de regresión logística (para el caso de datos no agrupados). La importancia en la distinción entre ambos modelos viene determinada tanto por razones teóricas como analíticas. Para el caso agrupado, la distribución condicional de la variable respuesta sigue una ley binomial $B(\pi, N)$.
   Para el caso no agrupado, la distribución condicional de la variable respuesta sigue una ley binomial con N=1, es decir, $B(\pi, 1)$. (Ato y Lopéz, 1996).
2. Las funciones de distribución acumulativa normal que dan lugar a los modelos probit.

- **La función logística**

La función que mejor se ajusta a las características arriba mencionadas tiene la siguiente forma:

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

Siendo Z una variable o una combinación de variables explicativas. Para facilitar la comprensión de la exposición se iniciara esta con la utilización de una sola variable independiente, según lo cual $Z = \beta_0 + \beta_1 x$, que sustituido en $\pi(x) = \frac{1}{1+e^{-z}}$ da como resultado la expresión:

$$\pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

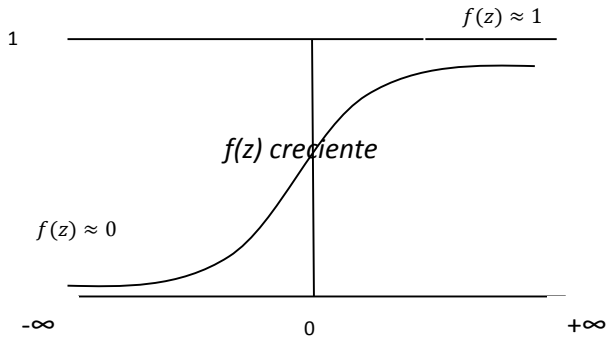
O de forma alternativa:

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

- **Características de la función logística**

La función logística tiene forma de "S", lo cual la hace idónea para la predicción de probabilidades, ya que sus valores siempre están comprendidos entre 0 y 1 (Kleinbaum, 1994).

Figura N° 2: Representación gráfica de la función Logística.



Fuente: Dobson (2002)

Donde:

Rango: $0 \leq f(z) \leq 1$

Si Z→+∞ entonces, $e^{-z} = \frac{1}{e^z} \to 0 \implies \frac{1}{1+e^{-z}} \longrightarrow 1$

Si Z→0 entonces, $e^{-z} = 1 \Rightarrow \frac{1}{1+e^{-z}} \longrightarrow 0.5$

Si Z→−∞ entonces, $e^{-z} = \infty \Rightarrow \frac{1}{1+e^{-z}} \longrightarrow 0$

Kleinbaum (1994) señala que la forma de la función logística es especialmente atractiva a los epidemiólogos, ya que si, por ejemplo, se trata de ver el efecto combinado de diferentes factores de riesgo Z sobre la probabilidad de padecer una enfermedad, esta es mínima para los valores bajos de Z hasta un determinado umbral, a partir del cual la probabilidad se eleva de forma rápida y, una vez Z ha crecido lo suficiente, la función permanece constante con valores altos, es decir próximos a 1.

2.2.5.2.1. Modelo Logístico Simple

Para una variable respuesta binaria *Y* y una variable explicativa *X*, el siguiente modelo de regresión permite predecir la respuesta binaria y_i de un sujeto *i* de la población a partir del valor de la variable predictora x_i. Los coeficientes β_0 y β_1 son los parámetros del modelo y ε_i es el término de error residual:

$$y_i = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} + \varepsilon_i$$

Teniendo en cuenta que la variable respuesta es binaria, la probabilidad condicional de éxito *P(Y=1|X)* sigue una función logística:

$$P(Y=1|X) = \pi(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Entonces:

$$\frac{\pi(x)}{1-\pi(x)} = \frac{\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}}{\frac{1}{1+e^{\beta_0+\beta_1 x_i}}} = e^{\beta_0+\beta_1 x_i}$$

Una transformación de $\pi(x)$ que va a resultar interesante, consiste en calcular el logaritmo de cada miembro de la igualdad, es decir, realizar una transformación logit, la cual:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

Como puede verse por la ecuación, el logit [*g(x)*] es el logaritmo neperiano de la razón de dos probabilidades.

La importancia de esta transformación viene dada porque el logit:

- No solamente es lineal en la variable *x* sino también en los parámetros.
- Aunque es lineal en la variable x, las probabilidades no lo son, en contraste con el modelo de probabilidad lineal ya visto, donde las probabilidades aumentan linealmente con *x.*
- Puede ser continuo y oscilar entre (-∞; +∞) dependiendo del valor de *x*.

En resumen, hemos visto que en un análisis de regresión cuando la variable de resultado es dicotómica (Hosmer y Lemeshow, 2000):

1. La media condicional de la ecuación de regresión debe ser formulada para ser delimitada entre cero y uno. Hemos afirmado que el modelo de regresión logística $\pi(x)$ satisface esta restricción.
2. La distribución binomial y no la normal describe la distribución de los errores y será la distribución estadística sobre el que se basa el análisis.
3. Los principios que guían el análisis mediante regresión lineal también nos guiará en la regresión logística.

- **Odds**

Se define como el cociente entre una probabilidad y la probabilidad complementaria. Indica cuantas veces más probable es que ocurra un evento respecto a que no ocurra. (Silva, 1995).Existe una relación simple entre probabilidades y odds. Si $\pi(x)$ es la probabilidad de un evento y O es la odds del evento, entonces:

$$O = \frac{\pi(x)}{1 - \pi(x)}$$

$$\pi(x) = \frac{O}{1 + O}$$

La utilización del odds se debe a que es más sensible para realizar comparaciones.

Por ejemplo, si una persona A tiene una probabilidad de 0.30 de donar, y una persona B tiene la probabilidad de 0.60, es razonable pensar que la probabilidad de la persona B es el doble que de la persona A. Pero si la probabilidad de al persona A es 0.60, es imposible que la probabilidad de la persona B sea el doble, es decir 1.20. En este sentido no hay ningún

problema si se trabaja con la escala del odds, ya que a una probabilidad de 0.6 le corresponde un Odds de $\frac{0.6}{0.4} = 0.5$, cuyo doble es 3. Volviendo este valor a probabilidad nos da un resultado de 0.75. (Ver anexo N° 1).

En el caso de la variable independiente dicotómica donde Y=1 y Y=0 las odds correspondientes son las siguientes:

Cuando *x=1* es $\frac{\pi(1)}{1-\pi(1)} = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}}{\frac{1}{1+e^{\beta_0+\beta_1}}} = e^{\beta_0+\beta_1}$

Cuando *x=0* es $\frac{\pi(0)}{1-\pi(0)} = \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = e^{\beta_0}$

Un inconveniente de la odds es que su rango oscila entre [0;+∞].

Si la odds>1, el evento es más probable que el no evento.

Si la odds<1, el no evento es más probable que el evento.

El valor de e^{β_1} es el factor por el que la odds cambia cuando la variable independiente se incrementa una unidad. Si β_1 es positivo, el factor será mayor que 1, lo que significa que la odds aumenta; si β_1 es negativo, el factor será menor que 1, en consecuencia la odds decrece. Cuando β_1 es 0 el factor es igual a 1, lo que hace que la odds permanezca constante.

Tal y como se ha visto el logaritmo de la odds, llamado logit será igual a:

$$g(1) = \ln\left(\frac{\pi(1)}{1-\pi(1)}\right) = \beta_0 + \beta_1$$

$$g(0) = \ln\left(\frac{\pi(0)}{1-\pi(0)}\right) = \beta_0$$

- **Estimaciones en el modelo de Regresión Logística Simple**

En un modelo de regresión logística simple, es decir, con una única variable explicativa, los dos parámetros desconocidos β_0 y β_1 son estimados usando el método de máxima

verosimilitud, el cual consiste en proporcionar la estimación que otorgue máxima probabilidad o verosimilitud a los datos observados.

En primer lugar se ha de construir la función de verosimilitud del modelo estimado, *l(B)*, la cual representa la probabilidad de reproducir los datos de la muestra a partir de dicho modelo, y donde B= (β_0 , β_1) es el vector de parámetros.

Supóngase un modelo de regresión logística simple y una muestra aleatoria formada por *n* observaciones de la variable binomial *Y* y de la variable independiente *X*, con probabilidades:

$$p(Y = 1|x) = \pi(x_i)$$

$$p(Y = 0|x) = 1 - \pi(x_i)$$

La probabilidad de la observación (x_i, y_i) correspondiente al sujeto i vendrá dada por la ley de Bernoulli:

$$P(Y = y_i, X = x_i) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

En el caso de una muestra con *n* observaciones independientes, la función de verosimilitud vendrá dada por el producto de las probabilidades según la expresión siguiente:

$$l(B) = \prod_{i=1}^{n} \pi(x_i)^{yi} \, [1 - \pi(x_i)]^{1-y_i}$$

La estimación de los dos coeficientes requiere maximizar la función de verosimilitud, o equivalentemente, maximizar su logaritmo:

$$\ln[l(B)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}$$

Derivando respecto a cada uno de los (β_0 , β_1) e igualando a cero obtenemos las ecuaciones de verosimilitud:

$$\frac{\partial \ln[l(B)]}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0$$

$$\frac{\partial \ln[l(B)]}{\partial \beta_1} = \sum_{i=1}^{n}[y_i - \pi(x_i)]x_i = 0$$

La solución de este sistema de ecuaciones no lineales en los parámetros β_0 y β_1 se obtendrá mediante un proceso iterativo basado en el teorema de Taylor, denominado método de Newton-Raphson (McCullagh y Nelder, 1989).

Las estimaciones máximo-verosímiles de los parámetros derivadas de las ecuaciones

$$\frac{\partial \ln[l(B)]}{\partial \beta_0} = \sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \;\; ; \; \frac{\partial \ln[l(B)]}{\partial \beta_1} = \sum_{i=1}^{n}[y_i - \pi(x_i)]x_i = 0,$$

Se suelen indicar como $\hat{B}$. Para la obtención del error estándar de las estimaciones, se calcula la inversa de la matriz del negativo de las segundas derivadas evaluadas en las estimaciones máximo-verosímiles, obteniéndose la matriz de varianzas-covarianzas de las estimaciones que permite calcular los intervalos de confianza.

Una consecuencia de la ecuación $\frac{\partial \ln[l(B)]}{\partial \beta_0} = \sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$ es la siguiente:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i)$$

Es decir, la suma de los valores observados de *y* es igual a la suma de los valores predichos o esperados.

El valor de verosimilitud *l* oscila entre 0 y 1, por lo tanto, el logaritmo neperiano de la verosimilitud,$\ln[l(B)]$, será un numero negativo que alcanzara el valor 0 en un hipotético modelo en el que reprodujeran los datos de forma exacta.

2.2.5.2.2. Modelo Logístico Múltiple

Cuando se dispone de una variable respuesta dicotómica y un conjunto de variables independientes, categóricas y/o cuantitativas, el modelo resultante es una regresión logística múltiple.

Sea una serie de *p* variables independientes definidas por el vector $x= (x_1, x_2, \ldots, x_p)$, medidas en una escala de intervalo y sea la probabilidad condicionada de que la variable respuesta *Y* tome el valor 1 igual a:

$$P(y = 1|x) = \pi(x)$$

Por lo que el logit del modelo de regresión múltiple vendrá definido por la ecuación:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Y en consecuencia:

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

- **Estimaciones en el modelo de Regresión Logística Múltiple**

Supóngase una muestra de n observaciones independientes del par ($\mathbf{X}_i$, y_i), *i=1, 2,…, n.* Al igual que el caso del modelo de regresión simple, el método de estimación para los parámetros del modelo será el de máxima verosimilitud, con el cual se obtendrán las estimaciones del vector $\hat{B}' = (\beta_0, \beta_1, \dots, \beta_p)$.

El número de ecuaciones de verosimilitud será igual a p+1, las soluciones se obtendrán calculando las derivadas parciales respecto a los p+1 coeficientes del logaritmo de la función de verosimilitud mediante el algoritmo de Newton-Raphson. Las ecuaciones de verosimilitud vendrán expresadas como:

$$\sum_{i=1}^{n} [y_i - \pi(\boldsymbol{X}_i)] = 0$$

.

.

.

$$\sum_{i=1}^{n} [y_i - \pi(\boldsymbol{X}_i)] x_{ij} = 0$$

para *j= 1, 2,…, p.*

2.2.5.3. Pruebas de Bondad de ajuste:

Una vez construido el modelo de regresión logística, tiene sentido comprobar que tan bueno es el ajuste de los valores predichos por el modelo a los valores observados. Existen diversas formas de medir la bondad de ajuste de un modelo de regresión logística. De forma global esta puede ser evaluada a través de medidas tipo R^2, de la tasa de clasificaciones correctas o a través de una serie de test estadísticos.

1. Ji-cuadrado de Pearson

Se trata de un estadístico que compara los valores observados *y* con los predichos $\hat{y}$ por el modelo, según la expresión:

$$\chi^2 = \sum_{j=1}^{J} \frac{\left(y_j - n_j\hat{\pi}_j\right)^2}{n_j\hat{\pi}_j(1-\hat{\pi}_j)} = \sum_{j=1}^{J} \frac{n_j\left(y_j - \hat{y}_j\right)^2}{\hat{y}_j\left(n_j - \hat{y}_j\right)}$$

Tiene la misma distribución asintótica que la devianza, es decir, una chi cuadrado con los mismos grados de libertad. Con lo cual, la Ho será rechazada para el nivel de significación **α** cuando $\chi^2 \geq \chi^2_{J-(R+1);\alpha}$ (para el modelo múltiple con R covariables), que es equivalente a que el p-valor del contraste sea menor que el nivel **α** fijado.

Este estadístico anterior puede calcularse como la suma de los cuadrados que fueron denominados por Hosmer como residuos de Pearson.

$$\chi^2 = \sum_{j=1}^{J} r_{j.}^2$$

$$r_j = \frac{y_j - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

Para poder aplicar este estadístico χ^2 tiene que verificarse que el número de observaciones para cada combinación de las variables explicativas sea grande, es por ello, por lo que estos métodos no se aplican en el caso de covariables continuas o modelos no agrupados de Bernoulli.

Existen dos índices que representan la proporción de incertidumbre de los datos que es explicada por el modelo ajustado y que son análogos al coeficiente de determinación en la regresión lineal, los cuales se destacan a continuación.

a) El índice de Cox y Snell:

Compara la verosimilitud del modelo con solo la constante $l(\beta_0)$ y la verosimilitud del modelo considerado $l(\beta)$:

$$R^2 = 1 - \left[\frac{l(\beta_0)}{l(\beta)}\right]^{2/n}$$

El índice de Cox y Snell tiene el inconveniente de no alcanzar el valor de 1 (100%) cuando el modelo reproduce exactamente los datos tal y como se puede comprobar a continuación:

$$\text{Si } l(\beta) = 1 \longrightarrow R^2_{máx} = 1 - [l(\beta_0)]^{2/n}$$

Por esta razón Nagelkerke (1991, citado en Domenéch, 1999) ha propuesto el índice corregido que vale 1 en caso de que el modelo explique el 100% de la incertidumbre de los datos, y que viene definido por la expresión:

$$R^2_c = \frac{R^2}{R^2_{máx}}$$

b) Prueba de bondad de ajuste de Hosmer y Lemeshow

Hosmer y Lemeshow (2000), debido a que en la mayor parte de los estudios el número *J* de combinaciones posibles es grande, proponen ordenar los *n* sujetos según las predicciones $\hat{\pi}_j$ y dividirlos en g=10 grupos de aproximadamente el mismo tamaño, a estos grupos los denomina deciles de riesgo.

Esta división en grupos de igual tamaño es especialmente adecuada cuando muchas de las probabilidades estimadas son pequeñas. Hosmer y Lemeshow (2000) demuestran que si comparan las probabilidades observadas y predichas de los 10 grupos con el estadístico ji-cuadrado de bondad de ajuste, este sigue una distribución con g-2=8 grados de libertad, en el caso de que la mayor parte de las frecuencias esperadas sean superiores a 5 y ninguna inferior a 1.

2. Test basados en probabilidades estimadas

Una vez estimados los coeficientes hay que proceder a ver cuáles de ellos son significativamente diferentes de 0. Para analizar si cada uno de ellos es cero se utiliza el test de Wald. Si el objetivo es comprobar que el conjunto de covariables elegidas explica el fenómeno a estudio, es decir, contrastar si todos los coeficientes son iguales a 0 o hay alguno distinto, se puede utilizar la razón de verosimilitud o el test de puntaje (score-test).

a) Prueba de razón de verosimilitud

Consiste en construir un test basado en la verosimilitud, para lo cual se calcula cuál es la lejanía o discrepancia que hay entre el modelo ajustado con la (s) variable (s) predictoras incluidas en el modelo $l(\beta)$ y el modelo que solo contiene la constante $l(\beta_0)$ según la expresión siguiente:

$$D = -2\ln\left[\frac{l(\beta_0)}{l(\beta)}\right] = -2\ln l(\beta_0) - [-2\ln l(\beta)]$$

La devianza (D) sigue una distribución ji-cuadrado con tantos grados de libertad como la diferencia en grados de libertad entre ambos. Como medida de la discrepancia o bondad de ajuste existente entre los valores empíricos y

ajustados, juega un papel similar en las estimaciones por máxima verosimilitud al de la suma de cuadrados de los residuos en las estimaciones por mínimos cuadrados (MacCullagh y Nelder, 1989).

Al resultado del cociente entre corchetes, se le denomina razón de verosimilitud**,** hay una razón matemática en el cálculo del negativo de dos veces el logaritmo, y se debe a la necesidad de garantizar la distribución ji-cuadrado del test de contraste D, al test resultante se le denomina prueba de la razón de verosimilitud.

En la valoración de la significación de una variable independiente se compara el valor de D con y sin la variable independiente en el modelo, el cambio en D se debe a la inclusión de la variable independiente en el modelo y se obtiene con la siguiente fórmula:

$$G = -2\ln\left[\frac{l(modelo\ \text{sin}\ la\ variable)}{l(modelo\ con\ la\ variable)}\right]$$

Bajo la hipótesis nula de que $\boldsymbol{\beta}$ es igual a cero, el estadístico G sigue una distribución ji-cuadrado con un grado de libertad. Si el test es significativo indicará que la variable independiente añade información al modelo.

b) El test de Wald

Wald, demostró que las distribuciones muestrales de las estimaciones máximo verosímiles de los parámetros, en el caso de muestras grandes, se distribuyen según la curva normal. Por lo tanto, la significación de los parámetros puede estudiarse mediante el estadístico $\boldsymbol{z = \hat{\beta}/SE(\hat{\beta})}$**,** el cual sigue una ley normal estandarizada, o mediante el estadístico de Wald, es decir el cuadrado de ese cociente, el cual sigue una ley de ji-cuadrado, con 1 grado de liberta, según fórmula:

$$Wald = \left[\frac{\hat{\beta}}{SE(\hat{\beta})}\right]^2$$

Este estadístico tiene un gran problema respecto a la falta de potencia de esta prueba cuando el valor del parámetro $\hat{\beta}$ se aleja de cero, es decir, al ser un valor cuadrático puede producir valores muy pequeños en presencia de coeficientes de regresión muy altos, por lo que diversos autores, como por ejemplo Ato y López (1996) recomiendan la utilización de la prueba de la razón de verosimilitud.

c) El test de puntaje o score-test

Este test no necesita cálculos iterativos, puesto que lo que hace es valorar el cociente entre la primera y segunda derivada del logaritmo de la verosimilitud en $\beta_1 = 0$ y cuyo resultado se compara con una ji-cuadrado con un grado de libertad.

2.2.5.4. Capacidad Predictiva del Modelo

La capacidad predictiva del modelo puede evaluarse a través de las tablas de clasificación, o lo que es lo mismo, a través de la tabulación cruzada entre los casos observados y los casos pronosticados por el modelo. En este tipo de tablas se observan dos índices, la sensibilidad y la especificidad. En resumen, para valorar el poder de clasificación del modelo primero se comprobara que la especificidad y sensibilidad tienen niveles aceptables y solo en el caso afirmativo se considerara el porcentaje total de clasificaciones correctas como un índice resumen de su poder de clasificación.

2.2.6. Árboles de Clasificación

Los árboles de clasificación son una clase de técnicas de minería de datos que tienen raíces en las tradicionales disciplinas estadísticas como la regresión lineal. Los árboles de clasificación también comparten raíces en el mismo campo de la ciencia cognitiva que produce las redes neuronales. Los árboles de clasificación más tempranos eran de detección de patrones y la formación de conceptos (Hunt, Marin, y Stone 1966).

Según Breiman (1984), es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir del conjunto de prototipos. Esta participación recursiva se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto y cada nodo hoja se refiere a una decisión (clasificación).

Según J.R.Quinlan (1986), es un conjunto de decisiones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta sus hojas.

Según Kumar et al. (2007), **se** define como una estructura en forma de árbol en la que las ramas representan conjuntos de decisiones. Estas decisiones generan sucesivas reglas para la clasificación de un conjunto de datos en subgrupos disjuntos y exhaustivos.

Según Zhao - Bhowmick (2006), se define como un modelo de clasificación que sirve para estructurar una serie de condiciones que llevarán a la decisión de asignar a un objeto una clase determinada.

2.2.6.1. Utilidad de los árboles de Clasificación

Los árboles de Clasificación son una forma de análisis de variables múltiples, nos permiten predecir, explicar, describir y clasificar. Un ejemplo de un análisis de variables múltiples es una probabilidad de venta o la probabilidad de responder a una campaña de marketing como resultado de los efectos combinados de variables múltiples de entrada, los factores o dimensiones.

El atractivo de los árboles de Clasificación se encuentra en su potencia relativa, facilidad de uso, robustez con una variedad de datos y niveles de medición, y la facilidad de interpretabilidad. Los árboles de Clasificación convierten los datos en bruto en un mayor

conocimiento y conciencia de los negocios, la ingeniería, y las cuestiones científicas, y que le permiten desplegar ese conocimiento de una forma sencilla.

2.2.6.2. Principales algoritmos de árboles de Clasificación

En la literatura han aparecido numerosos algoritmos de aprendizaje de árboles de Clasificación, entre los más populares se encuentran:

1. **ID3 (Iterative Dichotomizer 3):**
   Es un algoritmo que prefiere árboles sencillos frente a árboles más complejos ya que, en principio, aquéllos que tienen sus caminos más cortos hasta las hojas son más útiles a la hora de clasificar. En cada momento se ramifica por el atributo de menor entropía y el proceso se repite recursivamente sobre los subconjuntos de casos de entrenamiento correspondientes a cada valor del atributo por el que se ha ramificado.
2. **C5 (Commercial versión 5):**
   Es la última versión mejorada del algoritmo ID3 (hay una versión intermedia, la C4.5). Similar a CART, primero crea un árbol sobreajustado y luego lo poda para crear un modelo más estable. El objetivo del algoritmo es minimizar la tasa de error total de los nodos, asumiendo que el número de registros clasificados incorrectamente en un nodo dividido por el número de registros total del mismo es la mejor estimación de la tasa de error.
3. **Quest:**
   Es un método de cálculo rápido que evita errores de otros métodos, favoreciendo así predictores con varias categorías, trata de solucionar algunos problemas tradicionalmente asociados con CART mediante la utilización de métodos estadísticos que eviten el sesgo de CART a la hora de seleccionar atributos para ramificar el árbol de decisión.
4. **Chaid (Chi – cuadrado de detección Automática de Interacción):**
   Este método utiliza estadísticos de Chi – Cuadrado para identificar divisiones óptimas. La variable de destino puede ser nominal, ordinal, o continúa. No es un algoritmo binario, considera los datos perdidos como una categoría individual.
5. **Chaid Exhaustivo:**
   Este método es una modificación de CHAID que realiza un análisis más detallado de examinar todas las divisiones posibles para cada predictor, pero tarda más en calcular. La variable de destino puede ser nominal, ordinal o continua.

Este trabajo se centra en la metodología CART la cual se usa para la construcción de árboles de regresión y clasificación, y utiliza un algoritmo recursivo de partición binaria en cada nodo.

2.2.7. Árbol de clasificación con el algoritmo CART

Según Acuña (1999), CART es una metodología diseñada por Breiman, Friedman, Losen y Stone en 1984, como un algoritmo para construcción de árboles quienes los aplicaron a problemas de regresión y clasificación.

Es una técnica de extracción de datos y muestra el método como particiones secuenciales del conjunto de datos para maximizar las diferencias de la variable dependiente, ofrece una forma concisa de desarrollar grupos que son consistentes en sus atributos pero que varían en términos de la variable dependiente.

La idea básica es dividir los datos en dos subconjuntos, de modo que los individuos comprendidos dentro de cada uno de los subconjuntos sean más homogéneos que en el subconjunto anterior. Se trata de un proceso recursivo, que se repite hasta alcanzar el criterio de homogeneidad o hasta llegar a otro criterio de detención, pudiendo utilizar varias veces la misma variable predictora en distintos niveles del árbol.

2.2.7.1. Usos generales del Análisis basado en un árbol de Clasificación con el algoritmo CART

- **Segmentación:** Permite identificar individuos que puedan ser miembros de una clase particular.
- **Estratificación:** Permite asignar individuos u objetos a una categoría entre varias, tales como alto riesgo, medio riesgo, y grupos de bajo riesgo.
- **Predicción:** Permite crear reglas y utilizarlas para predecir eventos futuros. Predicción también puede significar, intentos de relacionar atributos de predicción para valores de una variable continua.
- **Reducción de datos y filtrado de variables:** Permite seleccionar un subconjunto útil de predictores de un gran conjunto de variables para su uso en la construcción de un modelo paramétrico formal.

- **Identificación de Interacción:** Permite identificar las relaciones que corresponden únicamente a subgrupos específicos y especificar éstos en un modelo paramétrico formal.
- **Fusión de categorías y discretización de variables continuas:** Permite recodificar categorías predictoras y las variables continúas con una mínima pérdida de información.

2.2.7.2. Ventajas y Desventajas del algoritmo CART

- **Ventajas:**

1. La metodología CART tiene como ventaja primordial su sencillez debido a que es un método de aprendizaje supervisado inductivo.
2. Este método no se enfrenta a los supuestos estrictos de normalidad multivariante y la igualdad de matrices de varianzas covarianzas, para la validación del árbol se recurre a la tabla de porcentajes de clasificación correcta (validación cruzada).
3. La metodología no se ve afectada por la presencia de valores extremos ni por datos perdidos.
4. Realiza selección de variables en forma automática, brindando una medida de importancia en forma natural.
5. Puede ser aplicado a cualquier tipo de variables predictoras: continuas y categóricas
6. Los resultados son fáciles de entender e interpretar.
7. Es invariante a transformaciones de las variables predictoras.

- **Desventajas:**

1. El proceso de selección de variables es sesgado hacia las variables con más valores diferentes.
2. Dificultad para elegir el árbol óptimo
3. La superficie de predicción no es muy suave, ya que son conjuntos de planos.
4. Requiere un gran número de datos para asegurarse que la cantidad de observaciones en los nodos terminales es significativa.

5. Ausencia de una función global de las variables y como consecuencia pérdida de la representación geométrica.
6. No toma en cuenta las interacciones que puede existir entre las variables predictoras.

2.2.7.3. Metodología CART

CART es una técnica exploratoria de datos que tiene como objetivo fundamental encontrar reglas de clasificación y predicción. Dado un conjunto de datos $D = (X, Y)$, donde Y es la variable a explicar y $X = (X_1, \ldots, X_p)$ es un vector de p variables que describe a los individuos, el objetivo de CART es predecir los valores de Y a partir de los valores observados de las variables X_i, $i = 1, \ldots, p$. Tanto la variable dependiente Y, como cada una de las variables explicativas X_i puede ser cuantitativa o cualitativa, esto dota a CART de una gran flexibilidad pues se puede aplicar en muchos contextos distintos.

En el caso en que la variable dependiente Y sea cualitativa, se dice que CART es un árbol de clasificación, y el objetivo es predecir la clasificación que le correspondería a un individuo con cierto perfil de valores en las variables explicativas. Por otra parte, si Y es cuantitativa, CART es llamado árbol de regresión y el objetivo es idéntico al de un modelo lineal, obtener una estimación del valor de Y asociado a cada nicho o perfil de predictores.

Además, esta técnica es utilizada para la selección de variables en el sentido que permite determinar cuál característica o conjunto de características es la que mejor define o discrimina a los grupos predeterminados. Los árboles de Clasificación de tipo CART, pueden verse como la estructura resultante de la partición recursiva del espacio de las variables explicativas (espacio de representación) a partir de un conjunto de reglas de decisión.

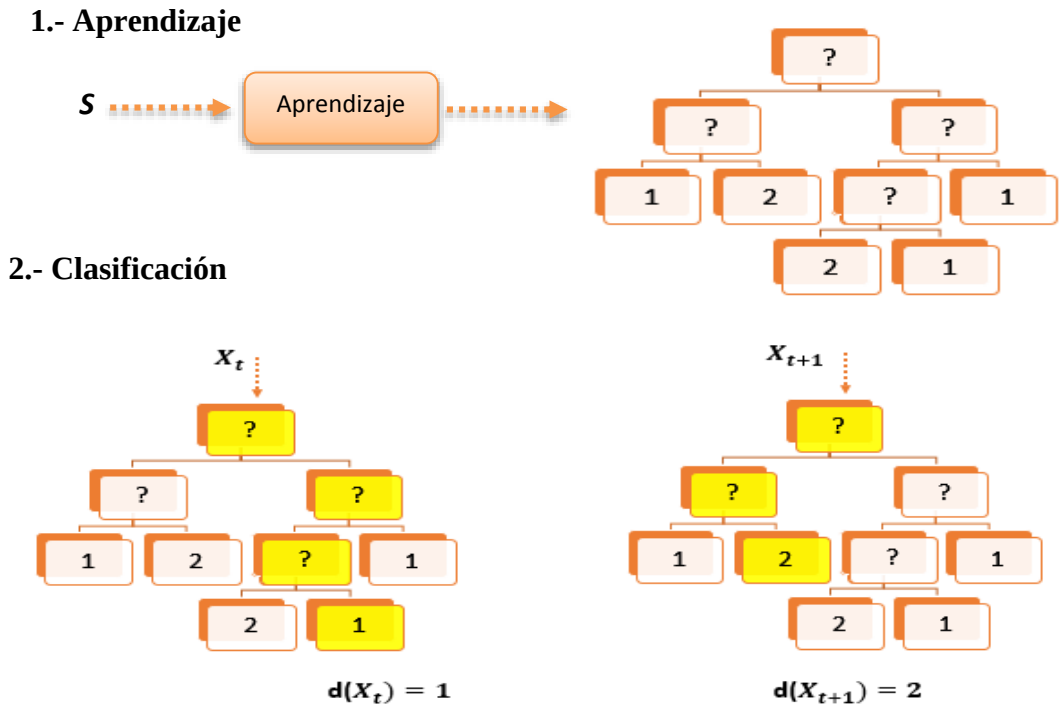
La manera en que se construye cada partición es lo que distingue a los distintos tipos de árboles, éstas son determinadas por un conjunto de decisiones sobre las variables explicativas. En CART las reglas de decisión son desplegadas en forma de árbol binario. Determinan en cada momento dos alternativas posibles, las mismas se suceden hasta que el árbol llega a su construcción final. El procedimiento es recursivo y se traduce en una organización jerárquica del espacio de representación.

2.2.7.3.1. Metodología a seguir

1. Aprendizaje: Consiste en la construcción del árbol a partir de un conjunto de prototipos, S. Constituye la fase más compleja y la que determina el resultado final. A esta fase dedicamos la mayor parte de nuestra atención.

2. Clasificación: Consiste en el etiquetado de un patrón, X, independiente del conjunto de aprendizaje. Se trata de responder a las preguntas asociadas a los nodos interiores utilizando los valores de los atributos del patrón X. Este proceso se repite desde el nodo raíz hasta alcanzar una hoja, siguiendo el camino impuesto por el resultado de cada evaluación.

Figura N°3: Aprendizaje y clasificación del algoritmo CART.



Fuente: Cortijo Bon, José , (Técnicas Supervisadas II,Aproximación No parametrica)

Un árbol de clasificación *T* representa una partición recursiva del espacio de representación, *P*, realizada en base a un conjunto de prototipos, *S*.

Además, existe una íntima relación entre **nodos** de *T*, **regiones** en *P* y **conjuntos** en *S*:

1. Cada nodo de *T* tiene asociado un subconjunto de prototipos de *S*.

2. El nodo raíz tiene asignado el conjunto completo.

3. Cada hoja, *t*, tiene asociada una región, R_t, en *P*. Así, si $\tilde{T}$ es el conjunto de nodos hoja del árbol *T*:

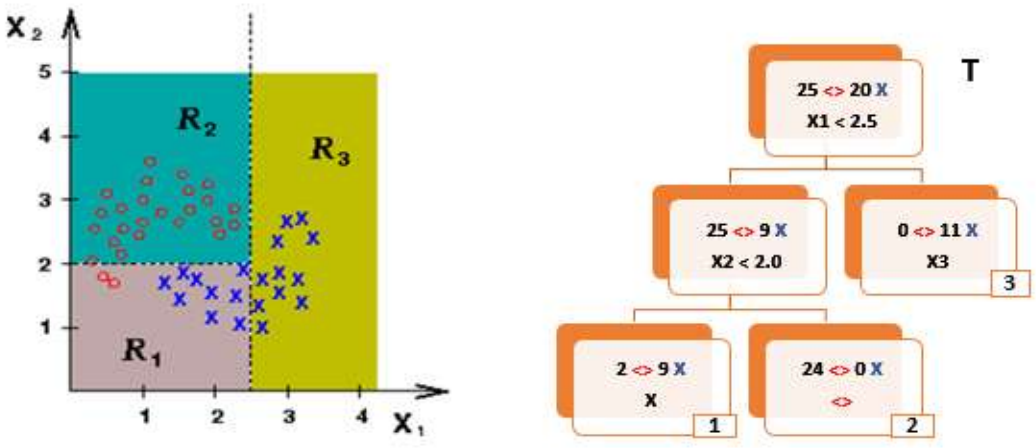
$$\bigcup_{t \in \tilde{T}} R_t = P$$

que se interpreta como que los conjuntos de prototipos asignados a los nodos hoja constituyen una partición de *P*.

4. Cada nodo interior tiene asociada una región en *P*, que es la unión de las regiones asociadas a los nodos hoja del subárbol cuya raíz es él.

5. La unión de los conjuntos de prototipos asignados a los nodos de un mismo nivel da como resultado el conjunto inicial.

Figura N° 4: Muestra gráficamente el significado de las propiedades del árbol.



Fuente: Cortijo Bon, José , (Técnicas Supervisadas II,Aproximación No parametrica)

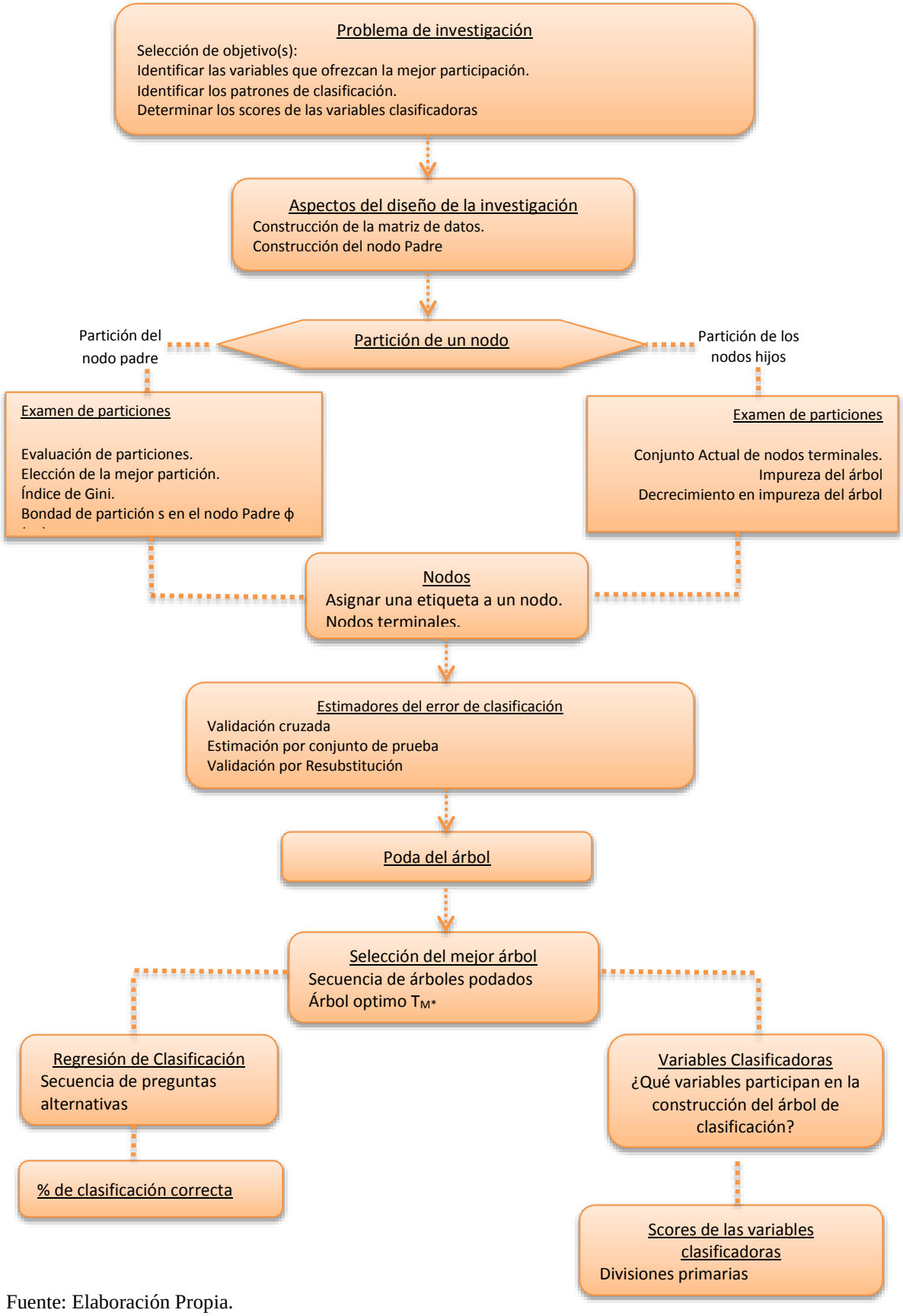
2.2.7.3.2. Construcción del árbol de Clasificación con el algoritmo CART

El árbol máximo es construido utilizando un procedimiento de partición binario, comenzando en la raíz del árbol, este árbol es un modelo que describe el conjunto de entrenamiento (grupo de datos original) y generalmente es sobreajustado, es decir, contiene gran cantidad de niveles y nodos que no producen una mejor satisfacción y puede ser demasiado complejo.

La construcción del árbol de decisión constituye la fase de aprendizaje. Entre los mecanismos de aprendizaje constituye el más complejo de los estudiados hasta ahora. Las líneas generales para su construcción pueden resumirse en los siguientes puntos, de acuerdo a un esquema recursivo:

1. El avance está basado en la partición de un nodo de acuerdo a alguna regla, normalmente evaluando una condición sobre el valor de alguna variable. Los prototipos que verifican la condición se asignan a uno de los dos nodos hijo (normalmente el izquierdo) y los restantes, al otro. Cuando un nodo se particiona, pasa a ser un nodo intermedio.
2. El caso base o condición de parada tiene como objetivo detener el proceso de partición de nodos. Cuando se verifica la condición de parada en un nodo, éste es un nodo hoja. Los prototipos asociados a un nodo hoja constituyen un agrupamiento homogéneo, por lo que al nodo se le asigna una etiqueta.

Figura N° 5: Diagrama para la construcción del árbol de Clasificación.



Fuente: Elaboración Propia.

2.2.7.3.3. Selección de las particiones

Una partición divide a un conjunto de prototipos en conjuntos disjuntos. En CART las particiones son binarias, resultado de evaluar una condición que tiene dos únicas respuestas: sí o no.

El objetivo de una partición es incrementar la homogeneidad (en términos de clase) de los subconjuntos resultantes, o lo que es lo mismo, que éstos sean más puros que el conjunto originario. Cada partición tiene asociada una medida de pureza, que se utiliza para:

1. Para la selección de la mejor partición.
2. Como criterio de parada, aunque no resulta muy recomendable.

2.2.7.3.4. Formulación de la regla de partición

En esta sección estudio cómo se formulan las preguntas (en CART) que dan lugar a las particiones. En primer lugar fijaremos el marco teórico de trabajo. Sea Q el conjunto de preguntas binarias de la forma:

$$\{¿\, X \in A?\}, A \subset P$$

El conjunto Q genera un conjunto de particiones, s, en cada nodo, t y cada nodo t se particiona en t_L (izquierdo) y t_R (derecho) de manera que:

- Los casos de t que verifican la condición $¿X \in A?$ se asignan a t_L,

$$t_L = t \cap A$$

- Los casos de t que no la verifican se asignan a t_R.

$$t_R = t \cap \overline{A}$$

2.2.7.3.5. Criterios de partición

Cada partición tiene asociada una medida de pureza y se tratará de incrementar la homogeneidad de los subconjuntos resultantes de la partición, esto es, que sean más puros que el conjunto originario.

- **Función de impureza y medida de impureza**

En primer lugar definiremos lo que se entiende por función de impureza. Una función de impureza es una función ϕ definida sobre *J*-uplas de la forma $(c_1, c_2,..., c_J)$ tales que:

1. $c_j \geq 0$ para $j = 1, 2,..., J$

2. $\Sigma_j c_j = 1$, con las siguientes propiedades:

a) ϕ tiene un único máximo en (1/J, 1/J, ..., 1/J).

b) ϕ alcanza su mínimo en (1,0,0,...,0), (0,1,0,...,0), ..., (0,0,0,...,1) y el valor mínimo es 0.

c) ϕ *es* una función simétrica de $c_1, c_2, ..., c_j$.

- **Medida de impureza de un nodo, i(t)**

La función de impureza es una medida que determina la calidad de un nodo, esta será denotada por *i(t).* Existen varias medidas de impureza (criterios de particionamiento) que nos permiten analizar varios tipos de respuesta.

Dada una función de impureza ϕ, definamos la medida de impureza de cualquier nodo *t*, $i(t)$,como:

$$i(t) = \phi (p(1| t), p(2| t),..., p(J| t))$$

Donde, $p(j| t)$ es la probabilidad de que un caso del nodo *t* (un prototipo asociado al nodo *t*) sea de clase *j*. Estas probabilidades pueden calcularse empíricamente como la proporción de casos de clase *j* en el nodo *t*:

$$p(j| t)=\frac{Nj(t)}{N(t)}$$

Dicho de otra forma, la medida de impureza de un nodo es el resultado de evaluar la función de impureza sobre ese nodo tomando las proporciones relativas de cada clase como los c_j. Observar que, por un lado,

1. $p(j|\, t) \geq 0$
2. $\sum_j p(j|\, t) = \sum_j \frac{N_j(t)}{N(t)} = \frac{1}{N(t)} \sum_j N_j(t) = 1$

a) La máxima impureza (respuesta mínima pureza) se obtiene cuando todas las clases están igualmente representadas en *t*.

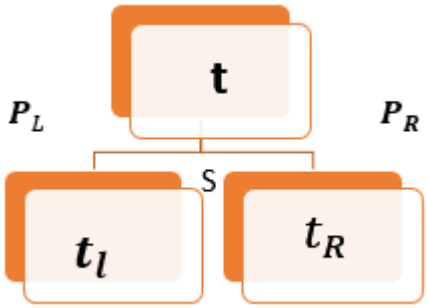
b) La mínima impureza (respuesta máxima pureza) se obtiene cuando en *t* sólo hay casos de una sola clase (máxima homogeneidad).

c) Cualquier permutación de los c_j produce el mismo resultado en el valor de impureza.

- **Bondad de una partición**

La bondad de una partición *s* en un nodo *t* debe estar relacionada con la impureza del nodo sobre el que se realiza la partición, *t*, y con la impureza de los nodos resultantes de la partición, t_L y t_R.

Figura N° 6: División de la partición.



Fuente: Cortijo Bon, José , (Técnicas Supervisadas II,Aproximación No parametrica)

La bondad de la partición *s* en un nodo *t*, ϕ *(s, t)*, se define como el decrecimiento en impureza conseguido con ella:

$$\phi\,(s,\,t)\;=\;\Delta\, i(s,\,t)\;=\;i(t)\,-\,p_{\mathrm{L}}\;i(t_{\mathrm{L}})\,-\,p_{\mathrm{R}}\;i(t_{\mathrm{R}})$$

Así, como conocemos cómo calcular i(t), podemos calcular ϕ (s, t) para cada partición s y seleccionar la mejor partición como la que proporciona la mayor bondad ϕ (s, t).

- **Criterios de medida de impureza**

Las tres funciones más comunes para la medida de la impureza de un nodo presentadas por Breiman, para los árboles de clasificación son el índice de información o entropía, el índice de Gini y el índice "Towing".

- **Índice de información o entropía**.

  Mide la entropía en un nodo *t* de acuerdo a la formulación clásica de la entropía, el cual se define como:

$$i(t) = -\sum_{j=1}^{J} p(j|\,t)\,\log p(j|\,t)$$

  Se asume que 0 log 0 = 0

  El objetivo es encontrar la partición que maximice Δ***i(t)*** en la ecuación

$$\Delta i(t) = -\sum_{j=1}^{k} p\big(j\big|\,t\big)\log p\big(j\big|\,t\big),$$

  Donde *j=1,....k* es el número de clase de la variable respuesta categórica y $p\big(j\big|\,t\big)$ la probabilidad de clasificación correcta para la clase j en el nodo t.

- **Índice de Gini**.

  Mide la diversidad de clases en un nodo.

$$i(t) = \sum_{\substack{i,j=1 \\ i \neq j}}^{J} p(i|\, t)\, p(j|\, t) = 1 - \sum_{j=1}^{j} p(j|\, t)2$$

  Este índice es el más utilizado. En cada división el índice Gini tiende a separar la categoría más grande en un grupo aparte, mientras que el índice de información tiende a formar grupos con más de una categoría en las primeras decisiones y por último,

- **Índice de "Towing".**

  A diferencia del índice de Gini, Towing busca las dos clases que juntas forman más del 50% de los datos, esto define dos "super categorías" en cada división para las cuales la impureza es definida por el índice de Gini. Aunque el índice towing produce árboles más balanceados, este algoritmo trabaja más lento que la regla de Gini. Para usar el índice de towing seleccione la partición *s*, que maximice

$$\frac{PLPR}{4} = \left[\sum_{j} |p(j|t_L) - p(j|t_R)|\right]^2$$

  Donde $\boldsymbol{t_L}$ y $\boldsymbol{t_R}$ representan los nodos hijos izquierdo y derecho respectivamente, $\boldsymbol{p_L}$ y $\boldsymbol{p_R}$ representan la proporción de observaciones en *t* que pasaron a $\boldsymbol{t_L}$ y a $\boldsymbol{t_R}$ en cada caso.

- **Criterio de parada**

La medida de pureza puede utilizarse para establecer un criterio de parada, el criterio está basado en el decrecimiento en impureza conseguido con una partición s:

Se trata de fijar un valor $\beta > 0$. Un nodo t será terminal si:

$$\boldsymbol{max_s}\{\Delta \boldsymbol{i}(\boldsymbol{s},\boldsymbol{t})\} < \boldsymbol{\beta}$$

En definitiva, se trata de detener el proceso de división cuando la mejor partición posible del nodo t (la que proporciona el máximo decrecimiento en impureza) produce un decrecimiento en impureza inferior al umbral.

Este procedimiento, aunque simple, no resulta totalmente satisfactorio en la práctica:

1. Si β es bajo, resulta muy "exigente"' para detener el crecimiento porque el máximo decrecimiento en impureza debe ser muy bajo. El resultado es que da lugar a árboles muy grandes ya que se realizan muchas particiones.
2. Si β es alto, resulta muy "permisivo"' para detener el crecimiento, dando lugar a árboles de menos altura. No obstante, aunque en un momento dado pueden encontrase nodos en los que $\boldsymbol{max_s}\{\Delta \boldsymbol{i}(\boldsymbol{s},\boldsymbol{t})\} < \boldsymbol{\beta}$ es pequeño, es posible que una posterior partición de sus descendientes proporcione mayores decrecimientos de impureza.

2.2.7.3.6. Estimadores de error

a) **$R^{cv}(T)$, Estimador por validación cruzada***:* Si el conjunto de entrenamiento tiene pocos prototipos no es aconsejable el uso del estimador por conjunto de prueba ya que reduce aún más el tamaño efectivo del conjunto de aprendizaje. Suele utilizarse validación cruzada con $V = 10$ conjuntos.

$$SE(R^{ts}(T)) = \sqrt{R^{ts}(T)\frac{1 - R^{ts}(T)}{|S^t|}}$$

b) $R^{ts}(T)$, Estimador por conjunto de prueba*:* Se calcula partiendo el conjunto de aprendizaje, *S*, en dos conjuntos disjuntos, uno destinado al aprendizaje, S^l, y el otro a prueba, S^t. Este estimador es aplicable si el conjunto de entrenamiento es lo suficientemente grande para ser particionado.

$$R^{ts}(t) = \frac{1}{|N_2|} \sum_{(p_{r,A_j}) \epsilon N_2} \Delta(P_r, A_j)$$

c) $R^s(T)$, Estimador por resusbstitución*:* Este estimador es el más sencillo de todos. A partir de todo el conjunto original de individuos N, compuesto por $|\boldsymbol{N}| = \boldsymbol{n}$ individuos de j clase, se construye el clasificador t. una vez construido el árbol, T, haciendo uso de la función indicadora se obtiene el número de individuos directamente clasificados, este valor dividido con el total de individuos proporciona el estimador de error por resubstitución.

$$R^s(t) = \frac{1}{n} \sum_{(p_{r,A_j}) \epsilon N_2} \Delta(P_r, A_j)$$

$$R^s(t)_{relativo} = \frac{R(t)}{p(t)_{nodo\_padre}}$$

Donde: $\boldsymbol{p(t)_{nodo\_padre}}$ es la proporción de individuos mal clasificados en el nodo padre.

2.2.7.3.7. Validación del Modelo y Cuantificación de la Bondad de Predicción

Una vez elegida y realizada la predicción con más de una técnica clasificatoria, es común que se proceda a evaluar y comparar sus resultados. Para estos efectos se suelen utilizar dos herramientas de comparación: las tablas de contingencia junto a la medición del grado de éxito logrado y las curvas Receive Operating Characteristic (ROC).

1. Tablas de contingencia y medición de aciertos

La manera más usual para medir la eficiencia en la clasificación es a través del porcentaje de acierto global de las predicciones. Existen formas de evaluar el nivel de acierto (Liu, Frazier y Kumar, 2007), como el acierto total dividido por el número total de casos.

Para su obtención es necesario conocer la tabla o matriz de contingencia o de confusión. Dicha tabla incluye tanto los aciertos como los errores para cada una de las clasificaciones, es decir, muestra los casos de clases predichas comparadas con los valores reales. Esta tabla es la principal y central fuente de información para evaluar la bondad de la predicción (Foody, 2002; Liu, Fraxier y Kumar, 2007).

Cuadro N° 3. Matriz de Confusión genérica

	Tipo A	Tipo B
Tipo A	Verdadero Positivo (VP)	Falso Negativo (FN)
Tipo B	Falso Positivo (FP)	Verdadero Negativo (VN)

Fuente: Elaboración propia

El verdadero negativo (VN) y el verdadero positivo (VP) son los aciertos en las predicciones y el falso positivo (FP) y el falso negativo (FN) son los errores. FP es predecir que ocurre A cuando en realidad ocurre B y FN es el error inverso. Todos éstos se pueden presentar en términos absolutos o como porcentaje, tanto del total de datos como del subgrupo al que pertenece la clasificación real. En la medida que VP y VN sean mayores, mejor es el desempeño de la técnica clasificatoria.

Esta tabla puede ser extensible a k clases. En este último caso la práctica habitual es identificar y separar los errores de clasificación de acuerdo a cuan distantes se encuentran de la diagonal (Koh, 1992; Bessis, 2002).

De la tabla de la matriz de confusión se extraen las siguientes medidas:

a) Tasa de Error:

$$\text{Error= \# errores / \# ejemplo= } \chi = \frac{|FN|+|FP|}{\text{N}}$$

Donde, $N = |FN| + |FP| + |VN| + |VP|$ es el total de ejemplos del conjunto de validación.

b) Exactitud (Accuracy): Es la proporción del número total de predicciones que son correctas.

$$\text{Accuracy= } \frac{|VN|+|TV|}{|FN|+|FP|+|VN|+|VP|}$$

c) Recall: Es la proporción de casos positivos que fueron clasificados correctamente

$$\text{Recall= } \frac{|VP|}{|VP|+|FN|}$$

d) Precisión: Es la predicción de casos positivos que fueron clasificados correctamente.

$$\text{Recall= } \frac{|VP|}{|VP|+|FP|}$$

e) Tasa de Falsos Negativos (FN Rate): Es la proporción de casos positivos que son incorrectamente clasificados como negativos.

$$FN= \frac{|FN|}{|VP|+|FN|}$$

f) Tasa de Falsos positivos (FP Rate): Es la proporción de casos negativos que son incorrectos clasificados como positivos.

$$FP = \frac{|FP|}{|VP| + |FP|}$$

2. Curvas ROC

Es una representación gráfica de la tasa de éxito (probabilidad de clasificar correctamente un individuo cuando dicho individuo está efectivamente presente) frente a la tasa de falsa alarma (probabilidad de detectar un individuo cuando efectivamente no está presente) para tareas de detección, con sólo dos resultados posibles (Si/ No, presente/ausente), según se varía el umbral o criterio para detectar a un individuo a lo largo de la escala de valores a partir de los cuales se hace la detección.

La curva ROC se basa en los conceptos de Sensibilidad y Especificidad. La Sensibilidad, es la probabilidad de clasificar correctamente a un individuo con el valor de interés objetivo del estudio y la especificidad, es la probabilidad de clasificar correctamente a un individuo sin el valor de interés objetivo del estudio. Una forma de ver estos resultados es por medio de una matriz cuadrada de confusión, la cual permite observar las clasificaciones observadas y las predichas por la metodología de clasificación.

Cuadro N° 4: Matriz de Clasificación

	Clasificador	
	+	**-**
+	**(VP)**	**(FP)**
-	**(FN)**	**(VN)**

Fuente: Kumar (2007)

Donde:

Casos mal clasificados

Casos bien clasificados (precisión del método de clasificación)

FP, error tipo I (α) muy costoso

FN, error tipo II (β) <5%, 20%>

El correcto desempeño de un algoritmo de clasificación de mide en base al área bajo la curva, la cual está formada por la sensibilidad y especificidad, esta área está comprendida entre 50% y 100%, donde el 100% representa un diagnostico perfecto y 50% una prueba sin capacidad de discriminación entre las categorías de una clase, por ende se busca mayor área bajo la curva.

2.2.7.3.8. Criterios de Comparación de Métodos

En este trabajo se desarrolló la comparación entre el Modelo de Regresión Logística y el algoritmo de Árbol de clasificación CART mediante los porcentajes de clasificación correctos con la tasa de clasificación que proporcionen cada uno de los métodos siendo el mejor modelo aquel que proporcione un porcentaje de clasificación correcto más alto (efectividad óptima), además la importancia de las variables predictoras serán calificadas con los scores respectivos de cada método.

III. MATERIALES Y MÉTODOS

Como se mencionó en la Introducción, el modelo que se desarrolló en este trabajo fueron Regresión Logística y Árboles de Clasificación usando el algoritmo CART como método de clasificación y segmentación aplicado en un caso particular de migraciones de clientes de un plan prepago a postpago.

3.1. MATERIALES

Libros de Investigación de Modelos de Regresión Logística, Técnicas Multivariadas, Minería de datos (Data mining), información proporcionada por los jefes de las oficinas del departamento de Estadística e Informática.

Equipos:

- Materiales de escritorio
- Computadora e impresora
- Software: Window 8, Ms Office Word 2012, Ms Office Excel 2012, Internet Explorer, R 2.14.0 , Rapid miner, SPSS 19 y Sql Server 2012.

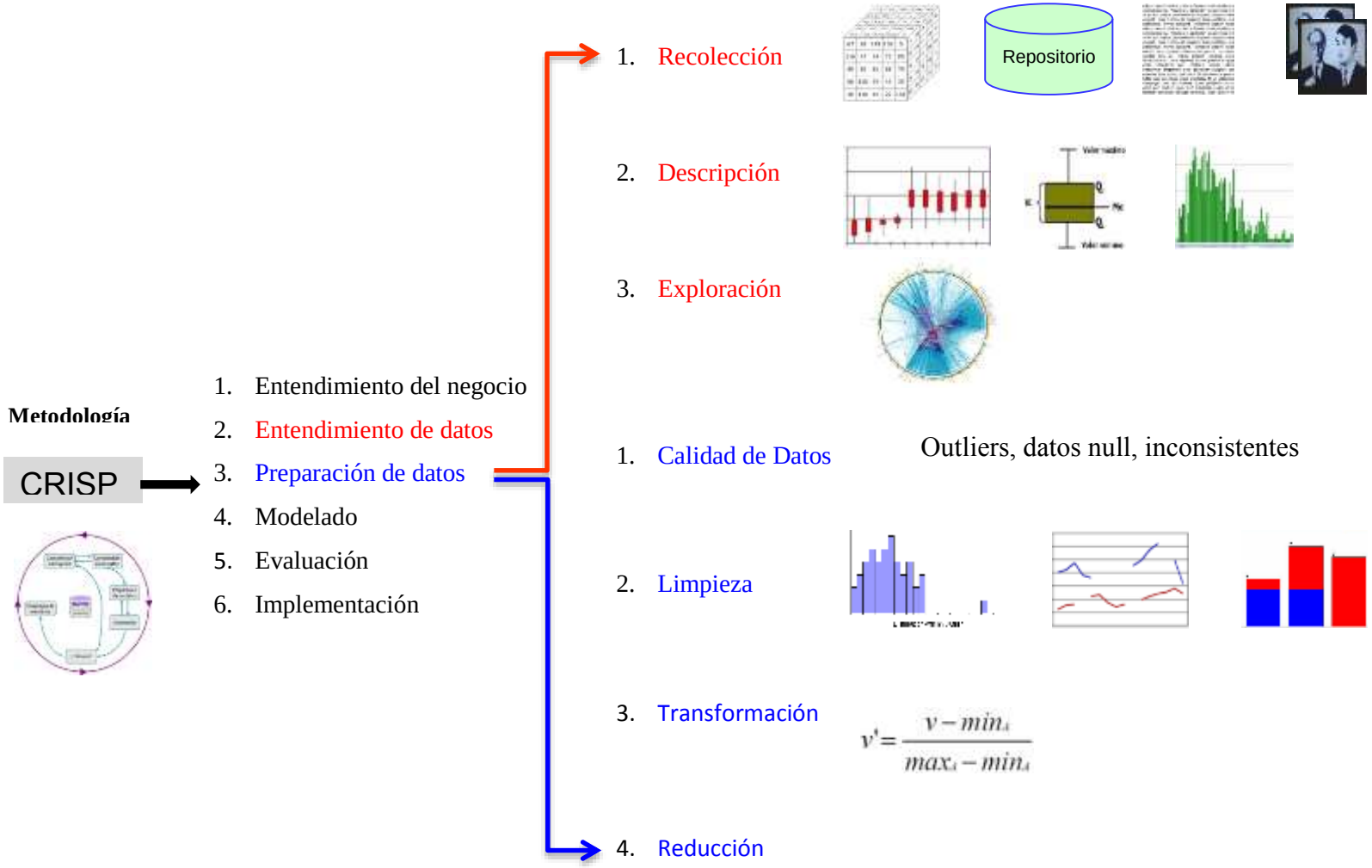
3.2. METODOLOGÍA

La metodología que se usó en el presente trabajo es denominada comúnmente como CRISP –DM muy conocida por sus aplicaciones en la minería de datos, ya que está descrita en términos de un modelo de proceso jerárquico. Como **metodología,** incluye descripciones de las fases normales de un proyecto, las áreas necesarias y una explicación entre las tareas. Como **modelo de proceso** ofrece un resumen del ciclo de vital de minería de datos.

La metodología CRISP-DM consta de 6 fases:

1. Entendimiento del Negocio.
2. Entendimiento de los datos.
3. Preparación de los datos.
4. Modelado.
5. Evaluación.
6. Implementación

Figura N° 7: Diagrama de la Metodología Crisp



3.2.1. Definición del Modelo a emplear

Para el análisis de clasificación y segmentación de la base de datos del Call Center, el modelo que se empleó fue el de Regresión Logística de respuesta binaria (éxito, fracaso) y Árboles de Clasificación usando el algoritmo CART, estos modelos verificaron las siguientes variables en estudio:

A continuación se detallara el método usado a profundidad.

3.2.1.1. **Entendimiento del Negocio:** (Objetivo comerciales: principales y específicos)

Objetivos:

- **Gestión operativa:**
  - Optimización de la gestión de llamadas.
  - Gestión de registros que generen llamadas efectivas.
  - Segmentación adecuada de la BD a gestionar.
- **Gestión de Ventas:**
  - Incrementar las ventas.
  - Aumentar la efectividad de la BD tramitada.
  - Incrementar la contactabilidad.
- **Gestión de Negocio:**
  - Gestión eficiente de los recursos.
  - Incrementar el margen de ganancias.

Objetivos de la Minería de Datos:

- Determinar patrones que permitan incrementar la tasa de contactabilidad y efectividad.
- Segmentar la base de cartera para conocer los perfiles del cliente.
- Determinar patrones de comportamiento en la base de gestión. “Best Time to Call”.

Criterios de éxito de Minería de datos:

- El % de recupero (dinero) a nivel total y por segmento: Cliente Nuevo y Antiguo.
- Incremento de los niveles de contactabilidad a nivel total y por cartera luego de la minería.

3.2.1.2. Entendimiento de Datos:

La base de información en estudio de clientes con líneas prepago a nivel nacional Lima, Callao y provincias de Telefónica del Perú, del mes de Agosto a Noviembre 2014.

Cuadro N° 5. Base de registros de Clientes

Base de Cliente	Base de Clientes	Base Gestión
CAR_201408 (Agosto)	223451	43401
CAR_201409 (Septiembre)	237873	51853
CAR_201410 (Octubre)	305755	50015
CAR_201411 (Noviembre)	266879	55427
TOTAL	**810507**	**200696**

Fuente: Elaboración Propia

- **Base de Clientes**:

  Es la base enviada por el cliente Movistar, contiene el uno a uno de registros que se van a gestionar en el mes respectivo.

- **Base Gestión:** Es la base que contiene los resultados de todas las llamadas realizadas en los respectivos meses. Se utilizó esta base para la Construcción de la Regresión Logística y el algoritmo CART.

- **Diccionario de Datos:**

Las variables que se escogieron de la base gestión fueron las siguientes:

Cuadro N° 6. Tabla de Variables

Variables	Descripción	Tipo de Dato
BAS_NRO_TELEF	Teléfono del cliente	polynominal
TIPO_RESULTADO	Resultado de la gestión final realizada al cliente. (Venta y No venta).	binominal
BAS_LOCALIDAD	Localidad del Cliente.	polynominal
BAS_SUBLOCAL	Localidad del Cliente.	polynominal
BAS_CODPLAN_A	Código del plan Prepago del Cliente.	integer
BAS_DESPLAN_A	Plan Prepago del Cliente.	polynominal
BAS_CODEQUI_A	Código del Equipo del Cliente.	polynominal
BAS_DESEQUI_A	Marca del equipo del Cliente.	polynominal
BAS_CODPLAN_OF1	Código del plan Post pago que se le ofrece al Cliente.	integer
BAS_DESPLAN_OF1	Plan Post pago que se le ofrece al Cliente.	polynominal
BAS_CF_PLAN_OF1	Monto del Plan que se ofrece al Cliente.	real
BAS_GAP_01	Salto entre el monto promedio de activación y monto del plan a ofrecer al cliente.	real
BAS_OTROS_02	Minutos de RPM del plan a ofrecer al Cliente.	integer
BAS_OTROS_03	Minutos para realizar llamadas a un Teléfono movistar.	integer
BAS_OTROS_04	Minutos para realizar llamadas a cualquier otro operador de telefonía.	integer
BAS_OTROS_05	Número de mensajes del plan a ofrecer al Cliente.	integer
BAS_OTROS_10	Monto promedio de activación que realiza un cliente en los últimos tres meses.	real
BAS_OTROS_21	Total de llamadas entrantes y salientes.	real
BAS_OTROS_24	Promedio de Tráfico de llamadas salientes a RPM.	real
BAS_OTROS_40	Último mes en que fue gestionado el Cliente.	integer

ANTIGUEDAD	Tiempo en años en que se gestionó un cliente por última vez.	integer
CLASE_PLAN	Clusterización del Plan de origen.	polynominal
NIVEL_5	Descripción del resultado del cliente gestionado.	polynominal
FECHA_HORA_CONTACTO	Fecha y Hora en el que se ha contactado al cliente.	nominal
FECHA_INICIO_CONTACTO	Fecha en la que inició la llamada al cliente.	nominal
FECHA_FIN_CONTACTO	Fecha en la que finalizó la llamada al cliente.	nominal
DURACION	Tiempo en segundos que se realizó la llamada.	integer
COD_GESTION	Mes actual de gestión.	integer
RESULTADO	Descripción del resultado que se obtuvo al realizar la llamada.	polynominal
HORA	Hora en la que se realizó la llamada.	integer

Fuente: Elaboración Propia.

3.2.1.3. Preparación de los datos

En esta fase se determinó, seleccionó y se extrajeron los datos de la fuente de información para el proceso de extracción de conocimiento. Se filtraron los datos de forma que se eliminan los valores incorrectos, no válidos, desconocidos según las necesidades y el algoritmo a usar. Esta fase fue soportada por herramientas estadísticas, herramientas de limpieza y otras.

Pasos:

1. Eliminación de registros inconsistentes.
2. Eliminación de atributos que no aportan información, redundancia de información y valores perdidos.
3. Discretización de atributos.

- **Estadísticos descriptivos:**

Se presentan los principales estadísticos descriptivos de las variables utilizadas en el trabajo de investigación.

Figura N° 8: Tabla de Frecuencia de Variables Categóricas.

Muestra el total de datos utilizados por cada variable en estudio.

Statistics

		TIPO_RESULTADO	LOG_GAP_01	CLASE_PLAN	ANTIGUEDAD	CLUSTER-opcion1
N	Valid	200696	200696	200696	200696	200696
	Missing	0	0	0	0	0

Fuente: Elaboración propia.

Variables Categóricas:

Figura N° 9: Tabla de Frecuencia de la Variable TIPO_RESULTADO.

TIPO_RESULTADO

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	CEF	14119	7,0	7,0	7,0
	CNE	186577	93,0	93,0	100,0
	Total	200696	100,0	100,0	

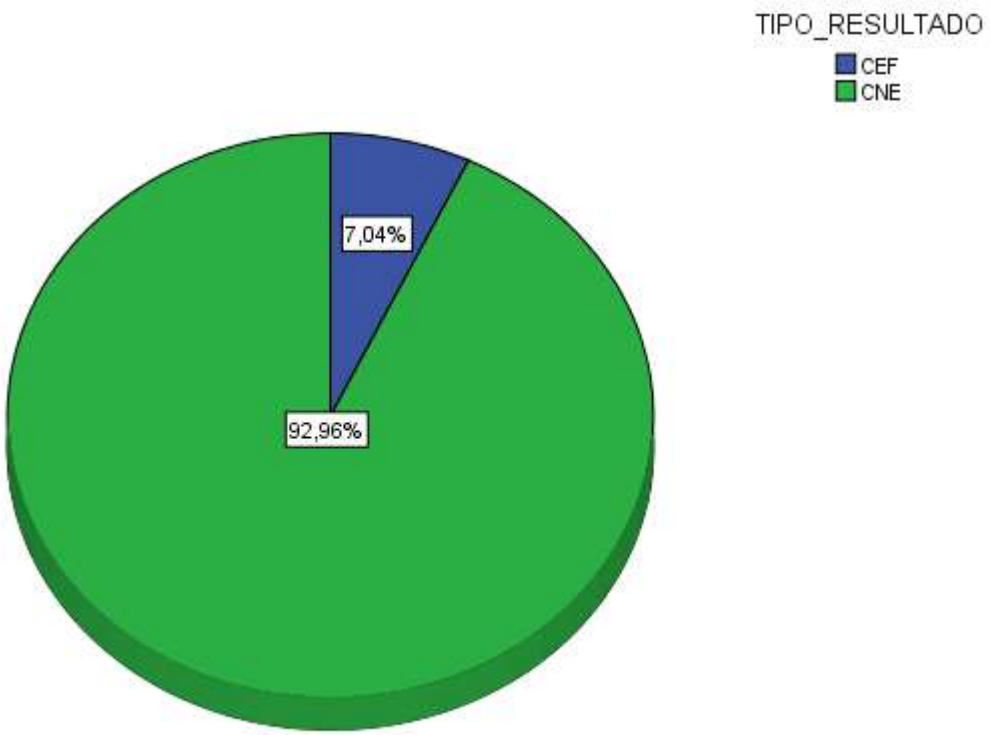
Fuente: Elaboración propia.

Donde:

CEF= Contacto Efectivo (venta)

CNE= Contacto No Efectivo (no venta)

Figura N° 10: Gráfico de pie de la Variable TIPO_RESULTADO.



Fuente: Elaboración propia.

De la Figura N° 10, el 7.04% de clientes es venta, mientras que el otro 92.96% de clientes es no venta.

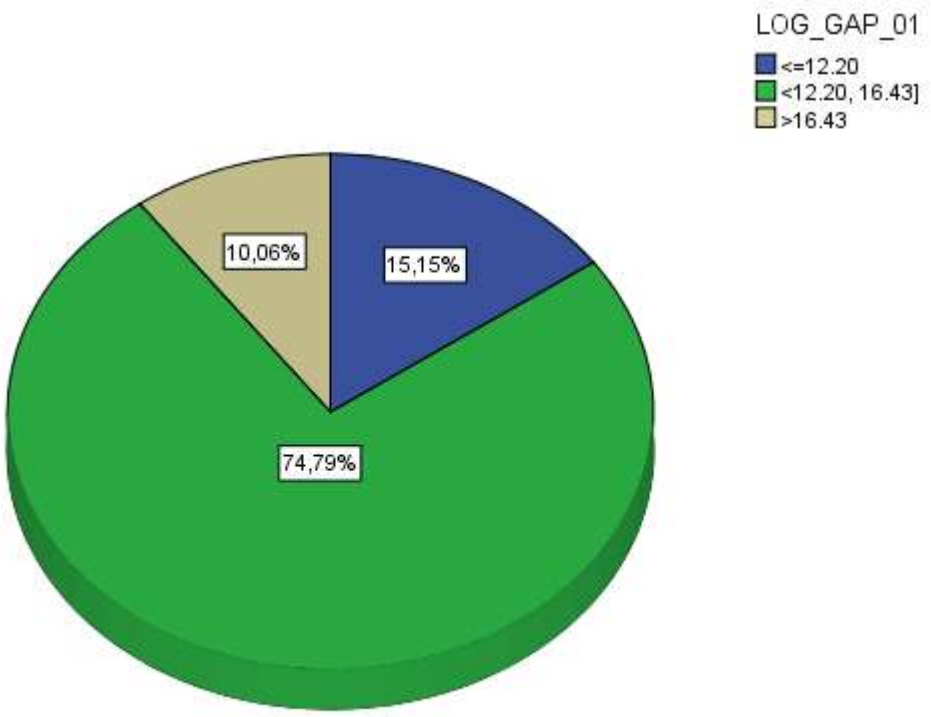
Figura N° 11: Tabla de Frecuencia de la Variable LOG_GAP_01.

LOG_GAP_01

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<=12.20	30410	15,2	15,2	15,2
	<12.20, 16.43]	150102	74,8	74,8	89,9
	>16.43	20184	10,1	10,1	100,0
	Total	200696	100,0	100,0	

Fuente: Elaboración propia.

Figura N° 12: Gráfico de pie de la Variable LOG_GAP_01.



Fuente: Elaboración propia.

De la Figura N° 12, el salto entre el promedio de activación y cargo fijo del plan a ofrecer mayor a S/. 16.43 soles representa el 10.06% del total de clientes, el salto entre el promedio de activación y cargo fijo del plan a ofrecer menor igual a S/. 12.20 soles representa el 15.15% del total de clientes y el salto entre el promedio de activación y cargo fijo del plan a ofrecer mayor a S/. 12.20 soles y menor a S/.16.43 soles representa el 74.79% del total de clientes.

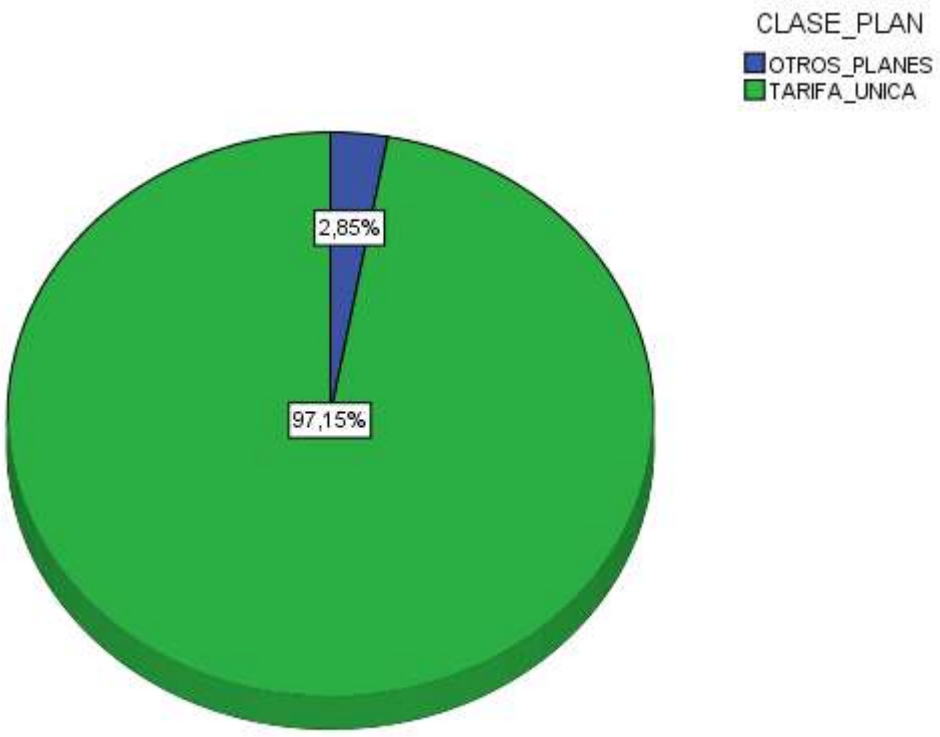
Figura N° 13: Tabla de Frecuencia de la Variable CLASE_PLAN.

CLASE_PLAN

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	OTROS_PLANES	5715	2,8	2,8	2,8
	TARIFA_UNICA	194981	97,2	97,2	100,0
	Total	200696	100,0	100,0	

Fuente: Elaboración propia.

Figura N° 14: Gráfico de pie de la Variable CLASE_PLAN.



Fuente: Elaboración propia.

De la Figura N° 14, el 97.15% del total de clientes tiene un plan de tarifa única, mientras el otro 97.15% tiene otros planes.

Figura N° 15: Tabla de Frecuencia de la Variable ANTIGUEDAD.

ANTIGUEDAD

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	187335	93,3	93,3	93,3
	2	13361	6,7	6,7	100,0
	Total	200696	100,0	100,0	

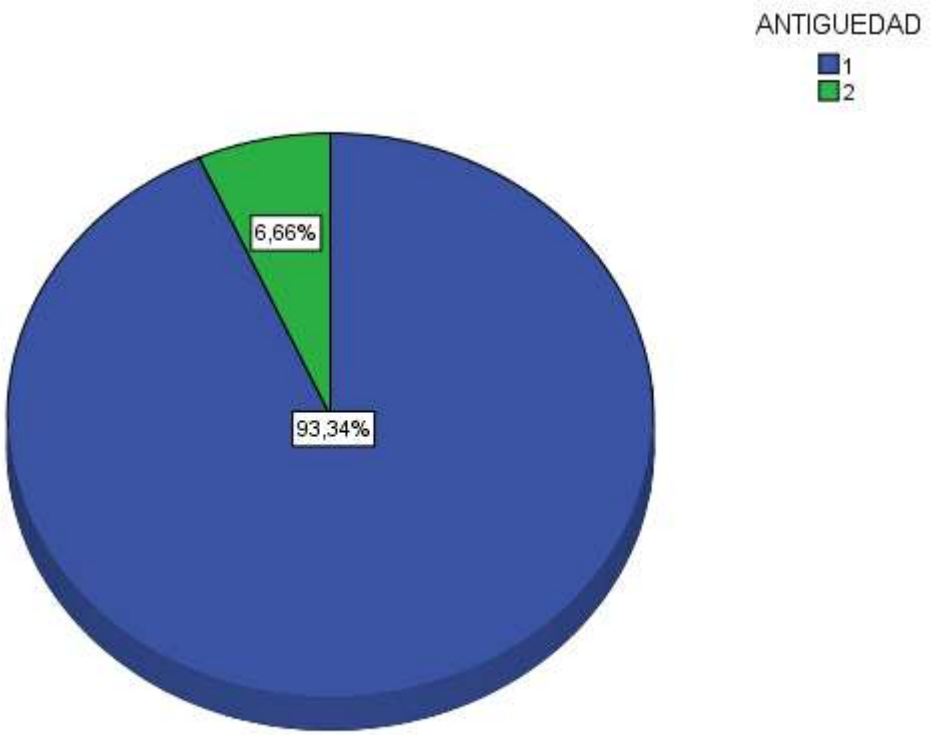
Fuente: Elaboración propia.

Donde:

1= Clientes gestionados por última vez en un periodo menor a un año.

2= Clientes gestionados por última vez en un periodo mayor a un año.

Figura N° 16: Gráfico de pie de la Variable ANTIGUEDAD.



Fuente Elaboración propia.

De la Figura N° 16 el 93.34% del total de clientes fue gestionado por última vez en un periodo menor a un año, mientras el otro 6.66% del total de clientes fue gestionado por última vez en un periodo mayor a un año.

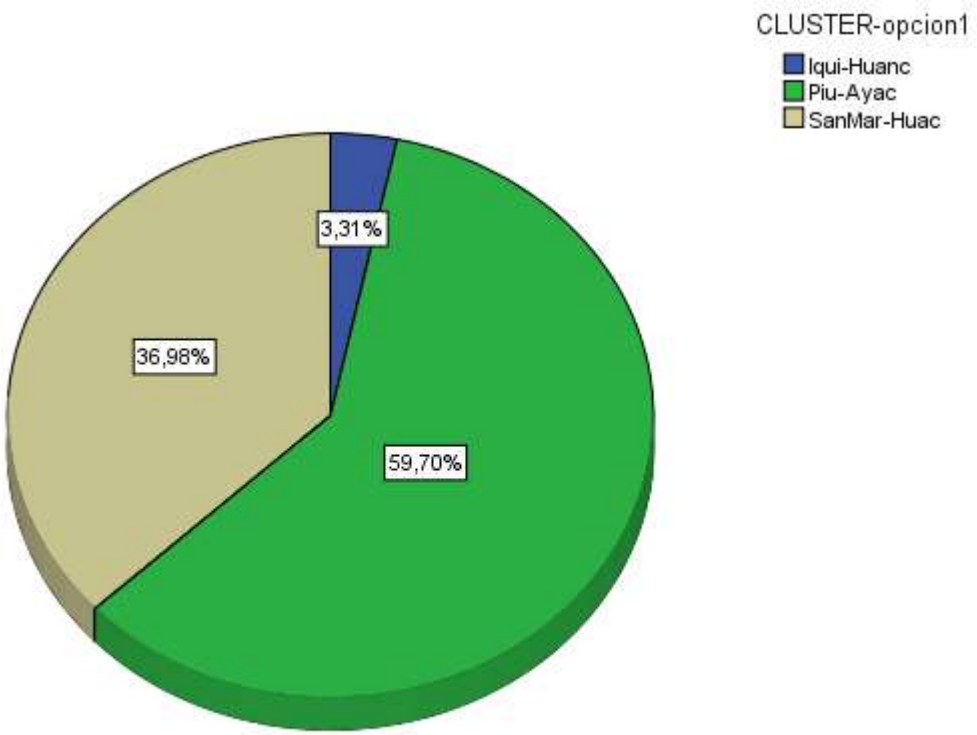
Figura N° 17: Tabla de Frecuencia de la Variable CLUSTER_Opción1.

CLUSTER-opcion1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Iqui-Huanc	6651	3,3	3,3	3,3
	Piu-Ayac	119822	59,7	59,7	63,0
	SanMar-Huac	74223	37,0	37,0	100,0
	Total	200696	100,0	100,0	

Fuente: Elaboración propia.

Figura N° 18: Gráfico de pie de la Variable CLUSTER_Opción1.



Fuente: Elaboración propia.

De la Figura N° 18 el 3.31% del total de clientes reside en Iquitos, Moquegua, Amazonas o Huancavelica, el 36.98% del total de clientes reside en San Martin, Chimbote, Cajamarca, Arequipa, Junín, Puno, Pucallpa, Tumbes, Huánuco, Huaraz, Tacna, Abancay, Cerro de Pasco o Huacho y el 59.70% del total de clientes reside en los departamentos de Ica, Pacasmayo, Cusco, Ayacucho, Piura, Trujillo, Chiclayo, Lima o Madre de Dios.

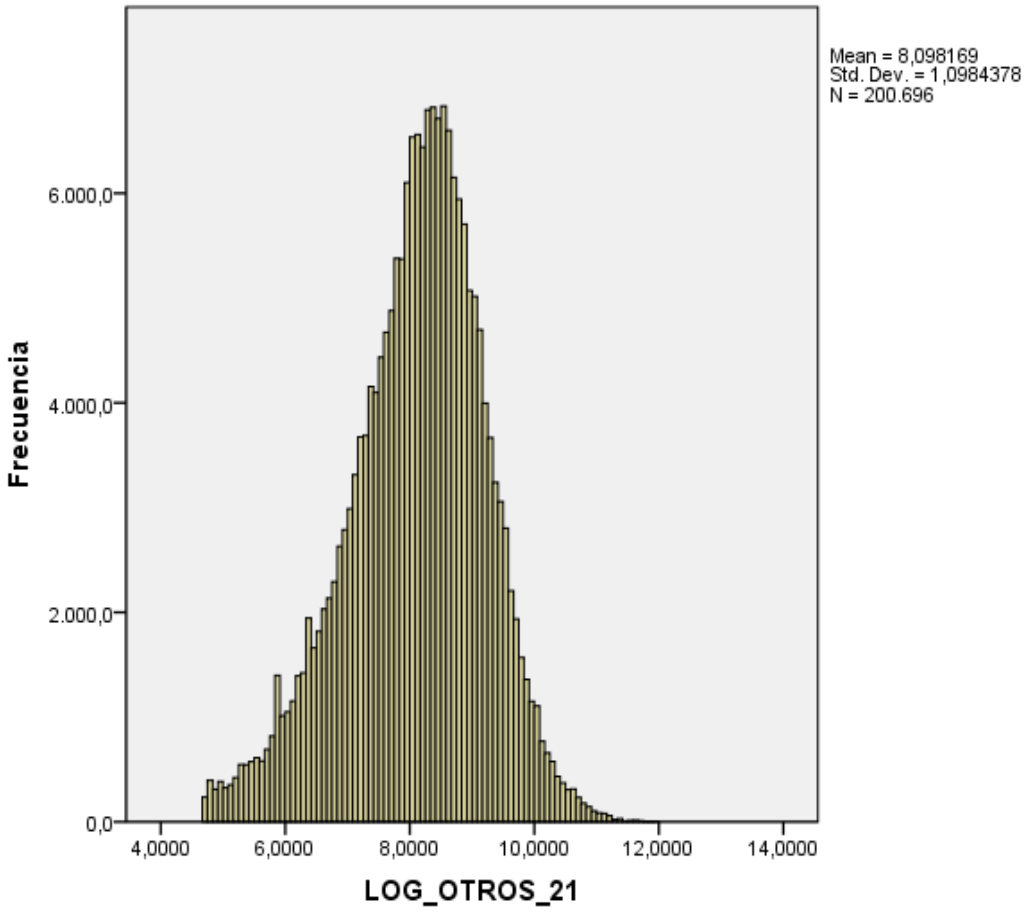
Variables Cuantitativas

Figura N° 19: Tabla descriptiva de Variables Cuantitativas.

	N	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance
LOG_OTROS_21	200696	4,7005	11,9564	1625270,074	8,098169	1,0984378	1,207
BAS_DESPLAN_OF1	200696	20	55	5898925	29,39	9,836	96,740
LOG_OTROS_24	200696	-2,6593	6,4586	368902,9402	1,838118	1,6077007	2,585
LOG_OTROS_10	200696	1,6094	3,8501	525099,4388	2,616392	,5674206	,322

Fuente: Elaboración propia.

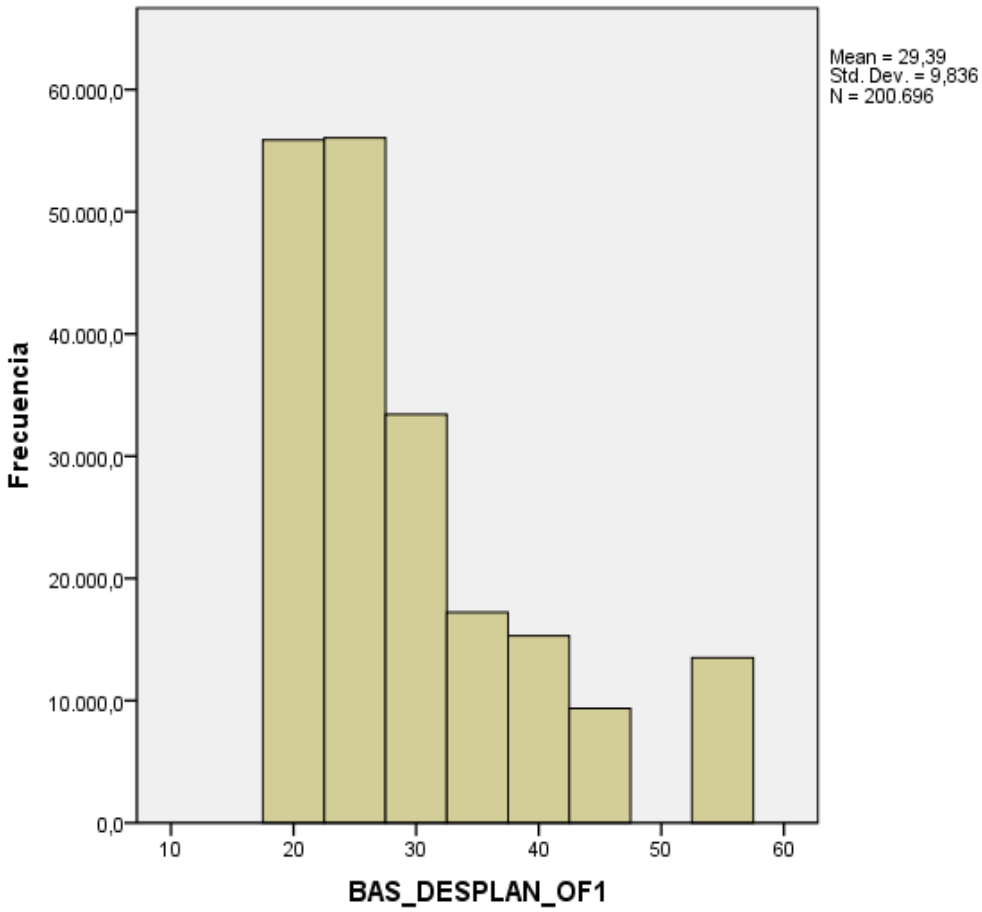
Figura N° 20: Histograma de la Variable LOG_OTROS_21.



Fuente: Elaboración propia.

La Figura N° 20, muestra que los datos se encuentran concentrados alrededor de la media.

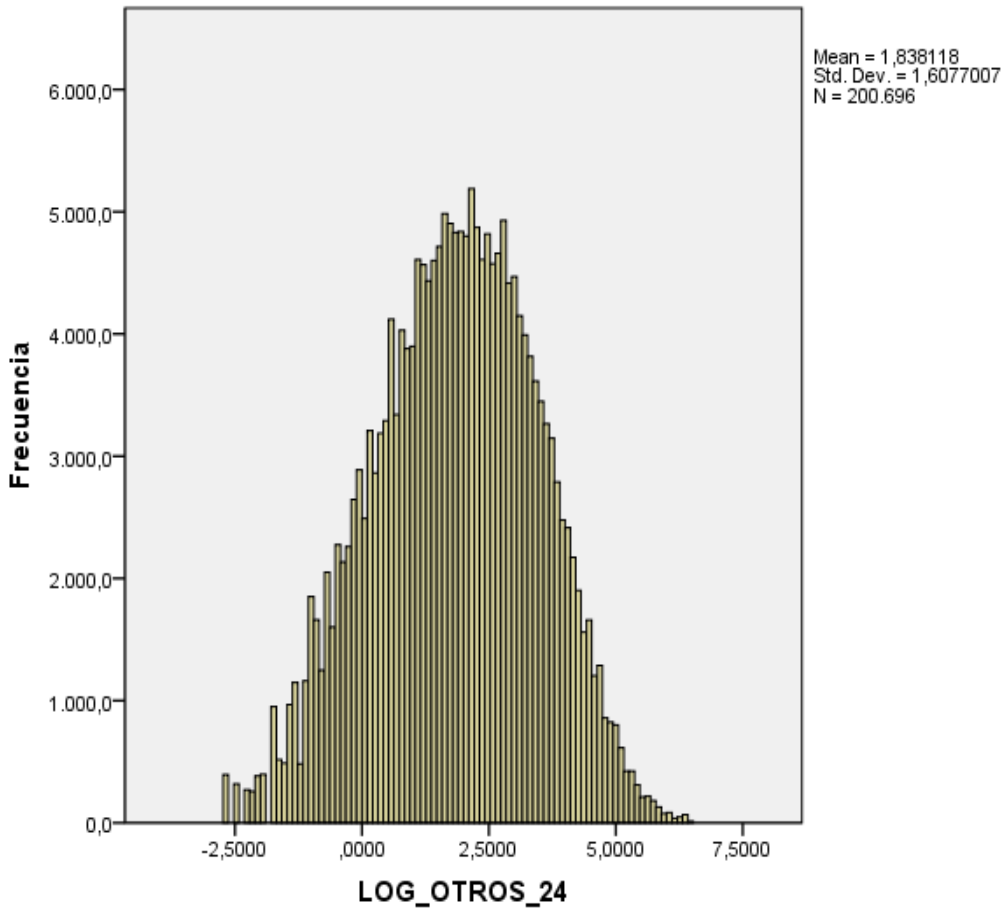
Figura N° 21: Histograma de la Variable BAS_DESPLAN_OF1.



Fuente: Elaboración propia.

La Figura N° 21, muestra que los datos se encuentran concentrados alrededor de la media.

Figura N°22: Histograma de la Variable LOG_OTROS_24.



Fuente: Elaboración propia.

La Figura N° 22, muestra que los datos se encuentran concentrados alrededor de la media.

3.2.1.3.1. Selección de Variables

En esta fase se procedió a identificar los atributos con mayor relevancia para el proceso de clasificación. Al inicio del proceso se consideraron 30 variables que fueron evaluadas para entender el comportamiento de las variables que pueden ser parte del modelo a construir de las cuales de acuerdo al conocimiento del negocio del Call Center se seleccionó aquellas que poseen información relevante y completa quedando finalmente 9 variables que pasaron por un filtro de calidad de datos, limpieza y trasformación de acuerdo a las características del modelo a utilizar.

Una vez identificada las variables en estudio se realizó dos métodos de selección de variables para cada técnica a usar (Regresión Logística y Árbol de Clasificación) que se describe a continuación:

1. **Eliminación de variables correlacionadas:**

   Si un par de variables está altamente correlacionada, es posible que con una de ellas sea suficiente para la construcción de un modelo, ya que se tendría redundancia de información. Una regla aceptable para definir una correlación alta en problemas reales, es corr > 0.5.

2. **Selección de variables por pesos:**

   Es posible asignar por cada atributo cuanto es su peso asociado al problema de clasificación particular. Según esto, varias técnicas como ganancia de información, test chi-cuadrado, coeficiente de correlación, entre otras se puede utilizar para definir cuanto es el peso de un variable dada, con respecto a la variable objetivo.

3.2.1.3.2. Análisis exploratorio de datos

Se realizó un análisis de cada variable de tal manera que se demuestre mediante una gráfica su distribución, considerando la variable respuesta como una segunda clasificación. Estos gráficos se presentaron como un histograma en el cual se diferenciaron la clasificación de la variable respuesta (Y), si es o no venta (Migar de un plan prepago a postpago).

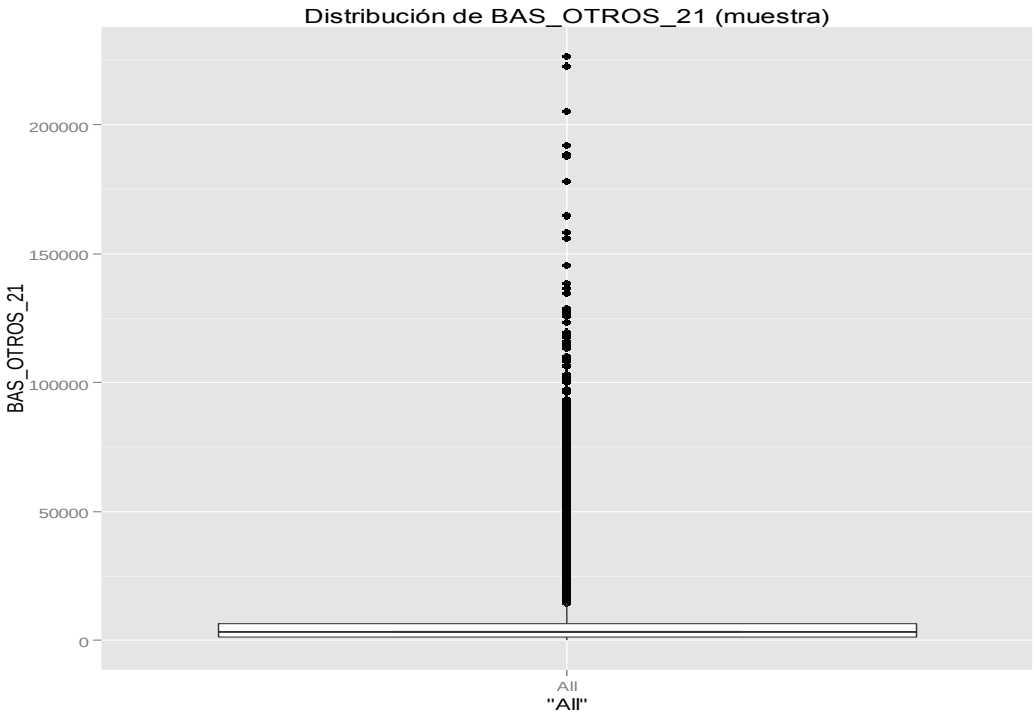
Primero:

Los datos fueron evaluados y analizados mediante una limpieza, la variable BAS_OTROS_21 (Total de llamadas entrantes y salientes), donde el total de llamadas se refiere al número de minutos que cada persona realiza al llamar o al contestar una llamada. Es una variable cuantitativa expresada en minutos.

- **On net**: Llamadas realizadas de móviles Telefónica a otro celular Telefónica.
- **Off net**: Llamadas realizadas de un teléfono celular de Telefónica a otros distribuidores de telefonía móvil (Claro, Entel, etc.)

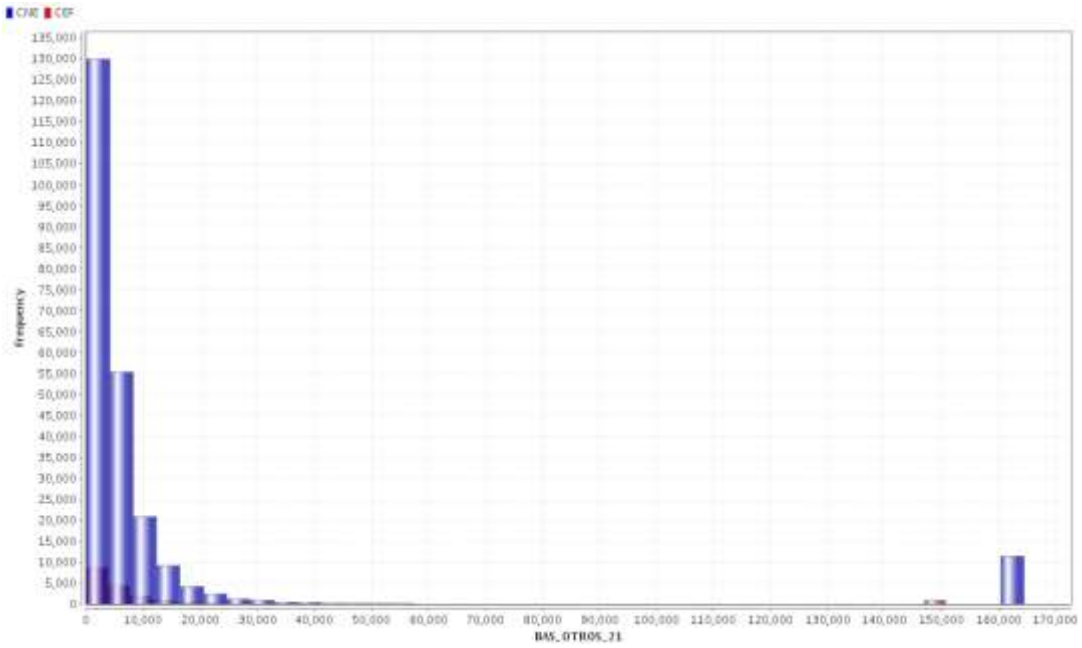
En esta variable se encontró un 0.04% de missing, por lo que fue necesario depurar los outliers y valores más visibles, ello fue verificado mediante un diagrama de cajas, luego se procedió a realizar una transformación logarítmica para obtener una mejor distribución de los datos.

Figura N° 23: Diagrama de cajas variable (BAS_OTROS_21).



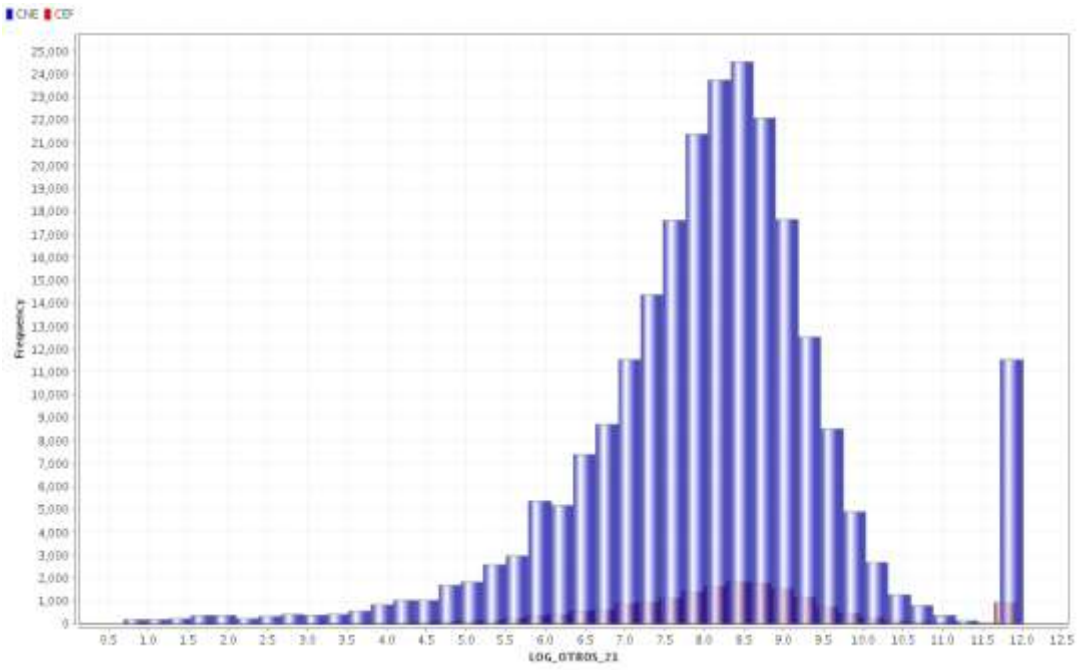
Fuente: Elaboración propia.

Figura N° 24: Gráfica sin considerar los valores más visibles variable (BAS_OTROS_21).



Fuente: Elaboración propia.

Figura N° 25: Gráfica después de la transformación logarítmica variable (LOG_OTROS_21).



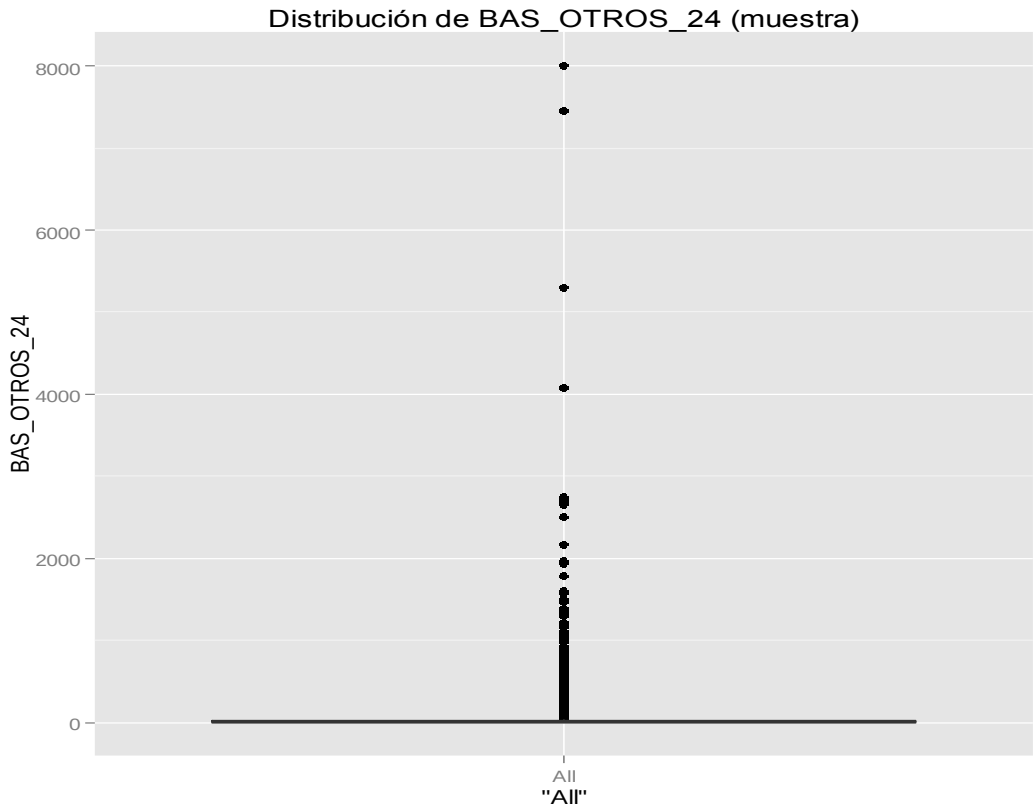
Fuente: Elaboración propia.

Nuevamente se eliminaron los outliers y los missing ya que se presentaron en una cantidad no muy significativa, más adelante se explicarán cómo trabajar con aquellos missing de las nuevas bases.

Segundo:

Se analizó y depuró la variable BAS_OTROS_24 (Tráfico de llamadas salientes a RPM).Es una variable cuantitativa expresada en minutos, presenta valores outliers y valores extremos

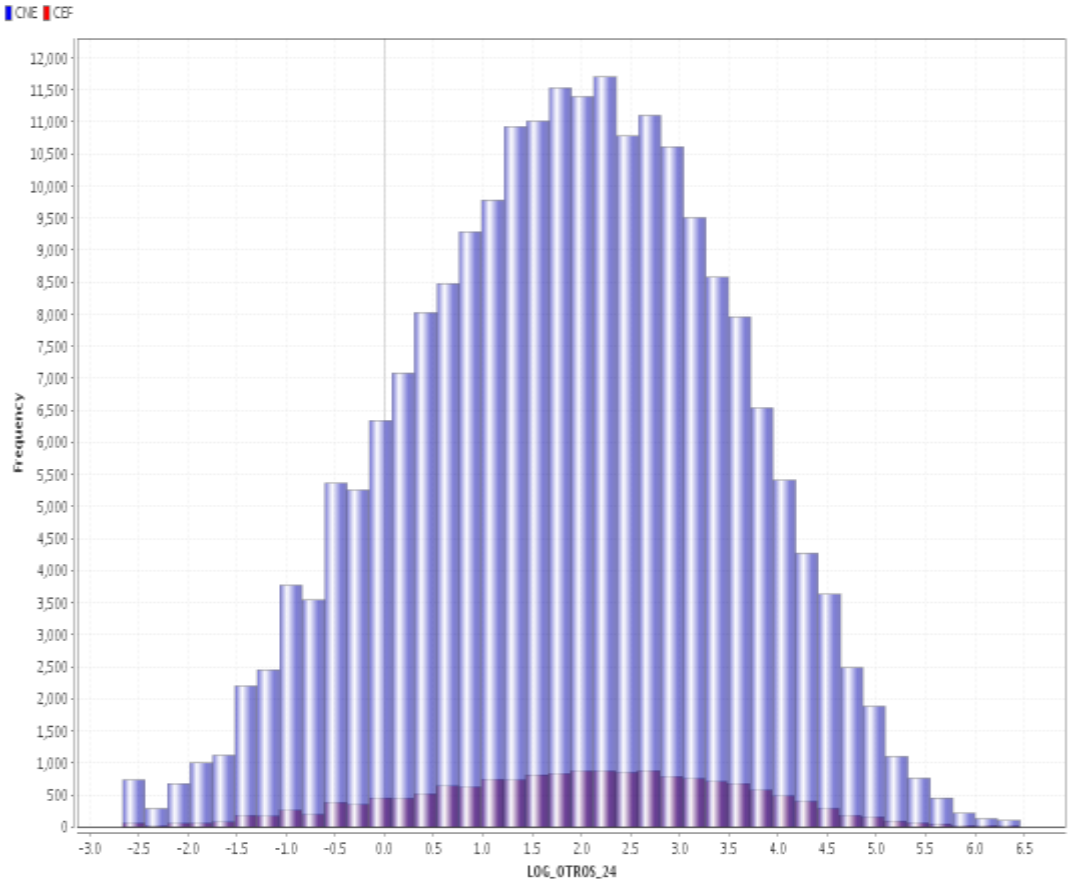
Figura N° 26: Diagrama de cajas variable (BAS_OTROS_24).



Fuente: Elaboración propia.

Se procedió a eliminar los outliers más visibles y aplicar una transformación logarítmica obteniendo como resultado la Figura N° 26.

Figura N° 27: Gráfica después de la transformación logarítmica variable (LOG_OTROS_24).

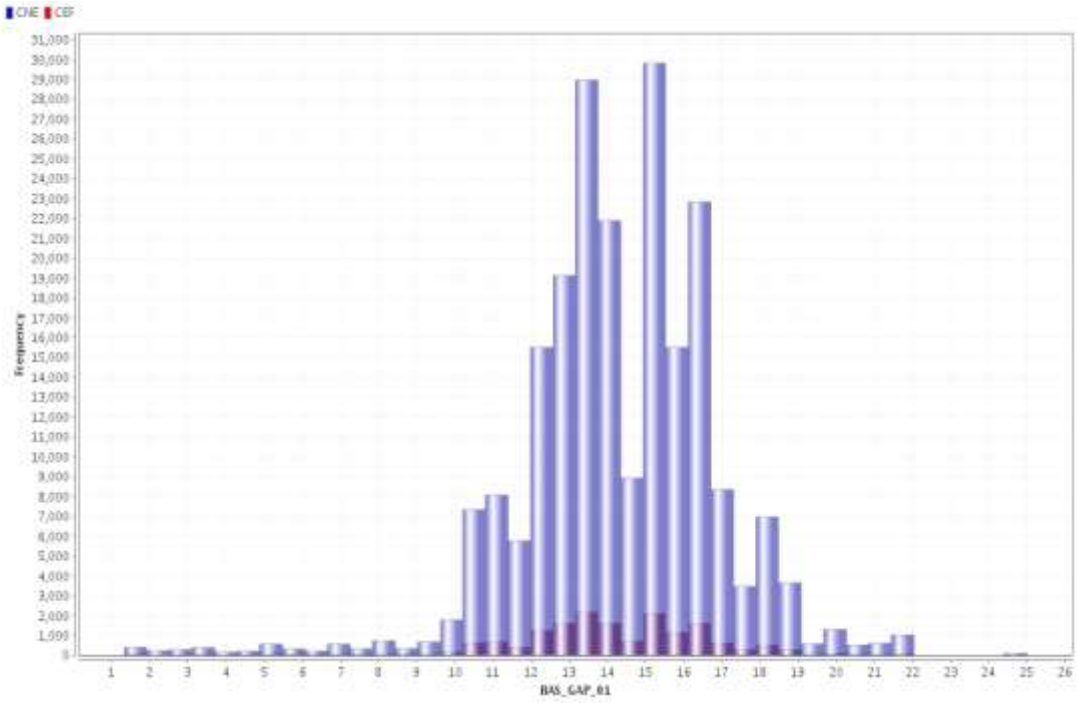


Fuente: Elaboración propia.

Tercero:

La siguiente variable que se evaluó fue BAS_GAP_01 (Es el salto entre el Monto promedio de Activación del cliente y cargo fijo del plan a ofrecer). Es una variable cuantitativa medida en soles.

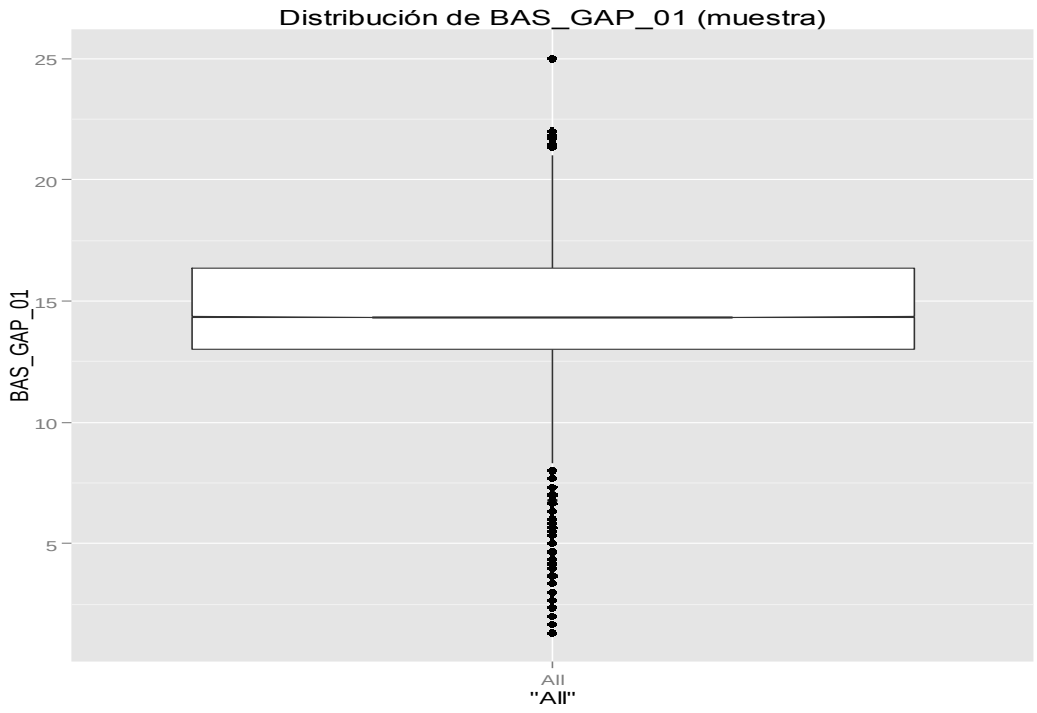
Figura N° 28: Representación gráfica de la distribución variable (BAS_GAP_01).



Fuente: Elaboración propia.

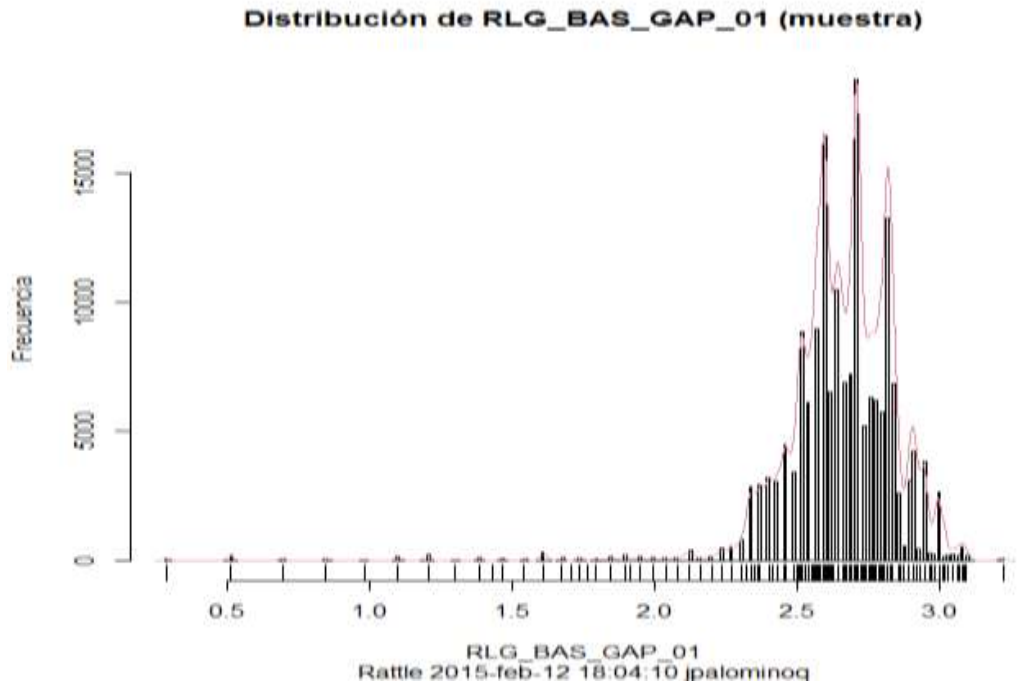
Esta variable presentó una tendencia a la normalidad, aunque al igual que el resto de variables cuantitativas presentó valores extremos y outliers, además no presentó una agrupación central completa y continua, en la parte central presentó una caída y luego sube.

Figura N° 29: Diagrama de cajas (BAS_GAP_01).



Fuente: Elaboración propia.

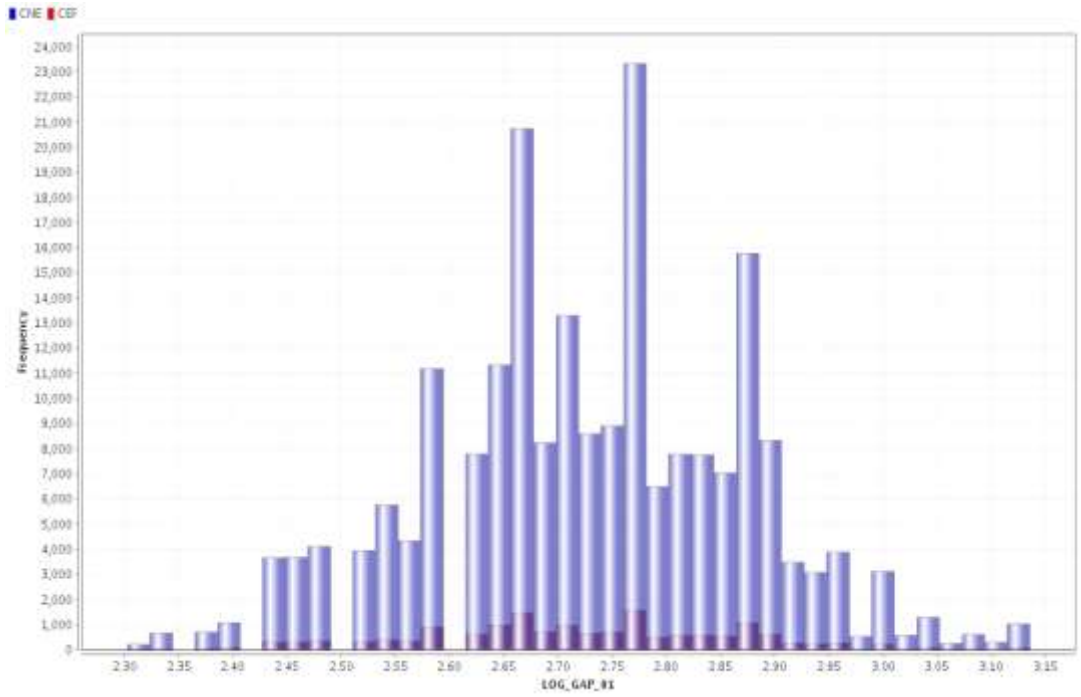
Figura N° 30: Grafica después de la trasformación logarítmica de la variable (BAS_GAP_01).



Fuente: Elaboración propia.

Con la transformación observamos que aún sigue presentando una tendencia a la normal; pero los valores no son continuos existen “huecos” entre uno y otro valor, además de picos altos y bajos, podemos confirmar que los datos de esa variable no son continuos. Depurando estos outliers la distribución queda de la siguiente manera:

Figura N° 31: Representación gráfica luego de depurar los outliers (LOG_GAP_01).



Fuente: Elaboración propia.

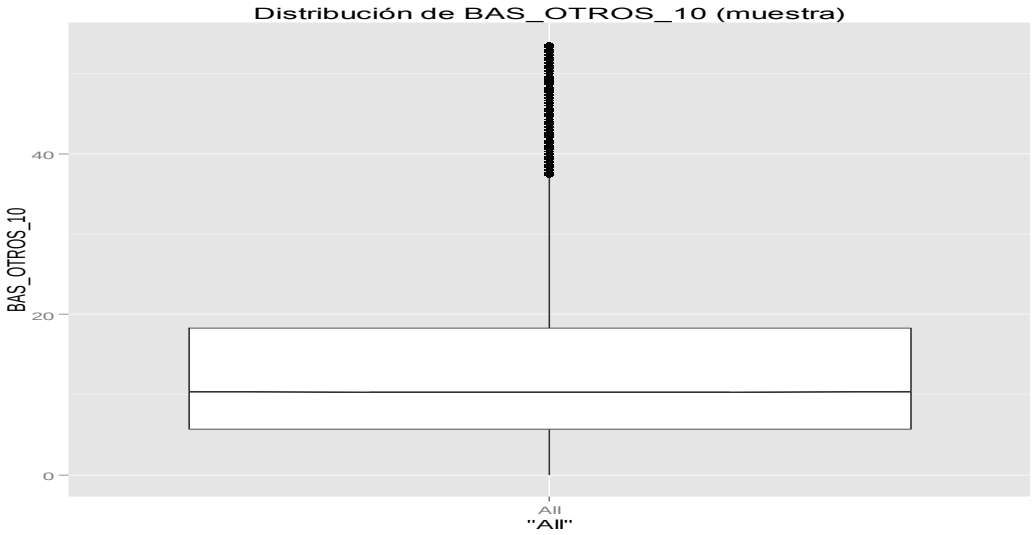
Cuando se tiene una variable cuantitativa como LOG_GAP 01 que no presenta tendencia a la normalidad optamos por realizar una discretización de la siguiente manera:

- Se verifica y comprueba los puntos de corte para agruparlos, con los quartiles, en este caso utilizamos el primer y tercer quartil.
- También podemos usar la discretización del rapid miner la cual separa en grupos proporcionados si la variable es plana, y si tiene tendencia a la normal entonces la divide en tres grupos de subida, ganancia y pérdida.

Cuarto:

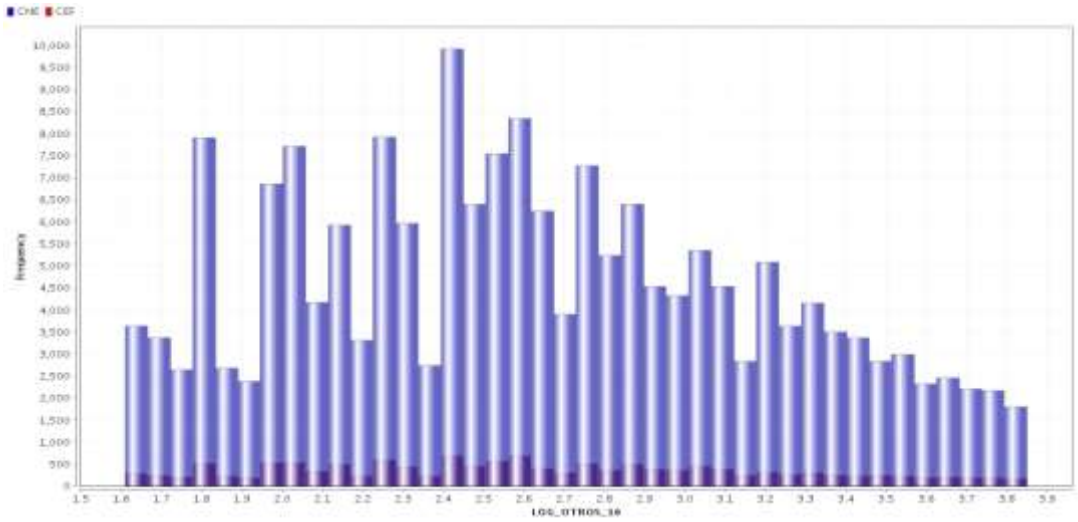
Terminando con las variables cuantitativas analizamos la variable BAS_OTROS_10 (Monto promedio de activación o recarga). Es el gasto promedio del consumo del cliente en los últimos tres meses. Es una variable cuantitativa expresada en soles.

Figura N° 32: Diagrama de cajas variable (BAS_OTROS_10).



Fuente: Elaboración propia.

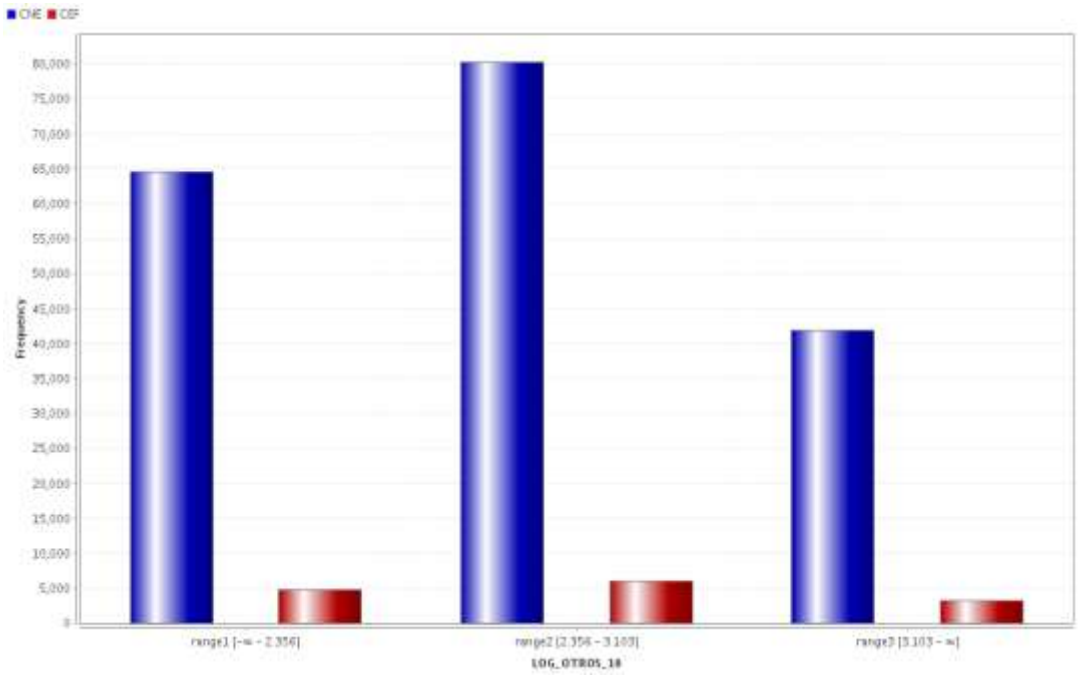
Figura N° 33: Gráfica después de la transformación logarítmica variable (LOG_OTROS_10).



Fuente: Elaboración propia.

Esta es una variable que no presenta una tendencia a la normal, se visualiza una distribución plana, por ello se optara por discretizarla.

Figura N° 34: Variable Discretizada (LOG_OTROS_10).



Fuente: Elaboración propia.

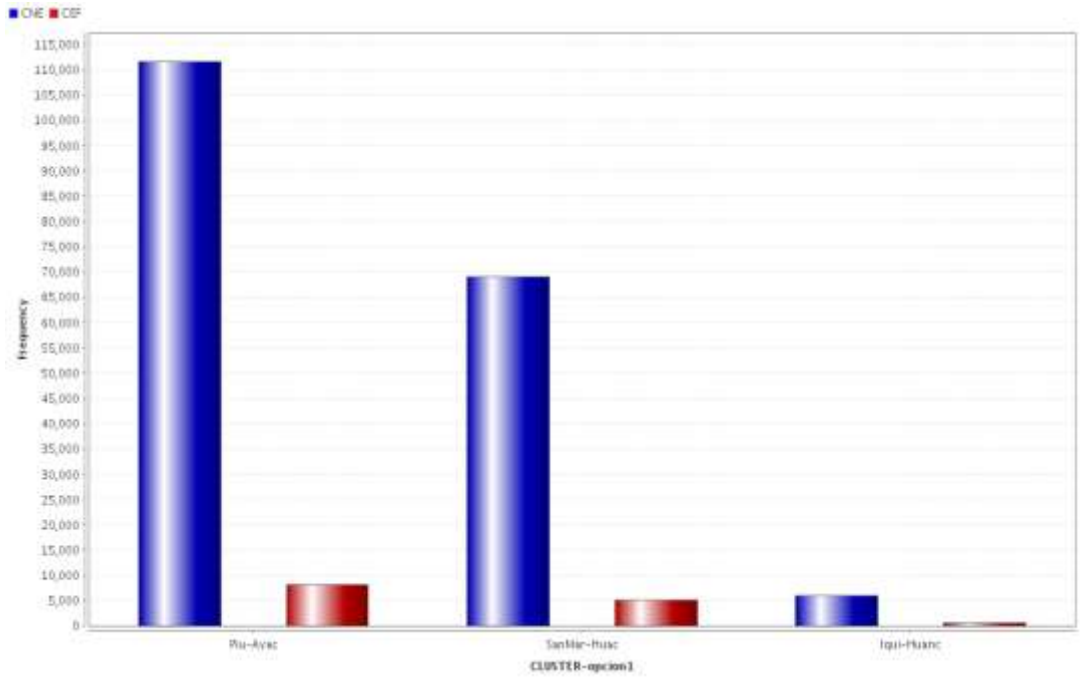
<u>Quinto:</u>

Continuamos analizando la variable BAS_LOCALIDAD que como su nombre lo indica, es la provincia donde reside actualmente el cliente, esta variable por ser cualitativa nominal y presentar diversas categorías se agrupara en base a sus efectividades mediante la técnica de clusterizacion.

La variable localidad es clusterizada mediante un clúster jerárquico en donde se obtuvieron finalmente 2 opciones de Cluster. (**Ver Anexo N° 3**).

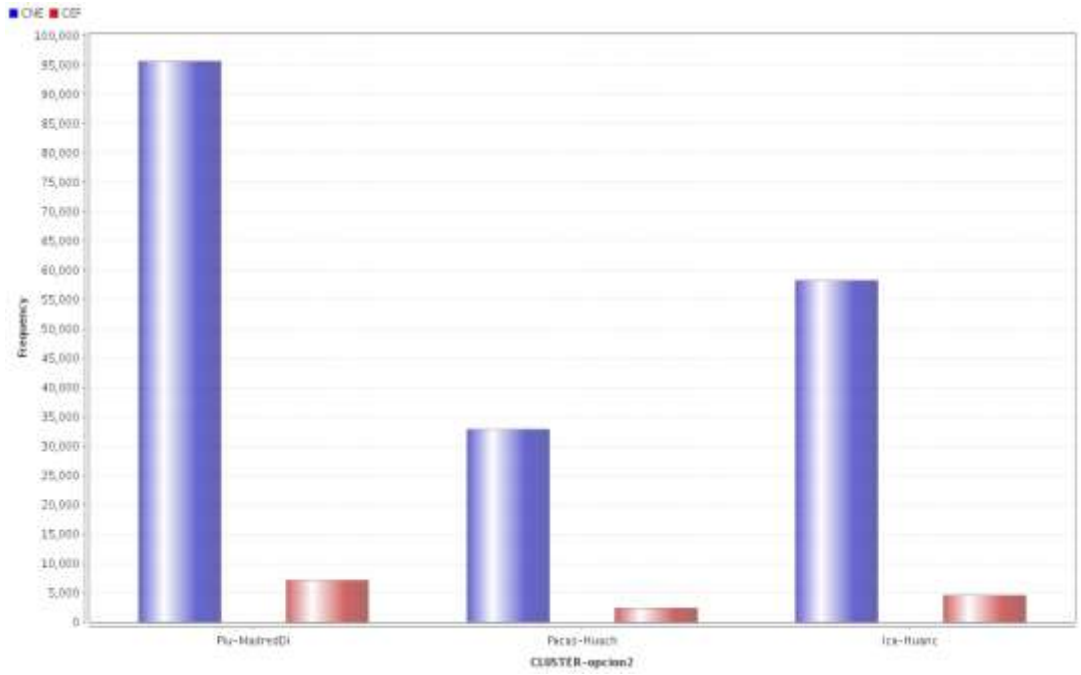
- CLUSTER_opción 1
- CLUSTER_opción 2

Figura N° 35: Variable CLUSTER_opción1.



Fuente: Elaboración propia.

Figura N° 36: Variable CLUSTER_opción2.



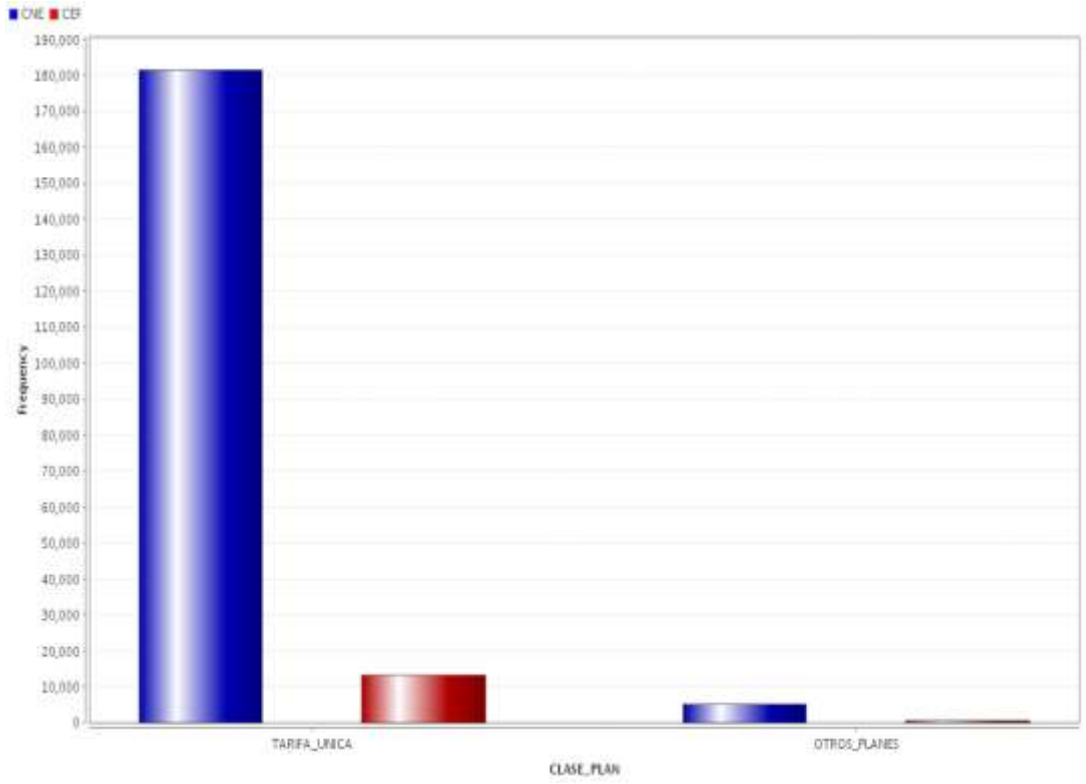
Fuente: Elaboración propia.

Sexto:

La variable CLASE_PLAN es el plan prepago que tiene cada cliente. Es una variable cualitativa nominal que fue dividida en 3 agrupaciones (TARIFA UNICA, TODO DIA 4 y OTROS PLANES); pero por la cantidad de base que presentaba y efectividades similares se tomó la decisión de agruparla en 2 categorías, debido a que una de ellas se encontraba en menor proporción y no aportaba mucho como una variable seleccionada para el modelo.

- Tarifa única Nacional (TUN)
- Todo día 4
- Otros planes (Otros)

Figura N° 37: Variable (CLASE_PLAN).



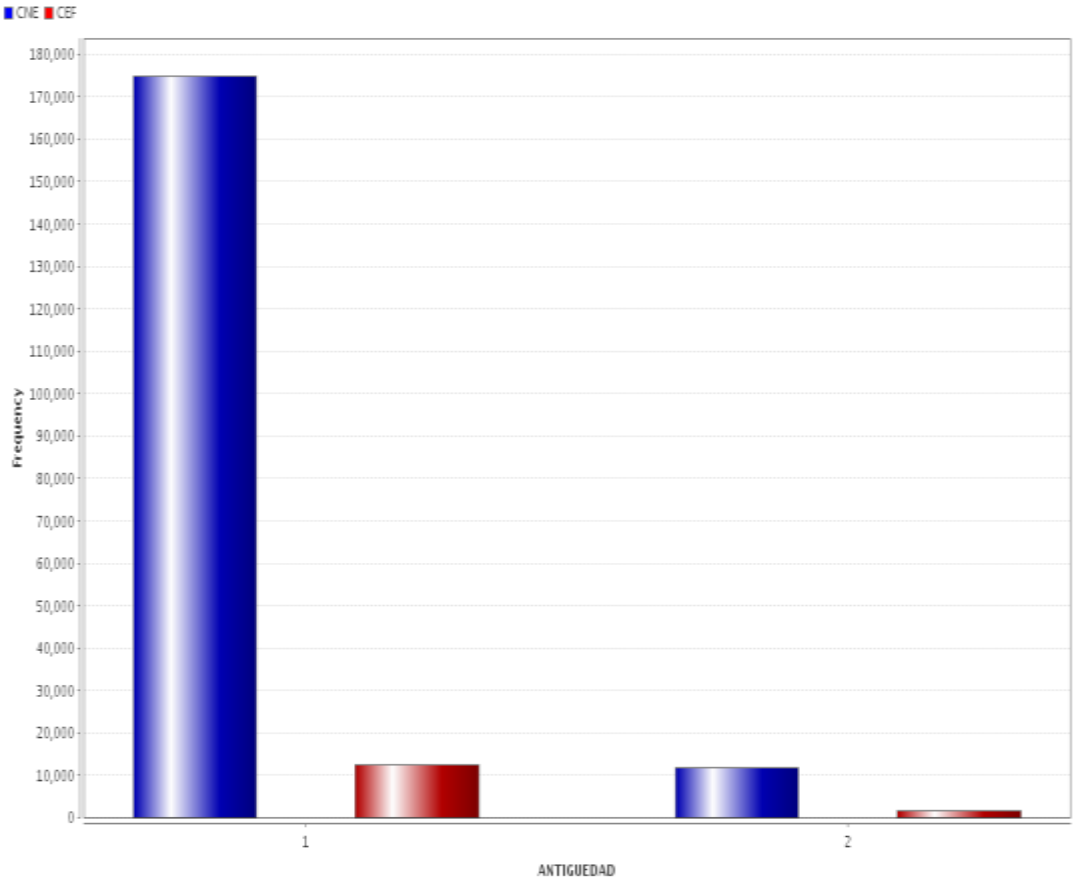
Fuente: Elaboración propia.

Séptimo:

La variable ANTIGUEDAD, es el tiempo en años en que se gestionó al cliente por última vez. Esta es una variable cualitativa ordinal, estaba dividida en cuatro categorías (1, 2,3 y 4), se tomó la decisión de agruparla en base a sus efectividades en dos grupos debido a que las categorías 3 y 4 se encontraban en menor proporción y no aportaba mucho como una variable seleccionada para el modelo

- **1=** Cliente gestionado por última vez en un periodo menor a un año.
- **2=** Clientes gestionados por última vez en un periodo mayor a un año.

Figura N° 38: Variable (ANTIGUEDAD).



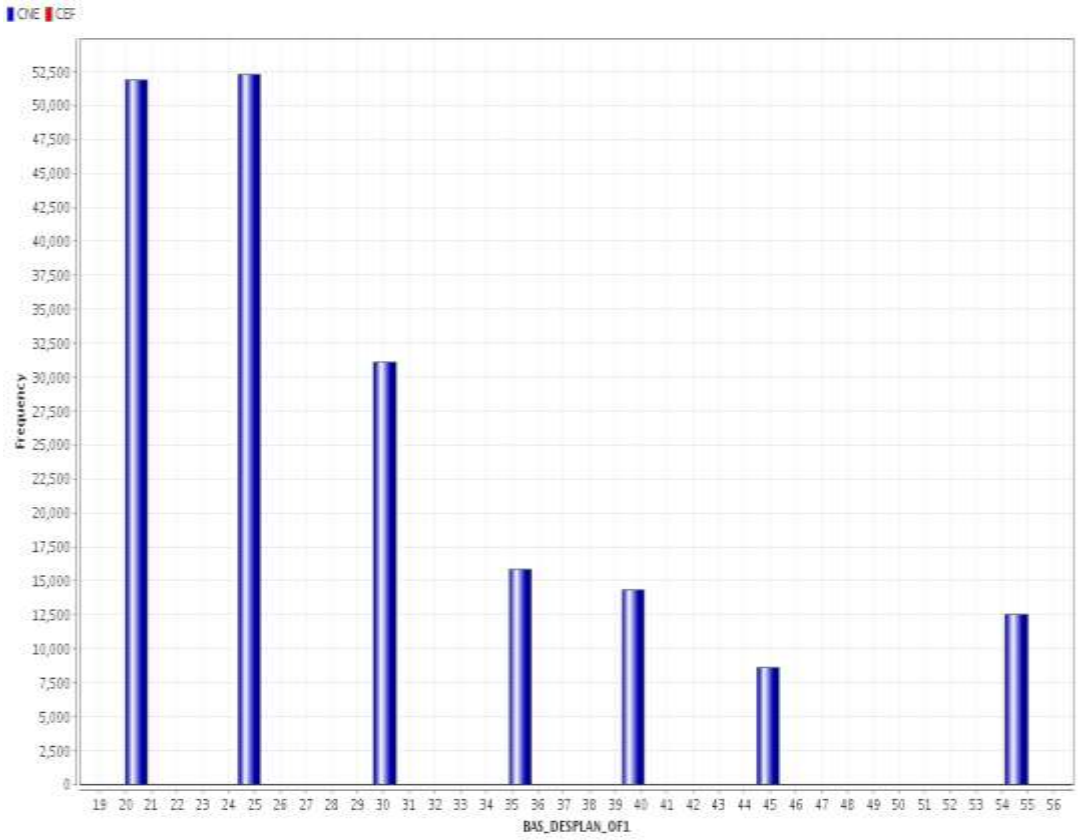
Fuente: Elaboración propia.

Octavo:

La variable BAS_DESPLAN_OF1, es cualitativa ordinal, y contiene el costo del plan postpago que se le ofrecerá al cliente y que pagara mensual.

19.99– 24.99 – 29.99 – 34.99 – 39.99 – 44.99 – 54.99

Figura N° 39: Variable (BAS_DESPLAN_OF1).



Fuente: Elaboración propia.

3.2.1.4. Modelado

En este trabajo se realizó una comparación entre el Modelo de Regresión Logística y el algoritmo de Árbol de Clasificación CART.

Se tomó la decisión de utilizar estas técnicas, ya que la variable dependiente es una binomial dividida como Tipo_Resultado = CEF (Contacto Efectivo venta) y Tipo_Resultado =CNE (Contacto No Efectivo no venta). Mientras que las variables predictoras son cualitativas y cuantitativas.

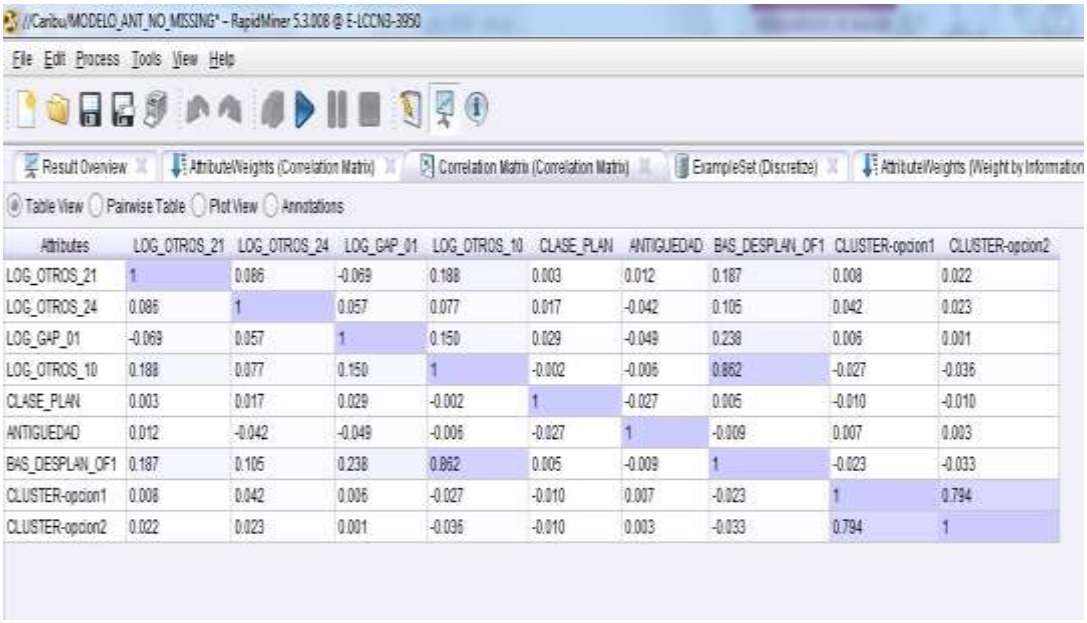
Modelo Aplicando Regresión Logística:

Se realizó un modelo para los Clientes Antiguos, es decir aquellos que ya fueron contactados anteriormente y además tienen información del Tráfico promedio de llamadas salientes a RPM (LOG_OTROS_24).

Selección de Variables:

Mediante la correlación de las variables se verifica cuáles son las que presentan una alta relación o dependencia entre sí.

Figura N° 40: Matriz de correlación del Modelo Logístico.



Attributes	LOG_OTROS_21	LOG_OTROS_24	LOG_GAP_01	LOG_OTROS_10	CLASE_PLAN	ANTIGUEDAD	BAS_DESPLAN_OF1	CLUSTER-opcion1	CLUSTER-opcion2
LOG_OTROS_21	1	0.086	-0.069	0.188	0.003	0.012	0.187	0.008	0.022
LOG_OTROS_24	0.086	1	0.057	0.077	0.017	-0.042	0.105	0.042	0.023
LOG_GAP_01	-0.069	0.057	1	0.150	0.029	-0.049	0.238	0.006	0.001
LOG_OTROS_10	0.188	0.077	0.150	1	-0.002	-0.006	0.862	-0.027	-0.036
CLASE_PLAN	0.003	0.017	0.029	-0.002	1	-0.027	0.005	-0.010	-0.010
ANTIGUEDAD	0.012	-0.042	-0.049	-0.006	-0.027	1	-0.009	0.007	0.003
BAS_DESPLAN_OF1	0.187	0.105	0.238	0.862	0.005	-0.009	1	-0.023	-0.033
CLUSTER-opcion1	0.008	0.042	0.006	-0.027	-0.010	0.007	-0.023	1	0.794
CLUSTER-opcion2	0.022	0.023	0.001	-0.036	-0.010	0.003	-0.033	0.794	1

Fuente: Elaboración propia.

Para analizar si las variables presentan correlación entre ellas se realizó la matriz de correlación fijando como punto de corte 0.5.

De la Figura N° 40 se descartó las variables BAS_DESPLAN_OF1 y CLUSTER_opcion2 por presentar una correlación mayor al punto de corte fijado en este caso 0.5.

Luego se procedió a verificar la importancia de los pesos que tienen cada una.

Figura N° 41: Matriz de pesos del Modelo Logístico.

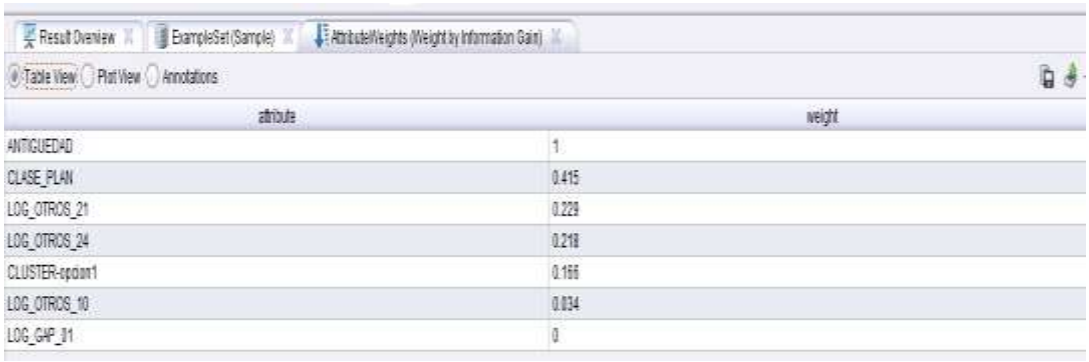

Result Overview | ExampleSet (Sample) | AttributeWeights (Weight by Information Gain)

Table View | Plot View | Annotations

attribute	weight
ANTIGUEDAD	1
CLASE_PLAN	0.415
LOG_OTROS_21	0.229
LOG_OTROS_24	0.218
CLUSTER-opcion1	0.166
LOG_OTROS_10	0.034
LOG_GAP_01	0

Fuente: Elaboración propia.

Analizando la matriz de pesos se descartó las variables LOG_GAP_01 y LOG_OTROS_10, por no contribuir significativamente con la variable en estudio.

- **Variables seleccionadas**

Luego de analizar la matriz de pesos de las variables, se procedió a seleccionar las variables que contribuyen significativamente con la variable en estudio, las cuales se muestran en el cuadro siguiente:

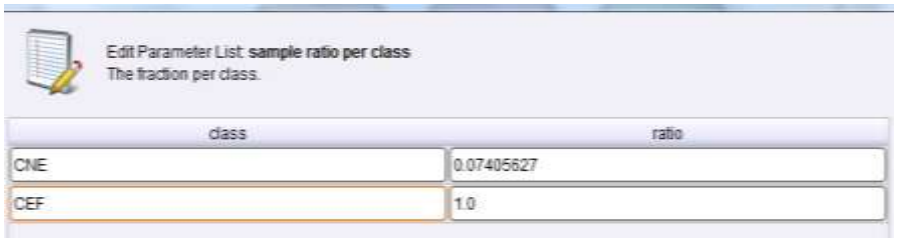
Cuadro N° 7: Variables Seleccionadas del Modelo Logístico.

Campo	Descripción
TIPO_RESULTADO (Variable Dependiente)	Resultado de la gestión final realizada al cliente. (Venta y No venta).
CLUSTER_LOCALIDAD	Localidad del Cliente.
LOG_OTROS_21	Transformación Logarítmica del Total de llamadas entrantes y salientes.
LOG_OTROS_24	Transformación Logarítmica del Promedio de Tráfico de llamadas salientes a RPM.
ANTIGUEDAD	Tiempo en años en que se gestionó un cliente por última vez.
CLASE_PLAN	Clusterización del Plan de origen.

Fuente: Elaboración Propia.

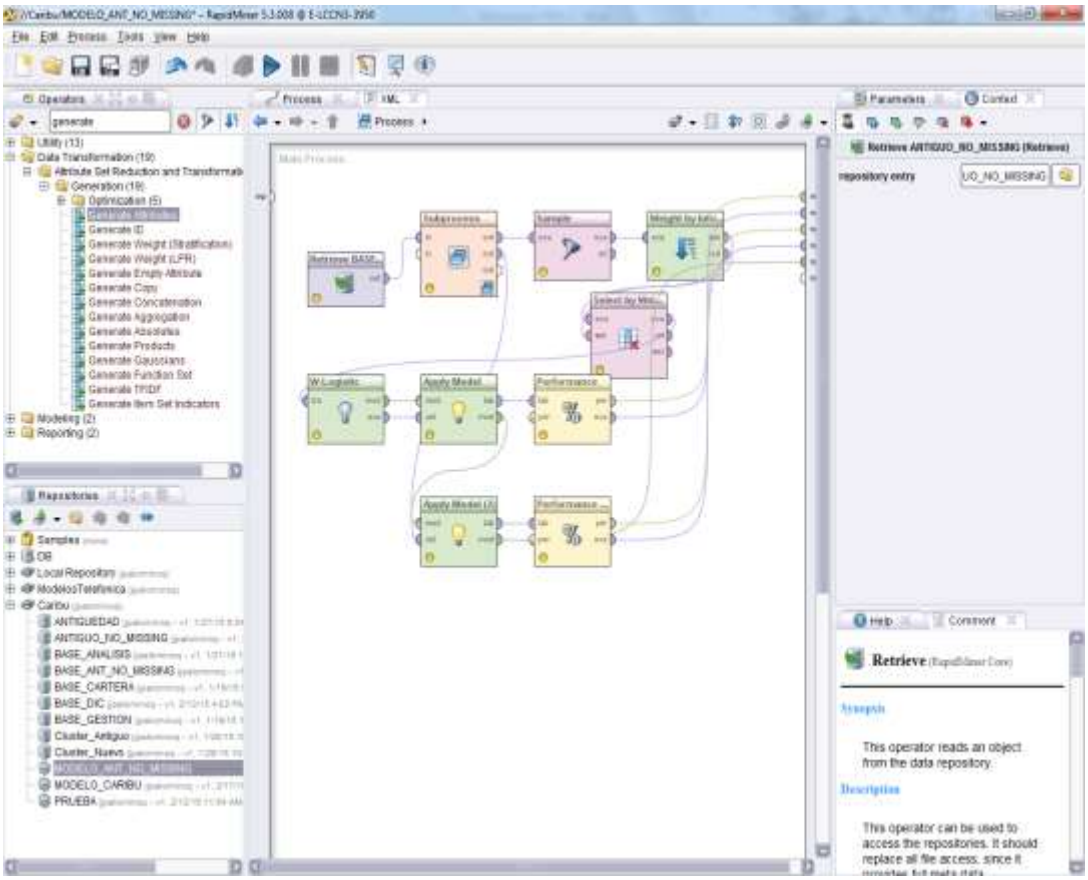
Con las variables seleccionadas se buscara alcanzar el objetivo, predecir que el 100% de clientes propensos a comprar un plan postpago sea venta, un resultado esperado y optimista. Para iniciar con el proceso de modelamiento se utiliza la caja SAMPLE, en la cual se balancea la base original en un ratio de 0.074 para las No ventas y 1.00 para las ventas. Tendremos un conjunto de datos de entrenamiento/prueba.

Figura N° 42: Balanceo de muestra del Modelo Logístico.



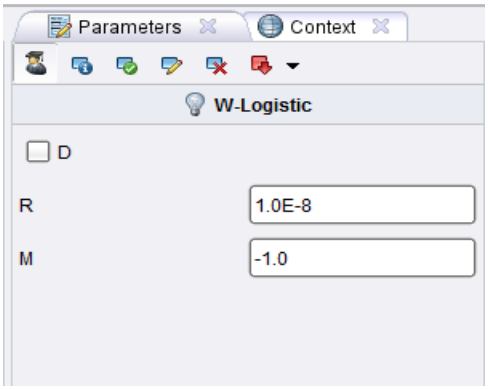
Fuente: Elaboración propia.

Figura N° 43: Procedimiento para generar el Modelo Logístico.



Fuente: Elaboración propia.

Figura N° 44: Parámetros del Modelo Logístico.



Fuente: Elaboración propia.

Modelo Logístico: (Clientes antiguos que si tienen un tráfico promedio de llamadas salientes a RPM)

$$logit\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

Figura N° 45: Coeficientes del Modelo Logístico.

Variables	Coeficientes	
Intercepto	β_0	0.4074
Log_Otros_24	β_1	-0.0514
Cluster_opcion1=Iqui-Huanc	β_2	-0.3241
Cluster_opcion1=SanMar-Huac	β_3	0.0233
Cluster_opcion1=Piu-Ayac	β_4	0.0264
Log_Otros_21	β_5	-0.0336
Antigüedad	β_6	-0.6014
Clase_Plan	β_7	-0.5992

Fuente: Elaboración propia.

Figura N° 46: Odds Ratio de las variables del Modelo Logístico.

Variables	Coeficientes
Log_Otros_24	0.9499
Cluster_opcion1=Iqui-Huanc	0.7232
Cluster_opcion1=SanMar-Huac	1.0236
Cluster_opcion1=Piu-Ayac	1.0268
Log_Otros_21	0.9669
Antigüedad	0.548
Clase_Plan	0.5493

Fuente: Elaboración Propia

Figura N° 47: Intervalos de Confianza del Modelo Logístico.

	95% C.I.for EXP(B)	
Variables	Lower	Upper
Log_Otros_24	0.936	0.964
Cluster_opcion1=Iqui-Huanc	0.621	0.804
Cluster_opcion1=SanMar-Huac		
Cluster_opcion1=Piu-Ayac	0.955	1.054
Log_Otros_21	0.946	0.988
Antigüedad	0.502	0.598
Clase_Plan	0.482	0.626

Fuente: Elaboración Propia

La variable Cluster_opción1=Piu-Ayac, tienen intervalos de confianza que cubre el 1, por lo que no tiene efecto alguno sobre la variable respuesta, es decir no evidencia una asociación entre las variables involucradas, por ende se concluyó que no debe considerar dicha categoría en el modelo.

Se procedió a correr nuevamente el modelo, quitando la variable Cluster_opción1, debido a que no tiene efecto alguno sobre la variable respuesta, obteniéndose los siguientes resultados:

Nuevo Modelo Logístico: (Clientes antiguos que si tienen un tráfico promedio de llamadas salientes a RPM)

$$\boldsymbol{logit}\left(\frac{\boldsymbol{\pi(x)}}{\boldsymbol{1-\pi(x)}}\right)=\boldsymbol{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4}$$

Figura N° 48: Coeficientes del Modelo Logístico sin la variable Cluster_opción1.

Variables	Coeficientes	
Intercepto	β_0	0.4075
Log_Otros_24	β_1	-0.0525
Log_Otros_21	β_2	-0.0322
Antigüedad	β_3	-0.6058
Clase_Plan	β_4	-0.5979

Fuente: Elaboración propia.

Figura N° 49: Odds Ratio de las variables del Modelo Logístico sin la variable Cluster_opción1.

Variables	Coeficientes
Log_Otros_24	0.9489
Log_Otros_21	0.9683
Antigüedad	0.5456
Clase_Plan	0.55

Fuente: Elaboración Propia

Figura N° 50: Intervalos de Confianza del Modelo Logístico sin la variable Cluster_opción1.

	95% C.I.for EXP(B)	
Variables	Lower	Upper
Log_Otros_24	0.935	0.963
Log_Otros_21	0.948	0.989
Antigüedad	0.500	0.596
Clase_Plan	0.482	0.627

Fuente: Elaboración Propia

Con los intervalos de confianza obtenidos, se corroboró la significación de las cuatro variables con respecto a la variable respuesta.

Modelo Aplicando Árboles de Clasificación:

Se trabajó con las mismas variables obtenidas en el modelo logístico.

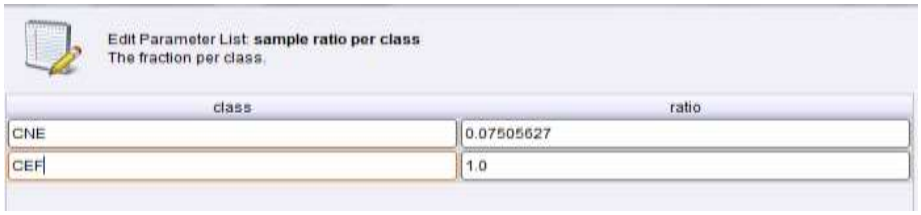
Cuadro N° 8: Variables Seleccionadas del Árbol de Clasificación.

Campo	Descripción
TIPO_RESULTADO (Variable Dependiente)	Resultado de la gestión final realizada al cliente. (Venta y No venta).
LOG_OTROS_21	Transformación Logarítmica del Total de llamadas entrantes y salientes.
LOG_OTROS_24	Transformación Logarítmica del Promedio de Tráfico de llamadas salientes a RPM.
CLASE_PLAN	Es el plan prepago del cliente.
ANTIGUEDAD	Tiempo en años en que se gestionó un cliente por última vez.

Fuente: Elaboración propia.

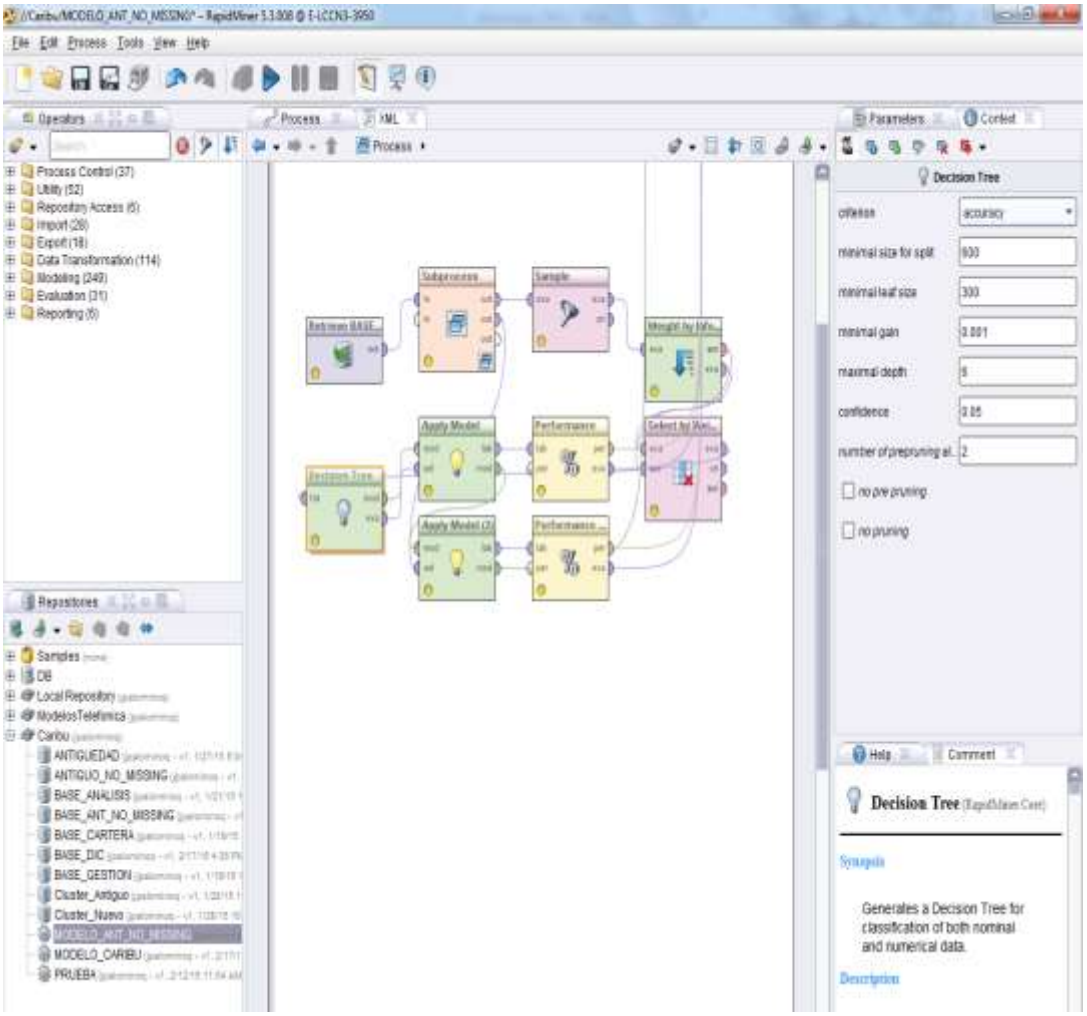
Con las variables seleccionadas se buscara alcanzar el objetivo, predecir que el 100% de clientes propensos a comprar un plan postpago sea venta, un resultado esperado y optimista. Para iniciar con el proceso de modelamiento se utiliza la caja SAMPLE, en la cual se balancea la base original en un ratio de 0.075 para las no ventas y 1.00 para las ventas. Tendremos un conjunto de datos de entrenamiento/prueba.

Figura N° 51: Balanceo de Muestras para la Aplicación del Árbol de Clasificación.



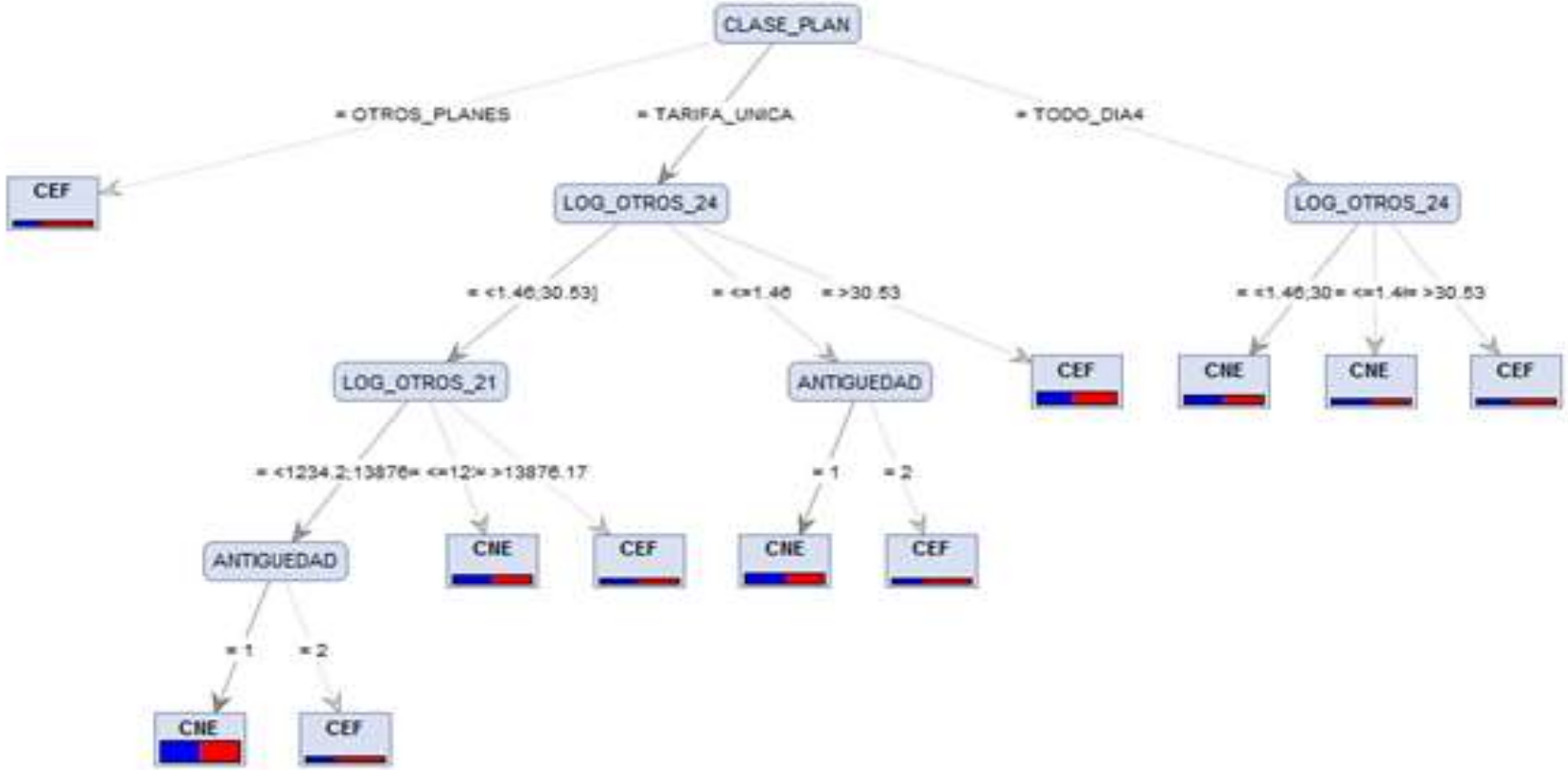
Fuente: Elaboración propia.

Figura N° 52: Procedimiento para generar el Árbol de Clasificación.



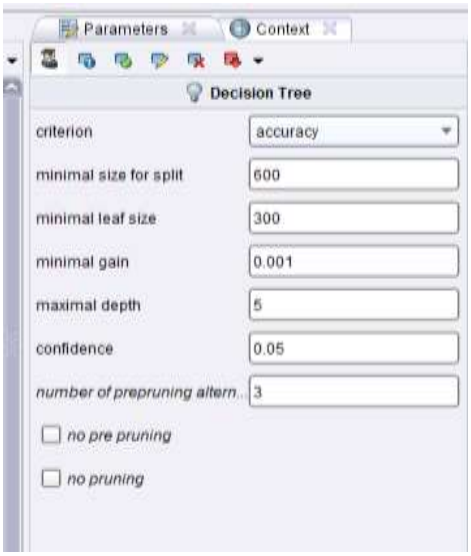
Fuente: Elaboración propia.

Figura N° 53: Gráfica del Árbol de Clasificación.



Fuente: Elaboración propia.

Figura N° 54: Parámetros del Árbol de Clasificación.



Fuente: Elaboración propia.

- Tamaño mínimo para realizar una partición: 600
- Tamaño mínimo de la hoja: 300
- Information Gain: 0.001
- β: 0.05

Árbol CART: (Clientes antiguos que si tienen un tráfico promedio de llamadas salientes a RPM)

Figura N° 55. Reglas definidas con el Árbol de Clasificación.

REGLAS	CEF	CNE	% CEF	NODOS	Probabilidad
CLASE_PLAN = OTROS_PLANES	648	5153	11.17%	1	0.640
CLASE_PLAN = TARIFA_UNICA LOG_OTROS_24 = <1.46;30.53] LOG_OTROS_21 = <1234.2;13876.17] ANTIGUEDAD = 2	596	4859	10.93%	2	0.616
CLASE_PLAN = TARIFA_UNICA LOG_OTROS_24 = <=1.46 ANTIGUEDAD = 2	292	2695	9.78%	3	0.588
CLASE_PLAN = TODO_DIA4 LOG_OTROS_24 = >30.53	375	4248	8.11%	4	0.547
CLASE_PLAN = TARIFA_UNICA LOG_OTROS_24 = >30.53	2231	26544	7.75%	5	0.539
CLASE_PLAN = TARIFA_UNICA LOG_OTROS_24 = <1.46;30.53] LOG_OTROS_21 = >13876.17	522	6357	7.59%	6	0.519
CLASE_PLAN = TARIFA_UNICA LOG_OTROS_24 = <1.46;30.53] LOG_OTROS_21 = <1234.2;13876.17] ANTIGUEDAD = 1	5035	71130	6.61%	7	0.487
CLASE_PLAN = TARIFA_UNICA LOG_OTROS_24 = <1.46;30.53] LOG_OTROS_21 = <=1234.2	1293	18364	6.58%	8	0.487
CLASE_PLAN = TODO_DIA4 LOG_OTROS_24 = <1.46;30.53]	1025	14586	6.57%	9	0.489
CLASE_PLAN = TODO_DIA4 LOG_OTROS_24 = <=1.46	245	3533	6.48%	10	0.469
CLASE_PLAN = TARIFA_UNICA LOG_OTROS_24 = <=1.46 ANTIGUEDAD = 1	1860	29191	5.99%	11	0.461

Fuente: Elaboración propia.

1. **Reglas de Clasificación para los Clientes CEF(Contacto efectivo venta)**

   Las características o el patrón que poseen los clientes CEF se puede apreciar en la Figura N° 35 y se resume en lo siguiente quedando como mejor patrón aquellos que superan el 7 % de Venta.

   - **Nodo 2:** Que la clase de plan del cliente sea tarifa única (TUN), que el promedio de tráfico de llamadas salientes a RPM es mayor a 1.46 minutos y menor igual a 30.53 minutos, que el total de llamadas entrantes y salientes es mayor igual a 1234.2 minutos y menor igual a 13876.17 minutos y que el tiempo que se gestionó al cliente por última vez es mayor a un año, representa el 10.93 % de los clientes venta.

   - **Nodo 5:** Que la clase de plan del cliente sea tarifa única (TUN), que el promedio de tráfico de llamadas salientes a RPM es mayor o igual a 30.53 minutos, representa el 7.75 % de los clientes venta.

2. **Reglas de Clasificación para los Clientes CNE (Contacto no efectivo no venta)**

   Las características o el patrón que poseen los clientes CNE se puede apreciar en la Figura N° 35 y se resume en lo siguiente:

   - **Nodo 8:** Que la clase de plan del cliente sea tarifa única (TUN), que el promedio de tráfico de llamadas salientes a RPM es mayor a 1.46 minutos y menor igual a 30.53 minutos, que el total de llamadas entrantes y salientes es menor igual a 1234.2 minutos representa el 6.58 % de los clientes no venta.

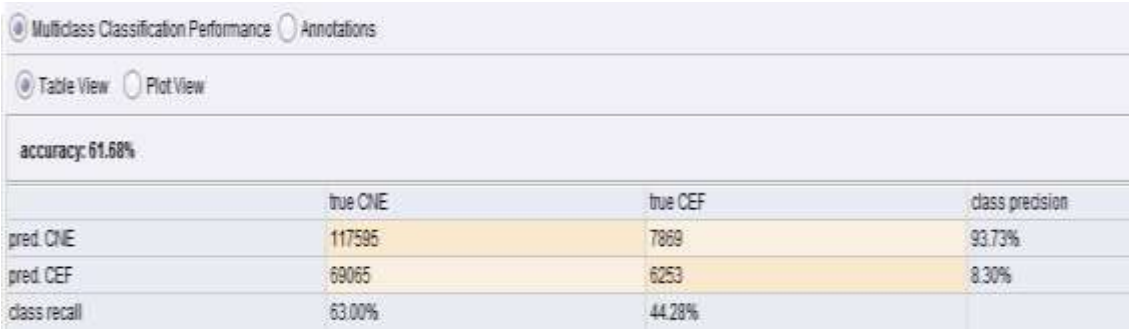
   - **Nodo 9:** Que la clase de plan Todo día 4, que el promedio de tráfico de llamadas salientes a RPM es mayor a 1.46 minutos y menor igual a 30.53 minutos, representa el 6.57 % de los clientes no venta.

   - **Nodo 11:** Que la clase de plan del cliente sea tarifa única (TUN), que el promedio de tráfico de llamadas salientes a RPM es menor igual a 1.46 minutos y que el tiempo que se gestionó al cliente por última vez es menor a un año, representa el 6.58 % de los clientes no venta.

IV. RESULTADOS Y DISCUSIÓN

Evaluación del Modelo Logístico:

Figura N° 56: Tabla de Clasificación del Modelo Logístico.



Multiclass Classification Performance | Annotations

Table View | Plot View

accuracy: 61.68%

	true CNE	true CEF	class precision
pred. CNE	117595	7869	93.73%
pred. CEF	69065	6253	8.30%
class recall	63.00%	44.28%	

Fuente: Elaboración propia.

Donde:

- **Pred CNE**: Predicción no venta.
- **Pred CEF:** Predicción venta.
- **True CNE:** valor verdadero no venta
- **True CEF:** valor verdadero venta

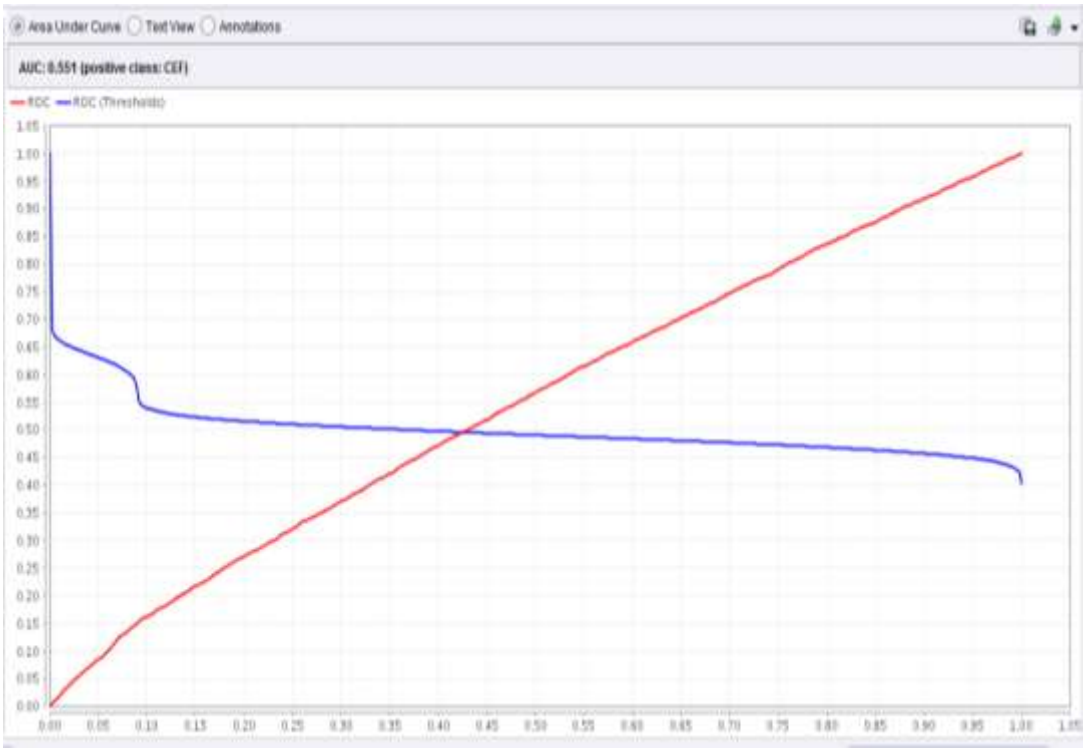
Interpretación: El modelo predice que el 44.28% de clientes son CEF verdaderos de toda la base gestionada.

Del total de 200696 registros, de las cuales el modelo detecto acertadamente al 61.68%, es decir detecto como ventas a quienes son TRUE CEF como PRED. CEF, y no ventas a quienes son TRUE CNE como PRED. CNE.

Entonces se sabe que el modelo comete errores del tipo I y II al detectar clientes propensos a migrar de un plan prepago a postpago (venta) cuando no lo son y viceversa, estos representan el 38.32% del total de registros utilizados.

Otro método para comprobar si el modelo es aceptable es el grafico ROC:

Figura N° 57: Gráfica de ROC del Modelo Logístico.



Fuente: Elaboración propia.

El desempeño de clasificación correcta para los individuos que son venta o no venta seleccionadas al azar es 55.1%.

Con un 55.1% de área bajo la curva se interpreta que existe una leve capacidad de discriminación entre los individuos que son ventas y no ventas.

Se tomó como validación la gestión del mes de Enero del 2015 y se comprobó que solo se consiguió alcanzar el 4.5% de ventas, mientras que el modelo predice el 35.2% de ventas a obtener.

Figura N° 58: Validación del Modelo Logístico.

Resultado Gestión	Predicción		Total	% Predicción
	CEF	CNE		
CEF	1633	3012	(4.5%) 4645	**35.2%**
CNE	8865	17355	(25.3%) 26220	**66.2%**
% Predicción	**15.6%**	**85.2%**		61.5%
Total	**10498**	**20367**	**23956**	**34.0%**

Fuente: Elaboración propia.

- **CEF:** Contactos efectivos (venta)
- **CNE:** Contactos no efectivos (no venta)

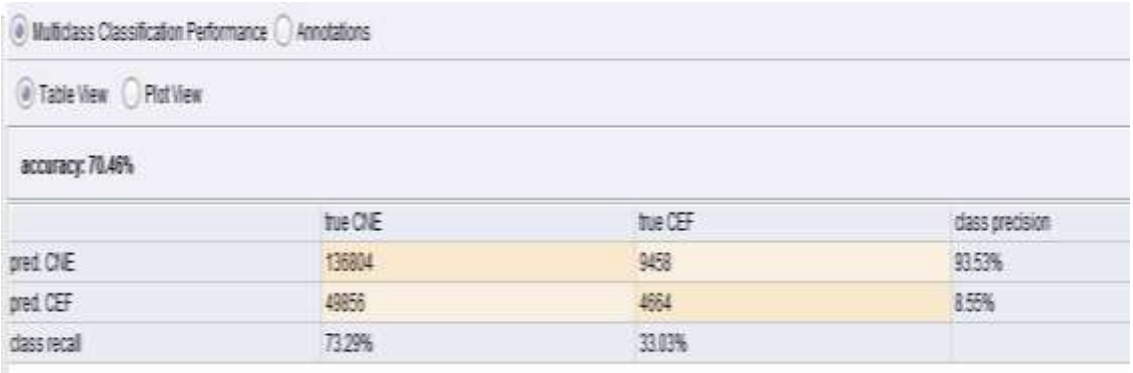
4.5% representa las ventas obtenidas en la gestión actual.

35.2% representa las ventas predichas con el Modelo.

El modelo predice para la base del mes de Enero 2015, que el 61.5% del total de predicciones son correctas.

<u>Evaluación del Árbol de Clasificación:</u>

Figura N° 59: Tabla de Clasificación del Árbol de Clasificación.



Fuente: Elaboración propia.

Interpretación: El modelo predice que el 33.03% de clientes son CEF verdaderos de toda la base gestionada.

Se observa que se trabajó con una muestra de 200696 registros, de las cuales el modelo detecto acertadamente al 70.46%, es decir detecto como ventas a quienes son TRUE CEF como PRED. CEF, y no ventas a quienes son TRUE CNE como PRED. CNE.

Entonces se sabe que el modelo comete errores del tipo I y II al detectar clientes propensos a migrar de un plan prepago a postpago (venta) cuando no lo son y viceversa, estos representan el 29.54% del total de registros utilizados.

Otro método para comprobar si el modelo es aceptable es el grafico ROC:

Figura N° 60: Gráfica de ROC del Árbol de Clasificación.

Fuente: Elaboración propia.

El desempeño de clasificación correcta para los individuos que son venta o no venta seleccionadas al azar es 53.9%.

Con un 53.9% de área bajo la curva se interpreta que existe una leve capacidad de discriminación entre los individuos que son ventas y no ventas.

Se tomó como validación la gestión del mes de Enero del 2015 y se comprobó que solo se consiguió alcanzar el 4.5 % de ventas, mientras que el modelo predice el 28.9% de ventas a obtener.

Figura N° 61: Validación del Árbol de Clasificación.

	Predicción		Total	% Predicción
Resultado Gestión	CEF	CNE		
CEF	1341	3304	(4.5%) 4645	**28.9%**
CNE	7177	19043	(25.3%) 26220	**72.6%**
% Predicción	**15.7%**	**85.2%**		66.0%
Total	**8518**	**22347**	**23956**	**27.6%**

Fuente: Elaboración propia.

4.5% representa las ventas obtenidas en la gestión actual.

28.9% representa las ventas predichas con el Modelo.

El modelo predice para la base del mes de enero 2015, que el 66.0% del total de predicciones son correctas.

- **Segmentación de las probabilidades de éxito de la Regresión Logística para predecir clientes futuros:**

Se Segmentaron las probabilidades de éxito en cuatro grupos usando los percentiles 0.25, 0.5 y 0.75, quedando finalmente los grupos de la siguiente manera:

G1: Prob>0.5091

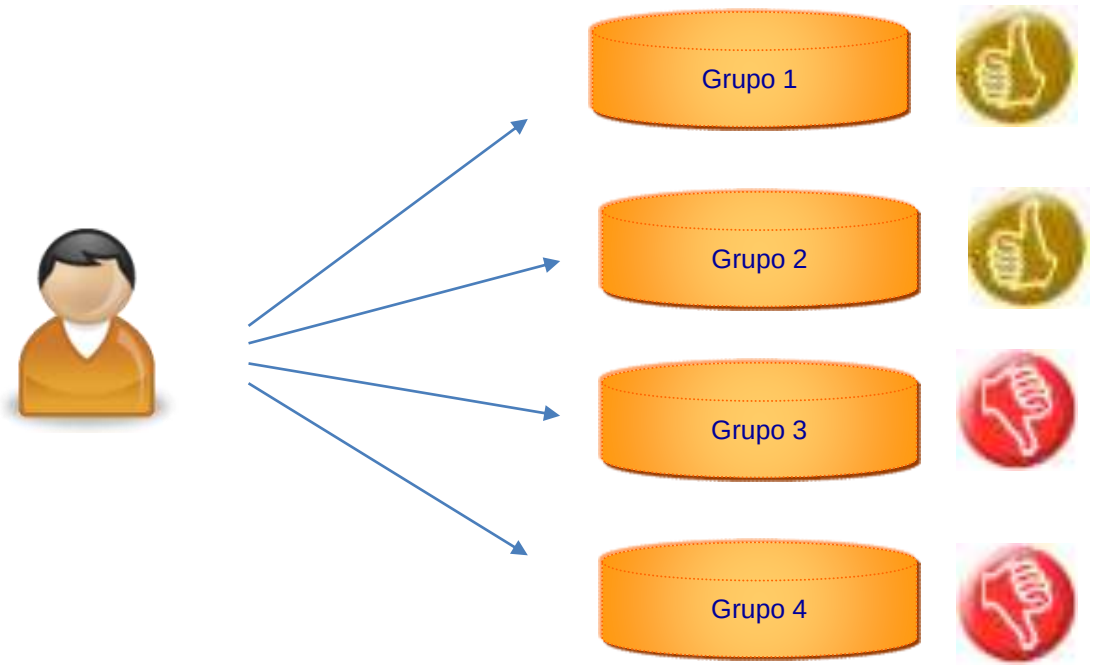
G2: 0.4881<Prob<0.5091

G3: 0.4699<Prob<0.4881

G4: Prob<0.4699

Con estas reglas definidas se garantiza que los clientes pertenecientes al G1 tendrán mayor probabilidad de ser venta (Migar de un plan prepago a postpago) y así sucesivamente quedando el G4 como el grupo con menor probabilidad.

Figura N° 62: Clasificación de un nuevo Cliente.



Fuente: Elaboración propia.

- **<u>Ejemplo en la predicción de un nuevo cliente:</u>**

Finalizado el Trabajo de investigación quedando como mejor modelo la Regresión Logística se procedió a predecir la pertenencia de un nuevo cliente a unos de los cuatro grupos mencionados líneas arriba.

Nuevo Modelo Logístico: (Clientes antiguos que si tienen un tráfico promedio de llamadas salientes a RPM)

$$logit\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Donde:

Variables	Coeficientes	
Intercepto	β_0	0.4075
Log_Otros_24	β_1	-0.0525
Log_Otros_21	β_2	-0.0322
Antigüedad	β_3	-0.6058
Clase_Plan	β_4	-0.5979

Con los valores obtenidos del modelo de Regresión Logística, se procedió a realizar una predicción a un nuevo cliente y determinar a qué grupo pertenecerían de acuerdo a sus respectivas probabilidades.

Cuadro N° 9: Predicción de un Cliente Nuevo.

Clase_Plan	Log_Otros 24	Log_Otros_21	Antigüedad
Tarifa_Unica	5.8 minutos	8 minutos	10 meses

Fuente: Elaboración propia

Donde:

Clase_Plan: Plan prepago del cliente.

- Tarifa_Unica =1
- Otros_Planes= 2

Log_Otros_24: Promedio de tráfico de llamadas salientes a RPM.

Log_Otros_21: Total de llamadas entrantes y salientes

Antigüedad: Tiempo en años en que se gestionó un cliente por última vez.

- Menor a un año= 1
- Mayor a un año= 2

Con los datos mencionados líneas arriba se procedió a hallar la probabilidad de éxito si el cliente es venta o no venta al momento de una llamada telefónica.

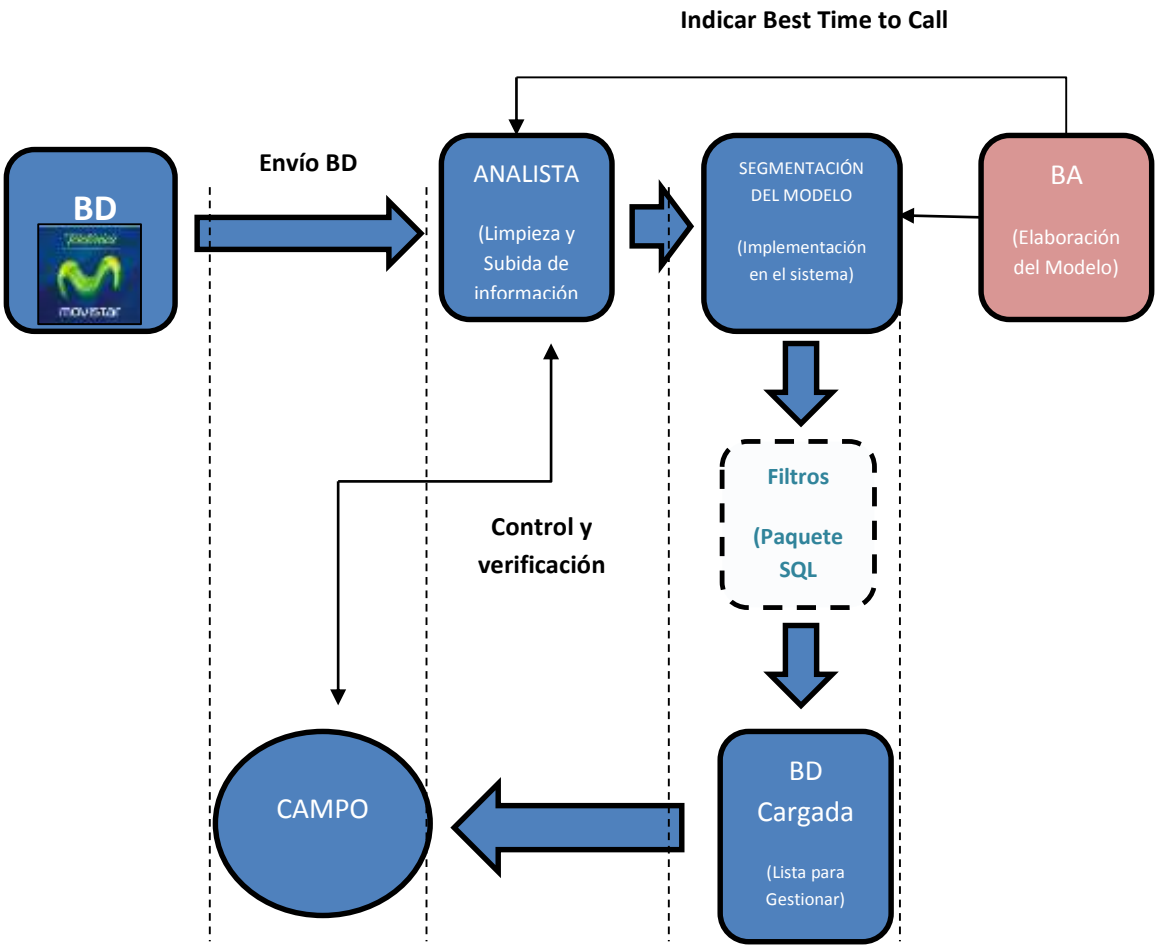
Cliente Nuevo:

$$\pi(x) = \frac{e^{[0.4075+5.8(-0.0525)+8(-0.0322)+1(-0.6058)+1(-0.5979)]}}{1+e^{[0.4075+5.8(-0.0525)+8(-0.0322)+1(-0.6058)+1(-0.5979)]}}$$**=0.5142**

El cliente nuevo que tiene un plan prepago de tarifa única (TUN), que realiza en promedio 5.8 minutos de llamadas salientes a RPM, que realiza un total de llamadas entrantes y salientes de 8 minutos y la última vez que fue gestionado fue hace 10 meses pertenecería al grupo **G1,** con una probabilidad obtenida de **0.5142** representa una mayor probabilidad de ser venta (Migrar de un plan prepago a postpago).

Implementación del Modelo de Regresión Logística:

Figura N° 63: Flujo de Implementación



Fuente: Elaboración propia.

1. Recepción de Base de datos enviado por Movistar Perú.
2. Realizar una limpieza de la base.
3. Aplicar el Modelo creado y segmentarlo en grupos para optimizar la gestión.
4. Crear filtros (campos o variables que se utilicen para filtrar grupos de clientes con alguna característica en particular).
5. Cargar la base de datos que esta lista para gestionar.
6. Se procede al inicio de la gestión con las debidas indicaciones de horarios para llamar (Best time to Call), obtenidos del histórico y utilizado para generar el modelo.

7. Controlar la gestión de llamadas con el fin de reintentar sobre los registros marcados como potenciales y lograr cumplir con el objetivo planteado.

Figura N° 64: Horarios para gestionar a los Clientes.

GRUPOS	8	9	10	11	12	13	14	15	16	17	18	19	20	21
R1	6.72%	9.68%	8.02%	8.56%	7.13%	5.35%	7.74%	8.93%	8.32%	6.31%	5.94%	6.39%	5.82%	5.10%
R2	6.91%	9.63%	7.74%	8.68%	7.48%	6.04%	7.59%	9.06%	8.45%	6.13%	5.70%	5.97%	5.59%	5.03%
R3	6.86%	9.75%	8.13%	8.61%	7.38%	5.74%	7.76%	8.87%	8.10%	6.21%	5.91%	5.66%	5.80%	5.21%
R4	6.65%	9.82%	7.96%	9.20%	7.88%	6.24%	7.37%	8.33%	8.02%	5.66%	5.58%	5.54%	6.18%	5.57%

Fuente: Elaboración propia.

Estos son los horarios para gestionar a los clientes antiguos que realizan tráfico de llamadas.

- El mejor horario para llamar entre las 8:00 a.m. y 9:00 am. para gestionar los grupos R1 y R3.
- El mejor horario para llamar entre las 9:00 am. y 11:00 am. para gestionar los grupos R1, R2, R3 Y R4.
- El mejor horario para llamar entre las 2:00 pm y 4:00 pm para gestionar los grupos R1, R2, R3 y R4.
- A partir de las 6:00 pm se puede llamar a cualquiera de los grupos, ya que son los horarios menos contactables.

V. CONCLUSIONES

- Como principal conclusión, el Modelo Logístico es mejor que el Algoritmo del árbol de clasificación CART para este trabajo, dado que la primera tiene una mejor tasa de clasificación, mejor desempeño de clasificación correcta para los individuos que son venta o no venta seleccionados al azar y la distribución de las probabilidades favorece la segmentación de 4 grupos para optimizar la gestión, mediante una validación el Algoritmo CART predice que el 28.9% será venta, mientras que el Modelo logístico predice 35.2% de ventas para una prueba aplicada a las gestión del mes de Enero del 2015.

- Mediante la validación del modelo usando la base de gestión del mes de Enero, para la aplicación del Algoritmo CART, se obtuvo una tasa de clasificación de 66.0%, mientras que con el Modelo Logístico el 61.5%, pese a ello se optó por aplicar este último; ya que presenta mejores indicadores y una mayor predicción de CEF (venta).

- Las reglas que se obtuvo con el Árbol de clasificación ayudarán a la gestión para tener una visibilidad del perfil del cliente por cada grupo segmentado.

- Se logró comprobar que para realizar un Modelo Logístico es mejor utilizar las variables cuantitativas, tal como son y no categorizarlas, ya que son más efectivas y aumentan su peso de importancia para ser seleccionadas.

- El grupo de clientes que realizan tráfico de llamadas a RPM tienen mayor probabilidad de ser venta y representan 35% aproximadamente de toda la base que se va a gestionar.

- Con este trabajo se concluye finalmente que a pesar de tener variables que se distribuyen con una tendencia a la normal que es lo que buscamos en todas las variables cuantitativas no se obtuvo un ROC óptimo, muchas veces las bases reales son duras y no aportan lo suficiente para mejorar este indicador. Pero a pesar de ello si se logró encontrar una buena tasa de clasificación.

VI. RECOMENDACIONES

- Se encontraron diferentes comportamientos para la variable Tráfico de llamadas salientes a RPM, un grupo de clientes presenta información, mientras que el otro grupo no, por ello se recomienda considerar dos modelos porque son poblaciones distintas y dentro de cada una el resto de variables si tienen tendencia a la normal y en la mayoría no existen missing.

- Se debe considerar aumentar el número de variables que aporten información suficiente para el Análisis, por ejemplo la edad, el tiempo que tarda en realizar una llamada a RPM, el promedio de monto que gasta para realizar una recarga presenta un histórico de 6 meses y no solo 3 meses, etc.

- Se debe considerar utilizar el BEST TIME TO CALL que significa el mejor horario para contactar al cliente, ya que existe una mayor probabilidad de contactarlo.

- Con el modelo planteado y por los resultados que arroja, se puede negociar con el cliente que provee las bases para dar la opción de ofrecer una segunda oferta que sea más factible para el cliente.

- Tratar de ofrecer planes postpago con un menor costo para aquellas personas que no realizan tráfico de llamadas a RPM.

- Mantener los datos brindados actualizados por parte del cliente que provee las bases.

- Utilizar las reglas del Árbol de clasificación como guía para la gestión, porque fue comprobado que dichas reglas si se cumplen a la actualidad.

- Para el caso del análisis de intervalos del odds ratio se recomienda realizar nuevamente una clusterización para la variable Cluster_Opción1, de tal manera que se obtenga grupos más homogéneos.

VII. REFERENCIAS BIBLIOGRAFICAS

- ABRAIRA SANTOS, A, PEREZ DE VARGAS LUQUE. (1996) Métodos Multivariantes en Bioestadística.

- ACUÑA FERNÁNDEZ EDGAR. (1999) Métodos de Análisis Discriminante, Universidad de Puerto rico en Mayagüez.

- ATO, M. y LÓPEZ-GARCÍA, J. J. (1996) Análisis estadístico para datos categóricos (Madrid, Síntesis).

- ARRENDONDO VIDAL TOMÁS. (2008) Modelos de Segmentación

- BESSIS, J. (2002). Risk Management in Banking. John Wiley & Sons, Itd. Inglaterra: west sussex.

- BREIMAN, L., J. H. FRIEDMAN, R. A. OLSHEN, AND C. J. STONE. (1984) Classification and regression tres. Belmont, Calif: Wadsworth.

- CORTIJO BON, JOSÉ FRANCISCO. Técnicas Supervisadas II, Aproximación No Paramétrica. (En línea).
  http://iie.fing.edu.uy/ense/asign/recpat/material/tema3_00-01/node26.html
  http://iie.fing.edu.uy/ense/asign/recpat/material/tema3_00-01/node26.html

- DANIEL PEÑA. (2002). Regresión y Diseño de experimentos.

- DOBSON, ANNETTE J. (2002). An Introduction to Generalized Linear Models Second edition.

- FOODY, G. M. (2002) Status of land cover classification accuracy assessment. Remote Sensing of Environment, 80.

- HOSMER, D. W. AND LEMESHOW, S. (2000) Applied Logistic Regression, Second Edition, Wiley, New York.

- HUNT, E.B., MARIN, J., & STONE, P.J. (1966). Experiments in induction. New York: Academic Press.

- JR. QUINLAN. (1986). Induction of Decision Trees.

- KLEINBAUM, D. G., KUPPER, L. L., MULLER, K. E. AND NIZAM, A. (1998) Applied Regression Analysis and Multivariable Methods, Third Edition, Duxbury, Pacific Grove, California

- KOH, H. (1992) The Sensitivity of Optimal Cutoff Points to Misclassification Costs of Type I and Type II Errors in the Going-Concern Prediction Context. Journal of Business Finance & Accounting.

- LÉVY MANGIN JEAN - PIERRE, VARELA MALLOU JESÚS. (2008) Análisis Multivariable para las Ciencias Sociales.

- LIU, C., FRAZIER, P. & KUMAR, L. (2007) Comparative Assessment of the measures of thematic classification accuracy. remote sensing of environment, 107, 606–616.

- LUIS CAYUELA. Universidad de Granada. Modelos lineales generalizados (GLM) (En Línea) http://www.uv.es/~lejarza/mcaf/glm2.pdf

- MCCULLAGH, P. y NELDER, J. (1989) Generalized Linear Models (2 ed.) (London, Chapman & Hall).

- O. Z. MAIMON AND L. ROKACH. Data mining and knowledge discovery handbook. Springer-Verlag New York Inc, 2005.

- Q. ZHAO AND S. S. BHOWMICK. (2006) Association rule mining: A survey, Nanyang Technological University, Singapore.

- RED DE REVISTAS CIENTÍFICAS DE AMÉRICA LATINA Y EL CARIBE, ESPAÑA Y PORTUGAL. ESCALAS DE MEDICIÓN EN ESTADÍSTICA. (En Línea). http://www.redalyc.org/articulo.oa?id=99315569009

- REFAEILZADEH, P. TANG, L. LUI, H. K. (2008) Fold Cross-Validation, Arizona State University.

- REGRESIÓN LOGÍSTICA (En línea) http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/CUALITATIVAS/LOGISTICA/regresion-logistica.pdf

- STEVENS, STANLEY. (1946) On the Theory of Scales of Measurement. Sciences New Series, Vol. 103.

VIII. ANEXOS

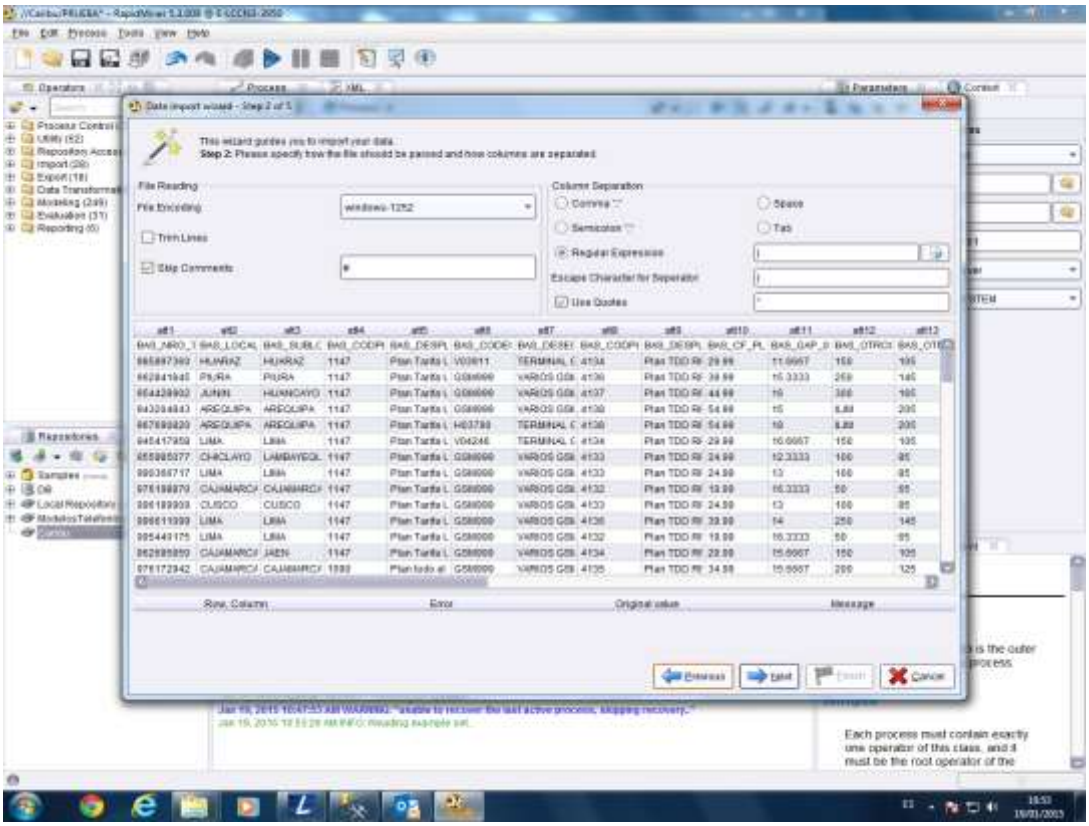
Anexo N° 1: Relación entre probabilidad y odds.

Probabilidad	Odds
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

Fuente: Elaboración propia.

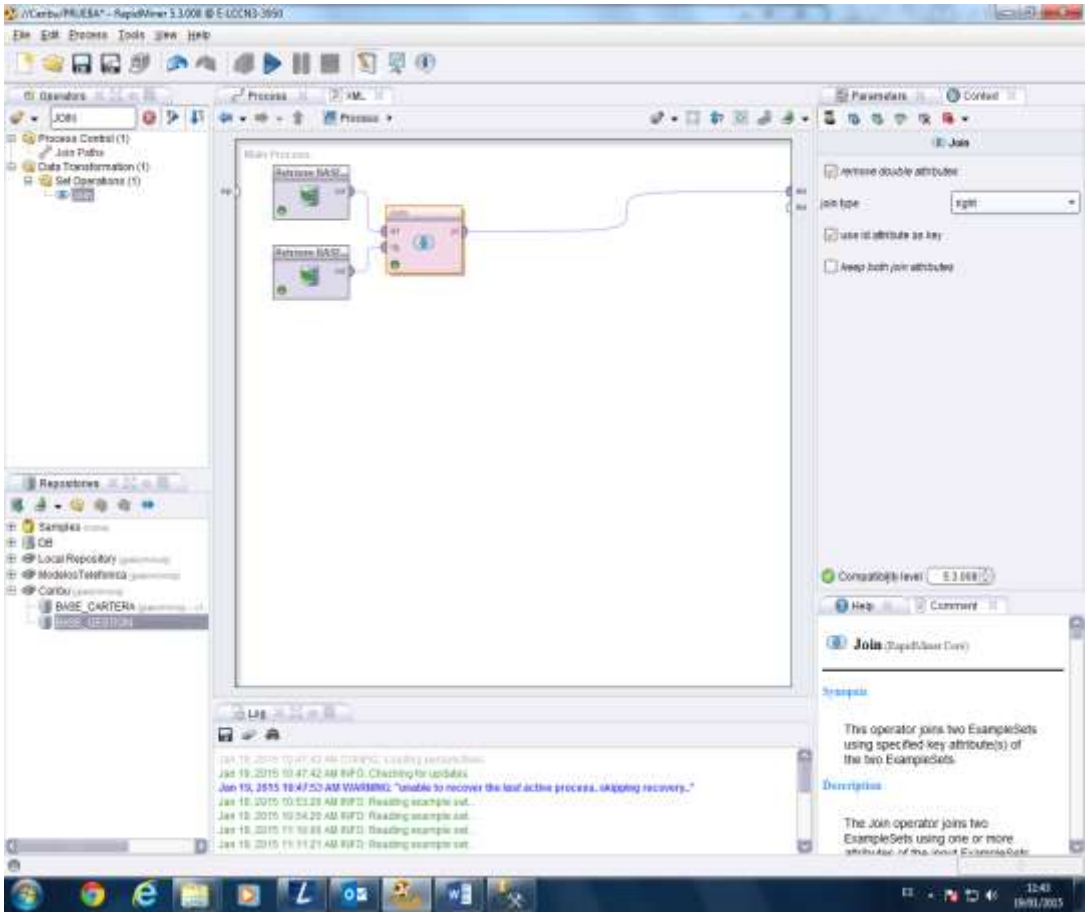
Anexo N° 2: Proceso en Rapid miner.

<u>Paso 1</u>: Subir la base.



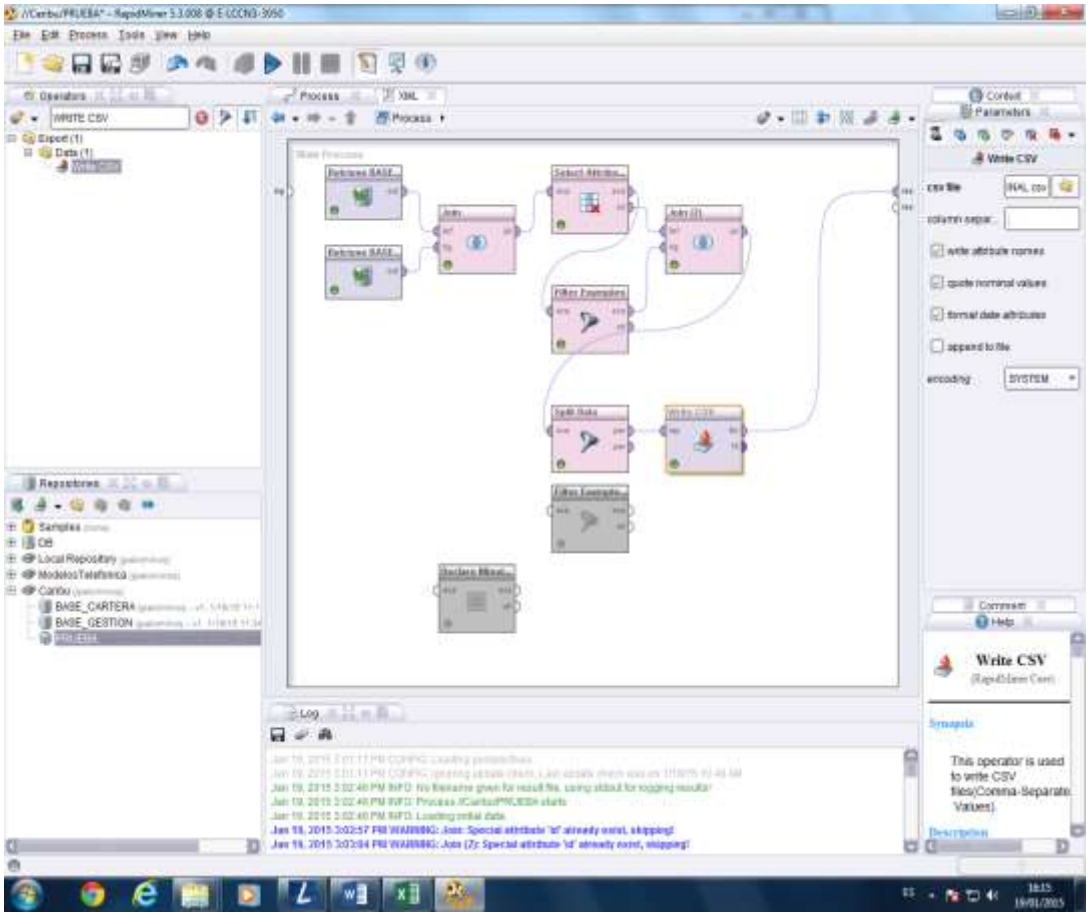
Fuente: Elaboración propia.

<u>Paso 2:</u> Unir las Bases.



Fuente: Elaboración propia.

<u>Paso 3:</u> Construir el Modelo.



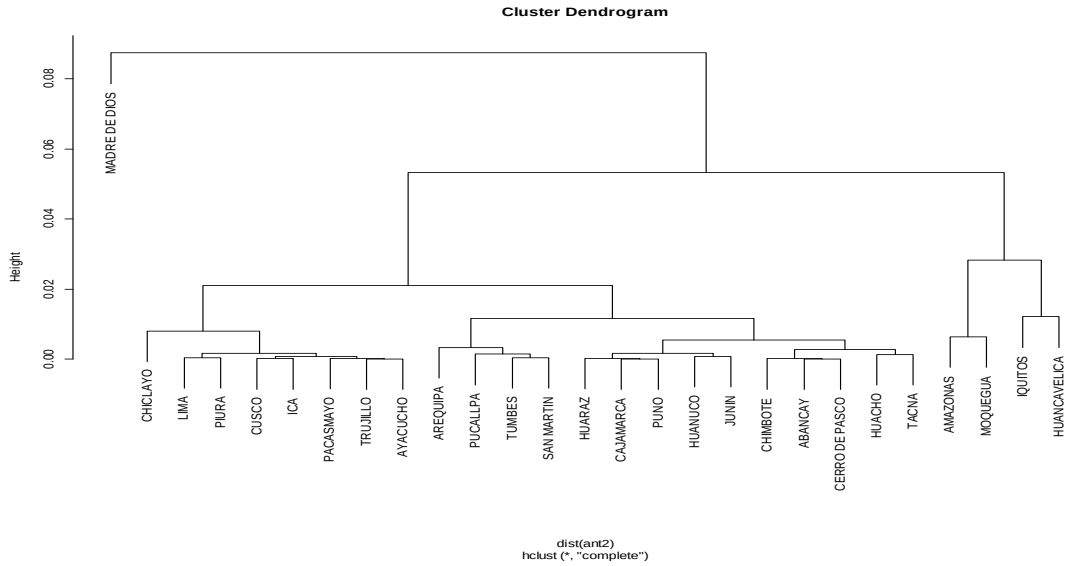
Fuente: Elaboración propia.

Anexo N° 3: Agrupación del Cluster.

LOCALIDAD	CLUSTER-opcion1	CLUSTER-opcion2
IQUITOS	Iqui-Huanc	Ica-Huanc
MOQUEGUA	Iqui-Huanc	Ica-Huanc
AMAZONAS	Iqui-Huanc	Ica-Huanc
HUANCAVELICA	Iqui-Huanc	Ica-Huanc
ICA	Piu-Ayac	Ica-Huanc
SAN MARTIN	SanMar-Huac	Ica-Huanc
CHIMBOTE	SanMar-Huac	Ica-Huanc
CAJAMARCA	SanMar-Huac	Ica-Huanc
AREQUIPA	SanMar-Huac	Ica-Huanc
JUNIN	SanMar-Huac	Ica-Huanc
PUNO	SanMar-Huac	Ica-Huanc
PACASMAYO	Piu-Ayac	Pacas-Huach
CUSCO	Piu-Ayac	Pacas-Huach
AYACUCHO	Piu-Ayac	Pacas-Huach
PUCALLPA	SanMar-Huac	Pacas-Huach
TUMBES	SanMar-Huac	Pacas-Huach
HUANUCO	SanMar-Huac	Pacas-Huach
HUARAZ	SanMar-Huac	Pacas-Huach
TACNA	SanMar-Huac	Pacas-Huach
ABANCAY	SanMar-Huac	Pacas-Huach
CERRO DE PASCO	SanMar-Huac	Pacas-Huach
HUACHO	SanMar-Huac	Pacas-Huach
PIURA	Piu-Ayac	Piu-MadredDi
TRUJILLO	Piu-Ayac	Piu-MadredDi
CHICLAYO	Piu-Ayac	Piu-MadredDi
LIMA	Piu-Ayac	Piu-MadredDi
MADRE DE DIOS	Piu-Ayac	Piu-MadredDi

Fuente: Elaboración propia.

Anexo N°4: Dendograma del Cluster.



Fuente: Elaboración propia.

Anexo N° 5: Códigos Cluster en R

```
>  ant<- read.delim("clipboard",T)
> ant2<-as.matrix(ant[,2])
> rownames(ant2)<-ant[,1]
> hca <- hclust(dist(neo2))
> hca <- hclust(dist(ant2))
> plot(hca)
>  ant<- read.delim("clipboard",T)
> ant2<-as.matrix(ant[2:3])
> rownames(ant2)<-ant[,1]
> hca <- hclust(dist(ant2))
> plot(hca)
>
```

Fuente: Elaboración Propia.

Anexo N° 6: Coeficientes del Modelo Logístico.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	CLASE_PLAN	-,599	,067	80,221	1	,000	,549	,482	,626
	ANTIGUEDAD	-,601	,044	182,706	1	,000	,548	,502	,598
	LOG_OTROS_21	-,034	,011	9,353	1	,002	,967	,946	,988
	CLUSTERopcion1			29,743	2	,000			
	CLUSTERopcion1(1)	-,347	,066	27,789	1	,000	,707	,621	,804
	CLUSTERopcion1(2)	,003	,025	,016	1	,901	1,003	,955	1,054
	LOG_OTROS_24	-,051	,008	46,201	1	,000	,950	,936	,964
	Constant	,408	,102	102,621	1	,000	1,506		

a. Variable(s) entered on step 1: CLASE_PLAN, ANTIGUEDAD, LOG_OTROS_21, CLUSTERopcion1, LOG_OTROS_24.

Fuente: Elaboración propia.

Anexo N° 7: Coeficientes del Modelo Logístico sin considerar la variable Cluster_opción1.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	CLASE_PLAN(1)	-,598	,067	79,951	1	,000	,550	,482	,627
	ANTIGUEDAD	-,605	,044	185,135	1	,000	,546	,500	,596
	LOG_OTROS_21	-,032	,011	8,553	1	,003	,968	,948	,989
	LOG_OTROS_24	-,052	,008	48,167	1	,000	,949	,935	,963
	Constant	,408	,100	102,060	1	,000	2,758		

a. Variable(s) entered on step 1: CLASE_PLAN, ANTIGUEDAD, LOG_OTROS_21, LOG_OTROS_24.